Complexity in Translation

An English-Norwegian Study of Two Text Types

Martha Thunes

Dissertation for the degree of doctor philosophiae (dr.philos.)

University of Bergen, Norway

2011

Contents

Preface

A notion that I will refer to as *the correspondence type hierarchy* plays a major part in this book. It is a classification model for translational correspondences, and its main principles were originally developed by Helge Dyvik of the University of Bergen, in connection with the research project "The Semantics of Mulitilinguality and Algorithms Related to Translation" (SMART), which he ran in collaboration with Jens Erik Fenstad, Tore Langholm, and Jan Tore Lønning. In 1993 I started working as a research assistant for that project, and it was my task to apply the type hierarchy to English-Norwegian parallel texts in order to collect empirical data. While I was doing so, the late Stig Johansson of the University of Oslo one day visited me and took interest in my analysis of translational correspondences. I gave him a copy of Helge's original definition of the classification model, and soon after Stig's colleague Hilde Hasselgård applied the type hierarchy in an English-Norwegian word order study. I myself integrated the model in my doctoral project, and after I had received a scholarship, Stig invited me to join a group of researchers in 1996/97, working under the heading "Contrastive Analysis and Translation Studies Linked to Text Corpora". Further development of the correspondence type hierarchy was my contribution to the group's work. On Stig's initiative, the group members wrote a book together, and, hence, documentation on the type hierarchy was published in 1998. Still, I did not know of others, apart from Hilde, who had applied this classification model until I fairly recently was contacted by the linguist Marco Antonio Esteves da Rocha of the Federal University of Santa Catarina, Florianópolis, who told me that he had used my article in teaching. It was highly inspiring to learn that students of his have applied the model to the language pair English-Portuguese, and that they have found it to be useful for the purpose of describing and analysing translational correspondences. In Florianópolis an approach based on the type hierarchy is implemented in an emerging doctoral project analysing Shakespeare sonnets and translations of them into Portuguese.

I have experienced that the type hierarchy easily gets into the blood of the analyst who works with it every day. As I was compiling data for my investigation, I had the habit of wondering, also when away from my desk, what type of correspondence it might be when I saw translationally parallel expressions in languages that I could understand. In the present contribution, the classification model is spelled out in detail, and I provide a description of it which conforms with the principles according to which it was originally defined. My motivation for applying the correspondence type hierarchy to English-Norwegian parallel texts has been to investigate to what extent it may be fruitful to try to automatise the translation of selected text types for this language pair. It is my view that as a classification model for translational correspondences, the type hierarchy is helpful, firstly, because it relies on linguistic criteria that are fairly easy to apply, and, secondly, because it is flexible — it can be modified according to the purposes of one's investigation.

Acknowledgements

The completion of the present product, my dissertation, has been a long and winding process, and it is now time to gratefully acknowledge the many good forces that have helped me.

I thank the Research Council of Norway for the doctoral fellowship (grant number 108126) which made it possible to start this project, and I am greatly indebted to the Centre for Advanced Study in Oslo, especially to its staff, for the truly enjoyable year I spent there. Further, I am equally grateful to the University of Bergen, who was my employer, and a highly supportive one during a period when I struggled with long-term illness. I acknowledge the very helpful assistance of a range of administrative people, formerly in the Department of Linguistics and Comparative Literature (now extinct), and more recently in the Department of Linguistic, Literary and Aesthetic Studies. Also, I am grateful to my previous employer Aksis (now Uni Digital) for allowing me periods of unpaid leave so that I could work on my thesis.

Moreover, I thank the many writers and translators who have produced the texts I have analysed. For making the texts available to me, I am indebted to the Norwegian Ministry of Foreign Affairs, the Norwegian Petroleum Directorate, and the English-Norwegian Parallel Corpus (ENPC) Project, in particular to Jarle Ebeling and Knut Hofland.

Before he sadly passed away last spring, I had the chance to thank Stig Johansson for the various ways in which he had contributed to my research. Firstly, he initiated the ENPC, which offers a goldmine of empirical data for contrastive studies of English-Norwegian, and which has been vitally important to my work. Secondly, he invited me to be a member, during the academic year 1996–97, of his research group "Contrastive Analysis and Translation Studies Linked to Text Corpora" at the Centre for Advanced Study at the Norwegian Academy of Science and Letters. This gave me a wonderful year, and it provided me with a sound basis for the empirical investigation of my project. Thirdly, thanks to Stig's initiative I was able to publish an article describing my work. In addition, Stig was, in relation to so many, a cherished colleague and friend, and while there was still time, I was lucky to be able to express my appreciation to him.

Next, I warmly thank Cathrine Fabricius Hansen, who has served as a secondary advisor, for valuable comments and moral support. I am also deeply indebted to my friend and colleague Bergljot Behrens, for advice, for feed-back on manuscripts, and, most importantly, for her enthusiasm and encouragement, which has been a strong source of inspiration.

Further, I am very grateful to Victoria Rosén, who has helped me with a native speaker's judgments of English expressions, and who has been a very supportive colleague during the final phase of this project.

As I could not have written this book without access to a library, I have truly appreciated the skilful staff of the University of Bergen Library, and in particular the excellent service of interlibrary loans, personified by wizard woman Kari Normo. I have even enjoyed the luck of having had two good friends among the librarians: in the early stages of my project Maya Thee worked as my patron saint, and later Jan Olav Gatland has taken care of my interests. Warm thanks to both of them.

Also, I want to express deep gratitude to Bjørn Tore Sund of the IT department of the University of Bergen. Thanks to him, it is still possible to run the Medley Lisp environment on our servers, which is a prerequisite for accessing the empirical data I have collected. Without his efforts it would have been difficult to complete my project.

Then, I want to thank heartily a group of people who have given me valuable input, or various kinds of help, big or small, during the long course of my project: Tone Aarland, Jan Aarts, Gisle Andersen, Øivin Andersen, Flávia Azevedo, Kristin Bech, Dagmar Čejka, Östen Dahl, Kjersti Fløttum, Nils Gilje, Sandra Halverson, Hilde Hasselgård, Torill Hestetræet, Torodd Kinn, Werner Koller, Randi Korneliussen, Maria Koptjevskaja-Tamm, Gunn Inger Lyse, Paul Meurer, Marco Antonio Esteves da Rocha, Antin Fougner Rydning, Ingrid Simmonæs, Koenraad de Smedt, Kjetil Strand, Arne Svindland, Marianne Thunes, and Filip Truyen.

Except for my participation in Stig Johansson's research group, the probably best part of my time as a doctoral student was to belong in a regular crowd of lunch mates, among which the majority were research fellows in linguistics or related studies. I thank them all for every healthy laugh we shared.

My year at the Centre for Advanced Study was an exceptionally good experience, and this was largely due to my fellow group members: Jan Aarts, Bengt Altenberg, Monika Doherty, Helge Dyvik, Jarle Ebeling, Cathrine Fabricius Hansen, Knut Hofland, and Stig Johansson. Quite soon a shared feeling evolved within the group, and it was something that brought much fun into everyday life at the Centre. It did not vanish as we parted at the end of that year, and although the entire group never met again, I have always felt that the feeling has been renewed on occasions of partial reunions. In gratitude to my fellow members I dedicate this work to the good, old "group spirit".

Working as a researcher easily leads to neglecting family and friends. I am very grateful to my mother who was always so full of understanding, and who told me not to have a guilty conscience when I was feeling bad about putting work first. After she passed away, my brother, sister, and father have taken over that function, never ever asking why on earth it has taken so long to finish. I heartily thank them for their loyal

support, and especially my father for being an extremely patient baby-sitter. I am also grateful to a large number of friends and relatives for their encouragement and sympathy.

I deeply thank my husband and favourite linguist, Helge Dyvik, for having carried the major burden of supervising this project. His assistance has included help on the grant proposal, extensive programming services, numerous discussions, secretarial assistance, and guidance through the writing process. In particular, I am deeply impressed by the way he has tailored software for parallel text processing to the specific needs of my investigation. Thanks to Helge, I have carried through. In our partnership we started out as colleagues, which we still are, and for more than a decade we have also been best friends. Although his contribution as a linguist has been invaluable, I cherish even more greatly his love, companionship, and day-to-day efforts as the wonderful father of our dear son.

Finally, endless thanks to Knut Helge, the other best friend in my life, for his generous hugs and kisses, for the way he cares about me, and for all the times he has said: “Good luck on your thesis, Mum!”

Bergen, May 2011

Martha Thunes

Abstract

The present study discusses two primary research questions. Firstly, we have tried to investigate to what extent it is possible to compute the actual translation relation found in a selection of English-Norwegian parallel texts. By this we understand the generation of translations with no human intervention, and we assume an approach to machine translation (MT) based on linguistic knowledge. In order to answer this question, a measurement of translational complexity is applied to the parallel texts. Secondly, we have tried to find out if there is a difference in the degree of translational complexity between the two text types, law and fiction, included in the empirical material.

The study is a strictly product-oriented approach to complexity in translation: it disregards aspects related to translation methods, and to the cognitive processes behind translation. What we have analysed are intersubjectively available relations between source texts and existing translations. The degree of translational complexity in a given translation task is determined by the types and amounts of information needed to solve it, as well as by the accessibility of these information sources, and the effort required when they are processed.

For the purpose of measuring the complexity of the relation between a source text unit and its target correspondent, we apply a set of four correspondence types, organised in a hierarchy reflecting divisions between different linguistic levels, along with a gradual increase in the degree of translational complexity. In type 1, the least complex type, the corresponding strings are pragmatically, semantically, and syntactically equivalent, down to the level of the sequence of word forms. In type 2, source and target string are pragmatically and semantically equivalent, and equivalent with respect to syntactic functions, but there is at least one mismatch in the sequence of constituents or in the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, but there is at least one structural difference violating syntactic functional equivalence between the strings. In type 4, there is at least one linguistically non-predictable, semantic discrepancy

between source and target string. The correspondence type hierarchy, ranging from 1 to 4, is characterised by an increase with respect to linguistic divergence between source and target string, an increase in the need for information and in the amount of effort required to translate, and a decrease in the extent to which there exist implications between relations of source-target equivalence at different linguistic levels.

We assume that there is a translational relation between the inventories of simple and complex linguistic signs in two languages which is predictable, and hence computable, from information about source and target language systems, and about how the systems correspond. Thus, computable translations are predictable from the linguistic information coded in the source text, together with given, general information about the two languages and their interrelations. Further, we regard non-computable translations to be correspondences where it is not possible to predict the target expression from the information encoded in the source expression, together with given, general information about SL and TL and their interrelations. Non-computable translations require access to additional information sources, such as various kinds of general or task-specific extra-linguistic information, or task-specific linguistic information from the context surrounding the source expression. In our approach, correspondences of types 1–3 constitute the domain of linguistically predictable, or computable, translations, whereas type 4 correspondences belong to the non-predictable, or non-computable, domain, where semantic equivalence is not fulfilled.

The empirical method involves extracting translationally corresponding strings from parallel texts, and assigning one of the types defined by the correspondence hierarchy to each recorded string pair. The analysis is applied to running text, omitting no parts of it. Thus, the distribution of the four types of translational correspondence within a set of data provides a measurement of the degree of translational complexity in the parallel texts that the data are extracted from. The complexity measurements of this study are meant to show to what extent we assume that an ideal, rule-based MT system could simulate the given translations, and for this reason the finite clause is chosen as the primary unit of analysis.

The work of extracting and classifying translational correspondences is done manually as it requires a bilingually competent human analyst. In the present study,

the recorded data cover about 68 000 words. They are compiled from six different text pairs: two of them are law texts, and the remaining four are fiction texts. Comparable amounts of text are included for each text type, and both directions of translation are covered.

Since the scope of the investigation is limited, we cannot, on the basis of our analysis, generalise about the degree of translational complexity in the chosen text types and in the language pair English-Norwegian. Calculated in terms of string lengths, the complexity measurement across the entire collection of data shows that as little as 44,8% of all recorded string pairs are classified as computable translational correspondences, i.e. as type 1, 2, or 3, and non-computable string pairs of type 4 constitute a majority (55,2%) of the compiled data. On average, the proportion of computable correspondences is 50,2% in the law data, and 39,6% in fiction.

In relation to the question whether it would be fruitful to apply automatic translation to the selected texts, we have considered the workload potentially involved in correcting machine output, and in this respect the difference in restrictedness between the two text types is relevant. Within the non-computable correspondences, the frequency of cases exhibiting only one minimal semantic deviation between source and target string is considerably higher among the data extracted from the law texts than among those recorded from fiction. For this reason we tentatively regard the investigated pairs of law texts as representing a text type where tools for automatic translation may be helpful, if the effort required by post-editing is smaller than that of manual translation. This is possibly the case in one of the law text pairs, where 60,9% of the data involve computable translation tasks. In the other pair of law texts the corresponding figure is merely 38,8%, and the potential helpfulness of automatisation would be even more strongly determined by the edit cost. That text might be a task for computer-aided translation, rather than for MT. As regards the investigated fiction texts, it is our view that post-editing of automatically generated translations would be laborious and not cost effective, even in the case of one text pair showing a relatively low degree of translational complexity. Hence, we concur with the common view that the translation of fiction is not a task for MT.

Outline

List of abbreviations

DEF	definite
EEA	European Economic Area
ENPC	English-Norwegian Parallel Corpus
LFG	Lexical-Functional Grammar
LPT	linguistically predictable translation
LGP	language for general purposes
LSP	language for special purposes
MT	machine translation
NLG	natural language generation
NLU	natural language understanding
SL	source language
ST	source text
TL	target language
TM	translation memory
WSD	word sense disambiguation

Abbreviations referring to authors and texts given in the list of primary sources:

AB	André Brink
AEEA	*Agreement on the European Economic Area*
BV	Bjørg Vik
DL	Doris Lessing
EFH	Erik Fosnes Hansen
Petro	*Lov om petroleumsvirksomhet*

Abbreviations for syntactic categories are given in tables 4.3–4 in chapter 4.

PART I
INTRODUCTION

1 Overview and background

1.1 The study in a nutshell

How complex is the translational relation between two languages, and to what extent may we expect that translation between that pair of languages can be done automatically? These topics constitute one of our primary research questions, and the present study attempts to answer this with reference to the language pair English-Norwegian, and by investigating two specific text types. In order to study the translational relation between two languages, it is necessary to examine its manifestations, and we have thus chosen an empirical approach where we analyse selected extracts of parallel texts as these constitute parts of the extension of the translational relation. By 'parallel text' we understand an original text paired with its translation into another language, and we have investigated human-translated texts since we regard the product of the bilingually competent human translator as a "gold standard" for translation. The extent to which our study can answer the questions raised initially is of course limited to the scope of our empirical analysis. That is, our results apply only to that part of the translational relation between English and Norwegian which is covered by the selected parallel texts. Furthermore, it is not our ambition to find out to what extent it is possible to achieve automatic translation in general; that is an issue far too wide for us.

In this project the translational relation is treated as a theoretical primitive, not to be defined in terms of other concepts. As will be explained in 2.3.1, we distinguish between the translational relation between two language systems and the translational relation between textual tokens of those languages.

The present study applies a method where translationally corresponding text units are classified according to a measure of the complexity of the relation between source

and target expression. In our analysis the basic unit of translation is the finite clause. The complexity measure is based on assumptions concerning a translator's need for information when producing the given target text, and this need for information is analysed in terms of how much information is needed, what types of information this involves, and the effort required in order to access and process them. We assume a scale of translational complexity, and on this scale we have identified four main types of translational correspondence. When a pair of translational units is analysed, it is assigned one of these four types, as a classification of the complexity of the translational relation between the two units. The four correspondence types are organised in a hierarchy, reflecting an increase in translational complexity. Thunes (1998) presents a pilot investigation of these matters, and the method of analysis applied in that study has been adopted, with some modifications, for the project reported on here.

The classification of correspondences involves no evaluation of translational quality as, for instance, in terms of the model by House (1997). Among the empirical data there are occasional instances of unsuccessful translations, but translational quality is by itself no element in the classification of correspondences. Moreover, our notion of translational complexity, being based on information sources for translation, is in principle independent of grammatical complexity, and of factors that may influence the ease or difficulty with which the translator comprehends the source text.[1] Translational complexity is also distinct from the notion of linguistic complexity, which will be discussed in 3.2.3.

In the present study the question of automatisation is directly linked with the notion of computability. We assume that automatic translation between two languages may be achieved to the extent that it is possible to compute the translational relation between those languages. We will discuss this with reference to our categorisation of translational correspondences, and in the light of the empirical investigation we will tentatively draw a borderline for the possibility of automatisation, a line to be drawn on the complexity scale that we apply to the

[1] Grammatical complexity in relation to translation is discussed by Izquierdo and Borillo (2000).

translational relation. Although the results of our analysis are most directly relatable to rule-based machine translation, we assume that the general issue of computability addressed here likewise applies to statistical machine translation, which is also dependent on the accessibility of relevant and sufficient information in order to predict correct target expressions from available translational correspondences.[2]

Of importance to automatic translation is the issue of text type, and two different text types, narrative fiction and law text, are represented in the analysed text material. The motivation behind this is to investigate whether the degree of translational complexity differs between the two text types, and this is another primary research question. It is an established view that the possibilities for automatising translation are better with respect to texts dealing with restricted semantic domains than with unrestricted texts (cf. 1.4.2.3). The chosen fiction texts represent unrestricted text types, whereas the law texts instantiate restricted text types. We do not intend to decide whether the subject areas dealt with in the selected law texts are true examples of restricted semantic domains, nor to find out whether those laws can rightly be said to be written in sublanguages of English and Norwegian. Our aim will be to focus on the difference in restrictedness between the two text types, and to discuss its impact on translational complexity.

1.2 Information typology

The present study is neither a cognitive nor a psycho-linguistic investigation of translation, and we do not investigate the procedure of human translation. Our approach is to analyse the product of translation, since we assume that an empirical investigation of parallel texts, as instantiations of the translational relation, may serve as a basis for studying translation competence. Thus, our investigation concerns external, intersubjectively available objects: pairs of source and target texts (cf. 2.2.4).

One important topic in the present investigation is the information that is accessible through the competence of translators, and we assume that analysing a

[2] Cf. the presentation of non-linguistic approaches to machine translation in 1.4.2.5.

translation in relation to its original may reveal the types of information included in translators' competence, as well as other types of information accessed by a translator in order to produce a specific target text. Process-oriented translation studies (cf. 1.4.1.3) have tried to develop cognitive models of what is referred to as *translation competence*.[3] That topic will not be pursued, but for the purposes of our study we may sketch a simple and intuitive conception of translation competence as a combination of the following:

(i) Competence in the source language (SL) as well as in the target language (TL), and knowledge of how these two language systems are interrelated.
(ii) Necessary background knowledge of various kinds.
(iii) The ability to assign an interpretation to the SL text by merging the information encoded in the text itself with the information present in the textual context and in the utterance situation.
(iv) The ability to construct a translation which will receive an interpretation in the TL context and utterance situation which is as close as possible to the interpretation of the original, given its purpose.

The various kinds of information that are accessible through translation competence are part of the information needed to produce a specific translation from a given SL expression. The present work aims to describe a typology of information sources for translation, and in this respect, the following main types provide a starting point:

(a) Purely linguistic information, some of which is encoded in the SL expression, and some of which is inherent in a translator's bilingual competence and knowledge of interrelations between source and target language systems.
(b) Pragmatic information from the textual context and the utterance situation of the source expression.

[3] Hurtado Albir and Alves (2009: 63–68) present an overview of different translation competence models; cf. 2.4.2. We discuss the knowledge of translators in 2.4.1.5.

(c) Various kinds of extra-linguistic background information.

In addition to these categories we apply a distinction between general and task-specific information sources. The general sources include information about source and target language systems and their interrelations, as well as information about the world (cf. (i) and (ii) above). These information types are given, and hence easily accessible, in any case of translation. The task-specific sources cover information about a particular piece of source text and the concrete task of translating it into a given target language.

The typology of information sources for translation is presented in 2.4.2 with subsections. Since we describe translational complexity in terms of the amounts and types of information needed to produce a given target expression, the information typology is developed for the purpose of analysing the degree of translational complexity in correspondences between expressions of two languages. In relation to the various information sources for translation, we will in chapters 2 and 3 consider two questions that are decisive for the complexity of translational correspondences: to what extent can the different kinds of information be represented in a finite way, and what is the amount of effort required in order to access and process them?

1.3 The correspondence type hierarchy

As mentioned in 1.1, our scale of translational complexity is captured by a hierarchy of four main types of translational correspondence. The origins of this hierarchy is found in Helge Dyvik's work on an experimental machine translation system, documented in Dyvik (1990, 1995). The four correspondence types will here be briefly presented in order to illustrate how this hierarchy is linked with a translator's need for information when producing a specific target text. We will refer to instances of correspondence types as *(translational) correspondences* or, alternatively, as *string pairs*, i.e. translationally related pairs of word strings. Our notion of 'translational correspondence' is in accord with that of Johansson (2007: 23), who uses the term *correspondences* about "the set of forms in the source text which are found to correspond to particular words or constructions in the target text." Furthermore, we

will use the term *correspondent* to refer to either of the units that constitute a translational correspondence. Hence, this term is neutral between original and translation. Moreover, *correspondent* may refer to entire units of translation, as well as to subparts of them.[4]

1.3.1 Four types of translational correspondence

In this section we present and illustrate the four correspondence types with reference to the finite clause, since it is, as pointed out in 1.1, the basic unit of translation in this study.[5]

The least complex correspondence type is labelled *type 1* and comprises cases of word-by-word translations where source and target string are identical with respect to the sequence of word forms. Cf. string pair (1):

(1a) Hun har vært en skjønnhet. (BV)[6]
'She has been a beauty.'
(1b) She has been a beauty,

Type 2 correspondences are somewhat more complex, since source and target string are not matched word by word, but every lexical word in the source expression has a target correspondent of the same lexical category and with the same syntactic function as the source word. Otherwise, there may be differences between source and target string with respect to the sequence of constituents and/or the use of grammatical form words; cf. string pairs (2) and (3):

(2a) Dessuten virket hun overlegen. (BV)
'Also looked she haughty.'
(2b) She also looked haughty.

[4] The notion of 'translational correspondence' is further discussed in 4.3.1.
[5] Our units of analysis are defined in 4.3.2.
[6] *BV* refers to the author Bjørg Vik; see the list of primary sources. When examples of translational correspondences are given, the source text is always given under (a) and the target text under (b). Punctuation is reproduced as given in the primary text.

(3a) Leiligheten var ufattelig rotete. (BV)
'Flat.DEF was unbelievably untidy.'[7]
(3b) The flat was unbelievably untidy.

In (2) source and target string differ with respect to constituent sequence: (2a) has a fronted adverbial (*dessuten*), followed by the verb *virket*, and then by the subject *hun*, whereas in (2b) the subject *she* is in the initial position, followed by the adverbial *also*, and then by the verb *looked*.[8] In example (3) the English definite article *the* in the translation is not matched by any word form in the source sentence.

In *type 3* correspondences, translational complexity is still higher as they involve greater structural discrepancies between source and target than correspondences of type 2 do: there is at least one structural difference violating syntactic functional equivalence between the strings, but there is no mismatch between original and translation on the semantic level; cf. string pair (4):

(4a) Hildegun himlet lidende mot taket og svarte med uforskammet høflighet: (BV)
'Hildegun rolled-eyes suffering towards ceiling.DEF and answered with brazen politeness'
(4b) Hildegun rolled her eyes in suffering towards the ceiling and answered with brazen politeness.

There are two main reasons why string pair (4) cannot be assigned a type lower than 3. Firstly, the Norwegian intransitive verb phrase *himlet* corresponds with the English expression *rolled her eyes*, which consists of a transitive verb phrase and a noun phrase (NP) functioning as direct object. But these expressions correspond semantically: the Norwegian verb *himle* ('roll one's eyes') describes the activity of rolling the eyes of the agent, and since this information is inherent in the lexical meaning of *himle*, the existence of the referent of the English NP *her eyes* is implied by the Norwegian verb phrase.[9] Secondly, the adverb phrase *lidende* ('suffering') in (4a) is

[7] The label *DEF* will be used as a shorthand for the grammatical feature *definite*.
[8] (2a) illustrates subject-verb inversion in Norwegian. The example is also discussed in 3.3.3.1.
[9] The mismatch between the verb phrases *himlet* and *rolled her eyes* may be described as a conflational divergence; cf. 1.4.2.3.

of a different syntactic category than the preposition phrase *in suffering* in (4b), and the English preposition *in* is not matched by any lexical unit in (4a). But the two expressions *lidende* and *in suffering* correspond semantically: both phrases modify the action described by the verb phrases *himlet* and *rolled her eyes*, and the verbs *lide* and *suffer* are denotationally equivalent.[10]

Finally, in *type 4* correspondences complexity is even higher: in such cases there are discrepancies between original and translation not only on the structural level, but also on the semantic; cf. string pair (5):

(5a) Her kunne de snakke sammen uten å bli ropt inn for å gå i melke-butikken eller til bakeren. (BV)
'Here could they talk together without to be called in for to go in milk-shop.DEF or to baker.DEF'

(5b) They could talk here without being called in to go and buy milk or bread.

In (5) there is a semantic difference between the corresponding expressions *for å gå i melkebutikken eller til bakeren* ('to go to the milk shop or to the baker') and *to go and buy milk or bread.* The italicised expressions do not denote the same activities, although we may infer from background information about the world that both activities may have the same result, i.e. the purchase of milk or bread.

A central aspect of the correspondence type hierarchy is the increase in the degree of translational complexity from type 1 upwards. A parallel to this increase in complexity is found in Vinay and Darbelnet's (1995) set of seven translation procedures, which are presented "in increasing order of difficulty", ranging from the simplest method of translation to the most complex.[11] Although this is an interesting similarity, the present correspondence type hierarchy is not related to Vinay and Darbelnet's classification of methods. Our type hierarchy is designed for the purpose of analysing existing correspondences between source and target texts, and must not be associated with the notion of translation procedures.

[10] Denotational equivalence between expressions of different languages is discussed in 6.3.2.

[11] The quotation is taken from Venuti (2000: 92), where an overview of the seven procedures is presented. Pages 31–42 of Vinay and Darbelnet (1995) are reprinted in Venuti (2000: 84–93).

We have applied the method to one language pair only, English-Norwegian, but in principle it is a language-pair independent approach. However, occurrences of the lower correspondence types require a certain degree of structural relatedness within a given language pair: if SL and TL are structurally unrelated, the lowest types may not be found. On the other hand, in the case of languages that are very closely related, such as Norwegian, Danish, and Swedish, the most complex types may be rare.

The basic principles of the correspondence type hierarchy were originally described by Dyvik (1993), and the hierarchy is further developed in Thunes (1998), where the notion of translational complexity is discussed in relation to information sources needed in translation. Another contribution made by the latter is that subcategories of the main correspondence types 3 and 4 have been identified and explored. A further development of the correspondence type hierarchy is here discussed in chapter 3, where the information processing structure of individual translation tasks is related to each correspondence type. Chapter 4 provides a new discussion of criteria for the identification of analysis units, and for the assignment of correspondence type to string pairs.[12]

In our analysis we assume that a translator's need for information is greater in translational correspondences of the higher types than in those of the lower types. If we consider a human translator, this may not seem so obvious: a bilingual person will simply produce a target text without paying much attention to the amount of information he or she uses when doing so, perhaps with the exception of those cases where the translator really needs to think twice, and possibly check with reference works etc., to create a target text. The increase in a translator's need for information from correspondence type 1 to 4 is easier to grasp if we imagine giving the translation tasks to an automatic translation system, and the discussion will be related to the PONS system (Dyvik 1990, 1995) since its design is the main source of inspiration for the correspondence type hierarchy.

[12] Cf. 1.5 for more information on how the content of this thesis is organised.

1.3.2 The background for the correspondence type hierarchy

The PONS machine translation system is endowed with information about source and target language systems and their interrelations; this may be seen as a model of the translator's bilingual competence. The first step of the translation task is to analyse the input, a procedure which is comparable to a translator's reading and understanding of the source sentence. The analysis provides the system with information about the syntactic structure of the input text, which is then compared with information about source and target language interrelations. Through this comparison, the PONS system is able to choose between three different modes of translation, according to the complexity of the translation task. In practice, the system identifies cases where the syntactic structure of the source text is matched by the target language and exploits this match for the purpose of target text generation.

If the entire structure of the input text has a match in the TL grammar, the system will translate word by word, thus producing a type 1 correspondence. In such cases generation of the target sentence requires information about the word order and syntactic structure of the source sentence, and about the translationally corresponding TL word forms.

In other cases the PONS system may find that the source sentence structure is matched by the target grammar except for at least one difference with respect to constituent sequence and/or the presence of grammatical form words. The system may then be said to translate constituent by constituent, and will produce a type 2 correspondence. In such cases the generation of the target sentence requires information about the syntax of the input text, about the syntax of the structurally deviating parts of the target text, and about the translationally corresponding TL word forms. In this way translation requires a greater amount of information than in type 1 correspondences.

In cases where the PONS system finds that with respect to the function and/or category of at least one lexical word, the syntactic structure of the source sentence cannot be matched by the target language, the system will produce a full semantic analysis of the input, and use a semantic representation of the source sentence as the basis for target text generation. The result will be a type 3 correspondence, and

generation of the output sentence requires semantic information about the input text together with structural and lexical information about the target language. Cases of types 2 and 3 have in common that solving the translation task requires information about how the target text will deviate structurally from the source text. But since type 3 correspondences exhibit greater structural discrepancies between source and target than type 2 correspondences do, the translation task requires a more thorough linguistic analysis than in the case of type 2, and hence the need for information is greater.

With respect to type 4 correspondences, we assume that they are not included in the set of translations that could be computed by the PONS system, since they are cases where purely linguistic information is insufficient, and the translation task requires additional information sources, such as extra-linguistic background information and discourse information derived from a wider linguistic context.

In our study the distribution of the four correspondence types within a body of parallel texts is meant to serve as an estimate of its degree of translational complexity, and this estimate may be seen as an indication of to what extent automatic translation is feasible within the investigated texts. That is, the complexity measurement may indicate how far it is possible to simulate human translation for the specific language pair, text types, and translational choices as instantiated by the analysed parallel texts. We will later argue that the limit of automatisation is defined by the limit of linguistic predictability in the translational relation, and it follows from the organisation of the correspondence type hierarchy that the distinction between the linguistically predictable and the non-predictable is drawn between types 3 and 4.[13] It should be emphasised that in the present project the question of automatisation is discussed without reference to the architecture of any particular machine translation system, although the analytical framework is inspired by the PONS design.

[13] Linguistic predictability in the translational relation is defined in 2.3.2. For details on correspondence types 3 and 4, see chapter 3.

1.3.3 Related contributions

Hasselgård (1996) employs a slightly modified version of the correspondence type hierarchy as defined by Dyvik (1993). In Hasselgård (1996) the method is used for classifying correspondences between translationally aligned sentences in a small-scale investigation of word-order differences between English and Norwegian. Adapted versions of the correspondence type hierarchy as presented in Thunes (1998) are used by Tucunduva (2007), Silva (2008), and Azevedo (in progress), all of which are studies where the model is applied for the purpose of analysing and describing translational correspondences in parallel texts. These contributions are concerned with the language pair English-Portuguese, and they study various types of text.[14]

A related approach is provided by Merkel (1999), who combines translation studies, natural language processing, and corpus linguistics in a study where the main theme is correspondence relations in parallel corpora. His contribution includes a model for describing various kinds of structural and semantic correspondences between translationally aligned sentences in a Swedish-English parallel corpus. The aim of the analysis is to find out to what extent the translations exhibit changes in structure, function, and content in comparison to the originals, and this, in turn, is done to investigate differences between text types and translation methods.[15]

Another approach is found in Macken (2010), who presents research on automatic alignment of translational correspondences below sentence level, i.e. words, phrases and chunks. This is relevant to the present study since the data compiled in our investigation also include a large number of correspondences involving sub-sentential units.[16] In Macken's project different alignment tools have been tested against a manually aligned Dutch-English reference corpus. Her presentation of various categories of sub-sentential translational correspondences contains many similarities to the correspondence type hierarchy as described in Thunes (1998), in particular

[14] I am indebted to Marco Antonio Esteves da Rocha, of the Federal University of Santa Catarina, for information on the studies presented in Tucunduva (2007), Silva (2008), and Azevedo (in progress).
[15] Cf. chapters 10, 11, and 12 in Merkel (1999).
[16] Cf. the presentation of extraction criteria in 4.3.2.

regarding the kinds of linguistic properties that are shared (or not shared) by translationally matched units (cf. Macken 2010: 33–36).

1.4 Relevant fields of research

The present study draws on insights from several disciplines: general and computational linguistics, translation studies, and corpus linguistics, to mention some. 1.4 with subsections will present a selection of topics from a few relevant fields, i.e. translation studies, machine translation, and parallel corpus linguistics. Since a key issue in our investigation is the division between linguistically predictable and non-predictable translations, and since this is related to the limit of automatisation, the discussion will give more weight to machine translation than to the other disciplines.

1.4.1 Translation studies

The very notion of 'translation' has so far not been commented on. The present study is limited to written translation, and by 'translation' we will understand the act of transferring a text from one language into another. Koller (1992: 81, referring to Wienold 1980) points out that translation belongs to a group of several kinds of text reproduction, all kinds involving an original text and a new version of it. In addition to translation, examples of such activities are popularisation, the writing of abstracts, and creating children's versions of literary works. The latter activities have in common with translation that they may be performed across languages, but translation differs from them in (at least) one important way, as translation does not allow any of the differences between original and version typically found in the other kinds of text-reproduction. Still, it is not unproblematic to define 'translation' whether by delimiting the concept of translation or by specifying its set of necessary and sufficient properties.

There is, however, an intuitive concept of translation, one that has intersubjective validity. Halverson (2000) shows that 'translation' is a prototypical concept: firstly, the concept displays "graded membership" in the sense that certain types of

translation seem to be more central members of the category than others, and, secondly, the concept has "fuzzy boundaries" in the sense that there are gradual slides, and not discrete leaps, from 'translation' to related concepts. In agreement with the prototypical view of translation we regard the following characteristics as *central* to the concept of interlingual translation:

(i) Taking into account differences between source and target language systems, the translated version will as far as possible convey the same meaning as the source text.
(ii) The sender of a translated text is identical to the sender of its original.[17]
(iii) Taking into account cultural differences between the source and target language communities, the recipient group of the translation is as parallel as possible to that of the original in the source language community.
(iv) The communicative function of the target text is as parallel as possible to that of the source text.

In relation to this list of characteristics, at least two reservations can be mentioned. Firstly, it follows from a prototypical view of translation that not all of the properties (i)–(iv) must be present in everything that can qualify as 'translation'. Secondly, we do not imply that if these four properties are present in a translation, it will necessarily be a fully satisfactory version of the original.

Although the study of translation may be traced back to antiquity, it is only after the Second World War that the field has become a substantial area of research. During this time translation researchers have tried to form theories explaining translational phenomena, and they have constructed models of the relationship between originals and translations, as well as models of the translation process. Theoretical frameworks like those of general linguistics and contrastive language analysis have been applied in order to define translation models. The heterogeneity of the field is illustrated by the fact that it is difficult to find a single cover-term for all

[17] However, in the view of Koller (1979, 1992), where translation is described as a bilingual communication process, the translator is regarded as the sender of the target text; cf. 1.4.1.1.

its branches. *Translatology*, *translation theory*, *translation studies*, or the German *Übersetzungswissenschaft* — none of these expressions can serve as a fully neutral label in the sense that all translation scientists would accept it as a cover-term.[18]

As stated in 1.2, our approach is to analyse the product of translation. There are basic differences between studying, respectively, the product and the process of translation. We may directly observe the translation product as a text available to our perception, whereas the translation process is not as easily observable. Special elicitation techniques are required to examine the mental processes behind the production of the target language text. Hence, the distinction between product- and process-oriented approaches is important when describing the field of translation.

Chesterman (2005) provides a critical review of the terms and concepts that have been used over the years in various studies of the changes that may occur when a source text is translated into a target text (cf. 6.2.1). In this connection he discusses the opposition between product and process orientations, and he observes that many translation researchers are not entirely "clear about whether the focus is on processes themselves or the results of processes" (2005: 19). To illustrate his point he gives several examples from various contributions, and presents a possible explanation for the confusion: many of the terms used to describe translational changes often have a linguistic form that is "ambiguous between a process reading and a result reading" (2005: 20).[19] It would require a larger study of the field to support this position, but the main points argued by Chesterman (2005: 17–22) seem indisputable: lack of terminological stringency across the field works against conceptual clarity, and it is necessary to start by defining the concepts in order to improve the terminology of translation studies.

In our view, the difference between product and process orientations can be perceived as a continuum rather than as a dichotomy. In 1.4.1.1–3 we will present a selection of approaches illustrating this. At one extreme there are models describing the product of translation in a declarative way, thus focussing on the relation between

[18] For this piece of information the author is indebted to Dagmar Čejka. However, according to Baker (1993: 234), *translation studies* is the most common term, and we will mainly use this expression when referring to the field.

[19] Chesterman's examples of such terms are *compression*, *omission*, and *compensation* (2005: 20).

original and translation. If such descriptions are truly declarative, they specify sets of relations holding at the same time between certain entities, and they may be interpreted as declarations of static facts about the entities involved. At the other extreme there are procedural models describing the translation process. A procedural approach implies that the object of study is described in terms of a set of operations that will produce that object, and hence the description is of a dynamic kind. In positions between the declarative and the procedural there are models describing the product of translation partly by paying attention to the steps leading from source to target text, and there are models describing the translation process, but to some extent in terms of the relation between source and target text.

Sections 1.4.1.1–3 are not intended as a full overview of the various directions within translation studies, nor as a historical outline. Our aim is to present a few contributions chosen as representatives of certain positions within the field, and in chapter 2 we return to the division between product and process orientation. For surveys of different theoretical approaches in translation studies, as well as information on the historical development of this area of research, see e.g. Venuti (2000), Gentzler (2001), and Munday (2008, 2009). Kittel et al. (2004, 2007) provide a more detailed reference work on translation studies, and Baker (2010) presents a state-of-the-art view of the field. Moreover, chapter 4 in Munday (2008) gives an overview of product- and process-oriented approaches, respectively.

1.4.1.1 Product-oriented approaches to translation

Among the topics of interest to product-oriented studies of translation there are phenomena such as particular features of translated texts, and relations between source texts and their translations. In such studies it is relevant to probe the texts by means of different linguistic analyses, i.e. analyses concerned with domains like syntax, semantics, discourse, textual macrostructure, and stylistics.

Starting at the end of the continuum mentioned, where we find clearly product-oriented approaches, we may discuss Werner Koller's explication of the concept of 'translational equivalence'. His work is representative of the so-called "equivalence tradition", one of the linguistically oriented approaches within studies of translation.

According to Koller (1992: 81, 215; 1995: 196), 'translation' is defined by means of 'translational equivalence': we have a proper instance of translation when there exists an equivalence relation between an original in the source language and a translated version in a target language.[20] His definition of translation is provided with a description of a set of different frames of reference under which translational equivalence may hold (1992: 214–216; 1995: 196–197). In that manner he decomposes the relation into five different equivalence types: denotational, connotative, text-normative, pragmatic, and formal-aesthetic equivalence (1992: 216). Each such type specifies properties with respect to which the source and target texts should be equivalent. Denotational equivalence pertains to the extra-linguistic state of affairs described by the source text, whereas connotative equivalence deals with the connotations conveyed by the expressions used in original and translation respectively, especially through choice of words, level of style, the use of particular sociolects or dialects, and the like. Text-normative equivalence is determined by text type-specific norms of language use, and formal-aesthetic equivalence by the formal aspects of source and target text. Finally, pragmatic equivalence pertains to the communicative function of the texts, to the recipient of the translation, and to her/his capacity of understanding the translated message.

The concept of 'translational equivalence' has been much debated, and Koller's view of it is not the only one. In general, 'equivalence' is always equivalence with respect to a set of given properties and is not in itself a gradable concept. Hence, problematic aspects of the notion of 'translational equivalence' arise from the fact that cultural differences, and differences with respect to grammatical and lexical structure between source and target language, often makes it impossible to achieve translational equivalence with respect to all desirable properties. In practice, then, the translation task is to create a target version that is equivalent to the original with respect to as many as possible of relevant properties, and the selection of relevant properties will depend on the purpose and communicative function of the source text.

[20] Translational equivalence, in the sense used in translation studies, is not an *equivalence relation* in the terms of formal logic.

Or in the words of Juliane House: "… the translator has to set up a hierarchy of demands on equivalence that [he] wants to follow" (1997: 26).

Koller focuses on the result of the translation process in relation to its starting point, and his view is thus directed towards phenomena which are available to inter-subjective investigation. He has also addressed the translation process, but, as noted by Krings (1986: 9), Koller (1979: 112) regards its investigation to be a task for psycholinguistics. Elsewhere he has presented translation as a bilingual process of communication: first, the source text is communicated from the original sender to the translator in the role of recipient; second, the translator transfers the source text to the target language, and, third, the target text is communicated from the translator, as a secondary sender, to the final recipient (1979: 123–125; 1992: 106–107).[21] However, Koller does not present this as a model of the translation process, but as an account of aspects of the translation situation.

Another important contribution among the product-oriented approaches is the work of Gideon Toury (1995) on norms in translation. In relation to the task of studying the norms that govern translation, he states explicitly that the norms themselves are not available for observation; it is only the products of norm-governed translation behaviour that can be studied in order to detect the norms (1995: 65).[22] However, Toury's work is not as purely product-oriented as Koller's account of translational equivalence. Since norms control the work of translators, they exist during the translation process, and the study of norms aims at revealing how they influence the production of target texts. Toury (1995: 88) describes this study as "an attempt to gradually reconstruct both translation decisions and the constraints under which they were made." In Toury's approach there are several points of relevance for the present investigation, but due to the elements of process orientation, it will not be discussed further here.

---

[21] Bhatia (1997: 204) also takes the view of the translator as a secondary sender, at least implicitly, when stating that translation "is an attempt to communicate someone else's message through another language."

[22] In 2.2.1 we will discuss the principled difference between behaviour and the products of behaviour.

1.4.1.2 An intermediate position

Koller's view of translation as a communicative process may lead over to other approaches intermediate to the extreme positions of product and process orientation. An example of these is Juliane House's model of translation quality assessment, as laid down in House (1997). House may be said to belong to the functionalist tradition within translation studies, in which the communicative purpose of translation is a central notion. Her model is based on pragmatic theories, and on cross-cultural studies of the language pair German-English.

House assumes that translation quality assessment requires a theory of translation, and that different theories will yield different views of translation quality and of its evaluation (1997: 1). In her theory the equivalence concept is central, and she holds equivalence, as a relation between source and target text, to be the fundamental criterion for translation quality evaluation (1997: 25, 29). Her equivalence concept pertains to the preservation of meaning, and she views it as a functional and communicative notion. With respect to translational equivalence she distinguishes three aspects of meaning: semantic, pragmatic, and textual meaning (1997: 30–31).

Another central ingredient of House's theory of translation is her distinction between overt and covert translation (1997: 29, 66–70). In the case of overt translation the product is presented to the target language recipient as nothing but a translation, and the original links to the source language culture are preserved. A typical example is translated foreign language literature. In the case of covert translations, for instance translated user instructions, the target text appears as an original text, so that the function of the translation in the target language community corresponds to that of the original in the source language community. In order to achieve this, covert translations are subject to what House describes as "cultural filtering", i.e. a process in which the translator must "transmute the original such that the function it has in its original and situational environment is re-created in the target linguaculture" (1997: 163).

House's method for translation quality assessment (1997: 36–45) involves three steps: First, the source text is subject to a detailed linguistic and pragmatic analysis in order to detect its function, or "textual profile", and the source text profile will be the

norm for the assessment of quality in the translation. Second, the same kind of profile analysis is applied to the target text, and, third, the textual profile of the translation is compared to that of the original, in order to evaluate the degree of match. A high degree of match will be the mark of good translation quality.

House's concern with translation quality assessment is by nature product-oriented, as it is impossible to evaluate translation quality without analysing the result of the translation process. However, her work also reveals a concern with the translation process: to some extent the cultural filter gives an account of what goes on during translation, or at least of certain consequences of the process. Moreover, her distinction between overt and covert translations is tied to the issue of translation strategy, as the two types of translation represent different tasks: in overt translation the translator must make as few alterations as possible, whereas in covert translation the translator must erase, or adapt, all traces of the source language culture or community (1997: 164).[23]

1.4.1.3 Process-oriented approaches to translation

In a process-oriented study of translation focus is directed towards the translator's activity during translation. Since this activity primarily takes place in the translator's brain, it is not sufficient to analyse the translation situation and the result of the translation process in relation to the source text. In order to discover the inside workings of this instance of a black box it is necessary to use the methods of psychology.

However, translation research offers examples of theorists who have created models of the translation process even if they have not carried out psycholinguistic studies of it. One of them is Eugene A. Nida, whose contributions from the 1950ies onwards were of great value to the development of modern translation studies. Nida

[23] Deliberately, we have so far not defined the notion of 'translation strategy', as it is not part of our object of study, but occasional references to it are inevitable when we discuss translation and its product. We will merely apply an intuitive understanding of the concept, and use the expression *translation strategy*, or *translation method*, to refer to the set of actions chosen, either deliberately or not, by the translator during the creation of the target text. See Palumbo (2009: 131–133) for a discussion of the notion, including an overview of relevant references. Within the field of machine translation, a special meaning is attributed to 'translation strategy'; cf. 1.4.2.4.

was strongly interested in translation activity, and his research was based on wide experience with Bible translation. His works were also rooted in descriptive and theoretical linguistics, as well as in anthropology.[24]

In Nida's model the translation process consists of three main stages: analysis, transfer, and restructuring (Nida 1975: 80–95). The analysis stage identifies relations of meaning and reference, as well as the connotative values of the source text. Thus, analysis yields a disambiguated version of the source text, which can be transferred to the target language at a level "deeper" than that of surface structure. It is Nida's opinion that transfer takes place at a level where languages exhibit a greater degree of similarity than at the surface. The transfer stage he describes as a process of redistribution, operating on structures of semantic features representing the source text, and this process will most likely modify the source text meaning. The process of restructuring is to a large extent determined by the target language system, and it involves both formal and functional aspects, the latter requiring that the translation is made equivalent to the original with respect to communicative effect.

Although Nida's model is a procedural description, it captures linguistic effects of the translation process rather than the nature of the process itself. Nida was, however, aware of the psychological aspects of translating, but at the time the field of psychology did not offer adequate methods for probing the cognitive activities of a translator at work. Nida carried out this research while behaviourism still held a strong position, and according to the behaviourist paradigm the processes inside our brain could not be investigated through truly scientific methods, since they could not be observed directly (see Lörscher 1991: 67). The behaviourists had thus renounced the method of introspection, which had been applied during the late 19th and early 20th century as a tool for the investigation of mental activity.

After the exit of behaviourist views, there has been a revival of the use of introspection in psychological research. The methodology aims at externalising internal data, thus making them available to intersubjective investigation, and the

[24] See the "Introduction" to Nida (1975).

means to this end is verbal reporting.[25] In the 1980ies the elicitation technique named Think-Aloud Protocols (TAPs) came into use among researchers concerned with the mental processes involved in the act of translating. The use of Think-Aloud Protocols is based on a psychological model in which human cognition, including translation, is understood as information processing, and a cognitive process is "seen as a sequence of internal states successively transformed by a series of information-processing steps" (Lörscher 1991: 71).[26] Moreover, the model assumes that we are able to monitor our own cognitive processes, and hence the act of "thinking aloud" will provide access to the steps of information processing. In TAP studies the informant, in this case a translator, is typically asked to report, unselectively, everything that goes through her/his mind when performing the translation task, i.e., literally, to think aloud, while the reporting is audio- or video-taped. Other actions, such as note-making and consulting reference works, are also documented. TAP studies involve substantial criticism of previous models of the translation process. E.g., Krings (1986: 8) is of the opinion that those models do not describe what he deems to be the real facts of the translation process. Rather, he views them as attempts at analysing the translation process in terms of categories external to the process, such as the categories of linguistic analysis.

Within process-oriented translation studies, Krings (1986) is worthy of attention. Jääskeläinen (1999: 40) describes it as the "first extensive published TAP study", and according to Palumbo (2009: 92), it is generally seen as the "beginning of the process-oriented research tradition in translation studies." On the basis of his empirical data Krings makes certain generalisations on the global course of a translation task (1986: 178–187). He splits the process into three phases, pre-processing, main processing and post-processing.[27] Moreover, he finds it necessary to distinguish between translation *from* the translator's first language (L1→L2) and translation *into* her/his first language (L2→L1), the reason being that he finds more

[25] On verbal reporting see Ericsson and Simon (1984, 1993), or Krings (1986: 63–64). Lörscher (1991: 69–76) presents an overview of the development of introspective methods in modern research on cognition, and in particular on language learning and translation.

[26] For information on the TAP method, see also Toury (1995: 234–238), Jääskeläinen (1999, 2000), and Jakobsen (2003).

[27] In Krings' words: "Vorlauf", "Hauptlauf", "Nachlauf".

similarities between the informants' strategy choices in translation from L2 into L1 than in translation from L1 into L2.[28] Also, in L2-to-L1 translation there are two main types of translation problems, i.e. reception problems and production problems, whereas in L1-to-L2 translation production problems are dominant and reception problems nearly absent.

We may look briefly at the three phases in Krings' model. Pre-processing basically involves reading through the source text. Some of the informants omit this phase in L2-to-L1 translation. Otherwise during this phase, there is generally great variation with respect to the efforts put into identifying, and possibly solving, translational problems. During the main processing phase all subjects perform the bulk of the work required by the translation task. At this stage there is more variation with respect to strategy choices in L2-to-L1 translation than in L1-to-L2 translation. In the latter case all subjects translate sentence by sentence, in sequence. Finally, the post-processing phase, if not omitted, involves correcting and completing the target text, typically in the way of proof-reading.

It is interesting to compare Krings' model with earlier models of the translation process. The earlier models typically comprise either two or three different stages in the process. In general, two-phase models contain an analysis stage and a reconstruction stage, and three-phase models comprise analysis, transfer, and synthesis.[29] There is, however, no isomorphy between Krings' model and earlier three-phase models. Although it may not be evident from our brief presentation of Krings' work, it is a fact that in each of the three phases he has identified there may occur elements of analysis, transfer, as well as synthesis, depending on the translator's strategy. Moreover, Krings' study shows that some translators do not perform any pre- or post-processing. On the other hand, in the earlier models of translating the three stages of analysis, transfer, and synthesis are discrete, and none of them are dispensable.

Above all, TAP studies have shown that there is great variation among translators with respect to translation strategies. Another interesting finding is the distinction

[28] In the case of Krings (1986), the informants' L1 is German and their L2 is French.
[29] Cf. Wilss (1977: 95f, 1978: 15f), cited by Krings (1986: 6).

between processes performed automatically by the translator and processes requiring conscious decision-making (cf. Jääskeläinen and Tirkkonen-Condit 1991). Although the method of verbal reporting has clearly been helpful, and TAPs represented a breakthrough in translation studies, there are also shortcomings in these techniques. Hurtado Albir and Alves (2009), who provide a comprehensive overview of process-oriented research on translation, mention several weak points (2009: 69): The major problem is that TAP studies document the informants' subjective view of their own activity, and not necessarily the correct facts about it. Moreover, the method is intruding in that the subjects are aware of being observed, and perform verbalisation along with translation. Also, TAPs do not reveal unconscious or automatic processes. In more recent years the methodological trend has been to combine verbal reporting with other techniques (cf. Hurtado Albir and Alves 2009: 70–71). These may include traditional ones like interviews and questionnaires, and more modern ones, such as measuring brain activity, and logging the keystrokes and eye movements of translators at work. Hurtado Albir and Alves (2009: 72–73) conclude that the empirical methods of process-oriented translation research still need refinement. As methods improve, interesting discoveries about the cognitive aspects of translation are sure to be made.

1.4.2 Machine translation

We will understand *machine translation* (MT), or *automatic translation*, as the use of a computer program to translate text in one natural language into another. Thus, the notion of machine translation does not include computerised bilingual dictionaries, since they apply to the translation of single words, possibly including multi-word expressions. On the other hand, it does include systems able to translate spoken language (speech-to-speech translation), but the present discussion of MT will primarily be limited to the translation of written text.

Jurafsky and Martin (2009: 898) divides the field into classic and modern machine translation, an opposition reflecting the important distinction between *rule-based MT* and *statistical MT*. In the former approach the translation procedure relies on information about source and target language and their interrelations, whereas in

the latter approach translations are computed on the basis of statistical information about existing correspondences in large bodies of parallel texts.[30] As indicated in 1.1, the results of our product-oriented study are in principle also relatable to statistical MT, but the following presentation will focus on the classic, rule-based approaches, since our principal interest, in relation to automatic translation, lies in the question of how far it is possible to simulate human translation by processing linguistic sources of information.

Machine translation started as a research field; commercial applications gradually appeared, and MT has grown into a quite heterogeneous field with a great variety of applications. Several authors have presented overviews of the field, and their different contributions show that machine translation systems can be described and categorised in various ways, depending on which aspect of the field the description is focussed on.[31] Some of these aspects will be presented in 1.4.2.2–5, while the remainder of this section will discuss the division between experimental and commercial MT systems, which may answer questions like: who builds MT systems, and where are they used?

Experimental translation systems are typically developed within research institutions, and for the purpose of investigating pure research issues, such as the testing of formalisms for computational language descriptions. Although the development of an MT system normally requires a team of researchers working together, experimental systems may be the work of one or only a few researchers. Such systems may also be used for educational purposes, especially in university courses on computational linguistics. Normally, experimental MT systems are limited with respect to the coverage of the grammars and vocabularies of the languages they are applied to. The PONS system, discussed in 1.3.2, is an example of an experimental MT system; it may be described as a development environment where the user creates his or her own lexicons and grammars for source and target language, thus

[30] The dichotomy between rule-based and statistical MT is also mentioned in 1.4.2.1, and it is further discussed in 1.4.2.5.

[31] See for instance Hutchins (1986), Lehrberger and Bourbeau (1988), Hutchins and Somers (1992), Dorr et al. (1998), Nirenburg et al. (2003). Chapter 25 in Jurafsky and Martin (2009) provides a more recent introduction to machine translation. Other possible information sources for updates on the field are the journal *Machine Translation* and proceedings from the conference series Machine Translation Summit.

experiencing how the encoding of linguistic information will enable the system to translate.

Commercial systems are developed for the purpose of reducing the amount of work needed by professional, human translators. Typically, they are developed by teams where different specialists, such as computational linguists, programmers, lexicographers, and terminologists, work on different modules that together constitute a translation tool. The overall motivation behind the design of the system will be cost effectiveness: a net profit must be the outcome when the expenses of development, which can be substantial, are measured against the eventual benefits from saving translators' work hours, and possibly also from selling the tool to other users. Thus, with respect to system design, operational efficiency will be more important than matters such as the soundness of theoretical assumptions underlying language descriptions encoded in the system. A prerequisite for the usefulness of a commercial system is that grammar and lexicon modules cover the vocabulary and set of constructions found in the texts to which the system is applied, and this normally means that such information modules are large and expensive to build. It is also common that commercial systems are designed for text types special to restricted, technical domains, since technical texts tend to exhibit a controlled vocabulary and limited inventory of sentence types, which means that such MT systems will not necessarily need broad-coverage grammars and lexicons. Typically, commercial MT systems have been developed by, or for, large multinational enterprises, of which IBM is a well-known example, and for the purpose of translating technical documentation. Some commercial systems have been available for decades, with new and improved versions appearing now and then.

It may seem as if experimental and commercial MT systems have belonged to separate camps with no mutual interests, but that is not true. There are many examples of system developers with experience from research institutions who have joined in the construction of commercial systems, and issues like efficiency, cost effectiveness, and broad coverage are clearly not uninteresting to developers working in the research sector, although they may not be the dominating research aims. Moreover, the German Verbmobil project (Wahlster 2000) is an example of coopera-

tion between research and science: in a large and prestigious project academic and commercial interests joined forces to develop a system for the translation of spontaneous speech.

In addition to experimental and commercial systems, in recent years certain MT applications have become available to everyone with access to the Internet. These tools are incorporated in search engines, so that if an information request identifies a document in a foreign language, the system can offer an automatic translation of that document. This will typically be a translation of low quality, but it may be sufficient for the user to decide whether it is worthwhile making further efforts to access the information contained in that document.

1.4.2.1 A brief historical overview

The earliest attempts at constructing mechanical systems for automatic translation were made in the first half of the 20th century (Hutchins 1986: 22), but with no success. After the Second World War the advent of modern computer technology paved the road for new attempts, and in the 1950ies machine translation was among "the first non-numerical applications of computers" (Hutchins 1986: 16). In the early years the major sources of motivation and funding behind MT development was found among military and intelligence authorities, notably in the United States and the Soviet Union. It was the era of the Cold War, and in many nations intelligence agencies were busy collecting information about enemy countries, so that there was a great demand for translating text produced in the languages of those states. During the war, computers had been used for coding and decoding military messages, and it is not surprising that in this context automatic translation was seen as a promising tool. MT activities were not only initiated in the US and Soviet Union, but also in Japan and certain Western European countries, as well as in Canada from the late 1960ies.

Early work on machine translation was strongly inspired by information theory, in the US especially by the work of the information theorist Warren Weaver, who argued that translation basically involved decoding the source language text into target language symbols (Weaver 1949). At the time, similar conceptions of

translation were also harboured by several translation researchers: to generalise, translation was seen as decoding the source text message and recoding it in the target language.[32] In the first generation of MT systems the encoding of linguistic information was based on shallow language descriptions. Roughly, the first systems could be seen as implementations of bilingual dictionaries with certain reordering rules for accommodating structural differences between SL and TL. The lack of linguistic sophistication in the early systems is understandable: theoretical linguistics did not yet offer linguistic models suitable for computational implementation, and the capacity of available computer technology put narrow limits on the amount of language information that could be encoded, and on how it could be done. Still, there were great expectations with respect to what would be achieved.

In the 1960ies the optimism vanished since there were still no really successful results of machine translation development. Even if computer technology was continually improving, there had been no substantial breakthrough, and MT researchers came to realise that certain fundamental problems related to linguistic issues had to be solved before better MT systems could be built. It became a widespread view that since natural languages are in so many ways ambiguous, it would be an unreasonable goal to achieve fully automatic, high quality translation of unrestricted text. As early as in 1960 the influential researcher Yehoshua Bar-Hillel explained why: “A human translator, in order to arrive at his high quality output, is often obliged to make intelligent use of extra-linguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge.”[33]

Then, in 1966 the famous ALPAC report appeared. It was presented by an evaluation committee appointed by the US state agencies that were the main sponsors of MT activities. The report brought MT into disrepute, and efficiently drained away research funding in the United States as it concluded that the field had so far been a

[32] Cf. the discussion in Koller (1992: 89–92) of early models of translation.
[33] The quotation is taken from Nirenburg et al. (2003: 62), where Bar-Hillel (1960) is reprinted.

failure, that there remained too many unsolved fundamental problems, and that human translation would anyway be more cost effective than developing automatic translation. After the ALPAC report US research environments turned their focus to artificial intelligence and fundamental issues in computational linguistics. In other countries the change was not so acute; work on MT development continued although it was not carried out on such a large scale as had been the case in the United States.

Even if perfection was not achieved, workable MT systems did appear on the market during the 1960ies, and the fact that they were actually used shows that there clearly was a need for MT as a supplement to human translation, even if it involved a considerable amount of revision by translators. An important market was the translation of technical documentation in industry.

In the late 1970ies the pessimism that spread during the sixties was slowly giving way to renewed, but careful, optimism. In 1977 the Canadian MT system METEO® was completed for the purpose of translating weather forecasts between English and French. The system was a success and in operation for about two decades. This achievement strengthened the view that machine translation was suited for texts with a controlled vocabulary and a limited set of possible syntactic constructions. Moreover, it fuelled new interest in MT development, and during the 1980ies research activities were increasing in a range of countries. Achievements made since the 1960ies in several fields of science now offered far better conditions for creating automatic translation. Computer hardware had improved greatly; new programming techniques had been developed, and formalisms more suitable to computationally implementable language descriptions had been developed within linguistics.

Thus, by the beginning of the 1990ies a range of different MT projects had appeared in many countries across the world, and, in comparison to early machine translation, systems were now of a quite different quality with respect to computational efficiency as well as sophistication in the treatment of linguistic phenomena. Also, research efforts were not any longer limited to languages with large numbers of speakers (like English, Russian, French, Japanese, etc.), but MT development was also carried out for small languages, such as those in Scandinavia. Moreover, multilinguality had become an important design issue: multilingual MT systems are

not limited to one language pair, but are constructed for translating between several languages, and should easily facilitate the inclusion of new language pairs. Hence, modularity was an important design issue, so that linguistic information was to a larger extent than before kept separate from the actual translation procedure in the systems. This was another way in which MT had come to differ from the earliest systems, where translation procedures generally were strongly dependent on the linguistic differences between specific pairs of languages.

In 1993 Sergei Nirenburg pointed out that machine translation had "recaptured its place as the single most important application of computational linguistics and natural language processing" (1993: v). Since then research funds have come from national governments as well as from commercial interests, and MT has retained an important, although today not dominating, position within the larger field of language technology. Here MT has had to compete over research grants with other activities like voice recognition, speech synthesis, word sense disambiguation, and the building of language resources.

Statistical approaches to machine translation emerged in the early 1990ies. While commercial systems were still rule-based, MT conferences during that decade became dominated by the discussion of statistical methods and the evaluation of their performance. Gradually, research efforts were directed mainly towards statistical MT, as it appeared to be highly promising. However, after 2000 there has been a growing awareness in the field that further improvement of performance requires that the statistical methods are augmented with some processing of linguistic information, an approach often described as *hybrid* (cf. Dorr et al. 1998: 35).

1.4.2.2 Degree of automation

One important aspect of rule-based machine translation systems has been degree of automation. Some MT systems have been fully automatic, whereas others have required interaction with a human user. E.g., Hutchins (1986: 19), and Sager (1994: 290) classify systems according to a scale ranging from fully automatic translation to human translation with no machine aids. In fully automatic translation (or *batch* systems) the user only needs to enter the source text and wait for the system to output

a translation. In interactive systems some kind of intervention is required from the human user during the translation process. This could amount to resolving linguistic ambiguities in the source text, or entering target words for certain SL words whose translations are unknown to the system, or also selecting the most appropriate target text when the system produces alternative translations. The operation of such interactive systems can be described as *human-aided* machine translation.

Other important kinds of human intervention in translation tools are known as *pre-* and *post-editing*, respectively. Pre-editing involves preparing the input so that the MT system is able to compute a translation given the linguistic information encoded in the system. The pre-editor must remove from the source text syntactic structures and lexical items which are not covered by the language descriptions of the system. Pre-editing may also involve inserting syntactic labels in the source text so that the system will be able to resolve linguistic ambiguities.

Post-editing of the output of an MT system means that a human who is competent in both SL and TL revises the target text according to demands on translation quality. This is really the same task as revising a draft version of a "manual" translation, but, as noted by King (1986: 6), there is great variation between human and machine translation with respect to the amount of post-editing needed and the types of errors made. When a considerable amount of post-editing is required, the phenomenon at hand may be described as *machine-aided* translation rather than as MT proper. Post-editing is still a current topic in machine translation, and the amount of necessary post-editing of the output has always been an important criterion in the evaluation of the performance of MT systems.

In relation to the degree of automation, there is perhaps one kind of tool used in machine-aided translation that is particularly relevant, i.e. the *translation memory* (TM).[34] This is defined by Palumbo (2009: 127–128) as "[a]n electronic database containing translated texts stored together with their originals," and the texts "are normally segmented into units one sentence long." Clearly, as Merkel (1999: 43) has observed, translation memory tools are particularly useful for maintaining consisten-

[34] Chapter 8 in Macken (2010) provides a survey of translation memory systems, and reports on an evaluation of the performance of two available TM systems.

cy in the translation of types of text with repetitive language, such as technical texts. The latter point is relevant to the dimension of text type, which will be introduced in chapter 4.

1.4.2.3 Challenges for automatic translation

As the history of machine translation shows, automatic translation is a greater challenge than merely decoding the source text and recoding it in target language symbols. Dorr et al. (1998: 4–12) have presented the challenges involved in MT building along two different dimensions, described as operational and linguistic considerations, respectively. Our primary focus will be on the latter kind, and the discussion in this section relates mainly to rule-based MT.

Among the *operational considerations* of machine translation, Dorr et al. (1998: 10) include "extension of the MT system to handle new domains and languages; handling a wide range of text styles; maintenance of a system once it has been developed; integration with other user software; and evaluation metrics for testing the effectiveness of the system." Operational considerations in MT building are of greater relevance to implementation issues than to the linguistic aspects of automatic translation. Hence, we will give more attention to the latter topic than to the former, since our interest lies with the question of automatisation independently of the architecture of any particular MT system.

However, among the operational issues there is some relevance to the present project in the topic of extending a system to new domains and languages. That is, the challenge can be said to be not only to extend, but to build, altogether, those information modules that will serve as lexicons and grammars for source and target languages in an MT system. Without such information sources the system cannot translate.[35] Another prerequisite for successful translation is that those information modules cover the lexical inventory and set of linguistic structures found in the input texts at hand. Realistic requirements in operative MT systems are lexicons with tens of thousands of entries, and grammars with hundreds of rules. MT system builders must

[35] Cf. our discussion of information sources for translation in 1.2, and in 2.4 with subsections.

collect this information from somewhere, and another prerequisite is a grammar formalism for the representation of lexical entries and grammar rules. Normally, creating such linguistic information modules involves a lot of manual work since it is impossible to convert traditional dictionaries and grammars into computational ones without major adaptations.

The following quotation indicates what a great challenge it is to build linguistic information modules for an MT system: "Providing the linguistic knowledge for an entire language is truly a staggering task. In fact, no single human language has yet been fully described in a form usable by computers" (Grishman and Kittredge 1986: ix). Now, about 25 years later, this is still true. One possible way of meeting the challenge is to tune an MT system for texts from a restricted semantic *domain*, and by this we normally understand a certain technical field, such as a specific trade, a branch of industry, a field of science, etc. The group of speakers associated with a restricted domain typically share some domain-specific knowledge which is not part of the common knowledge of the speakers of the entire language community.[36] Furthermore, in such domains only subsets of alternative meanings of certain ambiguous words will be probable, and texts dealing with restricted domains will normally share certain linguistic characteristics. More specifically, discourse related to a restricted semantic domain typically employs a limited set of preferred linguistic constructions, and a set of technical *terms*, whose meanings are unambiguous.

Such discourse can be tied to the concept of a *sublanguage*, a notion which was originally given a mathematical definition by Zellig Harris (1968).[37] Here, leaving the mathematical properties aside, we will emphasise the fact that a sublanguage is a well-defined subset of a given language. The meanings of its expressions are a subset of the meanings expressed by the general language, and it is regarded as a more manageable task to describe the grammar and lexicon of the sublanguage than of the general language. Thus, if an MT system is designed for a restricted semantic domain, it is not necessary to build lexicons and grammars for entire languages, as it is sufficient to cover the given sublanguages of SL and TL. It may be necessary to

[36] Cf. Kittredge (1987: 59).
[37] For information on this, see Kittredge and Lehrberger (1982: 1), and Kittredge (1987: 59–60).

describe constructions and lexical items which do not belong to the general languages, since they belong only to the source and target sublanguages, but that will be a limited task. The effect of tuning an MT system to a specific domain and sublanguage is to avoid many of the problems involved in achieving automatic translation of general text, problems we will mention in connection with linguistic challenges for MT. The disadvantage is that extending the system to other domains demands that new sublanguage lexicons and grammars must be created.

Linguistic challenges for machine translation are referred to by Dorr et al. (1998: 4–10) as *linguistic considerations* in MT development, and like Dorr et al. (1998: 4) we will divide them into problems related to source text analysis, to target text generation, and to the mapping between source and target language. Our main focus will be on types of analysis problems because identifying the correct interpretation of the input is crucially important to successful machine translation.

Analysis problems in automatic translation are, above all, caused by ambiguity in natural language expressions, i.e. the fact that more than one possible interpretation may be associated with a word, phrase, or sentence. One possible way of sorting the types of ambiguity that cause analysis problems is to divide them into lexical, structural, and referential ambiguity (cf. Thunes 1994: 4–6). In general language use ambiguity phenomenona are extremely frequent, whereas in sublanguage texts their incidence is lower, as indicated above. Ambiguity phenomena indeed highlight the difference between the human translator's ability to interpret a source text and the way in which an MT system is able to understand input text. The types of ambiguity that cause trouble in automatic translation are normally resolved effortlessly by humans, because we continuously make use of contextual and extra-linguistic information when reading a text. Thus, if a word, phrase, or sentence has more than one possible interpretation, we filter out all improbable alternatives to the intended interpretation by means of information surrounding the ambiguous expression. An MT system, on the other hand, normally works sentence by sentence and must rely on the information that is linguistically coded in the given input sentence, and the analysing system will try all possible readings of ambiguous expressions, and their combinations. This may yield a large number of possible interpretations, and in MT

systems it is difficult to simulate a significant amount of the kind of inferences used by the human translator, mostly subconsciously, when improbable interpretations are filtered out.

Lexical ambiguity covers phenomena like homonymy, homography, and polysemy. Here we shall not go into great detail, but mention a classic example of homonymy: the English noun *bank* has at least two meanings: 'river bank' and 'financial institution', respectively. Given a sentence like *They camped by the bank of the river*, a human reader with general world knowledge would never consider the second meaning of *bank*, but for an MT system it is a challenge to identify the intended meaning of the ambiguous noun *bank* in order to choose a correct target language equivalent. This is a problem especially since it is extremely rare that the translations of homonymous source words are homonyms, too. A possible way of handling this is to encode, in the lexical information associated with *bank*, the semantic conditions governing the proper use of the different meanings, and to do so in a principled way is a challenge for the designer of the lexicon of the MT system.

Lexical ambiguity frequently involves cases where a lexical item is ambiguous with respect to syntactic category, such as the English word form *increase*, which can be either a verb or a noun, thus constituting a pair of homographs. In automatic translation, such categorial ambiguity can be resolved by parsing the local syntactic context: e.g., if an article like *an*, or *the*, immediately precedes the word form *increase*, then the analysing system will be able to choose the noun reading. Lexical ambiguity is a kind of analysis problem that researchers have tried to amend by integrating automatic word sense disambiguation (WSD) in MT systems. In simplified terms, WSD methods work by estimating the probability of a given sense in relation to other words occurring in the context of the ambiguous word, thus exploiting the fact that different senses of a word tend to be used in different types of contexts.[38] However, Ide and Wilks (2006: 54) observe that WSD tools do not seem to improve the performance of MT systems substantially. One reason may be that although quite successful WSD tools have been developed, an even higher degree of

[38] For an introduction to WSD, see chapter 20 in Jurafsky and Martin (2009).

accuracy is required, since disambiguation errors during analysis can have quite damaging effects (cf. Ide and Wilks 2006: 65). Another reason may be that in systems where categorial ambiguities are anyway resolved by syntactic parsing of the input the usefulness of separate WSD modules is probably limited (cf. Ide and Wilks 2006: 55–56).

Structural ambiguity can be described as the phenomenon where an expression has more than one possible interpretation because the expression can be partitioned into phrases in more than one way. A standard example for illustrating structural ambiguity is (6):

(6) I saw the man with the binoculars.

(6) can be interpreted as the statement that the referent of *I* either saw a man by means of a pair of binoculars, or saw a man who was carrying binoculars. Choosing the intended interpretation requires extra information from the context in which the expression is uttered. For a human recipient it is trivial to access and use such information; for the analysis procedure in an MT system it is not, especially if the system works sentence by sentence and is unable to retain information from the linguistic context preceding each input sentence.

Such structural ambiguity is not necessarily a translational problem: if the target language is ambiguous in the same way, then the ambiguity must not be resolved before translating. (7) is a Norwegian translation of (6), and the possible syntactic analyses and interpretations of (7) are an exact parallel to those of (6):

(7) Jeg så mannen med kikkerten.
'I saw man.DEF with the binoculars.'

There is a fair degree of structural relatedness between English and Norwegian, which in this case helps the translation task. If the target language is Japanese, which is a structurally unrelated language, it is necessary to resolve the source sentence ambiguity because the two interpretations require different translations. The first reading of (6), 'I saw the man by means of the binoculars', can be translated as (8):

(8) Watasi wa booenkyoo de otoko o mita.
'I-TOPIC binoculars-INSTRUMENT man-OBJECT saw.'

In (8) the particle *de* marks the noun *booenkyoo* as an instrument in the described situation. The second reading of (6), 'I saw the man who was carrying the binoculars', can be translated as (9):

(9) Watasi wa booenkyoo o motte iru otoko o mita.
'I-TOPIC binoculars-OBJECT carrying was man-OBJECT saw'

In (9) the particle *o* marks the noun *booenkyoo* as an object of the verbal phrase *motte iru* ('was carrying').

Referential ambiguity occurs in cases where it is possible to assign more than one referent to an anaphoric pronoun. Example (10) may illustrate this:

(10) There is a ship on the harbour, and it is crowded with tourists.

In (10) there are two possible antecedents for the pronoun *it*: *a ship* and *the harbour*. Again, translation may require that the intended interpretation is found if the two different alternatives must be translated in different ways. That would be the case when translating (10) into Norwegian, as in (11) or (12), where the use of italics indicates the possible binding relations between antecedent noun phrase and anaphoric pronoun:

(11) Det ligger *et skip* på havnen, og *det* er fullt av turister.
'It lies a ship on harbour.DEF, and it (i.e. the ship) is full of tourists.'

(12) Det ligger et skip på *havnen*, og *den* er full av turister.
'It lies a ship on harbour.DEF, and it (i.e. the harbour) is full of tourists.'

In (11) the neuter gender of the noun *skip* ('ship') requires the neuter form of the anaphor *det*, while in (12) the masculine form of the anaphor *den* agrees with the masculine gender of the noun *havn* ('harbour').

The examples used to illustrate structural and referential ambiguity show that when these phenomena occur, the amount of information that is encoded in the linguistic expression itself is insufficient in order to choose one interpretation rather than another. For automatic translation it is a true challenge that information from a wider linguistic context, or even from background world knowledge, is necessary to resolve the ambiguities.[39]

Having discussed analysis problems for MT, we will look at *generation problems*, and concentrate on two main categories: first, problems created by lack of isomorphy between lexical distinctions in source and target language, and, second, problems arising when the target language obligatorily expresses grammatical distinctions absent in the source language. These are not the only kinds of problems for generation in MT, but the ones we would like to focus on.[40]

Dorr et al. (1998: 7) refers to the first type as the *lexical selection problem* in target text generation. It is a well-known fact that different languages carve up reality in different ways, and this has the consequence that lexical items in one language only rarely correspond one-to-one with lexical items in other languages.[41] Thus, the challenge for machine translation is that finding the correct target language equivalent for a given source word frequently involves making a choice within a set of possible candidates. E.g. the English verb *know* corresponds translationally with various Norwegian verbs, depending on the linguistic context. Appropriate translations of *know* in the sense used in *Do you know French?* are the verbs *kunne* and *beherske*, whereas in the case of *Do you know what time it is? know* corresponds with the Norwegian verb *vite*. Hence, we may say that *know* is translationally ambiguous. The semantic conditions governing these translational choices are fairly subtle and nontrivial to represent in a format usable in an MT system, and extra-linguistic information about the world may be needed to identify the appropriate target word in a given context. General language words, such as *know*, are normally polysemous, or semantically vague, and hence may cover various senses and have

[39] Cf. comments on *the resolution problem* in 2.4.2.2.
[40] In 3.3.2.2 the second type of generation problem is illustrated by morphological differences between English and Norwegian present tense verbs.
[41] Cf. the discussion in 6.3.2 of denotational equivalence between lexemes of different languages.

several possible translations.[42] Clearly, it is easier to manage the generation task if the input text is written in a sublanguage with a high frequency of technical terms. Typically, technical terms correspond one-to-one with terms in the target language, since it is a characteristic property of terms that they have been designed to be unambiguous.

The second type of generation problems is caused by a fact once formulated by Roman Jakobson: "Languages differ essentially in what they *must* convey and not in what they *may* convey" (1959: 236). Several grammatical categories, of which tense, number, and gender are typical examples but do not constitute an exhaustive list, are obligatorily expressed in certain languages while being absent in other languages. That is, the semantic distinctions expressed by these grammatical categories may be drawn in the other languages, too, but then by other means than grammatical markers. For instance, in English finite verb forms express either past or present tense,[43] while in certain East- and South-East Asian languages, e.g. Vietnamese, there is no tense-marking verbal morphology. When translating from Vietnamese into English, it is a problem to pick appropriate tense markers on finite verbs in the target text if the source text contains no explicitly expressed information to settle the choice. In practice, there will be contextual cues which a human translator will be able to interpret easily, but in automatic translation such information is normally not accessible. In such cases the challenge for MT lies in the fact that the amount of information that is linguistically expressed in the source sentence is insufficient for the generation of the target sentence.

Finally among linguistic considerations in MT development we want to mention *mapping problems*, i.e. problems related to the mapping between source and target language. This is a topic area where many researchers from, roughly, the 1980ies onwards, have tried out a multitude of sophisticated approaches for describing

---

[42] Insofar as automatic translation relies on successful word sense disambiguation, it is a harder problem to keep polysemous senses apart than to distinguish homographs with semantically unrelated meanings and which may even occur in separate domains. The reason is that there is a greater degree of overlap between the types of contexts that senses related through polysemy occur in than between those of homographs. Cf. Ide and Wilks (2006) on a discussion of what level of sense distinctions it is fruitful to aim at in natural language processing.

[43] Exceptions are imperative and subjunctive verb forms, which are marked with respect to the category of mood.

various kinds of linguistic phenomena that occur in the cross-linguistic setting. Interesting work has been done especially with reference to phenomena involving differences in predicate-argument structure between source and target text. Dorr et al. (1998: 8–9) discuss five different classes of such phenomena, among which we want to illustrate two types.

First, in the case of "thematic divergence" a verbal argument realised in one language as a syntactic subject corresponds translationally with an argument realised as a syntactic object in another language.[44] A simple illustration of the phenomenon is the English sentence *Writing pleases me* translated into Norwegian as *Jeg liker skriving* ('I like writing').

Second, there is the phenomenon referred to as "head-switching divergence", where lexical material realised as a main verb (i.e. a syntactic head) in one language corresponds translationally with lexical material realised as a subordinated verb in another language. A much used example of this is the correspondence between the German sentence *Peter schwimmt gern* ('Peter swims with-pleasure') and the English sentence *Peter likes to swim*.

In addition, Dorr et al. (1998: 9) mention structural, categorial, and conflational divergence as types of mapping problems. Structural divergence means that an argument has different syntactic realisations in source and target text, respectively. Categorial divergence covers cases where a given source word corresponds translationally with a target word of a different syntactic category, and in the case of conflational divergence a pair of translationally corresponding verbs differ with respect to the number of arguments that must be overtly expressed.[45]

The various kinds of mapping problems are easily solved by the human translator provided that he or she has sufficient knowledge about the relationship between source and target language. For the MT system developer the challenge is to identify and describe the divergence phenomena, and encode such descriptions in the linguistic components of the translation system. This can be implemented in a separate

---

[44] This has often been referred to as *argument switching*, which concerns divergences in the mapping of semantic arguments onto syntactic functions.

[45] The translational correspondence between the verb phrases *himlet* and *rolled her eyes* in example (4) in 1.3.1 is an example of conflational divergence.

component, a transfer module, which contains information about mapping relations between SL and TL.[46] In some cases source-target divergences of the types mentioned are associated with individual predicate-argument structures expressed by specific lexical items. If a certain type of divergence phenomenon pertains to several lexical items, then it is desirable to find a uniform description of the whole class of instances as this contributes to economy in the information modules. Moreover, an important question is whether the specific mapping relations apply whenever certain predicates are expressed in the source text. With respect to the English verb *please* (cf. above), it is not necessarily translated into the Norwegian verb *like*. The predicate expressed by *please* corresponds semantically with the predicate expressed by the Norwegian verb *behage*, and in that case there will be no head switching divergence as *please* and *behage* have isomorphic predicate-argument structures. *Behage* is, however, somewhat more archaic than the Norwegian verb *like*, and would not be an appropriate translation in any context. Then the problem for automatic translation is how to identify, in the source text, the conditions governing the choice between different possible mappings between SL and TL. To handle such challenges MT systems need to make correct choices between rather fine-grained sense distinctions. Citing Edmonds and Hirst (2002), Ide and Wilks (2006: 65) indicate that this can be achieved by integrating "additional knowledge and/or reasoning", which they regard as a task for computational lexicography and artificial intelligence, and not for word sense disambiguation.

From the perspective of theoretical linguistics, it is in itself an appealing task to account for such divergence phenomena through adequate grammatical descriptions, but in the context of machine translation, system developers will have to consider whether such efforts of grammar development are worthwhile. They are probably not if a given system is designed for a text type where the mapping problems are infrequent.

[46] Cf. the presentation of MT systems architectures in 1.4.2.4.

1.4.2.4 MT system architectures

In the presentation of machine translation we have several times referred to procedures and information modules, understood as components of MT systems. This section will briefly look at different types of MT system architectures, and we shall see that differences with respect to translation strategy are reflected by different ways of structuring the linguistic information encoded in an MT system. In this context the notion of 'translation strategy' covers the set of principles underlying the design of the translation procedure in an MT system, and it is commonly used for the purpose of classifying systems. There is a basic division between systems using *direct* strategies, and those using *indirect* strategies, and within the latter group a further distinction is made.

In direct MT systems translation is basically done by mapping the words in the input text directly onto words in the target language. The earliest systems, so-called *first generation systems*, used direct strategies, and, as already pointed out in 1.4.2.1, those systems could be seen as implementations of bilingual dictionaries with certain reordering rules for accommodating structural differences between SL and TL. Hence, in direct systems the encoding of linguistic information, as well as the implementation of translation procedures, were strongly dependent on the specific language pair, and the direction of translation, that each system was designed for. It has frequently been said that in direct systems the source text was analysed in terms of the target language, so that the target text could be generated directly from the result of the analysis.

In indirect MT systems translation is done by means of some sort of intermediate representation produced by a linguistic analysis of the input text. Such systems appeared as a response to the apparent failure of the direct technique, and are by some referred to as *second generation systems* (cf. Hutchins and Somers 1992: 71–72). Within indirect MT systems a distinction gradually evolved between the *transfer* strategy on the one hand and the *interlingua* strategy on the other.[47]

[47] Traditionally, the perhaps most common approach in MT system typologies has been the tripartite division into direct, transfer, and interlingua systems; cf. Hutchins and Somers (1992: 71–76), Dorr et al. (1998: 12–18).

Transfer-based MT systems are characterised by three separate stages in the translation process: analysis, transfer, and generation.[48] The first stage is a linguistic analysis of the input: by means of a grammar and lexicon describing the source language the system produces a representation of the meaning and structure of the source sentence. During the transfer stage this representation is changed so that it can eventually serve as the basis for target text generation. Necessary changes involve finding TL equivalents of the lexical items in the source text and transforming the input structure wherever it does not conform with the structural requirements of the TL grammar. Then, during the generation stage the information contained in the transformed representation of the input is used, together with information contained in the target language descriptions, to produce TL word forms and to arrange them according to correct TL word order.

The basic difference between interlingua systems and the transfer-based ones is that the transfer stage is dispensed with in interlingua systems. This can be done because the analysis stage "translates" the input text into an *interlingua expression* from which the target text may be generated. In the context of machine translation, an *interlingua* is a level of representation, in principle of a language-neutral kind, and in practice at least neutral between source and target language. The basic idea is that through linguistic analysis the information contained in the source text will be explicitly expressed in the format of an interlingua. Thus, the interlingua representation of the source text, together with target language descriptions, contains sufficient information for the system to produce an output sentence. In theory, an interlingual MT system does not need any bilingual information modules — not even a bilingual lexicon, provided that each monolingual lexicon is mapped onto the interlingua. Examples of interlinguas that have been used in MT systems are artificial logical languages, sets of (presumably) universal semantic primitives, and the artificial language Esperanto (cf. Hutchins 1986: 55). The PONS system (Dyvik

[48] Here we have omitted the initial stage of tokenisation, which involves reading the input text and identifying its word forms. This stage is, however, not peculiar to transfer systems, but necessary in any kind of automatic translation where the input text is syntactically parsed.

1990, 1995), when translating in mode 3 (cf. 1.3.2), uses situation schemata as an interlingua.[49]

The division between transfer and interlingua systems may be seen as a gradual one rather than as a discrete one. In a transfer system, the amount of work needed during the transfer stage depends on the depth of the linguistic analysis of the source text. If the analysis creates a sufficiently detailed, and sufficiently language-neutral, representation of the input, then it may contain enough information to serve as a basis for the generation of the output.

An important difference between direct and indirect MT systems is that in the latter type it is possible to keep linguistic information separate from the translation procedure, which makes it far easier to extend a system to new language pairs.[50] As pointed out in 1.4.2.3, it is a demanding task to build linguistic resources for MT systems, and it is an advantage if such information modules, once they have been compiled, may be reused. In this respect interlingua systems appear more attractive than transfer systems, since the interlingua strategy does not require any language-pair dependent components. Transfer systems, on the other hand, need bilingual lexicons as well as sets of transfer rules, and the latter may be not only language-pair specific, but also dependent on the direction of translation.

On the other hand, the interlingua strategy is not necessarily the most attractive approach to automatic translation, given the degree of complexity in the translation task. Interlingual translation requires a deep analysis of the input text, and this is computationally demanding. But actual translation does not always require great efforts. If there is a sufficient degree of structural similarity between source and target language, then it is sometimes possible to translate word-by-word, or almost word-by-word. Thus, there are cases where the direct translation strategy would be sufficient; those are included among what we have described as type 1 correspondences.[51] With respect to type 2 correspondences, the transfer strategy seems appro-

[49] The PONS situation schemata are based on Situation Semantics; cf. Barwise and Perry (1983), Fenstad et al. (1987).

[50] Cf. the remarks on modularity in 1.4.2.1.

[51] Cf. the brief introduction to the type hierarchy in 1.3.1. Quantitative results concerning the distribution of the four correspondence types within the analysed data are presented in chapter 5.

priate: at the transfer stage the structure of the source text is changed according to the TL grammar. The types of source-target divergences found in type 2 correspondences pertain to surface syntactic structure, which means that translation can be done by transfer at a "shallow" linguistic level. Moreover, as direct systems have been able to accommodate certain word order differences between SL and TL, it is possible that also type 2 correspondences could be handled by the direct strategy. Then, in cases where the translation task is more complex than in correspondences of types 1 and 2, transfer must take place at a deeper level, and it may be necessary to do a full semantic analysis of the source text in order to reveal sufficient information for target text generation. The experimental PONS system combines, in a sense, all three translation strategies — direct, transfer, and interlingua. The system demonstrates that deep analysis and interlingual translation is necessary only in certain cases, and that an interesting challenge is to find those instances of translation where either the direct strategy or shallow transfer is sufficient to produce an appropriate translation.

1.4.2.5 Linguistic vs. non-linguistic approaches

As mentioned in 1.4.2.1, a division emerged in the early 1990ies between linguistics based and non-linguistics based approaches to machine translation. This division applies to a dimension independent of that of translation strategy; it pertains to what kinds of information resources an MT system is equipped with, and in what ways those resources are designed.

Ever from the early days of machine translation and until about 1990 there was a general view that to achieve automatic translation it was necessary to use linguistic information, i.e. information about source and target language and about how SL and TL are interrelated. Such information sources can be seen as a parallel to the bilingual competence of a human translator (cf. 1.2 and 2.4.2). Until about 1990 the established view among MT researchers was not only that MT systems needed linguistic information, but also that such information should be given in language descriptions designed according to principles of linguistic theory. A great variety of approaches of this kind have been investigated, and they are presented as *linguistic-based research paradigms* by Dorr et al. (1998: 19–30).

It indeed caught some attention when researchers had implemented methods for automatic translation that did not use linguistic information. From about 1990 onwards several techniques of this kind appeared; they are presented as *non-linguistic-based paradigms* by Dorr et al. (1998: 30–35), and they cover what is referred to as *statistical MT* in 1.4.2. Non-linguistic translation systems have in common that they depend, either for their development or for their functioning, on the existence of large parallel corpora. That is, non-linguistic MT techniques use large parallel corpora as repositories of information about the translational relation between two languages. Another important prerequisite for the workability of these approaches is the development of efficient algorithms for the automatic alignment of words or word sequences.[52] Word alignment applies to translationally parallel texts of two different languages, and it involves identifying links between translationally corresponding word forms in the two texts. By using the information contained in such links it is possible to find recurring translational correspondences. To put it simply, non-linguistic MT systems compute translations on the basis of which translational patterns that are frequent in the parallel corpus used by the system. The key to identifying a target equivalent *b* for a given source expression *a* is the probability that *a* corresponds with *b* based on the actual correspondences in the parallel corpus.

An important reason why non-linguistic approaches have been developed is that even though linguistic methods have reached a high level of sophistication, there are large development costs involved when building linguistic-based MT systems, and it is not easy to combine computational efficiency and broad coverage in grammars and lexicons. On this background it is appealing to investigate what may be achieved by doing without linguistic information modules and by applying pure computer science to parallel corpora. Clearly, there are certain linguistic phenomena that are too complex to be handled by non-linguistic techniques (e.g. long-distance dependencies; cf. Dorr et al. 1998: 35), and now the trend is to integrate the two approaches in so-called hybrid MT design, so that the strengths of both techniques may be combined.

---

[52] An important contribution in this respect is Gale and Church (1993).

The question of automatisation which is implicit in the present study of translational complexity is not neutral in relation to the division between linguistic and non-linguistic approaches to MT. Our investigation relies on several assumptions regarding the types of information needed to produce a translation, and these assumptions have consequences for where and how we draw the limit of computability.[53] Although we have previously indicated that the results of our product-oriented study are in principle also relatable to statistical MT, it is the linguistic-based approaches that we see as relevant to our discussion of computability.

1.4.2.6 The scope of machine translation

After a history of more than 50 years there seems to be general agreement that MT will not replace human translation. It seems unrealistic that automatic systems will reach a level of perfection where they produce high quality translations of unrestricted text without any kind of human intervention. Thus, we cannot expect that post-editing of machine translation output will be dispensed with. On the other hand, MT systems have been applied for decades as translation tools, and this is because they have been useful, within their limitations. For years now it has been common to talk about *the translation industry*, and that expression indicates, firstly, how large the demand for translation is, in particular of the non-literary kind, and, secondly, that automatised tools are needed in order to meet that demand.

Thus, practice shows that, given certain conditions, computerised translation can be a very helpful tool for reducing the workload for human translators. For one thing, if there is a high degree of structural relatedness between source and target language, then the challenges involved in MT design are reduced. Moreover, researchers and developers have experienced that successful systems can be designed for so-called sublanguage texts. Examples could be maintenance manuals and similar kinds of technical documents, which are characterised by relatively precise and unambiguous language, often repetitive, and dominated by a limited set of syntactic constructions.

[53] This will be discussed in chapter 2.

Such texts are not attractive to human translators, and the task of translating them rather resembles what computers are particularly good at: to repeat tedious computations, and to do so with precision.

Although fully automatic high quality translation probably will remain an unattainable ideal, it is still the notion of fully automatic translation which is of relevance to the present project: when discussing to what extent it would be possible to simulate human translation as instantiated by the investigated parallel texts, we assume that the translation task is to be solved without any human intervention. This must be seen as a framework for posing research questions, and not as a norm for practical systems.

1.4.3 Parallel corpus linguistics

As our investigation of translational complexity applies to parallel corpus data, it is appropriate to pay some attention to the field of parallel corpus linguistics. And, as mentioned in 1.4.2.5, the availability and use of parallel corpora has also become highly important to machine translation research. The label *parallel corpus linguistics* is taken from Borin (2002), who identifies the field as a subpart of the larger domain of *corpus linguistics*.

1.4.3.1 Corpus linguistics

This field is defined as follows by McEnery and Wilson (2001: 2): "Corpus linguistics is not a branch of linguistics in the same sense as syntax, semantics, sociolinguistics, and so on. … Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description."[54]

In recent years this methodology has come to be regarded as an indispensible part of linguistic research, and, basically, it involves providing empirical resources in the shape of machine-readable and searchable corpora, together with systematic methods for using the corpora in order to investigate specific linguistic phenomena. Clearly, it

[54] For an overview of the field see, in addition to McEnery and Wilson (2001), Sampson and McCarthy (2004), McEnery et al. (2006), Renouf and Kehoe (2006, 2009).

is impossible to do linguistic research without testing theories against examples of actual language use. Earlier there used to be some antagonism between linguists who advocated corpus-based studies and those who claimed that corpus data would always be incomplete and were inferior to what might be gained from studying the intuitions of individual language users.[55] Over the years, large corpus resources have become available for many languages, and computational linguists have developed efficient tools for identifying and processing linguistic data in large corpora. Thus, there is now a general trend that investigations of linguistic phenomena are carried out, preferably, with the use of corpus data, since corpora are important repositories of information about language use. There is, however, always the possibility that even in a large corpus a certain linguistic phenomenon might have no manifestations; in such cases the problem is to interpret the absence of occurrences: it is accidental or a consequence of aspects of the language system? Still, such cases do not reduce the value of the data that are found.

The Latin word *corpus* means 'body', and as stated by McEnery and Wilson (2001: 29), any body of text is in principle a corpus. However, "… the notion of a **corpus** as the basis for a form of empirical linguistics differs in several fundamental ways from the examination of particular texts" (2001: 29). More specifically, the building of corpora as used in modern corpus linguistics is normally subject to certain demands, of which McEnery and Wilson (2001: 29–32) discuss four kinds. Firstly, a corpus for linguistic research should be representative in the sense that it must, as far as possible, cover a whole variety of a language. Hence, it will be unsatisfactory to include texts of for instance only one type, or texts produced by only one author, or by authors of only one sex. Secondly, a corpus is normally of finite size: once it has been compiled according to a certain plan, new texts are not added.[56] An example of a fairly large, finite corpus is the British National Corpus with about 100 000 000 running words. Thirdly, it has now become a standard requirement in corpus building that such resources are machine-readable. Otherwise, computerised research tools

[55] For a discussion of this, see chapter 1 in McEnery and Wilson (2001).

[56] There are some exceptions, in particular corpora where new texts are continually added in order to keep the corpus up-to-date on current language use; cf. McEnery and Wilson (2001: 30–31).

cannot be used. Fourthly, once a representative, finite corpus has been compiled and made available for a research community, it is in a sense unavoidable that it will be attributed the status of a standard reference. Because such resources are valuable repositories of linguistic data, and may be kept constant, they are excellent test beds for varying approaches to the description of linguistic phenomena. On the background of these four requirements, or characteristics, McEnery and Wilson (2001: 32) present a prototypical definition of a corpus in modern linguistics: "… a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration."

There is an important division between *annotated* and *unannotated* corpora. Unannotated corpora contain "raw" text, i.e. plain text with nothing added, whereas in annotated corpora labels signifying various types of linguistic information have been attached to specific word forms. Examples of such information types are parts of speech and syntactic functions. Corpus annotation may be done manually or by software. The field of natural language processing now offers a range of different applications for automatic linguistic analysis, among which corpus annotation programs are an important subclass. As pointed out by McEnery and Wilson (2001: 32), a significant difference between annotated and unannotated corpora is that in the case of the former the added labels make explicit linguistic information that is only implicit in unannotated text, and hence annotation increases the value of a corpus. However, it may also add some "noise": if the annotator, whether a human or a computer program, makes any wrong analyses, then errors are included in the corpus.

The present investigation is carried out using data taken from parallel texts, and as will be described in chapter 4, the result of our analysis is a manually annotated corpus of translationally corresponding strings extracted from running texts. Still, our empirical analysis has not been done with reference to corpora in the sense given above, and hence we shall not go deeply into the field of corpus linguistics.

1.4.3.2 The added value of parallel corpora

For language researchers working under a cross-linguistic perspective parallel corpora are an invaluable resource. Borin (2002: 1) applies the label of *parallel*

corpus linguistics to research on parallel corpora, and he states that the "prototypical kind" of parallel corpora "is that which consists of original texts in one language, together with their translations into another language" (2002: 1). This is in contrast to the phenomenon of *comparable corpora*, which are collections of original texts in different languages, but of the same, or similar, text type, so that the texts are functionally comparable (cf. Borin 2002: 3). Comparable corpora fall outside the focus of our interest, since they do not contain translational correspondences of the same kind as parallel corpora do, but they are clearly of great value to contrastive linguistic studies. Johansson (2007: 9) makes the point that the term *parallel corpora* has unfortunately been used to cover comparable corpora as well as parallel corpora in the prototypical sense given by Borin. To solve this problem Johansson refers to parallel corpora as *translation corpora* in order to keep them distinct from comparable corpora, and he adds the multilingual dimension by defining translation corpora as containing "original texts and their translations into one or more other languages" (2007: 9).

In the previous section we discussed the usefulness of corpora for linguistic research, and it is not difficult to see what is the added value of parallel corpora. A representative parallel corpus may of course provide empirical data for monolingual studies,[57] but primarily it serves as a repository of information about the translational relation between the source and target language texts included in it. We have already seen that large parallel corpora have been used to develop MT systems operating without linguistic information modules (cf. 1.4.2.5), and the great utility of parallel corpora in research on translation, manual as well as automatic, is obvious. In addition to (machine) translation research, Borin (2002: 1) mentions other examples of areas where parallel corpora have been put to use: translation training, language teaching, bilingual lexicography, and contrastive and typological linguistics. For the latter kind of studies, multilingual parallel corpora are especially useful.

[57] That is, preferably with reference to the original texts. It is generally agreed that target texts normally exhibit certain linguistic properties specific to translations.

With respect to the present project, it could not have been carried out without access to parallel texts.[58] Approximately one half of the empirical data are collected from texts included in the English-Norwegian Parallel Corpus (ENPC), documented in Johansson (1998, 2007), and Johansson et al. (1999/2002). The ENPC is described by Johansson (2007: 11) as "a bidirectional translation corpus consisting of original English texts and their translations into Norwegian, and Norwegian original texts and their translations into English." It includes fiction as well as general, non-fiction texts and has a total of approximately 2,6 million words (cf. Johansson 2007: 13). An important feature of the ENPC is that it is *sentence aligned*, which means that each sentence in the corpus is linked to a translationally corresponding sentence (if found) in the parallel text (cf. Johansson 2007: 14–16). Thus, the ENPC is also an example of an annotated corpus, and it provides a goldmine of empirical data for contrastive linguistic research.

A strong field of modern contrastive language studies has evolved along with the development of corpus-based methods for linguistic research. The value of contrastive studies is obvious: they provide information about systematic differences between specific language systems, and about the effects of those differences as manifested in parallel corpus data. Both kinds of information are highly useful in many other fields, such as translation, language teaching, and translator training.[59] We may quote Johansson (2007: 1) on the great value of modern text corpora, and in particular of multilingual corpora, as repositories of representative data about language use: by exploring such resources "[w]e can see how languages differ, what they share and — perhaps eventually — what characterises language in general."

1.5 Organisation

This thesis consists of five main parts, among which the present chapter constitutes the first one. The purpose of this chapter has been to state our research questions, to

[58] Cf. the list of primary sources.

[59] Describing the large field of contrastive linguistic research falls outside of the scope of this work. Concerning the language pair English-Norwegian, Johansson (2007) is a representative study within the field: it presents corpus-based contrastive investigations of a range of linguistic phenomena, and also provides a multilingual perspective by including German and Swedish in some analyses.

introduce our framework, and to present some important topics of disciplines which are relevant to this study.

Part II includes chapters 2 and 3, and covers the theoretical and analytical foundations of our investigation. In chapter 2 we argue for a product-oriented approach to the study of translation, before explaining principles for drawing the limit of computability, or linguistic predictability, in the translational relation. Then, the basic notions of information, knowledge, and informational content are discussed, and we present our typology of information sources for translation. Chapter 3 opens with an informal presentation of the information-theoretic concepts of computability, complexity, and related notions. Then we present some approaches to the description of linguistic complexity, and describe our own notion of translational complexity, as well as its relation to computability. The remainder of chapter 3 is a detailed description of the correspondence type hierarchy. The four types are presented as translation tasks in order to capture the information requirements of each type, and to relate the notion of translational complexity to the amount and types of information needed for solving a translation task, including necessary processing effort.

Part III contains chapter 4, which describes our empirical investigation. The chapter starts by presenting the analysed parallel texts, as well as the concerns lying behind the selection of texts. Further, the syntactic criteria for identifying units of analysis are presented and illustrated, before we discuss the principles governing the classification of extracted string pairs in terms of translational complexity. Also, chapter 4 describes several practical aspects of the recording of translational correspondences.

Part IV covers chapters 5 and 6, which present the results of our analysis, and discuss them in relation to the initial research questions. Chapter 5 focuses on the analysed pairs of texts, and we present the complexity measurements across all recorded data, as well for each direction of translation, for each text type, and for the individual text pairs. Text-typological differences revealed by the analysis constitute a central topic in the discussion of the results. Chapter 6 presents certain phenomena which are recurrent among the recorded data, and which involve some kind of semantic deviation between translationally corresponding units. These are sorted into

a set of subtypes within the main correspondence types. The discussion of the semantic subtypes shows how the line is drawn between, respectively, computable and non-computable translation, and it illustrates certain phenomena that are not included in the domain of linguistically predictable correspondences of the language pair English-Norwegian.

Part V consists of chapter 7, where certain conclusions are drawn. These are centred around three topics: our framework, the method, and the results of the study. Also, we indicate a possible extension of our analytical approach.

PART II
FOUNDATIONS

2 Theoretical assumptions

2.1 Overview

This chapter is divided into three main parts, which together present a theoretical basis for our study of translational complexity in selected parallel texts of English and Norwegian. The analysis is focussed on the relation between original and translation, and the first part of this chapter, 2.2 with subsections, argues for the choice of a product-oriented approach to translation.

With reference to tokens of parallel texts instantiating specific text types, the principal aim of our analysis is to find out to what extent it is possible to predict, or compute, a certain translation on the basis of a given source expression and otherwise accessible linguistic information, and without the aid of a human translator. For this purpose, the second part, 2.3 with subsections, presents principles for drawing the limit of computability in the translational relation between a unit in the source text and its correspondent in the target text.[1]

In the third part, 2.4 with subsections, the basic notions of information, knowledge, and informational content are discussed before we present our typology of information sources for translation.

2.2 An objectivist approach to translation

On the background of the discussion of different approaches to the study of translation (cf. 1.4.1 with subsections), a relevant distinction is one made by Karl R. Popper between the *products* of behaviour and production *behaviour*. Its relevance follows from the fact that translation is a kind of human behaviour which results in a

[1] The notion of 'computability' is discussed in 3.2.1.

product. Popper's distinction is part of his objectivist approach to knowledge, which we will present in 2.2.1, and in 2.2.2 the phenomenon of translation is discussed in the light of his approach. In 2.2.3 we will relate certain concepts, categories, and methods of translation studies to Popper's framework, and in 2.2.4 comment on the approach taken in our own investigation.

2.2.1 Popper's objectivist view of knowledge

The distinction between the *products* of behaviour and production *behaviour* is presented in the essay "Epistemology Without a Knowing Subject" (Popper 1979: 106–152), upon which the following exposition is based. The topic of his essay is epistemology, understood by Popper as "the theory of *scientific knowledge*" (1979: 108). He starts by making certain fundamental distinctions: he divides reality into three domains of knowledge, and he draws the line between objective and subjective knowledge. Then, starting from a discussion of biological behaviour in general, he presents a model of the growth of knowledge, in which scientific knowledge is a special case of objective knowledge, and the distinction between products and production behaviour plays an important role in the model of knowledge growth.

Popper describes the three different domains of knowledge as "worlds or universes": in his words, the first world is "the world of physical objects or of physical states"; the second world is "the world of states of consciousness, or of mental states, or perhaps of behavioural dispositions to act"; the third world is "the world of objective contents of thought, especially of scientific and poetic thoughts and of works of art" (1979: 106). Popper does not claim this to be the only possible way of dividing reality into domains, but he finds this approach to be convenient. He argues for the independent existence of the third world through two thought experiments, in both of which he imagines a scenario where all machines and tools created by man are gone, together with human skills and knowledge of building and using the tools. In the first case books and libraries still exist, so that after some time human civilisation may be rebuilt through man's capacity to learn. In the second case all books and libraries are also destroyed, so that there are no pools of objective

knowledge to learn from, and hence our civilisation cannot be rebuilt until the knowledge itself has been rediscovered.

Popper claims there are two different senses of knowledge (1979: 108–109): Subjective knowledge is something located in the mind of an individual; it is a state of mind, a second world object. According to Popper, it is knowledge in the subjective sense that has been the concern of traditional epistemology. Objective knowledge, on the other hand, exists independently of any particular knowing subject; it belongs to the third world, and consists of problems, theories, and arguments. Scientific knowledge falls within this domain, and hence it is third world objects that are of interest to the philosophy of science. Popper views the process of learning in humans as growth of subjective knowledge, and the second world as a medium between the physical first world and the abstract third world. He states that "all our actions in the first world are influenced by our second-world grasp of the third world" (1979: 148–149).

A prominent aspect of the third world is its autonomy, a point illustrated in several ways by Popper. For instance, he describes the content of a book as a third world object, and states that what makes it a book is something abstract, more specifically "its possibility or potentiality of being understood, its dispositional character of being understood or interpreted, or misunderstood or misinterpreted", and he claims that "this potentiality or disposition may exist without ever being actualized or realized" (1979: 116). In the same way, the abstract content of a book exists independently of its author, although there is (normally) not an arbitrary relationship between the book and its author.

Popper observes that although the third world has independent existence, it is a human creation (1979: 112–115). Objective knowledge is a product of human behaviour: it is a result of problem-solving and discovery carried out by humans in order to cope with the first (and possibly also the second) world. Moreover, the third world has an important "feed-back effect" upon our consciousness (1979: 112, 119, 147–148), and in that way the growth of objective knowledge is due to an interaction between humans and the third world.

Through a discussion of animal behaviour Popper arrives at his distinction between behaviour and the product, or structures, resulting from behaviour (1979: 112–114). The study of these structures gives rise to two types of problems: first, problems dealing with the methods used when producing the structures (e.g. the problems involved in a spider's act of weaving its web), and, second, problems dealing with the structures themselves (e.g. the problems related to the cobweb). Then, applying this distinction to human behaviour, especially to language and science, Popper takes an anti-behaviouristic and anti-psychologistic stance in stating that understanding the problems connected with the products is the basis for understanding the production problems. Moreover, he claims that "we can learn more about production behaviour by studying the products themselves than we can learn about the products by studying production behaviour" (1979: 114). If we relate this statement to Popper's conception of knowledge growth, we may see that the impact of objective knowledge on human behaviour can be greater than the impact of individual human behaviour on objective knowledge.

2.2.2 Translation in relation to Popper's theory

In our view it is highly interesting to discuss translation in the light of Popper's epistemological framework because of the two-sidedness of this phenomenon: translation consists of both a process and a product, and the two are mutually dependent. Having looked at Popper's theory, the return to translation brings forth the question of whether the study of the product of translation is basic to the study of the translation process, and the related question of whether it is fruitful to study the product of translation prior to a study of the translation process. Before trying to answer these questions in 2.2.4, we will here locate the objects involved in translation within Popper's different domains of reality, and then in 2.2.3 relate the different approaches to translation to Popper's framework.

The translator, as a physical object, naturally belongs to the first world. With respect to translation competence, we have in 1.2 described it as including the following components: knowledge of source and target language systems, and of how these systems are interrelated; background knowledge of various kinds; skill in

interpreting source language texts, which includes the recognition of pragmatic, stylistic and formal aspects of the texts, and skill in producing target language texts which satisfy relevant demands of equivalence.[2] The two skills mentioned are second world objects. With respect to the different knowledge components, their world status is not unique. The knowledge components have intersubjective existence insofar as they are independent of individual translators. Thus, as instances of objective knowledge they belong to the third world, on a par with the knowledge of a language system shared by the members of a language community. On the other hand, as components of the subjective knowledge of a specific translator the two skills belong to the second world of mental objects. The manner in which they are represented in the brain of an individual is a first world object.

The translation process, consisting of a series of information processing steps in the translator's mind, is a second world object, and so is each discovery, or creation, of a target expression in the translator's mind during the translation process. On the other hand, a particular translation strategy (such as the choice of resolving all reception problems before beginning to produce the target text), becomes a third world object if it is formulated and made intersubjectively available.[3] But as long as it remains an individual course of action, it is a second world object.

While the physical realisations of specific source and target texts belong to the first world, the product of an act of translation is, like the content of a book, a third world object, and so is the corresponding source text. After the product of the translation process is output, and thus in principle intersubjectively available, the relation between original and translation is an object of the third world. The set of translational interrelations between the source and target language systems is also a third world object, but holds between different types of entities than the translational relation between specific source and target texts do. While the former is a relation between linguistic types, the latter holds between linguistic tokens. This point is developed further in 2.3.1.

---

[2] The description given in 1.2 of a translator's ability to construct a target text has here been modified in accord with the discussion of translational equivalence in 1.4.1.1.

[3] Chesterman (1997: 91) makes a quite similar point regarding the world status of translation strategies.

2.2.3 Translation studies in relation to Popper's theory

Several different approaches to the study of translation were presented in 1.4.1.1–3. We will now return to some of the concepts, categories, and methods discussed in that connection in order to relate these to Popper's framework.

Starting with Koller's description of translational equivalence (cf. 1.4.1.1), we may observe that the equivalence relation, as a specification of the properties with respect to which original and translation should be equivalent, exists independently of individual text recipients, and is thus a third world object. However, not all of the properties involved belong to the third world. With respect to denotational equivalence, the extra-linguistic state of affairs described by the source text may be a physical object, a mental object, as well as a third world object, but the denotation relation between a linguistic expression and the described state of affairs belongs to the third world as a part of the language system. Both connotative effects, and pragmatic aspects, of source and target text are dependent on the subjective experience of, and understanding by, a text recipient. Connotative and pragmatic equivalence thus involve second world objects, although the links between certain linguistic expressions and specific connotative and pragmatic effects may belong to the domain of objective knowledge insofar as such links are shared by a community of language users. Text-normative and formal-aesthetic equivalence also pertain to third world objects, since the textual properties they involve exist independently of the individual language user.

Toury's notion of norms in translation was briefly commented on in 1.4.1.1, where we noted that according to Toury (1965: 65), the norms govern translation behaviour, but the norms themselves are not available for observation. If translation norms govern the production of translations, then they are included among the components of translation competence. Following the discussion in 2.2.2, it is our view that translation norms, as components of the subjective knowledge of a specific

translator, are mental objects of the second world, but insofar as the norms are shared by different translators, they are intersubjective entities of the third world.[4]

Turning to House's model of translation quality assessment (cf. 1.4.1.2), we may pass lightly over her concept of translational equivalence since it is, like Koller's, an objective relation between third world objects. With respect to her notion of a cultural filter involved in covert translation, this, too, belongs to the third world, as an over-individual entity. As regards the task of translation quality assessment, it applies to third world objects, i.e. source and target texts, while the evaluation itself takes place in the second world: the comparison of textual profiles is an instance of information processing in the mind of the evaluator. Once the evaluation is done, however, its result becomes an object of the third world, as a piece of objective thought content that may be discussed and criticised.

Finally, we may briefly consider the different models of the translation process (cf. 1.4.1.3). As we have seen, there is no isomorphy between Krings' three-phase model of the course of a translation task and earlier two- or three-phase models. Furthermore, in the earlier models, the three stages described as analysis, transfer, and synthesis are aspects of translation which have been abstracted away from the actual process, from that which happens in real time, and as abstractions made by translation researchers and integrated in theories of translation, they are third world objects. On the other hand, the three phases identified in Krings' model are psychological processes, and hence objects of the second world. Consequently, the modern process-oriented studies differ from the earlier approaches with respect to the world-status of the object of investigation.

2.2.4 The present approach

The approach taken in our analysis of translational complexity conforms with Popper's epistemological framework, and we adhere to his view that the third world creates "its own *domain of autonomy*" (1979: 118). Our empirical point of departure is the translational relation as instantiated by intersubjectively available parallel texts.

[4] The latter point is also made by Chesterman (1997: 78). See Chesterman (1997: 63–70) for a systematic overview of norms in translation.

Thus, our object of study is the product of translation, a third world phenomenon. The translational relation will be discussed further in 2.3 with subsections.

Since the translational correspondences studied in our investigation are correspondences between third world objects, they are themselves of the third world and hence create an autonomous domain. In our view this domain is a pool of information about a part of the extension of the translational relation between Norwegian and English, and we regard this domain as something we may learn from. We will even claim that this pool of information shows that it is fruitful to study translational correspondences in relation to source and target language systems and independently of the cognitive capacity and choice of strategy of individual translators.

With respect to the question raised in 2.2.2 of whether the study of the product of translation is basic to the study of the translation process, it is our opinion that the opposite cannot be true: the study of the process cannot be basic to that of the product. Product-oriented works like Koller's typology of equivalence and House's model of translation quality assessment demonstrate that it is possible to discover facts about the translation product without studying the process. We will even regard certain findings of process-oriented translation studies as supportive of the popperian view. For instance, Krings' description of the translation process is full of references to the *result* of the translator's activity, and it is difficult to imagine how to categorise the different phases of the translation process without relating them to the product. In other words, it seems unlikely that the described process itself, a second world object, can be isolated as an object of study without considering third world objects, the products. Also, as pointed out in 1.4.1.3, TAP studies have revealed a great degree of heterogeneity among translators at work. This implies that the product of translation is at least to some extent independent of translation method. On the other hand, it does not imply that the translation process is independent of its intended product.

Rather, in the case of translation it is the product and its relation to the original text which gives the process its identity: unless a certain psycholinguistic process creates a translation, it cannot be identified as a translation process. We do not claim that the study of the translation process is unimportant, but we believe that even in process-oriented investigations of translation it is useful to consider the relations

between the product and the source text, and that is our answer to the question, also raised in 2.2.2, of whether it is fruitful to study the product of translation prior to a study of the translation process. Translation research has accumulated substantial knowledge about the product, and this knowledge seems a most advantageous point of departure for further explorations into the translation process.

2.3 The translational relation

We regard translational relations as correspondence relations holding between languages as well as between linguistic items of different languages.[5] In 1.1 we have described the translational relation between parallel texts of two languages as constituting parts of the extension of the translational relation between that pair of languages. This indicates that relations of translation exist on two different levels. On the one hand, they exist on the level of linguistic usage, i.e. between items of situated language, ranging from single word utterances to entire texts. On the other hand, they exist on the level where language is seen in abstraction from usage, i.e. between units of language systems as well as between entire language systems. This distinction is the topic of 2.3.1.

We will follow Dyvik (1998, 1999, 2005), who treats the translational relation as a theoretical primitive. Thus, the concept is "not to be defined in terms of other concepts, but assumed to be extractable from translational data by interpretive methods" (2005: 27), and the translational relation between two languages can be seen as given since it has an empirical basis "in the ubiquitous activity of practical translation" (2005: 27). The activity of translation takes place in a (cross-linguistic) language community, and bilingually competent informants may share judgments concerning the appropriateness of specific translations of given source texts. Such convergence among language users with respect to the acceptability of translations provides an empirical basis for identifying translational correspondence relations as part of the extension of the translational relation. For language pairs where modern

[5] Toury (1995: 77), on the other hand, claims that "translation relationships … normally obtain first and foremost between TEXTUAL SEGMENTS, very often even small-scale, rather low-level linguistic items."

parallel corpora (see 1.4.3 with subsections) are available, there are now excellent opportunities for investigating such correspondence relations.

Dyvik (1998, 1999, 2005) argues that translation is an important source of knowledge about the semantics of natural languages. Due to its empirical basis "the translational relation emerges as epistemologically prior to more abstract and theory-bound notions such as 'meaning', 'synonymy', 'paraphrase' and 'inference'" (2005: 27). In particular, translation is a normal type of language use, as opposed to meta-linguistic reflection, and its results are intersubjectively available (cf. Dyvik 1998: 51).

This is further developed in Dyvik (1999: 217–218), where he discusses the difference between meaning properties and translational properties. The observable relations between pairs of source and target texts allow us to discover translational properties of words and phrases in the texts. Those properties provide a key to meaning properties since the words and phrases of a language have translational properties in common "only if they share meaning properties" (1999: 218). As translational properties are observable in cross-linguistic data, they are "epistemologically more accessible" than meaning properties, which have traditionally been analysed through methods with elements of subjective judgment (1999: 218). Thus, the epistemological status of translational properties supports treating the translational relation as a theoretical primitive. The translational relation between languages is "assumed to be extractable from translational data by interpretive methods" (Dyvik 2005: 27), which involve distinguishing aspects of the language *system* from those of language *use* in the translational relation between texts.

2.3.1 A phenomenon of *langue* or *parole*?

Thus, a relevant point in connection with the translational relation is the saussurean distinction between the language system seen in abstraction from actual language use, *la langue*, and the language when used as a means of communication, *la parole*. Again, we adhere to Dyvik, who points out that as a relation between situated texts the translational relation holds between items on the level of *parole* (1998: 51–52). This follows from the fact that the translation of a specific source text is shaped not

only by the linguistic expressions used in the original, but also by "the context of utterance, the purpose of the utterance, and various other kinds of background knowledge" (1998: 52). Thus, translational correspondences between texts may be determined not only by information about the source and target language systems and their interrelations, but also by additional information sources.

However, the translational relation can also be seen as a relation between languages, and then holding between items on the level of *langue*. Dyvik argues that studying the translational relation as a *langue* phenomenon implies that we "disregard translational choices that can be motivated only by reference to the particular text and its circumstances", and this is the basis for isolating "translational correspondence relations between the sign inventories of the two languages — relations between words and phrases seen as types rather than textual tokens" (1998: 52).

The type-token distinction is important in our empirical investigation. When we analyse the product of translation instantiated as translationally corresponding strings of words, we regard the corresponding strings as linguistic types (cf. 4.3.6.2), but since the activity of translation applies to situated texts, we cannot account for the relation between a specific string and its correspondent without paying attention to the factors governing language use. In particular, these factors determine the possible interpretations that may be assigned to the corresponding strings, which again influence the analysis of translational complexity in the string pair.[6]

2.3.2 Predictability in the translational relation

As indicated in 1.1, our investigation aims at finding out to what extent it is possible to automatise translation in the case of selected English-Norwegian parallel texts representing two specific text types. This presupposes viewing the translation task as a kind of computation.[7] The problem may also be described as the following: given a certain source language expression, how far is it possible to predict its target language correspondent? We assume that if we could have access to information

[6] This point is discussed further in 4.3.6.2.

[7] Cf. section 3.2.2, which comments of the topic of viewing different kinds of human language processing as instances of computation.

about all factors that may influence the choice of target expression, then we would be able to predict the translation.

Prior to translation a source text is located in a domain of discourse. When a translator has created a target text that is regarded as an optimal translation, he or she has been as well informed as possible regarding the choice of target expression. That is, all necessary information has been available to the translator in the given domain of discourse. Likewise, in order to achieve automatic translation this information must be represented in an accessible format prior to the translation task. Hence we assume that the translational relation is predictable insofar as the source text together with *a pre-structured domain of information* can provide all the information needed to produce the target text.[8]

Is it then possible that this pre-structured domain can contain information about all factors which, in addition to the source text, have an influence on the choice of target expression? As discussed in 2.3.1, the translation of a specific source text is determined not only by the source and target language systems and their inter-relations, but also by "the context of utterance, the purpose of the utterance, and various other kinds of background knowledge" (Dyvik 1998: 52). We will assume it is possible to describe language systems and their interrelations and to include representations of such information in the pre-structured domain — i.e. to capture the domain of translationally relevant *linguistic* information.[9] By this assumption we follow Dyvik (1998, 1999) where the notion of 'linguistic predictability' is used to distinguish the translational relation between situated texts from the translational relation between the sign inventories of two languages. Dyvik's point is that to identify the translational relation on the level of *langue* is to isolate "the linguistically predictable translations" between two languages (1998: 52).

---

[8] In 2.4.2.1–3 we discuss the information sources which we assume to be included in this pre-structured domain, as well as sources falling outside of it.

[9] This assumption may appear to be in conflict with the point made in 1.4.2.3 that, so far, no natural language has yet been fully described in a computer-implementable format. However, that this has not yet been done, does not mean that it is theoretically impossible to provide a full-coverage computational grammar for a given language. Our assumption is that it is *in principle* possible to describe all parts of a language system, given that all parts of it are known and that there exists a grammar formalism in which those parts may be represented.

When interpreting a given source text, a translator will also exploit relevant non-linguistic information that he or she has access to. Thus, we regard such non-linguistic information as included in the domain of discourse of the source text, and our question is then to what extent this, too, can be represented in a pre-structured domain of information. We will assume it is possible to describe the information contained in *restricted* semantic domains of the world. This has been achieved in artificial intelligence systems and in various systems for natural language processing, of which automatic translation is an example. In such systems, knowledge modules represent restricted domains of technical information.[10] On the other hand, we assume it is not a manageable task to capture information about all possible domains of the world. Granted unlimited storage possibilities, the amount of world information that could be captured might be theoretically unlimited, but in practice it is necessary to draw a limit in order to secure tractability of the pre-structured domain of information.[11] Moreover, as parts of the world are unstructured, how would information about those parts be formalised?

Thus, an important property of the pre-structured domain of linguistic information is that it is finite. To be *finite* basically means to have an end or a limit. If information is represented in a finite way, it is contained in, or derivable from, a limited structure, and hence we may assume that it is in practice a feasible task to find and identify a particular informational element contained in, or derivable from, this structure.

In the present study of translation, our point of departure is not a restricted domain, but the domain of general language. Although we want to investigate whether translational complexity varies between pieces of general language texts and samples of domain-specific texts, we have chosen to limit the pre-structured domain to information about the source and target language systems and their interrelations. We regard this a necessary and helpful restriction as it provides a principled delimitation of the pre-structured domain, and also puts a theoretical limit on the extent to which the translational relation is predictable. Our analysis of translational

[10] Cf. the discussion of restricted semantic domains and sublanguages in 1.4.2.3.

[11] *Tractability* in a technical sense is explained in 3.2.1. Here the word is used in a general sense. According to the *Longman Dictionary of Contemporary English* (3rd ed.), the adjective *tractable* means ‘easy to control or deal with’.

correspondences will demonstrate the consequences of this delimitation in relation to empirical data, and we shall see that the limit of predictability in the translational relation will be relative to certain presuppositions concerning the descriptions of the languages involved. In particular, this limit depends on where the division is drawn between linguistic and extra-linguistic information, to be described in 2.4.2.1.

As pointed out above, Dyvik (1998: 52) argues that to identify the translational relation between the sign inventories of two languages is to find the linguistically predictable correspondences of that language pair. Such sign correspondences are linguistically predictable because they hold between signs with shared meaning properties (cf. Dyvik 1999: 217). This should, however, not be understood as if our criterion for distinguishing between linguistically predictable and non-predictable correspondences is exclusively the presence or absence of shared meaning properties. Other properties than those related to meaning may also be shared in a linguistically predictable correspondence between an SL sign and a TL sign. E.g., syntactic properties may be shared between translationally corresponding phrases if source and target language are structurally related. The criterion of shared meaning properties specifies what must *at least* be present in a linguistically predictable correspondence.

The set of *linguistically predictable translations* of a source language sign, its LPT set, is the full set of target language signs sharing a maximum, given the TL, of meaning properties associated with the SL sign (cf. Dyvik 1998: 56–57). That is, since language systems are differently structured in terms of grammar and lexical inventory, we cannot, within the scope of general language, expect that all meaning properties associated with a given SL sign is present in each member of its LPT set.[12] Then, taking into account differences between the two language systems, the LPT set of a given SL sign is the set of TL signs exhibiting a maximum of the meaning properties of the former. In the case of specific translational correspondences, it is shared intuitions among bilingually competent language users which decide what properties are included in this maximum. Furthermore, to describe a given target

[12] This point is also made in 6.3.2 in connection with denotational non-equivalence in translation.

language expression as a linguistically predictable translation of a source expression means that the former is one of the members of the LPT set of the latter.

The LPT set of a given SL sign may have zero, one, or more than one, member. There may be cultural or other differences between source and target language causing the situation where there is no TL sign associated with the meaning properties expressed by the SL sign. In cases where the LPT set is empty, translators may solve the problem by paraphrasing the source expression, and *parole*-related factors such as the use of world knowledge or contextual information will contribute to the choice of target expression. Consider the following example, found among the recorded data:

(1a) Det var ikke skiføre lenger, (BV)
'It was not conditions-for-skiing longer.'
(1b) It was no longer possible to ski,

The Norwegian noun *skiføre* means 'conditions for skiing', and has no lexical correspondent in English. The source sentence (1a) describes the situation where it is impossible to ski because there are no longer suitable conditions for it. The English translation (1b) is a paraphrase of this, chosen on the basis of general world knowledge.

In cases where the LPT set has exactly one member, there is a one-to-one correspondence between source and target language sign. An example would be the relation between a technical term in the source language and its target language equivalent. In general language it is a more common situation that the LPT set includes more than one member, and in such cases translation involves making a choice between the alternative target expressions. However, it may depend on the circumstances which member will be the optimal translation among the predictable candidates (cf. Dyvik 1998: 56). Such *parole*-related factors may also motivate a translation which is not an LPT member.[13]

[13] This point is illustrated by several of the phenomena discussed in chapter 6. Cf. e.g. the analysis of example (21) in 6.3.1.3.

For the sake of illustration, consider the English noun *pencil*. In the general sense of 'writing instrument' its LPT set with respect to Norwegian is {*blyant*}. The following is an example where *pencil* is not translated into *blyant*:

(2a) Got a pencil?[14]
(2b) Har du noe å skrive med?
'Have you something to write with?'

The source text (2a) is found in a dialogue context: two characters are talking on the phone; one of them has important information to share with the other, and the question (2a) is uttered when the former person wants to make sure that the listener is able to write down the details contained in the information. In this context it is possible to choose the translation *Har du en blyant?* ('Do you have a pencil?'), but instead the translator has picked the semantically less specific expression *Har du noe å skrive med?* ('Have you something to write with?'). Thus, *pencil* corresponds with *noe å skrive med*. The chosen translation may be said to be pragmatically equivalent with the source text, as there is focus on the fact that the addressee needs a writing instrument, and not necessarily a pencil. In this sense the textual context has motivated the choice of a translation of *pencil* which falls outside its LPT set.

On this background we may draw a distinction between predicting translations and generating specific target texts. To *predict* the translation(s) of a given source expression is to identify its LPT set; i.e. to find the set of target expressions sharing a maximum of meaning properties associated with the original. To *generate* a specific translation from an original may involve accessing other information sources than the information expressed linguistically in the source text, and it may involve making a choice between several alternative translations, among which some may be linguistically predictable, and some may be not.

The distinction between linguistically predictable and non-predictable translation can be related to the notion of 'computability', which will be discussed in chapter 3.

[14] The example is taken from Sue Grafton's novel *"D" is for Deadbeat*; see the list of primary sources. The novel is included in the ENPC (cf. 1.4.3.2).

As a first approximation, a 'computation' may be defined as a step-by-step procedure for solving a task in a specific way, and, thus, a *computable* task is a task that can be solved by a specifiable procedure. In the beginning of this section, we presented the assumption that the relation between a source language expression and its translation is *predictable* provided that the source expression together with a pre-structured domain of information can provide the translator with the information needed to produce the target text. Moreover, we have restricted this domain to include information about source and target language systems and their interrelations. A translation task, then, is computable if an automatic translation procedure is able to produce the target text correctly by exploiting the pre-structured domain of linguistic information. In this sense, we regard the computable part of the translational relation as identical to the linguistically predictable part.

To sum up, our investigation of how far it is possible to automatise translation in selected English-Norwegian parallel texts is based on assumptions regarding the limit of linguistic predictability in the translational relation. We assume that the linguistically predictable part of the translational relation is limited to the level of correspondences between *langue* units, and that it is computable from the information contained in the source expression, together with pre-structured information about the source and target language systems and their interrelations.[15]

2.3.3 The notion of 'literal translation'

Through the notion of linguistically predictable translations, Dyvik (1999: 217) explains a further notion of 'literal translation': "...the meaning properties of a sign are precisely the set of properties we want to capture, if we can, in literal translation." Thus, literal translation covers predictable correspondences between signs of two different languages; it deals with LPT sets, and it does not cover translations involving *parole*-related factors. In the present approach *literal translation* and *linguistically predictable translation* are synonymous expressions.

---

[15] This topic is revisited in 2.4.2.1, discussing the distinction between linguistic and extra-linguistic information, and in 3.2.5, describing computability in relation to translation.

When relating literal translation to meaning properties care must be taken to avoid circularity. If literal translation is defined in terms of meaning properties, then the translational relation is no longer a primitive, and our task is to clarify why it is *plausible* to assume that literal translation and meaning properties are related in the manner described above. In this respect, we have in 2.3 cited Dyvik (1998, 1999, 2005), who argues that since bilingually competent informants may share judgments on the appropriateness of given translations, there is an empirical basis for identifying the translational relation. Hence, the literal translational relation can be assumed to be elicitable from informants without resort to meaning descriptions. Then we can use the relations, given our plausibility arguments, to describe meaning properties.

Literal translation in the sense described here must not be seen as related to the notion of 'literal translation' defined by Vinay and Darbelnet (1995) as a translation method: "Literal, or word for word, translation is the direct transfer of a SL text into a grammatically and idiomatically appropriate TL text in which the translators' task is limited to observing the adherence to the linguistic servitudes of the TL."[16] The product of literal translation in the sense of Vinay and Darbelnet matches types 1 and 2 in our correspondence type hierarchy, whereas types 1, 2, and 3 are included in Dyvik's concept of a literal translational relation. Then, we find a closer match between Dyvik's notion and the product of Newmark's (1981: 39) concept of literal, or semantic, translation, which he has defined as the translation method that "attempts to render, as closely as the semantic and syntactic structures of the second language allow, the exact contextual meaning of the original."[17] Chesterman (1997: 12) sums up the various understandings of literal translation by observing that they have in common an emphasis on "closeness to the original form."

For the purposes of the present study, *literal translation* refers only to Dyvik's concept, which primarily serves to describe the relation between source and target text, and must not be associated with translation methods.

---

[16] The quotation is taken from Venuti (2000: 86). Pages 31–42 of Vinay and Darbelnet (1995) are reprinted in Venuti (2000: 84–93).

[17] Cf. Palumbo (2009: 49, 70, 167).

2.4 Information sources for translation

The topic of information sources for translation was introduced in chapter 1. Section 1.2 presented a tentative overview of our description of the types of information needed to produce a specific translation from a given source text, and information sources for translation were briefly mentioned in the context of automatic translation (cf. 1.4.2.3–4). The basic notions of information, informational content, and knowledge will be discussed in 2.4.1 with subsections, before we present our typology of information sources for translation in 2.4.2 with subsections.

2.4.1 Basic notions

In the preliminary version of the typology presented in 1.2 two important, basic notions are 'information' and 'knowledge'. These are concepts used by laypersons as well as by specialists. In non-technical discussions among laypersons these notions are normally taken for granted, as concepts that we all have an intuitive understanding of, whereas within a certain field of study, such as linguistics, information theory, or philosophy, the same concepts may be used in specific, technical senses. Our understanding of these, and related, notions are presented in 2.4.1.1–5.

2.4.1.1 Information

There is similarity, but also important differences, between Popper's concept of objective knowledge, which exists without a knowing subject, and the notion of information found in communication theory, i.e. information existing independently of any interpreting, cognitive agent. The work of Shannon and Weaver (1949) is commonly recognised as the origins of communication theory (also referred to as *information theory*). Our notion of 'information' is borrowed from this science, and the present discussion is based on Dretske (1981), whose project is "an attempt to develop an information-based theory of knowledge" (1981: 3), an attempt to apply the insights of communication theory in order to develop "a genuine theory *of information* as this is understood in cognitive and semantic studies" (1981: 4).

Within communication theory 'information' is understood as "an objective commodity, something whose generation, transmission, and reception do not require or in any way presuppose interpretive processes" (Dretske 1981: vii). Rather, what defines information are relations holding between distinct states, events, and structures (1981: x). In contrast to cognitive and semantic studies, communication theory treats information as a purely quantitative notion: the theory deals only with *amounts* of information, not with informational *contents* (1981: 3). Thus, information is either present or not; it is something that can be measured. Unlike notions like beliefs and propositions, information cannot be either true or false: its existence requires truth, and as Dretske points out, this property of information (in the technical sense) has the consequence that 'false information' or 'mis-information' are inconsistent concepts (1981: 45).

Further, "the amount of information associated with, or generated by, the occurrence of an event (or the realization of a state of affairs)" is measured in terms of "the reduction in uncertainty, the elimination of possibilities, represented by that event or state of affairs" (1981: 4). Thus, the emergence of a state or occurrence of an event for which there is an overwhelming probability represents very little information, whereas an unexpected state or event represents a relatively large amount of information (1981: 8–9).

Dretske points out that "*any* situation may be regarded as a *source* of information" (1981: 9). The focus of communication theory is on information sources, on measuring average amounts of information available from such sources; the theory does not aim to describe particular pieces of information, which would be of interest in semantic studies (cf. Dretske 1981: 10–11, 47, 52–53). Thus, although we want to exploit the information concept, we do not share the focus of communication theory, as our analysis will deal with particular pieces of text.

Dretske observes that communication theory has by some been viewed as "a theory of *signal transmission*, a theory about those physical events (signals) that, in some sense, carry information" (1981: 40). This yields a mathematical theory of information which describes statistical and other properties of signals, but, as he further points out, "[a] genuine theory of information would be a theory about the

content of our messages, not a theory about the form in which this content is embodied" (1981: 40). The distinction between, on the one hand, the signal as a physical event governed by probabilities, and, on the other hand, the informational content carried by the signal emphasises Dretske's view that the study of information involves not only those properties of information that can be accounted for in terms of quantitative measures, but also properties pertaining to the content of a particular piece of information. Studying the latter falls, as we have seen, outside the scope of communication theory.

Although 'information' in the ordinary, non-technical sense may be viewed as a semantic notion, Dretske warns against merging it with the concept of 'meaning': "... there is no reason to think that every meaningful sign must carry information or, if it does, that the information it carries must be identical to its meaning" (1981: 42).[18] He thus keeps 'meaning' strictly apart from the communication-theoretic concept of 'information' (1981: 41–44), and in his view meaning is a product manufactured from information (1981: vii). It may seem that communication theory, with its quantitative focus, cannot contribute to the study of meaning. However, Dretske argues that it is misguided to assume that "*meaning* is the *only* semantically relevant concept" (1981: 46). Information, as "[a] commodity capable of yielding knowledge", is also semantically relevant, and for that reason Dretske finds it fruitful to apply the insights of communication theory also when studying the semantic aspects of information (1981: 46). The information concept is relevant to the present project because our focus is on the various pieces of information that contribute to the selection of a given translation, and not merely on describing the meaning of the corresponding source expression.

2.4.1.2 Informational content

An important part of Dretske's project is to exploit the insights of communication theory in order to give an account of 'informational content'. The basic difference

[18] For instance, if a small child says to his parent "I have a tummy-ache", then the meaning of that signal is that he has a tummy-ache. However, if it is the case that the child has no tummy-ache — only happened to utter this sentence to get attention — then the signal does not carry the information that he has a tummy-ache.

between the amount of information a signal carries and its informational content is that the latter cannot be quantified. While it makes sense to ask whether a certain signal carries more or less information than another signal, it does not make sense to ask whether the informational content of that signal is larger or smaller than the content of the other signal (cf. Dretske 1981: 47–48). This illustrates how a study of the semantic aspects of information necessitates a shift from the communication-theoretic focus on average amounts of information to a focus on particular pieces of information.

To phrase it in very general terms, informational content is the information that something is the case. Dretske uses *message* as a synonym of *informational content* (see e.g. 1981: 55), and in his notation informational content is the information "that *s* is *F*", where "*s* is *F*" is used as a shorthand for some state, event, or structure, the lowercase *s* indicating an information source (cf. 1981: 66). His explication of informational content involves describing what conditions must be satisfied when a signal *r* carries the information that *s* is *F* (1981: 63–65). Firstly, the signal cannot carry a smaller amount of information than the amount generated by the state of affairs described as "*s* is *F*". This is a purely quantitative condition, and it illustrates the point made by Dretske (1981: 60) that to communicate a specific informational content, i.e. to convey a particular message, requires that *all* the information behind that message, and nothing less, must be transmitted. The second condition on informational content states that the signal *r* cannot carry the information "that *s* is *F*" unless *s* really is *F*, and the third condition states that *r* must carry the same information as that generated by *s*'s being *F*. The latter two restrictions are of a qualitative kind, or, in the words of Dretske, they "together constitute ... the semantic conditions on information" (1981: 64).

In addition to these three conditions, the informational content carried by a signal *r* is influenced by information already available to the recipient from other sources than *r*, in particular information about the conditions governing the probability of the informational content carried by *r*. When measuring the amount of information generated by some source, information is needed about the set of alternative possibilities existing at the source, the absolute probability of each of these possibilities,

and the probability of each of these possibilities relative to conditions governing the transmission of information from the source (cf. Dretske 1981: 43, 53–56). Hence, the amount of information gained by the recipient of a signal is influenced by information already available to the recipient with respect to the probabilities of the alternative possibilities, and in that way background information may determine the informational content that is transmitted by a specific signal to the recipient.

This may be illustrated by a simple example: if we already have the information that it is daytime, then receiving the signal of twelve bell strokes will tell us it is noon, because that is a far more probable state of affairs than the alternative of midnight. Thus, background information plays a part in Dretske's eventual definition of informational content (1981: 65): to say that a signal *r* carries the informational content "that *s* is *F*" means that there is a maximal probability for *s* being *F*, given *r* and available information concerning the possibilities existing at the information source, and that there would not have been such a maximal probability without the signal *r*. Thus, if *s* being *F* is the cause of the signal, then the signal has the informational content that *s* is *F* provided that there is no possible alternative cause of the signal, given available information about the possibilities. That is, something contains information about its cause only if other causes are impossible: frozen water tells us that the temperature in that water is below zero degrees Celsius, because temperatures above zero cannot cause water to freeze.

What is here referred to as "background information" is in Dretske's definition labelled "*k*" and described as "what the receiver already knows (if anything) about the possibilities that exist at the source" (1981: 65). In 2.4.1.3 we shall see that he conceives of knowledge as something existing within the mind of the cognitive agent. Thus, Dretske may seem to imply that informational content is dependent on the state of mind of the recipient — on how the signal is interpreted by the recipient. We prefer to regard this as an inaccuracy in his description of informational content, and we have chosen to read *k* as 'background information'. Elsewhere Dretske stresses that the conditional probabilities of the possibilities existing at the information source are objective features, that they are not determined by how likely the recipient believes each possibility to be, and that the amount of information carried by a signal

is independent of how much information the recipient is able to gain from it (1981: 55–57). Thus, background information influences the informational content of a signal regardless of whether the signal has been absorbed by the recipient or not, and informational content exists, like information, as an objective commodity, independent of interpretive processes.

In the present study we will relate the notion of informational content to the analysis of translational correspondences. More specifically, the concept will be applied when we describe semantic divergences between translational units in chapter 6. E.g., the discussion will show that differences with respect to amounts of information may have the effect that source and target text do not convey identical messages, and that a certain expression may carry different messages depending on whether specific background information is available or not.[19]

2.4.1.3 Knowledge

Dretske presents the traditional conception of 'knowledge' as "justified true belief" (1981: 85), and points out that as long as the notion of 'justification' is left unanalysed, this is not a satisfactory account. In his approach 'justification' is linked to information: the true belief that something is the case (*s* is *F*) counts as knowledge only if it is caused by the information that *s* is *F* (1981: 86). As described in 2.4.1.1, information, according to Dretske, requires truth, so that 'false information' becomes an inconsistency, and thus the causation of a belief by information amounts to a justification of that belief.

A consequence of this account is that instances of true belief do not necessarily count as 'knowledge'. It is possible to form a true belief without having received information supporting the belief. For instance, if Mary takes a look in the fridge and perceives some round fruits of red and yellow colour in a semi-transparent plastic bag, she may believe there are nectarines in the fridge. But the plastic bag contains apples, and she has mistaken the apples for nectarines. However, as there happen to be nectarines, too, in the fridge (hidden in a paper bag), her belief is true. But she has

[19] The former point is relevant to the discussions in 6.3.1 with subsections, and the latter point is illustrated by the analysis of example (28) in 6.3.2.3.

not received any information about the nectarines, and thus she does not have the knowledge that there are nectarines in the fridge.

Dretske underlines that his account of knowledge is intended as a description rather than as a definition of 'knowledge', as an explication of what 'knowledge' *is* rather than of what it *means* (1981: 91–92). To see knowledge as information-caused belief is to understand knowledge as a property of individual minds, as a state of mind of information-receiving cognitive agents.

There is a clear difference between Dretske's account of knowledge and Popper's concept of objective knowledge:[20] while the latter exists independently of particular knowing subjects, the former is understood as a state of mind of the individual. In Popper's terms, Dretske's 'knowledge' is a second world object, whereas 'objective knowledge' belongs to the third world. 'Knowledge' as described by Dretske corresponds, at least partly, with Popper's notion of 'subjective knowledge' (cf. Popper 1979: 108). Popper's 'objective knowledge' is of a more abstract kind than Dretske's 'knowledge': objective knowledge, being a result of human activity, presupposes past or present knowledge states in humans, but cannot be reduced to such knowledge states. Objective knowledge exists in the form of *shared content* of different knowledge states (perhaps caused in different ways) in human minds, and we have to ascribe a sort of intersubjective existence to this shared content in order to account for human interaction with it. Through this intersubjectivity the popperian 'objective knowledge' becomes a more abstract object than Dretske's 'knowledge', and it may seem as if Dretske, when viewing 'knowledge' as a cognitive object, does not draw the distinction made by Popper between the content of knowledge and how knowledge is represented in the mind of the individual.

2.4.1.4 Knowledge and information compared

It is clear from the preceding discussion that knowledge and information are different commodities, and a further comparison of these concepts is relevant for our later discussion of information sources for translation. Although we want to adhere to the

[20] Cf. the discussion of 'objective knowledge' in 2.2.1.

conception of 'information' as given by communication theory, Dretske's information-based description of 'knowledge' does not quite suit our purposes, since it is understood as a state of mind, and our object of study is a third world phenomenon (cf. 2.2.4).

We have previously stated that our investigation conforms with Popper's epistemological framework, and an important similarity between Popper's concept of objective knowledge and the notion of information found in communication theory has already been pointed out in 2.4.1.1: objective knowledge exists independently of the knowing subject, and information exists whether there is any interpreting agent or not.

There are also differences between the two notions, and a few of these could be mentioned. First, we have seen that Popper views objective knowledge as a product of human activity; the creation of objective knowledge requires knowledge acquisition in humans (cf. 2.2.1). Conversely, human activity is not a prerequisite for the creation or existence of information (although, of course, some information is information about humans and their activities).

Second, in Popper's concept there is focus on knowledge *content*, whereas information, as we have seen, is a quantitative notion. Objective knowledge is described as contents of thought, commodities that cannot easily be measured in the way that information is measured in terms of reduction in uncertainty.

A third difference between information and knowledge pertains not only to the popperian 'objective knowledge', but also to Dretske's 'knowledge': Dretske makes the point that knowledge and information belong to different "orders of intentionality" (1981:171–175). Physical structures and signals represent intentional states of the lowest order. When a signal carries information about its source, it occupies an intentional state relative to the source (cf. Dretske 1981: 172). While signals exhibit low-order intentionality, beliefs, knowledge, and meaning represent higher-order intentional states. According to Dretske (1981: 172) it is the ability to occupy higher-order intentional states that distinguishes information-processing systems with cognitive attributes from those that are unable to perform cognition. He explains this in a way which highlights the *selective* character of knowledge (higher-order) as

opposed to information (lower-order). It is not possible for a system exhibiting low-order intentionality to carry the information that *p* without also necessarily carrying all information that follows from *p*, either analytically or by natural law.[21] For instance, the information that a given amount of water freezes necessarily includes the information that the water is expanding. This property of information is described by Dretske as "nesting" (1981: 71, 179). However, it is possible for a system with cognitive attributes (e.g., a person) to have the *knowledge* that *p* without necessarily having the knowledge of everything that follows from *p*. Dretske's example is that it is possible to know that the solution to an equation is 23 without knowing that the solution is also the cube root of 12167 (1981: 173).

Then, how is the selective character of knowledge related to Popper's concept of 'objective knowledge'? It seems clear that the property of knowing *p* without knowing everything that follows from *p* pertains to the cognitive agent rather than to the objectivised knowledge content. Moreover, in arguing for the separate existence of objective knowledge, Popper makes the point that a theory may have consequences which nobody has discovered yet (1979: 116). That is, the content of a theory comprises everything that follows from it, whether anybody has realised it yet or not. This indicates that objective knowledge does not have the same selective quality as subjective knowledge; and it indicates a further similarity with 'information' in the technical sense. It seems that if we may assume the existence of the objective knowledge that *p*, Popper would also assume the existence of at least all analytic consequences of *p*. Still, Popper's 'objective knowledge' is distinct from the concept of information: because objective knowledge originates in subjective knowledge states in human minds, objective knowledge inherits a higher order of intentionality than that of information.

2.4.1.5 The knowledge of translators

As previously accounted for, our object of study is the product of translation, which, in our view, may serve as a reflection of translation competence.[22] In 1.2 we presen-

[21] For the sake of convenience "*p*" is used, like "*s* is *F*", as a shorthand for some state, event, or structure.
[22] Cf. 1.2 and 2.2.4.

ted our conception of translation competence as a combination of several factors: knowledge of source and target language systems, and of how these systems are interrelated, various kinds of background knowledge, and skills in interpreting and producing text in context. The mentioned skills involve knowledge of the pragmatic factors governing the interplay between linguistic forms and textual contexts.

It was pointed out in 2.2.2 that when these types of knowledge belong to a particular translator, they fall under the notion of subjective knowledge. Similarly, the skills mentioned are also second world objects and cannot be common objects of knowledge. However, when we, in this study, analyse translational relations between texts, we observe the product independently of its production, and we assume that a certain set of translational correspondences may be produced by different translators and by various translation strategies. We think it is safe to assume this because, as already pointed out in 2.3, different bilingually competent informants may share judgments concerning the appropriateness of specific translations of given source texts. Further, this assumption presupposes the existence of objective knowledge about translational relations between texts, knowledge which can be shared by different translators and which can be exploited by different translation methods. We aim to study this objective knowledge content insofar as it is detectable by analysing translational correspondence relations in our empirical data, and we will mainly disregard the possibly varying strategies or mental procedures of individual translators, although the recorded data can to some extent indicate differences concerning translators' preferences.[23] These strategies are of course legitimate and worthwhile objects of study in other contexts. Our focus is on the objective knowledge of *translators*, not on *the* translator's knowledge, and this is what we refer to when stating in 1.2 that our study is neither a cognitive nor a psycho-linguistic investigation of translation.

Then it is our task to try to find out more about the content of the objective knowledge presupposed by translational relations between texts of two languages.[24] We are interested in what is *implied* in the knowledge of translators: we do not

[23] The latter point will be illustrated by discussions in chapters 5 and 6.
[24] Cf. the description in 2.4.2.2 of given, general information sources for translation.

assume that actual translators use all available knowledge in every translation task, but we assume that given the existence of this knowledge there is the potential for performing the amount of analysis and inference required by each translation task. This resembles the property of information described by Dretske as "nesting" (cf. 2.4.1.4): embedded in the knowledge of source and target language and their interrelations is the knowledge required to analyse a particular piece of source text and produce a linguistically predictable translation of it. It may also be compared to the work of a grammarian: the grammarian explores and systematises what is involved in the knowledge of a given language, without assuming that the individual language user, whose knowledge the grammarian describes, is able to produce the same kind of systematisations. Explicating what is involved in a given body of knowledge is not the same as making claims about the inferences actually made by people having the knowledge. Our study of the objective knowledge of translators is a similar kind of explication, and, in line with the view taken in 2.2.4, we think that such explication can and should have its empirical basis in the observed products of the knowing subjects, which are, in our case, actual translations.

2.4.2 Typology of information sources

Sections 2.4.2.1–3 present a typology of information sources for translation, defined for the purpose of measuring translational complexity in terms of how much and what kinds of information are needed in translation. It is intended as one possible way of describing information sources for translation, and the typology is motivated by the nature of our object of study. The classification is not done according to criteria related to the cognitive equipment of individual translators, as our approach is to draw distinctions reflecting the types of information sources we assume are relevant in order to account for the observable relations between originals and their translations.

As presented in 1.1 and 1.3–1.3.2, translational complexity in our approach is associated with the need for information in translation tasks. In chapter 3 the structure of translation tasks will be described in terms of how much, and what kinds, of information are needed in order to translate. For those purposes the information

typology will be applied, as well as in the discussion of the empirical results in chapters 5 and 6. Given the analytical framework to be described in chapter 3, it will not be possible to quantify the need for information in mathematical terms; it can be analysed only insofar as each of the four correspondence types represents an upper and a lower bound on the required amount of information within its class.[25] Moreover, in chapter 3 the need for information is related to two questions raised in 1.2: to what extent can the various information sources for translation be represented in a finite way, and what is the amount of effort required in order to access and process them? With respect to the issue of finiteness, it is appropriate, in this chapter, to consider whether the various information types are included in the pre-structured domain of linguistic information introduced in 2.3.2 as defining the limit on predictability (and, hence, also on computability) in the translational relation.

In the information typology, distinctions are drawn along three different dimensions. Along the first dimension we assume a division between linguistic and extra-linguistic information sources. Previously, in 2.3.2, the limit of predictability in the translational relation is associated with a pre-structured domain containing information about the source and target language systems and their interrelations. Thus, extra-linguistic information is not included in this domain. It is, however, debatable to what extent it is possible to distinguish between purely linguistic information and world information, and it is especially difficult to draw a line between the linguistic and the extra-linguistic when we enter the fields of semantics and pragmatics, which will be discussed in 2.4.2.1.[26]

Second, we assume a division between general and task-specific information sources. General information is given prior to the translation activity; it includes information about source and target languages and their interrelations, and various types of information about the world. Task-specific information comprises the

[25] This point is explained towards the end of section 3.2.4.

[26] In 1.2 we have indicated a preliminary tripartite division into (a) purely linguistic, (b) pragmatic, and (c) extra-linguistic information sources. In 2.4.2.1 we will argue that pragmatic information may occur in the linguistic as well as in the extra-linguistic domain of information, thus advocating a binary main division between linguistic and extra-linguistic information sources.

different kinds of information associated with a particular piece of source text and the concrete task of translating it into a given target language.

Third, we distinguish between mono- and bilingual information sources: monolingual information includes information about source and target language respectively, and the information coded linguistically in the source text. Bilingual information deals with how the two languages correspond translationally.

It should be noted that we do not assume that every one of these three dimensions is necessarily crossed by each of the other two. That is, we do not assume that the domain of information sources for translation has a geometric structure like that of a cube with three axes crossing each other. In particular, the distinction between mono- and bilingual sources is only relevant within the domain of linguistic information.

As we have made clear in 1.2, as well as above, the study of translation competence is not part of our investigation, although the information that is accessible through the competence of translators is naturally included in the typology of information sources for translation. There are certain points of relatedness between our typology and models of translation competence that have been developed within process-oriented translation studies. Hurtado Albir and Alves (2009: 63–68) provide an overview of such models. In general, translation competence models have in common that they are divided into components, and that they distinguish between knowledge modules and skills, or abilities. Further, certain distinctions seem to be shared by several of them, in particular the opposition between linguistic and extralinguistic knowledge, and the division between general and specialised skills. According to Hurtado Albir and Alves (2009: 64), most of these models still lack empirical validation.

The three dimensions of our information typology are the concern of sections 2.4.2.1–3. It is not our ambition to provide exhaustive descriptions of these dimensions, but rather to clarify the distinctions we want to draw along them, since these distinctions are exploited in the empirical analysis of translational correspondences. Moreover, we do not assume that each and all of the information types to be discussed are available in any case of translation, although some of them, such as information about SL and TL and their interrelations, are necessarily required.

2.4.2.1 Linguistic versus extra-linguistic information sources

In our framework, this is an opposition between information derived from the source and target language systems and information about the extra-linguistic world. With respect to translational complexity, the division between linguistic and extra-linguistic information sources is closely related to the limit of predictability in observed translational correspondences in parallel texts. As previously discussed in 2.3.2, we assume that given a specific source expression, it is possible to predict a translation insofar as information about the factors that determine the translation is available in a pre-structured domain of linguistic information. Further, we argued that language systems and their interrelations can be described in a finite way,[27] and that these are the information sources included in the pre-structured domain. On the other hand, we have pointed out that to include extra-linguistic information about the world in the pre-structured domain will yield intractability, and that there must be a principled limit on the amount of information it may contain.[28]

Thus, granted that the domain of extra-linguistic information is infinite, we assume that linguistic and extra-linguistic information will show different properties in relation to translational complexity. More specifically, we assume that the degree of complexity is higher in translational correspondences involving extra-linguistic information than in cases involving purely linguistic information (cf. 1.3.1–2). But we do not *a priori* assume that processing information about the extra-linguistic world will be more complex than processing linguistic information, simply because it is non-linguistic. Intuitively, it seems reasonable to assume that there can be pieces of linguistic information which lead to greater complexity, and are harder to make representations of, than certain pieces of information about the extra-linguistic world. It also seems reasonable to assume that there can be many instances of extra-linguistic information which may readily be represented in a finite way.

---

[27] In this context we disregard the phenomenon of type 0 grammars, a class of formal grammars which are assumed to be finite, but for which there exists no known procedure for distinguishing the set of structures generated by a grammar of this kind from structures that cannot be generated by it (cf. Partee et al. 1990: 519–520.) Natural languages are generally seen as falling outside of this class, as a language user is normally able to decide whether a given expression belongs to the language or not.

[28] *Intractability* in a technical sense is explained in 3.2.1. Here the word is used in a more general sense. According to the entry for *intractable* in the *Longman Dictionary of Contemporary English* (3rd ed.), "an intractable problem is very difficult to deal with or find an answer to."

As we distinguish the linguistic from the extra-linguistic sources of information present in the discourse domain of a given source expression, the source text is considered on the level of *parole*. Within the *linguistic information sources* for translation there is, firstly, the information supporting the translator's knowledge of source and target language systems and their interrelations. Secondly, these sources include the information that is linguistically encoded in the source expression. This covers information about the situation type described by the source text, information about the linguistic structure of the source expression, as well as information about relations of reference holding between expressions in the source text and extra-linguistic entities. The latter is derivable from the source language expression as it is interpreted in a specific context. Thirdly, the linguistic sources also include information available in the linguistic context of the source string.

The *extra-linguistic information sources* for translation comprise general background information about the world, information about particular technical domains, information about textual norms, and information derivable from previous translation training and practice. They also cover information about the utterance situation of the source text, and about the translation situation. These types may include elements such as information about the sender, about the purpose(s) of original and translation, about temporal and geographical location, etc. Another extra-linguistic information source may be information derived by applying different kinds of background information in common-sense reasoning about facts described by the SL text. It may appear surprising that information about textual norms is regarded as extra-linguistic; we will argue below that this is a consequence of the way in which we distinguish between linguistic and extra-linguistic information.

The fact that we have listed different types of information sources classified respectively as linguistic and extra-linguistic does not imply that it is always clear where to draw the line between them. However, there are certain kinds of information that we regard as purely linguistic. Traditionally, a language system is seen as a structure divided into four levels: phonology, morphology, syntax, and semantics. At each level the language system specifies an inventory of units, or building blocks, together with a set of rules for how these units may be combined. In addition, the

language system includes a lexicon, which is an open set of lexical units, and each such unit contains information from all the four different levels of the language system. Descriptions of the phonological, morphological, and syntactic structures of a language appear as plausible examples of purely linguistic information.

With respect to semantic phenomena, on the other hand, extra-linguistic pieces of information are not always easily distinguished from the linguistic, and it seems difficult to find a principled way of doing so. Considering a lexical unit, such as *apple*, it seems reasonable that information about its meaning falls within the domain of linguistic information. But how are the meaning properties of *apple* identified? Knowing the meaning of *apple* implies knowing that apples are a kind of fruit, normally round, which is good to eat, and it may also include knowledge of what different colours apples may have, how they taste, etc. All these pieces of knowledge are supported by information available from the extra-linguistic world, but it is not necessary to have all this information about apples in order to understand the meaning of the word *apple*. In our opinion neither the language system itself, nor the extra-linguistic world, can offer a definitive principle for sorting the meaning properties of a lexeme from extra-linguistic properties associated with its denotata; there is no *a priori* basis for a sorting of that kind.

But the fact that a conceptual distinction cannot be drawn in a unique way *a priori* does not imply that it is meaningless.[29] We have argued that the linguistic domain is limited, and that this determines the limit of predictability in the translational relation. In our study of translational correspondences the division between the linguistic and the extra-linguistic is often a question of distinguishing between semantic information derived from the language system and extra-linguistic information sources that also contribute to the interpretation of a given source text. This depends on how

[29] Pustejovsky (1995: 232–233) arrives at a similar position in a discussion of how to draw the border between "linguistic or lexical knowledge" and "commonsense knowledge". In his view this is a continuum rather than a dichotomy, but he still finds it fruitful to maintain the distinction because there are "clear cases of paradigmatic linguistic behaviour that are better treated as language specific knowledge, rather than in terms of general inferencing mechanisms."

the given language system is delimited, and thus we relate the distinction to the way in which language systems are conceptually individuated.[30]

Since there is no objective answer to where the limit is drawn, there is an element of choice here. The choice will be influenced by the purpose for which the language description is meant to be applied, and by empirical facts about language use. Also, there are restrictions on what may be conceived of as a language system. As it is unmotivated to include large amounts of world information in the semantic component (cf. the discussion of *apple*), there is an upper bound on this, and a lower bound follows from the fact that there must be a reasonable amount of language users sharing a certain inventory of signs as the means of communication within their community. Given these constraints, a certain textual token may be seen as an instance either of general language or of a certain sublanguage, possibly depending on the purpose of the analysis.

Hence, the distinction between linguistic and extra-linguistic information must be recognised as relative to certain chosen presuppositions concerning the descriptions of the language systems involved. A translation example from a text dealing with a restricted domain may illustrate this relativism.[31] In the *Agreement on the European Economic Area* (*AEEA*) the English expression *competent authority* corresponds translationally with the Norwegian expression *vedkommende myndighet*. An example of the correspondence is shown in (3):

(3a) *The competent authority* shall take the necessary decisions within the framework of its internal rules. (*AEEA*)
(3b) *Vedkommende myndighet* skal treffe de nødvendige beslutninger innen rammen av sine interne regler.

When analysing the correspondence with respect to translational complexity, we treat the expressions as system units, or signs (cf. 4.3.6.2). The target sentence (3b) is not glossed, since we regard it as semantically equivalent with the source sentence (3a),

[30] This is in accord with Dyvik (2003: 9), who points out that the distinction "between instances of literal and instances of non-literal translation … must be drawn relative to the delimitation of the languages (general languages, sublanguages etc.) in which we assume that the texts are composed."
[31] The relativism is also discussed in chapter 6; cf. the analyses of (13) in 6.3.1.2, and of (20) in 6.3.1.3.

except for the pair of NPs in italics. *The competent authority* is translated as *vedkommende myndighet* ('the authority concerned'), and these two expressions deviate with respect to denotation: in the English text the property of having competence is attributed to *authority*, whereas in the Norwegian text the property of being concerned is attributed to *myndighet* ('authority').[32] Seen as system units, then, we do not regard *vedkommende myndighet* as a linguistically predictable translation of *competent authority* since certain meaning properties are not shared. However, this NP correspondence is recurrent among the data compiled from the *AEEA* and its Norwegian translation, which raises the question whether it is after all a linguistically predictable correspondence within the domain dealt with in the agreement text. Expressions of general language frequently acquire specialised meanings in texts pertaining to restricted, technical domains. If it is the case that within the domain of the *AEEA*, 'authority concerned' is one of the identifiable meanings of the expression *competent authority* when considered in isolation and independently of context, then the Norwegian expression *vedkommende myndighet* is a literal, linguistically predictable translation. This is an analysis which relies on the assumption that the *AEEA* is written in a domain-specific sublanguage with its own specialised vocabulary, so that the use of certain expressions in that sublanguage will be regulated by other conventions than those governing the use of general English. As those conventions will be shared by a community of sublanguage users, they are part of a language system, and we may assume that the expression *competent authority* is here a term-like lexical unit in English, and hence the italicised NP correspondence in (3) is predictable from linguistic information available prior to the translation task.

However, at least one instance of *competent authority* in the *AEEA* is not translated as *vedkommende myndighet*. In Article 58 *the competent authorities* is translated as *de kompetente organer* ('the competent institutions/bodies'):

(4a) With a view to […] , *the competent authorities* shall cooperate in accordance with the provisions set out in Protocols 23 and 24. (*AEEA*)

[32] This pair of NPs illustrates denotational non-equivalence between translationally corresponding, and co-referential, noun phrases; cf. 6.3.2.3.

(4b) *De kompetente organer* skal samarbeide i samsvar med bestemmelsene i protokoll 23 og 24 med sikte på […] .

We will not regard (4) as a counterexample indicating that *vedkommende myndighet* is after all a non-predictable translation of the phrase *competent authority* in the domain of the *AEEA* text. Rather, we will analyse *kompetent organ* as a member of the set of linguistically predictable translations of *competent authority*. We regard the italicised NP correspondence in (4) as a case where not only the translational relation between the phrases as units is linguistically predictable, but where also each lexical component within the target expression *de kompetente organer* is a predictable translation of its correspondent in the source expression. The Norwegian lexeme *organ* may not at first glance seem a plausible member of the LPT set of the English lexeme *authority*, but within the given textual domain this is a recurrent lexical correspondence.[33] Thus, with respect to the restricted domain of the Agreement text, both phrases *vedkommende myndighet* and *kompetent organ* are predictable Norwegian translations of the English phrase *competent authority*.

On the other hand, if we analyse (3) and (4) in relation to the domain of general language use, we will conclude that both translations of *competent authority* are cases falling outside the linguistically predictable. This presupposes an analysis where the expression *competent authority* is not treated as a unit of the language system, and where we assume that when it is translated into Norwegian, the choice of target expression is determined by information about the world. In this particular case such information may be derived through the following inference based on world knowledge: an authority concerned with making certain decisions is required to have the necessary competence for that task, and will hence be the competent authority.

Bhatia (1997) presents a genre-based approach to legal translation which may support the choice of ascribing information about these lexical correspondences to the extra-linguistic domain. A technical field, or specialist discipline, of which the law is an example, is associated with what Bhatia (1997) refers to as a "disciplinary culture". Members of a specialist discipline communicate by using specialist genres,

[33] In the *AEEA* and its Norwegian translation this lexical correspondence is elsewhere found in the recurrent pair of compound nouns *surveillance authority – overvåkningsorgan*.

and these genres are shaped by conventions determined within the disciplinary culture, described as *generic conventions* by Bhatia (1997). He explains the necessity of learning these conventions for anyone who wants to produce, or translate, texts in these genres (1997: 206–208). With respect to the legal discipline, these conventions are described as "expectations about the way in which language operates in legal contexts, but such expectations are never explicitly stated anywhere but in legal culture" (Bhatia 1997: 208). Information about such conventions of the legal culture is derived from technical knowledge, and although it pertains to the linguistic form of law texts, it belongs to the domain of extra-linguistic information.[34] In our view, the information that *competent authority* corresponds translationally with *vedkommende myndighet*, as well as with *kompetent organ*, is an example of a convention specific to the genre in which the different language versions of the Agreement text are written.[35]

Bhatia's notion of generic conventions may clarify what we understand by information about textual norms, identified above as a subtype within the extra-linguistic information sources for translation. Textual norms, or conventions, control or influence *parole*-related factors such as lexical choices, style, and textual structure. We regard information about such norms as extra-linguistic since information about the characteristic features of specific genres, or text types, is not part of a language system: textual norms are distinct from the conventions that constitute a language system and are shared by the members of the language community.[36] But as this information type deals with linguistic usage, we want to keep it apart from world information, whether general or technical. The distinction is motivated since information about textual norms may account for other aspects of a linguistic expression than those determined by information about facts of the world. In general terms, this is a distinction between information about extra-linguistic states of affairs, and information about norms controlling the use of language describing those states of affairs. With respect to law text, the division is clear: the former kind of information

[34] Cf. the discussion of norms in law texts in 5.4.2.1.
[35] This point is also illustrated by example (20) in 6.3.1.3.
[36] The notions of 'genre' and 'text type' are discussed in 4.2.1.2, where we explain why we prefer to speak of *text type*.

is derived from the legal domain, whereas the latter type is derived from the domain of law writing. The distinction may apply also in non-technical settings, as there are numerous contexts, written as well as spoken, where ordinary language users follow shared conventions governing their linguistic behaviour (e.g., dinner conversation, the writing of personal letters, etc.).

Thus, the notion of information about norms controlling language use is a very wide category, which may be refined by identifying types of norms included in this kind of information source. One possible subdivision is between norms applying to texts of general language and those that control specialised, technical language.[37] Norms of the first kind will be shared knowledge among general language users, whereas the second kind will be known by specialists within technical fields. Another distinction may be drawn between norms that influence the characteristics of various text types, and norms that govern the translation of the same types. The latter kind of norms is acquired by translators through translation instruction and practice. We assume that they largely correspond with the concept of norms in translation (cf. 2.2.3), although that notion may include more than textual norms.[38] As regards text-type specific norms, these may be different in, respectively, SL and TL, since text type characteristics are not always identical across languages.[39] Hence, the source text author is subject to the norms applying to the given text type in the source language, and the translator likewise to the corresponding textual norms of the target language.

It may seem arbitrary to relate the distinction between the linguistic and the extra-linguistic to the delimitation of language systems when the latter issue is, as we have seen, to some extent a matter of choice. In particular, as the distinction plays an important part in our analysis of translational complexity, it may seem as if the outcome of that investigation is determined by the way in which we choose to delimit

---

[37] Cf. the definition of *language for special purposes* (LSP) in 5.4.2.3.

[38] This indicates a certain degree of overlap between information about norms governing translation, and information derivable from previous translation training and practice. The latter type is identified above as a separate subtype within the extra-linguistic information sources. We return to this point in 2.4.2.2. Toury's (1995) translation norms have previously been mentioned in 1.4.1.1 and 2.2.3.

[39] This is e.g. shown by Nordrum's (2007) study of how English nominalisations are translated into Norwegian and Swedish in texts of popular science. Her results indicate that the norms of this text type are language-specific, since one fifth of the analysed English nominalisations were found to corresponds with finite constructions in the Norwegian and Swedish texts. The language-specificity of textual norms is also illustrated in the discussions of examples (17) and (20) in 6.3.1.3.

the languages represented in our empirical data. But arbitrariness may be avoided. Firstly, it is a prerequisite in our analysis to be consistent with respect to the chosen presuppositions concerning the description of the languages. Secondly, arbitrariness can be avoided if the conceptual individuation of language systems is based on empirical facts about language use. Such facts are available through text corpora, dictionaries, and linguistically competent informants, and enable us to conceive of what information it is *reasonable* to include in a language system, given the purpose of its description. In particular, when working with empirical data we find that it is frequently quite possible to determine whether extra-linguistic information has contributed to an interpretation, and subsequent choice of translation, or not. To illustrate this, we may again consider an example discussed in 1.3.1, repeated in (5):

(5a) Her kunne de snakke sammen uten å bli ropt inn for å gå i melke-butikken eller til bakeren. (BV)
'Here could they talk together without to be called in for to go in milk-shop.DEF or to baker.DEF'

(5b) They could talk here without being called in to go and buy milk or bread.

The example has previously been used to illustrate semantic divergence in a translational correspondence: the expressions *for å gå i melkebutikken eller til bakeren* ('to go to the shop selling milk or to the bakery') and *to go and buy milk or bread* do not denote the same activities, but both activities may have the same result, the purchase of milk or bread.[40] Otherwise, we consider string pair (5) to be a linguistically predictable correspondence. In the case of the Norwegian sequence *for å gå*, the English sequence *to go* is a linguistically predictable translation, and the pair of substrings *for å gå – to go* is a correspondence between system units, derivable from information about the lexicons and grammars of SL and TL and about their interrelations. Then, the Norwegian NPs *melkebutikken* and *bakeren* have no direct translational matches in the English sentence. Suggested LPT sets (cf. 2.3.2) in English of the Norwegian nouns *melkebutikk* and *baker* are given in (6):

[40] Cf. the discussion of example (5) in 1.3.1.

(6) *melkebutikk*:{*dairy*, *dairy shop*, *milk shop*}
baker: {*baker*, *baker's*, *baker's shop*, *bakery*, *bakery shop*, *bakehouse*, *bakeshop*}

Thus, one literal translation of the Norwegian expression *for å gå i melkebutikken eller til bakeren* could be *to go to the milk shop or to the baker's*, but the translator has chosen the non-literal translation *to go and buy milk or bread*. We assume that through general world knowledge the translator will have been aware that the story from which (5a) is extracted takes place in a time when milk and bread were normally sold through specialised shops in Norway, while, at least in a certain part of the English-speaking world, milk would typically be delivered at people's homes. Thus, background information provides the motivation for disregarding the linguistically predictable *go to the milk shop* as an optimal translation of *gå i melkebutikken*. Then, applying common-sense reasoning to the described facts of the world makes it seem obvious that the purpose of going to the places described in (5a) would be to buy milk and baker's products, and this is the information that gets the focus in the chosen English translation: *to go and buy milk or bread*.

Example (5) thus illustrates the distinction between meaning and context-induced interpretation. The pre-structured domain of linguistic information available prior to translation contains information about the meaning properties of the words in the source text, and is thus the basis for identifying predictable translations. But the pre-structured domain is only a subset of the discourse domain of a source text, and, as (5) shows, extra-linguistic information present in the source text context may induce an interpretation which disfavours the use of a linguistically predictable translation.

In the discussion of examples (3)–(5) we have several times referred to reasoning, or inferencing, about extra-linguistic pieces of information. Such matters fall within the field of pragmatics, which concerns the relationship between linguistic expressions and the situations in which they occur, and studies how discourse-related factors influence the interpretation of linguistic expressions.[41] Pragmatic phenomena are of interest to our investigation of translational correspondences as translation

[41] Huang (2007: 2) defines pragmatics as "the systematic study of meaning by virtue of, or dependent on, the use of language." Leech (2008: 88) defines it as "the study of meaning in speech situations."

applies to situated texts, and is typically done to serve a communicative purpose. It is not an aim to make pragmatic factors in translation the centrepiece of our study, but to consider certain relevant phenomena. In particular, we are interested in how the information available to discourse participants influence the production and interpretations of situated expressions, since a text or an utterance is the product of information processing performed by the sender, and its interpretation is the result of information processing on the part of the recipient.[42] To interpret a source expression prior to translation involves finding its propositional content, and identifying its illocutionary force, or type of speech act performed. The notion of 'proposition' is normally associated with sentences; it designates "what a sentence says about the world" (Allwood et al. 1977: 20).[43] A speech act is "the type of action the speaker intends to accomplish in the course of producing an utterance" (Huang, 2007: 102). Type of speech act, or illocutionary force, is commonly attributed also to written statements.[44] In the task of interpreting a situated expression, pragmatic factors contribute to finding the propositional content as well as to identifying the speech act, and an important part of our analysis of translational correspondences involves comparing the respective interpretations of source and target text (cf. 4.3.6.2).

How are pragmatic factors then related to the division between linguistic and extra-linguistic information sources for translation, or to what extent is pragmatic information part of the language system? This pertains to how far the interaction between discourse participants is expressed through linguistic conventions shared by the members of a language community. For instance, in English it is a convention that both the imperative and the interrogative may be used to express the speech act of requesting something, as illustrated by (7) and (8), respectively:

---

[42] Cf. the discussion in 2.4.1.2 of how available background information may determine the informational content of a specific signal transmitted to a recipient.

[43] Löbner (2002: 23–24) defines the proposition of a sentence as its "descriptive meaning", i.e. the set of situations it may refer to, but this does not capture the difference between sentence and utterance. The notion of 'proposition' is also commented on in 6.3.2.

[44] According to Huang (2007: 106), the most infuential approach to the classification of speech acts is the "neo-Austinian typology of speech acts", based on Searle (1975). In this taxonomy, there are five main categories of speech acts: assertives, directives, commissives, expressives, and declaratives, and each main category is further divided into subtypes. E.g., typical examples of directives are advice, orders, questions, and requests; cf. Huang (2007: 106–108).

(7) Please close the door!

(8) Would you close the door?

Given an appropriate context, such as the situation where some people are having a conversation in a room where a door has been left open to a noisy corridor, a similar request could be made by uttering the indicative sentence in (9):

(9) Excuse me, I find that noise on the corridor a bit disturbing.

If someone utters (9) in that context, an addressee would most likely infer that the speaker wants some action to reduce the disturbance, such as closing the door, and the speaker's intention would probably be exactly to achieve that. The relevant difference between, on the one hand, examples (7) and (8), and, on the other hand, (9), is that the piece of information through which a speech act is performed, is available in the linguistic expressions in (7) and (8), whereas in (9) it is not linguistically encoded, but derivable from the extra-linguistic context of the utterance. These examples illustrate that pragmatic information may be linguistically encoded and it may be not, partly depending on the speaker's choice of expression, and partly on the extent to which a language system exhibits conventionalised ways of encoding pragmatic constraints on the use of language in context.

To sum up, linguistic information sources for translation firstly include information about the source and target language systems and about their interrelations, seen in abstraction from the utterance situation of the source text. These sources constitute the pre-structured domain of information which defines the limit of predictability in the translational relation, and which is a subset of the wider domain of discourse in which the source text is located. Further, the linguistic information sources include the information coded in the source text expression, i.e. information about the situation type described by the source text, about the linguistic structure of the source expression, and about reference relations derived by interpreting the source text in context. They also cover information available in the linguistic context of the source expression. The extra-linguistic information sources for translation include general

and technical information about the world, information about textual conventions, information about the utterance situation of the source text, and information derived by reasoning about facts described by the source text.

2.4.2.2 General versus task-specific information sources

In 2.4.2 we have described the division between general and task-specific information sources for translation as a division between information available prior to the translation activity and information associated with a particular piece of source text and the concrete task of translating it into a given target language.

Thus, *general information sources* exist independently of specific translation tasks, and through the distinction between linguistic and extra-linguistic information they can be divided into information about source and target language and their interrelations, and information about the world derivable from the translator's background knowledge. The former corresponds with the pre-structured domain of linguistic information discussed in 2.3.2 and 2.4.2.1. General, extra-linguistic information sources cover information available through the general world knowledge of ordinary language users, as well as information about restricted, technical domains, which is required in the translation of technical texts. They also include information about textual norms, and information derivable from previous translation training and practice.[45]

As mentioned in 1.2, the information needed to produce a specific translation from a given source expression includes the types of information that are accessible through translation competence. Thus, the given, general information sources correspond with a translator's competence. In 2.4.1.5 we have argued for the existence of objective knowledge about translational relations between texts. The fact that translational relations hold between texts of two languages presupposes knowledge of how source and target languages are interrelated. Thus, we abstract away from individual translators and assume that prior to any translation activity,

[45] This is only one suggested way of dividing world information into subcategories. For one thing, restricted domains of information need not be technical. E.g. within a group of persons who have a certain "history" together, knowledge about shared experiences will constitute a restricted domain that may serve as a frame of reference influencing the interpretation of utterances made within that group.

there is a certain body of knowledge functioning as a pool of given information. Although we have pointed out similarities between information and objective knowledge, we have argued that these are not the same notions (cf. 2.4.1.4) and would thus avoid viewing the objective knowledge of translators as information. But since objective knowledge has the potential for being made intersubjectively available, it is our opinion that the objective knowledge of translators functions as an information source for translation. By regarding it as something that supplies given information, we assume that it is accessible when required for specific translation tasks, and as translations cannot be produced without a necessary amount of previously acquired knowledge, the objective knowledge of translators must exist prior to a translation activity. This is not to say, of course, that an individual translator possesses a constant body of knowledge which must exist before that translator is able to produce any translations — the knowledge of a translator normally grows through practice.[46]

In 2.4.2.1 we pointed out that there is some degree of overlap between two of the notions identified among the general, extra-linguistic information sources, i.e. information about textual norms, and information derivable from previous translation training and practice. The notions are clearly interconnected as a translator may acquire knowledge about the former through translation practice. Still, we keep the distinction, since textual norms apply to texts of individual languages independently of translation. Moreover, given our product-oriented approach, it is not relevant in the present study, whether information about textual conventions that have contributed to the choice of specific translations is derivable from a translator's general knowledge of text types, or from experience with translation.

Task-specific information sources for translation are available, or derivable, only in connection with specific translation tasks. These, too, may be sorted according to the distinction between linguistic and extra-linguistic information (cf. 2.4.2.1). Task-specific, *linguistic* information sources cover the information coded in the source language expression, as well as information available in its linguistic context. The

[46] Cf. Popper's view of knowledge growth, described in 2.2.1.

former includes information about the situation type described by the source text, about the linguistic structure of the source expression, and about relations of reference holding between expressions in the source text and extra-linguistic entities. The latter are derivable when the source text is interpreted relative to a specific utterance situation. The division between the information within the source expression and the information contained in its context reflects the fact that the information encoded in a linguistic expression is normally insufficient to determine the intended interpretation of a given utterance of that expression. Kay et al. (1994: 20) describes this interpretation task as "the resolution problem": in order to determine the intended interpretation it is necessary to merge the linguistically encoded information with information derived from the context, or utterance situation, in which the expression is located.[47] With respect to accessibility, we assume that the information coded in the SL expression is easier to access than contextual information: the former is directly available through general knowledge of the source language, whereas the derivation of the latter requires a greater amount of processing effort.

Task-specific, *extra*-linguistic information is derived from world knowledge possessed by, or given to, the individual who interprets, and translates, the source text. Pieces of task-specific, extra-linguistic information have been mentioned in connection with examples (3), (4), and (5) in 2.4.2.1, in order to illustrate how the task of interpreting a source expression may involve reasoning about the facts described in the source text, or in its context. Such reasoning may thus supply information which is not linguistically encoded in the source text. Furthermore, task-specific, extra-linguistic information includes information related to the utterance situation of the source text, such as information about the sender, about the purpose of the source text, and about its spatial and temporal location. It may also cover information about various aspects of the translation situation itself, such as information about the purpose of the translation, which is not necessarily the same as the purpose of writing the original.

---

[47] This is described by Huang (2007:5) as "linguistic underdeterminacy": "… the linguistically encoded meaning of a sentence radically underdetermines the proposition the speaker expresses when he or she utters that sentence."

A certain understanding of the notion of 'translation task' lies behind the present description of task-specific information sources. A translation task may involve translating anything from a single lexical item, or a sentence, to an entire document, such as a handbook or a novel.[48] As stated above, the very characteristic of task-specific information sources is that they are available only in connection with specific translation tasks, and this sets them apart from the general information sources, which are given prior to the translation activity. However, information about the sender, location, and purpose of the source text pertains to the text on a macrolevel, and it will thus be given prior to a concrete translation activity in the case where the task is to translate a subpart of a larger document for which the mentioned information types are known to the translator. Still, we do not find it appropriate to regard these types as general information, as they are associated with specific texts, and are not derivable from translation competence as such.

Perhaps the most important difference between general and task-specific information sources pertains to accessibility: we assume that information available prior to translation is easier to access than information that must be derived during the translation task. In chapter 3 this topic is developed further in discussions of the efforts required to solve translation tasks.

2.4.2.3 Mono- versus bilingual information sources

The third dimension identified in our typology cuts across only a subset of the other information types. Firstly, with respect to the opposition between linguistic and extra-linguistic information, it does not make sense to classify information about the extra-linguistic world as either mono- or bilingual.[49] Secondly, the distinction between general and task-specific information is relevant in the case of monolingual information sources, but not in the case of the bilingual, which we will comment on below.

Monolingual information sources for translation may be divided into those that are given prior to the translation task, and those associated with the translation of a

[48] The notion 'translation task' is further discussed in 3.2.4 and 3.3.1.1.

[49] On the other hand, translation competence models may include components described as "intercultural", or "bicultural" (cf. Hurtado Albir and Alves 2009: 65, 66), but the present typology applies to information, not to knowledge modules.

specific piece of text. Thus, general, monolingual information sources include information about source and target language systems, respectively; they are located, as discussed in 2.4.2.1, in the pre-structured domain of linguistic information. Task-specific, monolingual information sources, on the other hand, fall outside the pre-structured domain; as explained in 2.4.2.2, they cover the information coded in the source language expression, as well as information available in its linguistic context.

We assign only one type of information to the category of *bilingual* information, i.e. information about how source and target language are interrelated with respect to grammars and lexicons. It is our view that bilingual information for translation is located on the level of *langue* — it covers relations between linguistic signs — and this is a consequence of our delimitation of the finite, pre-structured domain of linguistic information (cf. 2.3.2). Since this is determined by the delimitation of language systems, and the distinction between mono- and bilingual information applies only to the linguistic domain, then bilingual information is limited to the correspondence relations between source and target language systems. Thus, we assume that there are no task-specific, bilingual information sources for translation, only general, bilingual information, which, together with general, monolingual information, constitute the pre-structured domain of linguistic information.

At one point we need to make an exception from our principle that the distinction between mono- and bilingual information does not apply to the extra-linguistic domain. As regards textual norms, we explained in 2.4.2.1 that they are not part of language systems, and hence information about textual norms are, in our approach, classified as extra-linguistic. However, since this is information about language use, and since the realisations of textual norms are language-specific, it makes sense to treat information about the textual norms of, respectively, source and target language as monolingual information, and information about how corresponding norms of the two languages differ, can be seen as bilingual.

2.5 Summary

As the present project investigates relations between translationally corresponding texts, a product-oriented approach is necessary. In this chapter, 2.2 with subsections

is a discussion of Karl R. Popper's distinction between the products of behaviour and production behaviour, and its relevance to the study of translation. Following Popper (1979), we have argued that with respect to translation, the study of its products is primary to the study of the translation process, in particular because it is the product and its relation to the original text that gives the process its identity.

The main objective of the present project is to investigate to what extent it is possible to automatise translation in selected English-Norwegian parallel texts instantiating two specific text types. In 2.3 with subsections we have, in accord with Dyvik (1998, 1999, 2005), described a principled limit on predictability in the translational relation. The notion of 'translational relation' covers correspondence relations between language systems as well as between texts and utterances of different languages. We assume that the linguistically predictable part of the translational relation exists on the level of correspondences between *langue* units, and that it is computable from pre-structured information about the source and target language systems and their interrelations. Then, with reference to specific original texts and their translations, the computability issue is a question of to what extent the translational correspondences contained in that body of parallel texts fall within the set of linguistically predictable correspondences between the given source and target language. In line with Dyvik (1999), we have defined 'literal translation' to be the same as 'linguistically predictable translation'.

For the purpose of developing a typology of information sources for translation, we have discussed certain basic concepts in 2.4.1 with subsections. 'Information', in the sense of communication theory, is a purely quantitative notion, something that is either present or not, and it exists independently of interpretive processes (Dretske 1981). 'Informational content', or the message carried by a specific signal, is of a different kind: it is determined by the existence of the information to be transmitted, and by the amount of information carried by the signal; it demands that the information transmitted is identical to the information generated at the source, and it is influenced by background information available to the recipient of the signal (Dretske 1981). Further, 'knowledge' is described by Dretske (1981) as information-supported belief, an account which makes knowledge a property of individual minds.

In our investigation of translation, we have rather put emphasis on Popper's notion of 'objective knowledge' (cf. 2.2.1). Objective knowledge exists in the form of shared content of different knowledge states in different human minds, and hence it may in principle exist independently of individual knowing subjects. Thus, in 2.4.1.5 we have argued that since different bilingual informants may share judgments concerning the appropriateness of specific translations of given source texts, we assume the existence of objective knowledge which can be shared by different translators and which can be exploited by various translation strategies. Moreover, we regard the objective knowledge of translators as a pool of information that is available prior to translation.

Our typology of information sources for translation is presented in 2.4.2 with subsections. The information sources are sorted along three different dimensions, each containing a binary division. Firstly, we distinguish between linguistic and extra-linguistic information; secondly, between general and task-specific information, and, thirdly, within the linguistic domain, between mono- and bilingual information. Figure 2.1 presents an overview of the information typology.

The most important distinction in the typology is that between linguistic and extra-linguistic information as it is associated with the limit of computability in the translational relation. The pre-structured domain of information about the source and target language systems and their interrelations, which defines the linguistically predictable part of the translational relation, is a subpart of the linguistic information sources for translation. In 2.4.2.1 we have tied the limit of the linguistically predictable to the delimitation, or individuation, of language systems, and we have further argued that the conceptual individuation of a language system relies on empirical facts about language use, and the delimitation of the relevant language community, together with certain chosen presuppositions regarding the purpose for which the description of the language system is meant to be used.

<table>
<tr><th></th><th colspan="2">linguistic</th><th>extra-linguistic</th></tr>
<tr><td>task-specific</td><td colspan="2">• information coded in the source text, i.e. information about the described situation type, about the linguistic structure of the source string, and about reference relations holding between expressions in the source text and extra-linguistic entities
• information available in the linguistic context of the source text</td><td>• information derived by reasoning about facts described by the source text, or in its context
• information about the utterance situation of the source text
• information about the translation situation</td></tr>
<tr><td rowspan="2">general</td><td>monolingual</td><td>bilingual</td><td rowspan="2">• general background information about the world
• domain-specific technical information
• information about textual norms
• information derivable from translation training and practice</td></tr>
<tr><td>• information about the SL
• information about the TL</td><td>• information about interrelations between source and target language systems</td></tr>
</table>

Figure 2.1. A summary of the typology of information sources for translation. The shadowed boxes indicate what is included in the pre-structured domain of linguistic information.

3 Analytical framework

3.1 Overview

This chapter is divided into two main parts. In the first part, 3.2 with subsections, we start by presenting, in informal terms, the information-theoretic concepts of computability and complexity, as well as certain related notions. Next, the relevance of complexity theory for studies of natural language is discussed before various approaches to the notion of 'linguistic complexity' are presented. On that background we introduce our own framework for describing translational complexity, and, finally, discuss the notion of 'computability' in relation to translation.

In the second part, 3.3 with subsections, the correspondence type hierarchy is presented in detail. Each correspondence type is described in terms of (i) the linguistic characteristics of the relation between source and target string, (ii) the amounts and types of information needed to translate, and (iii) the processing effort required by the translation task.

3.2 Computability and complexity

As mentioned in 1.1 and 1.2, the notions of 'computability' and 'complexity' are central concerns in the present investigation. So far in our discussion, 'computable' has been understood as 'solvable by a specifiable procedure' (cf. 2.3.2), and with respect to 'complexity', it has been introduced in the sense of 'translational complexity' and linked with the translator's need for information.[1]

Information science provides the formal tool of computational complexity theory for the purpose of measuring the inherent complexity of computable tasks. For

[1] Cf. 1.1, and 1.3–1.3.2.

several reasons this tool cannot be applied to our investigation, but we want to give an informal description of the information-theoretic notions of 'computability' and 'complexity' in order to reach an intuitive understanding of the concepts. This is meant to throw some light on the motivation behind our concept of translational complexity.

3.2.1 An informal look at the information-theoretic concepts

'Computability' is a property of tasks: if a certain task can be solved by a specifiable procedure, then it is a computable task. This means it is possible to write a procedure leading step by step from an initial state to a final state, and in the final state the task is solved. This kind of procedure is called an *algorithm*. Thus, an 'algorithm' is a well-defined sequence of steps which always gives a result, i.e. the final state of a computation.[2] Likewise, a 'computation' is a step-by-step procedure solving a certain task according to the specifications of an algorithm.

In the context of computational complexity theory 'complexity' is a mathematical property which concerns the amount of time and space needed to solve a computable task. Thus, computability is a prerequisite for complexity measurements. We assume that any computation requires certain resources used by the computing device (be it a computer or the human brain), and these resources are processing time and memory space. Computational complexity is a measure of the rate at which a specific computation consumes time and space (van de Koot 1995: 41). In the following we will refer to computable tasks as *problems*, i.e. problems to be solved.

Barton et al. (1987: 7) point out the difference between problem complexity and algorithm complexity. These aspects are in principle independent of each other: it is normally possible to write different algorithms for solving the same problem, and there are cases where the same algorithm may be used to solve different problems. Different algorithms written for the same problem may be more or less efficient than each other, and for this reason the complexity of problems is not measured with reference to specific algorithms.

---

[2] Cf. van de Koot (1995: 40–41).

The task of complexity analysis is to study the structure of a problem, i.e. what Barton et al. (1987: 4) refer to as the *information processing structure* of a problem. This structure determines how the problem can be solved in the most efficient way, i.e. how to specify an algorithm that does not consume larger amounts of time and space than necessary. For instance, given the task of looking up a certain word in a dictionary, this can be done by an inefficient algorithm where the search starts from the beginning of the dictionary, and checks every entry word until the given search word is found. A more efficient algorithm would exploit the fact that a dictionary is an alphabetically sorted list. This could be done by splitting the list into two equal parts, and then checking the beginning and end of each part in order to find out whether the search word is contained in the first or the second part. The search then continues in the relevant half of the dictionary, and the algorithm repeats the splitting into halves until the search word is found. This algorithm is called *binary search*, whereas the former algorithm can be described as search by brute force, or as *exhaustive search*. For obvious reasons a binary search will be more efficient than the exhaustive method unless the search word happens to be located in the beginning of the dictionary. An exhaustive search would be necessary only if the search space is an unstructured list.

Computational complexity theory works not only by analysing the complexity of specific problems, but also by comparing new problems to problems for which the complexity is known. Such comparisons group problems into so-called *complexity classes*, which are classification schemes based on measurements of problem complexity.[3] In this context complexity theory makes a distinction between, on the one hand, problems at type level and, on the other hand, instances of given problems.[4] Such instances can be seen as specific computations where the problem is to be solved for a given input. The amounts of time and space needed to solve a problem are correlated with its size, i.e. the length of the input to a computation. In general, the longer the input the greater the need for time and space. In complexity theory

[3] For more information on this, see chapters 1 and 2 in Barton et al. (1987).

[4] See e.g. van de Koot's description of the technique for comparing problems with respect to complexity: "Take a problem of known complexity… Then construct an efficient mapping from instances of the problem of known complexity to instances of the new problem…" (1995: 45).

such correlations can be expressed by mathematical functions relating the size of the problem to the required amounts of time and space, and for the purpose of sorting problems into complexity classes, it is particularly important to know the *order of growth* of these functions (van de Koot 1995: 41). This refers to the rate at which the consumption of time and space increases when there is a growth in the size of instances of specific problems, and this rate serves to group problems into complexity classes. Different classes will have different rates, and classes of harder problems will have higher growth rates than classes of less hard problems.

The sorting of problems into complexity classes provides a precise measure of how hard it can be to solve a problem in typical instances as well as in so-called worst cases. Given a certain complexity class, no algorithm can do better than a certain level of performance. Problem complexity thus delimits the efficiency of optimal algorithms: it is impossible to write algorithms which take less time in the worst case than the amount of processing time required by the inherent complexity of the problem (Barton et al. 1987: 8).

A few main classes of complexity could be mentioned. First, there is a class of problems where the increase in the consumption of processing resources is proportional to an increase in the problem size. That is, if the input grows in size by some integer n, then the growth rate of the processing effort is also n. In mathematical terms such problems are solvable in linear time, and represent the least hard problems. Second, there is a class of problems where the growth rate can be described as $c \cdot n^k$, where c and k are constants, and n is the size of the problem instance (van de Koot 1995: 41). If the input grows in size by n, then the increase in processing time is proportional to $c \cdot n^k$. Such problems are solvable in polynomial time, and within this complexity class there is a considerable degree of variation: cases where the constants c and k have low values are much less demanding to solve than cases where their values are high, and a sharp increase in the value of n will also give a high growth rate. Third, there is a class of harder problems where the growth rate is expressed as $c \cdot k^n$ (van de Koot 1995: 41), so that if the input grows in size by n, then the increase in processing time is proportional to $c \cdot k^n$. These problems are solvable in

exponential time, which means that only the smallest increase in the problem size causes a great increase in the consumption of processing resources.

The mentioned classes are only a few main categories; several other classes and subclasses have been identified in complexity theory. An important result of this classification is the division between *tractable* and *intractable* problems. We will here not go into the mathematical properties behind this distinction (see Barton et al. 1987: 8–10), but only roughly indicate that tractable problems are solvable within polynomial time, whereas the intractable ones are "problems for which only exponential solution algorithms are known" (1987: 9). In practice, this means that tractable problems are solvable in reasonable time on an ordinary computer, while intractable problems are not.

Closely related to the distinction between tractable and intractable problems is the notion of an 'efficient algorithm'. Algorithms working within the limits of polynomial time are regarded as *efficient*, while algorithms exceeding this upper limit on processing time are not. Keeping in mind the fact that algorithm complexity is independent of problem complexity, it is appropriate to mention the following point made by Barton et al. (1987: 10): "… if a problem is efficiently solvable at all, it will in general be solvable by a polynomial algorithm of low degree."

3.2.2 The relevance of complexity theory for natural language

Human language processing may be seen as instances of computation. In language comprehension the human brain processes the input in order to construct an interpretation of it, and speech production involves producing an output for the purpose of expressing some intended meaning. These are types of computation where the brain uses time and memory to process information drawn from linguistic knowledge as well as from knowledge of the world. Such processes fall within the study of psycholinguistics, and shall not be dealt with here.

Viewing human language processing as computation makes it natural to apply the tools of computational complexity theory to natural languages, and this is clearly a way of gaining insight into the challenges mastered by the human language ability. A different avenue of research is to study frameworks for language descriptions, i.e.

grammar formalisms, in terms of computational complexity. Van de Koot (1995: 38) mentions several publications dealing with the application of complexity theory to natural language and linguistic theory.

With respect to complexity analysis of natural languages, we may briefly present a few contributions found in the literature. Because it is difficult to investigate the algorithms used by the human brain for language processing, little is known about the very processes.[5] Hence, Barton et al. (1987: 2) point out that in order to study human language processing, computational complexity analysis is a most appropriate tool since it is independent of solution algorithms as well as of computing devices. In the case of human language processing we have access to its input and output, i.e. natural language, but we do not have direct access to the inside workings of the processes themselves. That is, however, no problem for complexity analysis since it pertains to problem structure (cf. 3.2.1) and may be applied to available linguistic data. Barton et al. (1987: 4) hold the view that "…there is every reason to believe that natural language has an intricate computational structure that is not reflected in combinatorial search methods."[6] In other words, they believe that if the outcome of a natural language problem is intractability, then it is likely that some of the structure of the problem has not been detected.

Van de Koot (1995: 39) observes the following with respect to the complexity of natural languages: "The picture that emerges … is that natural language computations are computationally intractable … But … they have the useful property of being on the verge of tractability: their solutions are hard to find but easy to check once found." Van de Koot (1995: 39) makes an interesting point about the usefulness of applying complexity analysis to natural language computations: studying the computational complexity of language problems such as comprehension makes it possible to "relate language computations to other computations whose structure we understand and provide a design target for language algorithms (i.e. for algorithmic characterizations of language computations)." This means that complexity analysis of natural

[5] Cf. the discussion of process-oriented translation studies in 1.4.1.3.
[6] Combinatorial search is search by brute force, i.e. to try all possible combinations; cf. 3.2.1.

language problems may facilitate the design of algorithms for natural language processing within the field of human language technology.

The application of complexity analysis to grammar formalisms also deserves to be mentioned. It is regarded as necessary requirements of grammar formalisms that they, on the one hand, have the sufficient means for expressing all possible structures in any language, and that they, on the other hand, rule out the description of structures that do not belong to any possible language. These requirements are concerned with what is referred to as the *generative capacity* of grammar formalisms, or their generative power. A consequence of the view of Barton et al. (1987: 4) that natural language has a computational structure that does not require combinatorial search methods, is that a grammar formalism should not be able to express linguistic structures for which the comprehension task would require exponential solution algorithms. If that is the case, then the formalism is too "powerful"; it will generate more than the structures found in natural languages. According to Barton et al. (1987: 4), complexity analysis can be used, then, to identify the parts of a grammar formalism that allow the generation of linguistic structures more complex than natural ones. In other words, complexity theory can be used to weed out over-generation in grammar formalisms, and this is a very useful tool for the study, and development, of linguistic frameworks for language description.

3.2.3 Linguistic complexity

Having discussed the information-theoretic notion of 'complexity', we now move on to complexity in the context of language. This is not a computational concept; it may (although not exclusively) be related to learning rather than to processing, and it may be related to subjective experience, whereas computational complexity is an objective property of a computable task.

In the structures of natural languages we may intuitively perceive different degrees of complexity, for instance in connection with language learning. A language with a rich inflectional system may be experienced as very complex by a foreign language learner if his or her mother tongue is a language with little morphology. E.g., a Norwegian learner of German will normally find it challenging to acquire

command of four different types of grammatical case in German, since Norwegian exhibits no grammatical case distinctions apart from an opposition between nominative and accusative in pronouns. And for a Norwegian the Finnish case system, needless to say, seems overwhelmingly complex with its fourteen different kinds of grammatical case.

An interesting challenge for linguists, then, is to find a principled way of describing linguistic complexity. This opens up for questions such as what kinds of phenomena, or which aspects of language, are involved in linguistic complexity, and how can a given structure be described as more, or less, complex than other structures it may be compared with? In 3.2.2 we have presented computational complexity theory as a tool suitable for studying complexity in natural language, but we have not mentioned that such studies cannot be done without certain prerequisites. Since complexity analysis is, basically, to analyse the structure of a problem for the purpose of finding a solution to it, it is necessary to transform language into a computational problem before it can become the object of complexity analysis. Firstly, this requires a computationally implementable formalism for language description, and, secondly, a grammar written in that formalism for the relevant language. Thirdly, it is necessary to reduce the study of structures in that language to the task of deciding whether the structures belong to the language described by the given grammar. This is commonly called a *recognition* problem,[7] and it presupposes a conception of language as the set of strings, or expressions, generated by a certain grammar. Complexity analysis is then applied to the recognition problem, and the result of the analysis serves eventually as an estimate of the complexity of the given language structures. In other words, using complexity analysis to investigate linguistic complexity involves the construction of search tasks performed on the basis of formal grammars.

As a consequence, there are many language researchers who want to study linguistic complexity, but do not apply computational complexity analysis. The field may not be seen as relevant, or the necessary prerequisites may be lacking. But irrespective of method, such studies anyway need a clear understanding of

[7] See e.g. van de Koot (1995: 46ff).

complexity, and several researchers have made contributions in this respect. One example is Dahl (2004), whose approach to linguistic complexity is based on information theory, in particular the principles of signal transmission. His project is to study, from a diachronic perspective, how the complexity of language systems evolves and is maintained. In the context of language, Dahl describes the notion of 'complexity' as "[not] synonymous with "difficulty" but as an objective property of a system — a measure of the amount of information needed to describe or reconstruct it" (2004: 2). With respect to the information-theoretic notion of 'complexity', he says informally that "the complexity of an object would … be measured by the length of the shortest possible specification or description of it" (2004: 21).

Dahl points out that available background information may have consequences for how long the shortest possible description needs to be, and this calls for a distinction between absolute and relative complexity (2004: 25–26). The complexity of an entity *relative* to a certain amount of available information is measured by the length of the specification needed, in addition to the available information, to describe the entity. *Absolute* complexity, on the other hand, pertains to the total amount of information needed to describe the entity.

In his approach to complexity in languages, Dahl describes a language as consisting of resources and regulations: resources are the building blocks of linguistic expressions, while regulations determine how the resources are used correctly (2004: 40–42). He then argues that measuring the complexity of the resources of a language is distinct from measuring the complexity of its regulations. The former pertains to the size of the inventories that are included in the resources, whereas the latter relates to the complexity of the expressions of the language, and is understood by Dahl as *system complexity* (2004: 42–43), which is the central concern of his study of complexity in languages.

Miestamo (2006) presents other approaches to studies of linguistic complexity, and his context is the study of language typology, and in particular studies where languages are compared with respect to the complexity of specific domains of grammatical functions, such as aspect, tense, negation, or definiteness. He, too, applies an

opposition between absolute and relative complexity in natural language, although his explanation of relative complexity is somewhat different from that of Dahl (2004):

> "The absolute (or theory-oriented) point of view looks at complexity in terms of the number of parts in a system, or in information-theoretic terms (Shannon 1948) as the length of the description a phenomenon requires (cf. Dahl 2004). The relative (or user-oriented) point of view pays attention to the users of language and defines as complex what makes processing, acquisition or learning more difficult." (Miestamo 2006: 346)

Miestamo (2006: 348–349) argues that to be able to study linguistic complexity in general terms, it is not sufficient to investigate complexity in relation to one group of language users and not another. The reason is that it appears to be arbitrary which group(s) of language users experience(s) a certain linguistic property as difficult, or complex (cf. the above example of grammatical case). Hence, it is problematic to describe linguistic complexity relative to language users, and Miestamo's conclusion is that the absolute, or theory-oriented, approach is more fruitful in order to achieve precise criteria for the description of linguistic complexity.

Like Dahl (2004), Miestamo applies an informal interpretation of the information-theoretic notion of 'complexity' when he defines a complex phenomenon as "something requiring a longer description than a less complex phenomenon" (2006: 349). He argues that this definition provides an objective criterion for cross-linguistic comparisons with respect to specific grammatical properties, and in Miestamo (2006) it is applied to a typological study of the functional domain of negation. More specifically, it is used to study the relationship between grammatically expressed functions and their formal encoding, and Miestamo observes that "a language where more grammaticalized distinctions are made in a given functional domain, requires a longer description for that functional domain than a language where less distinctions are made" (2006: 349). This is the standard of measurement for cross-linguistic comparison, and it is a quite separate topic whether it is difficult to comprehend or

acquire the linguistic distinctions expressed in a given domain. The latter illustrates the relative point of view in relation to linguistic complexity.

3.2.4 Translational complexity

Our approach to complexity in translation is an attempt at creating a fairly precise frame of reference for the characterisation of translational correspondences, although it is not an aim to find a mathematically exact description of complexity in relations of translation. We cannot apply the tools of computational complexity analysis, as the prerequisites needed for that (cf. 3.2.3) are not available in our project. Like Dahl (2004) and Miestamo (2006), we adhere to an absolute point of view in our description of complexity: translational complexity is analysed in terms of a quantifiable, objective commodity (information), and independently of the competence of the translator who has produced a specific translation. Our approach differs somewhat from those adopted, respectively, by Dahl (2004) and Miestamo (2006) in the sense that they measure linguistic complexity as *length of description*, whereas our analysis of translational complexity is based on *amounts and types of information* needed in translation, and in several respects our investigation rather resembles the techniques of computational complexity analysis. However, since we are not in a position to quantify information in exact terms, we have to rely on more intuitive notions.

Firstly, complexity analysis applies to computable tasks, described as problems (cf. 3.2.1), and in a similar way we want to characterise the degree of complexity in translational correspondences by viewing them as *translation tasks*, i.e. the task of producing a particular target expression on the basis of the information encoded in the given source expression together with other information sources. This could more precisely be seen as describing the complexity of a specific solution to a translation task.[8] Then, the notion of a 'search task' becomes a common denominator between computational problems and translation tasks. An algorithm solving a problem carries out a search task, i.e. the search for the solution, or the final state of the computation. Likewise, a translation task can be regarded as a search task: the search for the

[8] This is a narrower notion of 'translation task' than the one presented in 2.4.2.2. The concept is further discussed in 3.3.1.1.

information needed to interpret the source text correctly, and for the information needed to produce the target expression.

Next, complexity analysis uncovers the structure of a problem for the purpose of finding a solution to it, and this analysis is independent of possible algorithms that may solve the problem. As a parallel to this, we study the linguistic relation between source and target expression in translational correspondences irrespective of how the translation has been produced. Given our view of translation as a search task, this is an analysis of the *structure* of the search task. In this context we may regard the source expression as the initial state of the translation task, and the target expression as its final state. Since we do not have direct access to information about the translation process (cf. 1.4.1 and 1.4.1.3), what we can do, in order to measure the complexity of a specific translation task, is to study the relation between the initial and the final state by analysing how source and target expression correspond (or not) with respect to linguistic properties. On the basis of syntactic and semantic correspondence relations between source and target string (divergences included), the structure of a translation task can be analysed in terms of the types and amounts of information needed to solve it. We emphasise that this approach to translational complexity is no attempt to describe aspects of the translation process or any possible algorithm producing the translation. Rather, it aims at capturing the necessary requirements for solving the translation task.

Further, there is a parallel between our analysis and complexity theory with respect to the measuring of complexity. In complexity theory this is measured in terms of the consumption of resources, i.e. processing time and memory space, needed by a computation (cf. 3.2.1). These factors reflect the *processing effort* required by the computation. In our analysis translational complexity is measured in terms of the need for information in a translation task, and the processing of this information is a kind of computation that will consume time and space. As already made clear, it is, however, not possible in our study to calculate numerically the precise amounts of processing resources needed for specific translation tasks, but we do distinguish between classes of translational correspondences according to certain assumptions regarding the consumption of resources, or processing effort. These

assumptions may be said to concern the *weight* of the search task and pertain to two closely related topics: (i) the extent to which the various types of information needed to carry out a translation task can be represented in a finite way, and (ii) the amount of effort required in order to access and process them. Naturally, the processing of easily accessible information requires a smaller effort than the processing of less easily accessible information. E.g., information that can be looked up as easily as information presented in a table, is directly accessible, while the accessing of information that must be derived through linguistic analysis of an expression, or through inferencing, involves a greater effort. These topics will be revisited in the presentation of the correspondence type hierarchy later in this chapter.[9]

Finally, our sorting of types of translational correspondences resembles the way in which complexity theory groups computable tasks into complexity classes. As explained in 3.2.1, complexity classes are distinguished on the basis of the rate at which the consumption of resources grows in proportion to an increase in the size of the computable task. The growth rate indicates how hard it is to solve the problems of a given class in the best, typical, and worst cases, respectively. The presentation of the correspondence type hierarchy will show that the four types are distinguished in terms of a lower bound on how easy it can be to produce the target text in a translation task, i.e. which types of information, and how large amounts of them, the translator must *at least* have access to within a certain class of correspondences. The four types are also distinguished in terms of an upper bound: each correspondence type is associated with a certain set of sufficient information types, so that translation tasks requiring information types in addition to the set associated with a given correspondence type, are of a more complex type. In the present approach differences in the degree of translational complexity cannot be measured *within* the different correspondence types. Still, it is our hope that the distribution of the different correspondence types in a body of parallel texts may reflect the degree of translational complexity in an interesting way.

---

[9] These two topics have previously been mentioned in 1.2 and in 2.4.2. The issue of finiteness concerns the pre-structured domain of linguistic information which defines the limit of predictability in the translational relation; cf. 2.3.2. The issue of processing effort has so far been mentioned in general terms only in connection with the information typology presented in 2.4.2 with subsections.

3.2.5 Computability in relation to translation

The notion of 'computability' is defined in 3.2.1 as a property of tasks: a task that can be solved by a specifiable procedure is computable. By *computing a translation* we understand creating a target text by means of software for automatic translation and without the aid of a human translator. Previously, in 2.3.2, we have drawn a principled limit on computability in the translational relation between two languages: this relation is computable insofar as it is *linguistically predictable*, i.e. to the extent that the translation of a given source text can be predicted, as one possibility, by means of the information encoded linguistically in the original, together with pre-structured information about the source and target language and their interrelations. These interrelations represent the translational relation on the level of system units, which is distinct from the translational relation on the level of linguistic usage (cf. 2.3.1). As explained in 2.3.2, there are instances of translation that cannot be accounted for by pre-structured linguistic information alone; hence we assume that the extension of the translational relation on the level of *langue* is a subpart of the extension of the translational relation on the level of *parole*. This entails the further assumption that there is a part of the translational relation that cannot be automatised, since we regard the *langue* relation as the limit on computability.[10] With reference to automatic translation, we thus understand the *computable* part of the translational relation to be the same as the *linguistically predictable* part of it, and computable translation tasks are solvable within the pre-structured domain of linguistic information, as previously explained in 2.3.2.

In 3.2.2 we have argued for viewing human language processing, of which translation is one kind, as computation. The translational relation on the level of *parole* covers what bilingually competent humans are able to translate, which means that in relation to human translation, we will by 'computable' understand 'translatable'. In the same way as computability is a prerequisite for computational complexity measurements, translatability is a prerequisite for analysing translational complexity. To avoid confusion, we will in the context of translation restrict the

[10] I.e., we assume that only the linguistically predictable translations are computable.

notion of 'computability' to the linguistically predictable part of the translational relation, and use the notion of 'translatability' in connection with translation tasks that humans are able to solve. Although we find it appropriate, in the context of complexity analysis, to view human translation as a partly computable task, we will not in general speak of the human activity of translation as *processing* or *computation*, as such analogies do not contribute to keeping a clear distinction between human and automatic translation.

The notion of 'translatability' can hardly be mentioned without evoking the question whether there is a limit to translatability — whether any translation task can be solved by sufficiently competent human translators. There may be cultural differences, as well as differences between source and target language systems, which may force the translator to paraphrase a given source expression to the extent that the target expression appears as a rewriting rather than a translation. It is then a question of definition how much a target text may diverge from the source while still functioning as a translation.[11] The discussion of the translatability issue falls outside the present project, but has received considerable attention within translation studies. Our focus is on delimiting the computable within the translatable.

3.3 Translational correspondence types

Our empirical investigation, to be described in chapter 4, is a classification of translationally corresponding strings into four different types, introduced in 1.3 with subsections. In simplified terms, type 1 correspondences are cases of a full linguistic match, structurally as well as semantically, between source and target string; type 2 correspondences allow minor mismatches on the structural level, but none on the semantic; in type 3 correspondences there can be major structural divergences while there is still a semantic match, and in type 4 correspondences there are semantic as well as structural mismatches between source and target string. Instances of each correspondence type are identified through syntactic and semantic criteria, and the types are related to each other in a hierarchy, reflecting an increase, from type 1 to 4,

[11] Cf. the prototypical view of 'translation' in Halverson (1998), discussed in 1.4.1.

in the amount of information necessary to produce the target expression. This need for information is in turn an indication of the degree of translational complexity in the translation task. Given the assumptions of our analytical framework, correspondence types 1, 2, and 3 fall within the limit of computability in the translational relation, whereas type 4 correspondences are not computable as they involve information sources not included in the pre-structured domain of linguistic information. In 3.3.2–5 with subsections, the correspondence type hierarchy is presented in detail.

3.3.1 General aspects of the classification of translational correspondences

Certain topics involved in the classification of translational correspondences are relevant to the whole set of correspondence types. To avoid repetition in the presentations of each type, these topics are discussed in 3.3.1.1–4.

3.3.1.1 The notion 'translation task'

The notion of 'translation task' is important in our approach to translational complexity. The concept is used in chapters 1 and 2 in a general sense which covers the task of translating anything from a single lexical item, or a sentence, to an entire document, such as a handbook or a novel.[12]

Section 3.2.4 introduces a more precise sense, where a 'translation task' is understood as the task of producing a particular target expression by means of various sources of translator's information together with the information encoded in the source expression, given its relevant interpretation. This can be seen as the specific task of translating a textual occurrence of an expression a of a certain source language (L_1) into expression b of a given target language (L_2), i.e.: $a_{L1} \rightarrow b_{L2}$

It should be noted that we have decided to keep the subtask of source text disambiguation apart from the translation task. The reason is that our analysis of complexity pertains to translation tasks only, and hence we do not consider the problem of source text disambiguation. By the *relevant interpretation* of a_{L1} we here mean the interpretation which lies behind the chosen translation b_{L2}. How the

[12] The general notion is commented on in 2.4.2.2.

translator has identified the relevant interpretation falls outside the scope of our analysis.

There is also a somewhat different, and extensionally wider, sense of 'translation task', namely the task of translating a textual occurrence of a certain source expression a_{L1}, given its relevant interpretation, into a specific target language L_2. In such general translation tasks the source expression a_{L1} corresponds with a set of possible target expressions (T_{L2}), i.e.: $a_{L1} \rightarrow T_{L2}$. Thus, whereas translation tasks of the first kind involve correspondences between specific, single expressions of L_1 and L_2, translation tasks of the second, more general, kind involve correspondences between specific expressions of L_1 and sets of translations in L_2, from which the translator makes a motivated choice.

Notably, what we have aimed at in the analysis of translationally corresponding string pairs is to measure the complexity in a collection of concrete translation tasks (i.e. string pairs) where the chosen target expression is only one of a set of possible translations in L_2. Thus, the complexity measurement applies to specific translation tasks $a_{L1} \rightarrow b_{L2}$, and the analysis of each string pair is an attempt to describe the complexity of the selected task *solution* in relation to the source expression a_{L1}, given its relevant interpretation. We do not consider the complexity of the translation task that is not solved yet; that would amount to analysing the complexity of the general translation task ($a_{L1} \rightarrow T_{L2}$), which has a set of possible solutions.

In the presentation of the correspondence type hierarchy each type is related to the specific notion of translation task. The purpose of describing the hierarchy in terms of translation tasks is to explicate the information requirements of each type, i.e. the types and amounts of information needed in order to produce the chosen solutions to specific translation tasks. Still, it should be noted that when we refer to the task of producing a particular translation from a given source text, the notion of a 'task' cannot be related directly to the translator's situation. In many cases the translator might have chosen less complex (i.e. literal) solutions, which means that the task of translating the given source expression in general may be simpler than the solution actually chosen. Hence, what we aim to describe is the complexity and information requirements of a specific solution to a translation task. In the correspondence type

descriptions we will discuss complexity measurement, or type identification, as part of the task. By this we do not mean the identification of the simplest possible type of solution to a given task, but the identification of the complexity type (1, 2, 3, or 4) of the solution that has been chosen by a translator.

Although the orientation of the present study is different from that of Toury (1995), there is an interesting parallel between our approach to translational correspondences and his notion of 'coupled pairs'.[13] Toury (1995: 77) defines 'coupled pairs' as correspondences between specific translation problems in the source text (i.e. tasks to be solved), and their solutions in the target texts. In his view, such coupled pairs should be the starting point for the description of translational phenomena, and he emphasises that in coupled pairs, source problems and target solutions "should be conceived of as determining each other in a mutual way" (1995: 77).

As stated in 3.2.5, our investigation aims at delimiting the computable part of the translational relation within the domain of the translatable. Information about source and target language and their interrelations defines the linguistically predictable, or computable, set of correspondences between SL and TL. When we analyse the degree of translational complexity in selected parallel texts, we describe how translation tasks are solved by certain translators, and the classification of translational correspondences is thus meant to reflect the complexity in the task of generating automatically the translations that some humans have produced.

3.3.1.2 Criteria for distinguishing and describing correspondence types

In the present approach a set of three criteria is used to distinguish between the four types of translational correspondences. The first criterion pertains to the linguistic characteristics of the relation between source and target string, characteristics which show the degree to which there exist implications between relations of equivalence between source and target string. The second criterion concerns the amounts and

[13] As previously observed in 1.4.1.1, Toury describes his study as "an attempt to gradually reconstruct both translation decisions and the constrains under which they were made" (1995: 88), and this is his motivation for identifying units of comparative analysis.

types of information needed to produce the translation, and may be conceived of as the *structure* of the search task involved in translation. The third criterion deals with the processing effort required by the translation task, which may be seen as the *weight* of the search task.[14] As explained in 3.3.1.1, each correspondence type is to be described in terms of the notion of translation task, which will be decomposed into the three subtasks of source text interpretation (or analysis), complexity measurement, and target text generation. The subtask of complexity measurement will be referred to as *type identification* in the presentations of the four correspondence types.

Firstly, translational correspondences are classified in terms of the linguistic properties of the relation between source and target strings: these properties are the criteria through which tokens of each correspondence type may be identified. If there is some degree of structural similarity in a given language pair, then there will be a certain set of linguistic structures in the source language sharing properties with translationally corresponding structures in the target language. Information about such correspondence relations is included in the general information about interrelations between source and target language (cf. 2.4.2.2). I.e., we assume that information about how constructions in the two languages are translationally related (or unrelated) to each other is information available prior to translation. In cases exhibiting a high degree of structural relatedness between the source and target expression, the translation task is easy to solve, while it is harder in cases where original and translation are structurally unrelated.

The presentation of the correspondence type hierarchy will show that in cases where similar structures of respectively SL and TL are translationally matched, there will exist relations of equivalence between source and target string, and, also, implications between such equivalence relations.[15] These relations of equivalence concern different linguistic levels: syntax, semantics, and pragmatics. The discussion of the correspondence types will illustrate that in cases where source-target equivalence with respect to syntax implies equivalence also with respect to semantics and pragmatics the degree of translational complexity is low, and that as translational

[14] Cf. 3.2.5, where we have previously commented on the structure and weight of translation tasks.
[15] Dyvik (1999: 229–230) describes translational complexity in terms of such implications.

complexity increases, such implications exist to a lesser degree. The domains within which such implications hold are assumed to be delimited by information about the translational relationship between source and target language. In this context 'equivalence' should not be taken as identical to the notion of 'translational equivalence' discussed in 1.4.1.1, but rather be understood as 'linguistic matching relations'. We have nevertheless chosen the expression *equivalence* since we regard it as more precise than *match*, and in this context it may be understood as equivalence between original and translation with respect to specific linguistic properties.

Secondly, each correspondence type will be characterised with respect to the amounts and kinds of task-specific information required to translate source language strings. According to the discussion of translational complexity in 3.2.4, this may be interpreted as an analysis of the structure of the search task involved in translation. The search task is twofold: there is, first, the search for the information needed to interpret the source text correctly, and, second, the search for the information required for producing the target expression. Solving the first subtask, interpretation, involves using the information encoded in the source string together with information about the source language system. From the perspective of computing the translation, there is also a need for an intermediate subtask of complexity measurement, i.e. diagnosing the degree of complexity in the translation task. This requires information about the linguistic structure of the source string and information about the interrelations between the source and target language systems.[16] The final subtask, generation of the target string, requires information about the interrelations between source and target language, as well as information about the target language in isolation.

With respect to these subtasks, we shall discuss how the need for information is correlated with the degree of translational complexity in the different correspondence types. The decomposition of the translation task into analysis, complexity measurement, and generation should not be taken as assumptions concerning how a translation task is solved by a human translator; rather, it is a description of an idealised,

[16] This topic is discussed further in 3.3.1.4.

minimal procedure on which possible translation algorithms may be based, i.e. a description of the information processing structure (cf. 3.2.1) of the task of computing a translation.

Thirdly, each correspondence type is characterised with respect to what is described in 3.2.4 as the weight of the translation task, i.e. the amount of required processing effort. For each correspondence type the necessary information sources are for this purpose viewed in relation to two topics: the extent to which they can be represented in a finite way, and the amount of effort required in order to access and process them. The decomposition of the translation task into three subtasks is relevant also for these topics as the amount of required effort varies not only among the types of translational correspondences, but, as we shall see, also among the subtasks.[17]

3.3.1.3 The notion 'necessary information'

The string pairs we have analysed are produced by human translation, and hence they represent translation tasks solvable by bilingually competent language users (insofar as each source string does have a corresponding target string). A subset of these string pairs represent computable tasks, which can be solved, given certain assumptions, by pre-structured linguistic information sources alone. In the classification of string pairs into correspondence types it is an aim to identify the information sources that are *at least* necessary in order to compute (or, if not computable, to produce "manually"), each target string.[18] With respect to the subset of computable translation tasks, we will argue that in some cases it is not necessary, in order to generate the target text, to analyse the source text further than to the level of syntax (types 1 and 2), whereas in other cases a semantic analysis of the source text is also required (type 3).

Again, this must not be interpreted as a way of conceptualising the translation process. It is not plausible that a human translator, after having read a text string in the source language will consider only its syntactic structure, and disregard its semantic content as well as accompanying contextual information, because he or she

[17] Cf. 3.3.1.4, as well as the discussions of each correspondence type.
[18] Cf. the remarks in 3.2.4 on how the correspondence types are distinguished from each other.

is aware that a target string with a structure identical to that of the source string is an appropriate translation. Rather, a competent translator must continually pay attention to the meaning and context of the source text, and in cases where he or she chooses a word-by-word translation, that is done especially because it seems appropriate after having considered the meaning and context.

But from the perspective of a system for automatic translation, it is a formidable task to process the various types of information associated with even a very short text. To analyse the syntactic structure of a limited source string is, on the other hand, a computationally tractable problem, and hence a good starting point for identifying the degree of complexity of the given translation task. If the system is able to decide that the target language offers syntactic structures matching those found in the source string and with corresponding compositional semantic properties, then we assume that an efficient strategy for automatic translation is to refrain from analysing the meaning and context and simply proceed to identifying the corresponding target words and generating the translation directly from the source string.[19] Thus, our attempt at identifying *the necessary information sources* for translation in relation to each correspondence type is a way of describing how the complexity of chosen translation task solutions is determined by how much and what kinds of information that must *at least* be available in order to produce them.

3.3.1.4 The need for general information sources

As discussed in 2.4.2.2, we assume for each correspondence type that certain general information sources are available prior to the translation activity, i.e. information about source and target language and their interrelations, and various kinds of extra-linguistic background information. These sources exist independently of specific translation tasks, but constitute an important part of the total amount of information needed to solve a given task. As already indicated in 3.3.1.3, the different correspondence types vary with respect to how much of the given information sources are required. Granted that types 1–3 represent translation tasks solvable within the pre-

[19] This is one of the central design principles of the PONS system (Dyvik 1990, 1995). Cf. the description in 1.3.2 of the different modes of translation in that system.

structured domain of linguistic information, it is only in type 4 correspondences that other information sources are needed to produce the target text. Moreover, with regard to the types 1–3, we shall see that syntactic and morphological information is sufficient in types 1 and 2, while in type 3 semantic information is also required in order to compute translations.

As discussed in 2.3.2, linguistic and extra-linguistic information sources differ in the sense that the former represents a limited domain, whereas information about the world is an open-ended domain. Thus, given the scope of general language, it is theoretically possible to represent information about source and target language and their interrelations in a finite way, while there is no principle available in order to determine which pieces of world information to include in information modules for automatic translation systems. In cases where translation requires the processing of given, general *world* information, we assume that, in general, this is not a problem that the computer can solve: the information is not available in the pre-structured domain of linguistic information, and hence not accessible. It is only within artificially delimited domains that world information can be made accessible in finite ways.[20] For the human translator, on the other hand, it is hardly an effort to make use of general, extra-linguistic background knowledge.

Given, general *linguistic* information sources are needed in all translation tasks. The need for general linguistic information can be discussed in relation to the division of the translation task into three subtasks: analysis, complexity measurement (or type identification), and generation (cf. 3.3.1.2). Insofar as each of these subtasks requires the processing of given, general *linguistic* information sources, we assume that this is a challenge that the computer can handle, since the information is finite and directly accessible as it is given prior to translation. The amount of processing effort will be determined by the amount of general linguistic information that is needed and the complexity of the task of retrieving it, and on this point there are differences between the subtasks.

[20] Cf. the discussion of restricted semantic domains in 1.4.2.3.

Irrespective of the type of translational correspondence, the subtask of analysis requires syntactic parsing of the source string, and the parsing problem is solved by using information contained in the representations of the source language lexicon and grammar. The amount of information that must be accessed is correlated with the length and linguistic complexity of the source string. The first step in parsing is to recognise word forms, and we assume that the information structure representing the SL lexicon is organised by base forms, so that for each inflected word morphological analysis is necessary to identify the lexeme it belongs to. Thus, recognising uninflected word forms requires smaller computational resources (i.e. a smaller number of calculations) than identifying word forms with inflection. Subsequent to word recognition, information about each lexical item can be merged with information about possible syntactic structures of the source language in order to create a parse of the source string, i.e. a representation of its syntactic structure. With respect to the amount of processing effort required by parsing, several researchers have studied the computational complexity of parsing problems, e.g. Barton et al. (1987) and van de Koot (1995). We will not go more deeply into parsing and the topic of its computational complexity, since the complexity of the translation task solution is determined by the relation between source and target string, and not by the complexity of parsing problems, because the parsing task is common to all four types of translational complexity. Thus, the basic amount of parsing needed for all kinds of source strings does not contribute to distinguishing between degrees of translational complexity; it is only parsing tasks associated with certain translational correspondence types, such as the retrieval of semantic information in type 3, which can be seen as contributing specifically to translational complexity.[21]

With reference to the typology of information sources for translation (cf. 2.4.2 with subsections), it is worth noting that while the input to the analysis step is general, given linguistic information (together with the word forms of the source

[21] As regards type 3, cf. 3.3.4.2–3 on this topic. Otherwise, due to the tendency that the high degree of structural relatedness found between source and target strings in types 1 and 2 is more likely to occur in short and structurally simple expressions than in longer and more complex ones, it is the normal case that parsing requires a smaller effort in correspondences of the two least complex types than in more complex cases. This is, however, a contingent aspect of the string pairs, and is in principle independent of the factors that contribute to the degree of translational complexity.

string), the output of the analysis step — i.e. an interpretation of the source string — can be seen as task-specific linguistic information. It is a representation of the information which is linguistically encoded in the source string, and it includes the described situation type as well as information about the linguistic structure of the source string.

The second subtask in the computing of a translation is to measure the complexity of the translation task, and this is done by combining the task-specific linguistic information given in the interpretation of the source string with general information about the interrelations between source and target language systems. These are interrelations between translationally corresponding elements of the lexicons of the two languages as well as between translationally corresponding structures described by rules of the respective grammars of SL and TL.[22] Here it is relevant to explain how information about such interrelations can be made directly accessible to the subtask of complexity measurement.

In the computational perspective it is rational, for a given language pair, to calculate such interrelations once, and store them, so that that information is available, and directly accessible, prior to any translation task. This amounts to computing a comparison of the language descriptions representing respectively SL and TL. We assume that when the interrelations between source and target language are calculated, it is possible to reveal not only between which elements of the two languages there exist translational correspondences, but also to what extent there are relations of equivalence between the corresponding elements, i.e. to determine the linguistic properties which are shared by source and target elements. Once such interrelations between two language systems have been calculated, information about them can be associated with individual elements of the lexicons and with individual rules in the grammars of respectively source and target language.[23] Such information may be seen as describing the translational properties of the individual lexemes and rules in the source language *with respect to the given target language*. Thus, informa-

[22] As regards lexicon information, we assume that interrelations between the word inventories of two languages normally apply to lexemes, and not to word forms, as more than one word form may be associated with one lexeme in languages with inflection.

[23] This approach has been tested in the PONS system for automatic translation (Dyvik 1990, 1995).

tion about source-target interrelations is directly accessible when monolingual information about lexical elements and grammatical structures is processed in order to interpret, or analyse, the source text in a given translation task.

This means that when a translation is computed, the subtask of analysis provides the bilingual information needed to diagnose the complexity of the translation task. The underlying principle is that information about how SL and TL are interrelated entails information about translational correspondences between specific linguistic elements in the two languages, so that identifying a particular lexeme or a particular syntactic structure in a source text will provide direct access to information about translationally corresponding elements in the given target language and information about linguistic properties shared by source and target elements. This kind of information is the basis for measuring the complexity of a given translation task, and we assume that the effort involved in accessing and processing such information is comparable to the effort required by the computable task of table lookup.[24] The presentations of each type of translational correspondence will provide further details on how the subtask of complexity measurement (or type identification) is solved.

Finally, with respect to the need for general linguistic information, the subtask of generation requires information retrieved from the representations of the target language lexicon and grammar. The amount of necessary information, as well as required processing effort, will be commented on in connection with each type of translational correspondence.

3.3.2 Type 1 correspondences

In 1.3.1 correspondences of type 1 are described as “word-by-word translations”, and they represent the least complex class of translational correspondences. With respect to the language pair English-Norwegian such cases are not very frequent, and the frequency would be higher in language pairs with a greater degree of structural relatedness between SL and TL.[25] The example given in 1.3.1 is here repeated in (1):

---

[24] Intuitively, table lookup demands very small computational resources. It requires no derivations or processing other than reading off the table the information that is available for a given search key.

[25] The proportion of type 1 correspondences within the recorded data is given in table 5.1 in 5.2.1.

(1a) Hun har vært en skjønnhet. (BV)
'She has been a beauty.'
(1b) She has been a beauty,

3.3.2.1 Linguistic characteristics of type 1

Type 1 correspondences are cases where translationally matched structures of respectively source and target language are so similar that there is equivalence between source and target string with respect to the sequence of translationally corresponding surface word forms. For such a string pair to count as a type 1 correspondence, some further requirements need to be fulfilled. Firstly, the strings must be syntactically equivalent, i.e. equivalent with respect to the assignment of syntactic functions (subject, object, etc.) to constituents.[26] Secondly, the syntactic structures have to be compositionally equivalent in the sense of having corresponding properties with respect to compositional semantics: predicates and arguments must be contributed by corresponding constituents. Such compositional equivalence will in the normal case be a consequence of syntactic functional equivalence. Finally, the strings have to be pragmatically equivalent in the sense of being used to perform corresponding pragmatic functions, or speech acts, in the given texts.

These requirements specify to what extent source and target string must exhibit corresponding linguistic properties in order to be classified as a type 1 correspondence. Word-by-word correspondences do not qualify as type 1 unless they also correspond syntactically, semantically, and pragmatically in the way described here. Within a domain of type 1 correspondences delimited by these requirements, there will hence exist relations of implication that can be exploited in the translation process: the fact that there is a type 1 correspondence between source and target string includes the fact that the existence of syntactic equivalence implies semantic equivalence, and that semantic equivalence further implies pragmatic equivalence.[27]

---

[26] There may be differences of detail in the phrase structure trees, as motivated by differences between SL and TL. Further development of this point requires illustration by means of language descriptions implemented in specific grammar formalisms, which we will not do. Anyway, such differences must not violate the requirement of syntactic functional equivalence.

[27] These relations of implication must not be understood as causal relations, but rather as material implications of the form "if *a* is true, then *b* is also true."

As explained in 3.3.1.2, this information is entailed in general, given information about the interrelations between source and target language, and it can be exploited so that the translation task is solved simply by translating word by word and without doing a deep linguistic analysis of the source string.[28]

Given the extent to which linguistic properties are shared between original and translation in type 1 correspondences, in particular the sharing of semantic properties, it follows that type 1 correspondences are included among the linguistically predictable translational correspondences, as described in 2.3.2. That is, a target string corresponding to the source string according to type 1 requirements is a member of the LPT set of the source string.

3.3.2.2 The structure of the translation task in type 1: information sources

Since type 1 correspondences are included among the linguistically predictable translational correspondences, a translation task of type 1 is solvable within the pre-structured domain of linguistic information. It may appear, from a computational point of view, that in type 1 the translation task merely involves replacing the word forms in the source string with the translationally matching word forms of the target string.

However, interpreting the source string is an initial, indispensable subtask, especially since it is required to determine that the given translation task is a type 1 case. The point was made in 3.3.1.3 that because a deep linguistic analysis of text is computationally resource-intensive, we assume that a rational strategy for computing a translation is to determine the amount of work required to generate the target text, i.e. to measure, or diagnose, the degree of translational complexity. In order to identify a translation task as a type 1 case, it is necessary to compute a syntactic analysis of the source string. This is the task of parsing, described in 3.3.1.4, and it is solved by processing the information encoded in the source string together with given, general information about the source language system. From the perspective of computing the translation, it is not necessary to process all information available

[28] Cf. the discussion in Dyvik (1999: 229–230).

about the words in the source string; what is needed is sufficient morphological and syntactic information in order to identify all lexemes, here including function words, contained in the source string and to derive the constituent structure of the source string. Analysing the semantic structure of the source expression is, for instance, not necessary; all that is required for translation is the information that in the type of construction found in the source string, there is syntactic equivalence, which within the domain of type 1 correspondences implies semantic equivalence, between source and target string. The result of the analysis step, i.e. task-specific information about the lexemes and the constituent structure of the source string, is, in the subsequent step, the key to identifying the translation task as an instance of type 1.

Next, the subtask of type identification is solved by processing given, general information about the interrelations between source and target language systems. Identifying a translation task as a type 1 case involves checking whether the following two requirements are met. Firstly, every lexical item in the source string must have a target language correspondent with syntactic and semantic properties matching those of the source item. Secondly, in the target language there must be a structure which is equivalent to that of the source string with respect to the linear order of constituents and the assignment of syntactic functions to constituents. In 3.3.1.4 we have explained that prior to the computing of translations, source-target interrelations may be calculated, and information about them may be associated with individual elements of the lexicons and grammars of the two languages, so that the result of the analysis task will provide direct access to information about any TL elements matching the lexemes and structures identified in the source string. Moreover, we argued that through calculating source-target interrelations it is also possible to identify the linguistic properties which are shared by translationally corresponding elements of the two languages. Consequently, when the lexemes of the source string have been identified, it is possible to decide whether each of them has TL correspondents with shared syntactic and semantic properties. Likewise, when the constituent structure of the source string has been derived, it is possible to decide whether it corresponds with a TL structure matching the source structure according to the requirements of type 1 cases. Thus, the outcome of the type identification task is

in practice given by the result of the analysis task, and the information needed to solve type identification is the amount of bilingual information present in the constituent structure derived for the source string. If the outcome of type identification is that all type 1 requirements are met, then the translation task conforms with the characteristics of type 1, and the target text can be generated on the basis of information about the lexemes and the constituent structure (phrase-internal structures included) of the source text.

The final step in the translation task, generation of the target string, involves, in a type 1 correspondence, a search for the target language word forms to replace the words of the source string. The sequence of word forms in the target string is already given by the word order of the source string, which is at this point directly accessible from its constituent structure. Information about lexical correspondence relations between SL and TL has already been accessed, and, in the case of lexemes without inflection, this information is sufficient to identify the correct target word forms, but in cases where more than one inflectional form exist further information is required to identify the appropriate word forms.

In cases where the source and target languages instantiate the same morphological categories, such as number on nouns, the target word form is determined on the basis of morphological information already identified in the source text analysis. E.g. since the Norwegian noun form *skjønnhet* in the source sentence (1a), given in 3.3.2, expresses the morphological feature "singular", the English singular form *beauty* is generated.

A different situation holds if there is a conflict between morphological features expressed by an SL word form and features expressed by its TL correspondent. If the consequence is that the two word forms are not semantically equivalent, then the correspondence violates the demands of type 1 on source-target equivalence with respect to linguistic properties.[29] However, it may be allowed within type 1 that corresponding word forms exhibit morphological differences which do not affect denotational properties, i.e. which do not influence the semantic translational

[29] E.g., number differences affect the denotational properties of the corresponding word forms; this is discussed in 6.3.2.1.

properties of the expressions involved. This type of mismatch may be illustrated by gender marking with reference to the language pair Norwegian-Swedish: both languages have obligatory gender marking, and sometimes translationally corresponding nouns of respectively Norwegian and Swedish exhibit different genders. E.g., for the Norwegian neuter noun *skjørt* ('skirt') the linguistically predictable translation into Swedish is the masculine noun *kjol*. If this kind of lexical correspondence occurs in a type 1 correspondence, there is a conflict between the gender information associated with the source string lexeme and that associated with the target language correspondent. The latter piece of information is available after the SL lexeme has been identified and information about its TL correspondent is retrieved from information about the lexical interrelations between source and target language. In such cases the diverging morphological property of the TL lexeme must, for the purpose of generation, overrule the morphological property of the SL lexeme. Due to TL-specific requirements of gender concord, the TL gender marker must also overrule that of the source string if the structure contains any adjectives or determiners governed by the noun in question. Information about such requirements becomes available through analysis of the constituent structure of the source string since the result of the analysis step provides information about the linguistic properties of translationally corresponding elements in the target language.[30]

It should be added that what distinguishes type 1 from type 2 with respect to the generation task is that whereas type 2 requires the retrieval of the corresponding TL syntactic rules, type 1 only requires the determination that such rules exist (cf. 3.3.3.2). The reason is, simply, that when a translation task is identified as a case of type 1, then it follows from the defining criteria of type 1 correspondences that the generation of the translation can be based directly on the constituent structure of the source string.

Thus, type 1 does not necessitate accessing information about the target grammar, although there are some exceptional cases where it may seem necessary to process syntactic information about the target language. To illustrate, the translation of the

[30] This point will be developed below in the discussion of the translation of present tense verbs from Norwegian into English.

Norwegian present tense verb form *har* in (1a) requires, firstly, that the English verb *have* is identified as the translational correspondent of the Norwegian verb *ha*, and, secondly, that the present tense, singular, third person form *has* is chosen among the various inflectional forms of *have*. The Norwegian source word *har* is marked as a present tense form; this provides temporal information to restrict the search to the set of present tense forms, *have* and *has*, in the English verb paradigm. The marking of number and person is obligatory in English present tense verbs, whereas Norwegian verbs are unmarked with respect to both categories. The source word *har* thus carries no information to settle the choice between *have* and *has*. This problem can be solved by using information about the English grammar rule of subject-verb agreement, together with information about the syntactic structure of the source string. In contrast to Norwegian, English requires agreement between the verb and its subject with respect to the grammatical categories of number and person. According to the syntactic structure identified for the source sentence, the subject (*hun*) carries the grammatical features singular and third person, and hence the singular, third person verb form *has* must be chosen in the translation. In this manner this *appears* to be a case of consulting the target language syntactic rules, contrary to the assumptions of type 1.

However, the type of information required here does not really pertain to syntactic structure (i.e. constituent order and hierarchy), but only to constraints among syntactic elements identifiable by function (i.e. the subject and the verb). Hence these constraints can straightforwardly be assumed to have been retrieved in connection with the calculation of source-target interrelations prior to translation (cf. 3.3.1.4, and above): if it is possible to establish a translational correspondence between the specific Norwegian syntax rule which describes the sentence structure of the given source string and an English sentence rule matching the Norwegian one according to the requirements of type 1, then it is also possible to retrieve the information that subject-verb agreement is included in the corresponding English rule and to associate this information with the Norwegian rule. I.e., this information is included in the set of translational properties, with respect to English, which are associated with the Norwegian sentence rule after the interrelations between the two languages have been calculated, and hence it is not necessary to retrieve information about the target rule

one more time in order to solve the generation task. We find it motivated to include, within type 1, cases like the given example when the criteria for type 1 are otherwise met.

Thus, the generation step in translation tasks of type 1 requires different linguistic information sources in order to identify the correct target word forms. In general, these sources include correspondence relations between the lexemes of SL and TL, morphological information derived from the word forms of the source string, information about the syntactic structure (which is derived from the source string and, in type 1 correspondences, shared with the target string), and information about morphological restrictions in the target language. All of these sources need not be required in a given string pair.

3.3.2.3 The weight of the translation task in type 1: processing effort

The structure of the translation task in type 1 correspondences, presented in 3.3.2.2, provides the basis for characterising the weight of the translation task in terms of required processing effort. As explained in 3.2.4 and 3.3.1.2, processing effort pertains to finiteness and amount of required effort. Like the description of the structure of the translation task, the analysis of processing effort can be related to the three subtasks, i.e. source text analysis, type identification, and target text generation.

In type 1 correspondences each of these subtasks is assumed to be computable on the basis of the linguistic information encoded in the source string together with the given, general linguistic information discussed in 2.4.2.2. This settles the question of finiteness as these information sources are available in a finite domain, and can be represented in a finite way. What then remains to be considered is the amount of effort required in order to access and process the necessary information.

Firstly, we have seen that source string analysis requires sufficient lexical, morphological, and syntactic information about the source language to identify all lexemes in the source string, and to derive its constituent structure. This is the task of parsing the source string, and the processing effort involved in syntactic parsing has already been commented on in 3.3.1.2.

Secondly, with respect to the subtask of type identification we have previously discussed how the necessary information about SL-TL interrelations is directly accessible after the analysis of the source string has been done. As stated in 3.3.2.2, identifying type 1 cases involves checking whether two specific requirements are met. Concerning the first requirement, we assume that the computational effort involved in verifying that each source lexeme has a target language correspondent with matching syntactic and semantic properties is comparable to the effort involved in looking up information in a table.[31] With respect to the second requirement, we also assume that it is not more complex than the task of table lookup to check whether the syntactic structure identified in the source string is associated with information about a translationally corresponding structure in the target language which is equivalent to that of the source string with respect to the assignment of syntactic functions to constituents. Thus, we assume that the subtask of type identification can be solved in linear time, since the required number of calculations is proportional to the size of the translation task, i.e. the length of the input (cf. 3.2.1).

Thirdly, with respect to the subtask of generating the target string, we have seen that in type 1 correspondences this involves identifying the correct target word forms to replace each word form in the source string. In general, due to the characteristics of type 1 correspondences the generation task does not involve the computing of any linguistic structures, since the sequence of target words is identical to that of the source string. Although the identification of correct word forms in the translation may require accessing different types of information, all these types are, as has been argued in 3.3.2.2, directly accessible after analysis has been done. For this reason we assume that the computational complexity of each replacement of a source word with its target correspondent is comparable to that of table lookup, and consequently we assume that also the subtask of generation is solvable in linear time.

[31] The complexity of table lookup is commented on in 3.3.1.4.

3.3.2.4 Summary of type 1 correspondences

Type 1 correspondences represent the lowest degree of translational complexity on the scale ranging from type 1 to 4. Relations of equivalence hold between source and target string on the levels of syntax, semantics, and pragmatics, and these equivalence relations must hold with respect to linguistic properties that influence the meanings of the two strings and which are obligatorily expressed in respectively source and target language. Moreover, there exist implicational relations between these equivalence relations: in a type 1 correspondence syntactic equivalence between source and target string implies semantic equivalence, which again implies pragmatic equivalence.

We assume that translation tasks conforming to the characteristics of type 1 correspondences are computable as they fall within the domain of the linguistically predictable translation tasks. Solving them requires the following information sources: firstly, sufficient information about the source language to identify all lexemes in the source string and to derive its constituent structure; secondly, sufficient information about the interrelations between source and target language to find out that each source string lexeme has a syntactically and semantically matching TL correspondent, and that the source string structure likewise has a match in the target language; thirdly, information about the word order of the source string in order to generate a target string where the sequence of words is identical to that of the source string, and sufficient information about morphological restrictions in cases where the lexical interrelations between SL and TL are not enough to identify the correct word forms in the target string.

With respect to processing effort, we assume, firstly, that all types of information required to solve the translation task can be represented in a finite way. Secondly, we assume that analysing the source string is, in type 1 correspondences, potentially the most resource-intensive part of the translation task since it involves syntactic parsing of a natural language expression, whereas we assume that the processing effort required by, respectively, type identification and the generation of target word forms to be proportional to the size of the translation task. Thus, the latter subtasks are assumed to be solvable in linear time, while we assume analysis to be a heavier task, but due to the tendency to low syntactic complexity in type 1 correspondences, we

expect it, normally, to be computationally tractable. The conclusion is that the translational complexity of type 1 cases is basically determined by the complexity of the given parsing task.

3.3.3 Type 2 correspondences

Type 2 correspondences are translationally somewhat more complex than type 1: it is not possible to translate word by word, but the degree of complexity is low enough to allow translation "constituent by constituent", as in examples (2) and (3), previously given in 1.3.1:

(2a) Dessuten virket hun overlegen. (BV)
'Also looked she haughty.'
(2b) She also looked haughty.

(3a) Leiligheten var ufattelig rotete. (BV)
'Flat.DEF was unbelievably untidy.'
(3b) The flat was unbelievably untidy.

As in the case of type 1 correspondences, string pairs of type 2 are not frequent with respect to the pair of languages English and Norwegian, and, as type 1, it is a phenomenon caused by a high degree of structural relatedness between original and translation.[32]

3.3.3.1 Linguistic characteristics of type 2

As mentioned in 1.3.1 and 3.3, the four types of translational correspondences are related to each other in a hierarchy reflecting an increase, from type 1 to 4, in the degree of translational complexity. A consequence of this hierarchical structure is that once we have described the least complex type in the hierarchy, this description can serve as a basis for characterising the second least complex type. Thus, type 2 correspondences are subject to the same restrictions as those applying to type 1 (cf.

[32] The proportion of type 2 correspondences within the recorded data is given in table 5.1 in 5.2.1.

3.3.2.1), except for two deviations from those constraints: in string pairs of type 2 there may be differences between source and target string with respect to the sequence of constituents, and/or with respect to the occurrence of function words. The first kind of deviation is illustrated by example (2) above: (2a) has a fronted adverbial (*dessuten*), followed by the verb *virket*, and then by the subject *hun*,[33] whereas in (2b) the subject *she* is in the initial position, and followed by the adverbial *also*, and then by the verb *looked*. The second kind of deviation is illustrated by example (3): in (3a) there is no word form matching the definite article *the* in (3b), and this is due to a grammatical difference between English and Norwegian: definiteness in nouns is in English marked by the definite article *the*, while in Norwegian it is marked by noun suffixes, which in the case of singular masculine nouns like *leilighet* ('flat') has the form *-en*. Example (3) is a minimal instance of a type 2 correspondence as the string pair exhibits only one linguistic deviation that violates the requirements of type 1 while being allowed within type 2.

Thus, in type 2 correspondences the structures of respectively source and target are not so similar that there is equivalence, through the entire string pair, with respect to the sequence of translationally corresponding surface word forms. Still, in type 2 cases there is near-equivalence on the level of syntax, and the same syntactic requirement as was described with respect to type 1 must be fulfilled: source and target string have to be equivalent with respect to the assignment of syntactic functions to constituents. Correspondences of types 1 and 2 have in common that they are syntactically congruent in the sense defined by Johansson (2007: 202): "Translations which preserve the syntax of the original are termed *syntactically congruent* translations."

In order to clarify the distinction between types 1 and 2, we will add that in type 2 correspondences every source string lexeme *with semantic content* must have a translational correspondent in the target string which is equivalent to the source lexeme with respect to both lexical category and syntactic function. In this connection the relevant distinction is between lexical words and function words, i.e. between seman-

[33] This is due to the verb-second restriction which applies in Norwegian when a non-subject appears sentence-initially.

tically heavy and semantically light lexemes. The use of function words is predictable from information about the grammatical structure of a language, and the requirements of type 2 correspondences are not violated by source-target deviations with respect to the occurrence of function words. Otherwise, the further requirements described for type 1 correspondences must also be fulfilled in type 2. I.e., the syntactic structures of respectively source and target string have to be equivalent with regard to compositionally derived semantic properties, and the two strings need to be pragmatically equivalent (cf. 3.3.2.1). Type 2 correspondences are, like type 1, included among the linguistically predictable translational correspondences.

In the same way as was described for type 1, we may observe implications between equivalence relations on different linguistic levels: we assume that information about how source and target languages are interrelated includes information about what sets of constructions of the two languages which correspond translationally according to the requirements of type 2. If a string pair is identified as a type 2 correspondence, there is syntactic near-equivalence between source and target string, and within the domain of type 2 correspondences this implies also semantic equivalence, which in turn implies pragmatic equivalence, between the two strings. Like in the case of type 1 correspondences, information about these implications can be exploited to solve the translation task without doing a deep linguistic analysis of the source string.

3.3.3.2 The structure of the translation task in type 2: information sources

Since type 2 correspondences are linguistically predictable, a translation task of type 2 is solvable within the pre-structured domain of linguistic information. The structure of the translation task is similar to that of type 1 correspondences, but somewhat more complex since it involves computing certain minor structural differences between source and target string.

The initial subtask of analysing the source string involves the same kind of parsing task as the analysis step in type 1 correspondences does, and it requires the same types of information as discussed in 3.3.2.2. Hence, the subtask of parsing will

not be commented on further, since it does not contribute to distinguishing between the degrees of translational complexity in types 1 and 2, respectively (cf. 3.3.1.2).

With respect to the subtask of type identification, we have previously explained in 3.3.1.4 and 3.3.2.2 that its solution is implicit in the result of the analysis task. Like in the case of type 1 correspondences, the information needed to solve type identification is the amount of bilingual information present in the constituent structure derived for the source string. We will illustrate type identification with reference to string pairs (2) and (3), given in 3.3.3.

In the case of (2), the analysis of the source string will reveal that (2a) is a main clause of indicative form with a fronted adverbial and subject-verb inversion. We assume that the Norwegian sentence rule which accounts for the constituent structure of the source string is associated with the information that in the translationally corresponding English sentence structure, the subject precedes the verb, while the sequence of other constituents matches that of the Norwegian structure. We also assume it to be available information that the assignment of syntactic functions to constituents in the English structure is identical to that of the Norwegian sentence. Furthermore, a result of the analysis step is that each lexeme in the source string is associated with information about the translationally corresponding target language lexemes, and in this case this information will reveal that each lexeme in (2a) is linked to English lexemes with matching syntactic and semantic properties. The conclusion is that the task of translating (2a) into the English sentence (2b) is in accord with the requirements of type 2, as specified in 3.3.3.1.

With respect to string pair (3), the analysis of (3a) will reveal that the Norwegian string can be translated word by word into English except for the noun phrase *leiligheten* (‘the flat’). The analysis of (3a) will identify *leiligheten* as a definite NP, and the syntactic rule which accounts for this NP will be associated with the information that in the translationally corresponding structure in English the definite article *the* precedes the noun. Since this is a function word, the correspondence conforms with the requirements of type 2.

As pointed out in 3.3.3.1, (3) is a minimal example of a type 2 correspondence, since it exhibits only one kind of source-target divergence which exceeds the

restrictions on type 1 while being allowed within those of type 2. This illustrates the point that in our approach the translational complexity of a given translation task, or in a given string pair, is determined by the degree of complexity of the most complex subpart of the task (cf. 4.3.6.1).

Generation of the target string requires a constituent structure in order to compute the linear sequence of surface word forms — this holds for all four types of translational correspondences. It has previously been explained in 3.3.2.2 that with respect to the subtask of generating the target string, types 1 and 2 differ in the sense that while generation in type 1 cases can be based directly on the constituent structure of the source string, generation in cases of type 2 requires also some processing of syntactic information specific to the target language. But to the extent that syntactic structure is shared between the source string and the corresponding rules of the TL grammar it is unnecessary to derive again syntactic structure already identified by the analysis of the source string.

What must be computed for the purpose of generation is that part of the constituent structure, the subtree, which is specific to the target language. To achieve this, it is necessary to retrieve the information given by the relevant syntactic rule(s) of the target language grammar, and it follows from the analysis task which rule (or rules) it is necessary to access information about. These TL grammar rules also provide the necessary information in cases of source-target divergences concerning the occurrence of function words: either the generation of the target text requires introducing a function word not found in the source string, or a certain function word occurring in the source string is *not* matched by a function word in the target string, and these facts will follow from syntactic information about the target language.

Otherwise, the task of identifying the correct target word forms requires the same kinds of information as are needed in type 1 cases (cf. 3.3.2.2). Given the restrictions on type 2 correspondences, the words in the target string will either be TL-specific function words or words which correspond translationally to the lexemes identified in the source string according to the same restrictions as those applying to lexical correspondences in type 1 cases.

With respect to example (2) above, the generation task involves processing the English sentence rule which is translationally linked with the Norwegian sentence rule describing the constituent structure of (2a), and hence generation necessitates a reordering of the verb and the subject in relation to the sequence found in the source string. Otherwise, the constituent structure does not need to be changed. With respect to example (3), generation requires the processing of the English NP rule which is translationally linked with the Norwegian NP rule describing the definite noun phrase *leiligheten*, and, as pointed out above, this will produce the target expression *the flat*.

The subtask of generation is the point where solving the translation task demands a larger amount of information in cases of type 2 than in those of type 1. In tasks of type 2 the need for information in analysis and type identification is on the same level as in tasks of type 1. With respect to generation, the two types have in common that information about the constituent structure of the source string must be available, but in type 2 cases generation also requires information about how source and target must be structurally different and about how the correct target structure is derived.

3.3.3.3 The weight of the translation task in type 2: processing effort

As previously explained, processing effort concerns the extent to which necessary information sources can be represented in a finite way, and the amount of effort required in order to access and process them (cf. 3.2.4 and 3.3.1.2). Translation tasks of type 2 are, like those of type 1, assumed to be computable on the basis of the linguistic information encoded in the source text together with the general linguistic information sources given prior to translation (cf. 2.4.2.2). As argued in 2.3.2, the latter information sources are available in a finite domain, and can be represented in a finite way, so that in this respect translation tasks of type 2 exhibit the same properties as those of type 1 do.

Also with respect to the amount of effort needed in order to access and process the necessary information sources, the requirements of type 2 are mostly the same as those of type 1, but differ on one point, reflecting how the two types vary with respect to the structure of the translation task. In 3.3.3.2 we have argued that the subtasks of analysis and type identification require the same types and amounts of

information in cases of type 2 as in those of type 1, and thus we assume that accessing and processing these information sources requires the same amount of effort in both types. Hence, the effort required by analysis in translation tasks of type 2 is determined by the complexity of the parsing task involved in the derivation of the constituent structure of the source string (cf. 3.3.1.4).

With respect to the subtask of type identification, we assume that the necessary information about SL-TL interrelations is, as in the case of type 1, directly accessible after the analysis of the source string, and, as explained in 3.3.2.3, the computational complexity of each checking operation involved in type identification is comparable to that of table lookup. We thus assume that the subtask of type identification is solvable in linear time, and this is common to translation tasks of respectively types 1 and 2. What distinguishes them at this point is that certain translational properties associated with the lexemes and structures of the source string are of different kinds in the two types, but this difference has no consequences for the effort involved in type identification.

Concerning the subtask of generating the target string, it is explained in 3.3.3.2 that this is the point where translation tasks of type 2 require a larger amount of information than tasks of type 1 do, since it is necessary to change one or more subparts of the constituent structure of the source string into TL-specific structure. Insofar as the structure of the target string is shared by the source string, this structural information is directly accessible once the source string has been parsed, and the effort involved in retrieving it is comparable to that of table lookup. Computing the necessary structural changes in the target string involves processing the information available in the relevant grammar rules of the target language, and then substituting target-specific structure(s), or subtree(s), for certain part(s) of the original constituent structure. The information needed in order to generate the correct target structure is easily, if not directly, accessible, and the complexity of the task is modest. Each such substitution is an isolated step in the sense that it is independent of possible other substitutions within the same translation task, and for that reason we assume that in type 2 also the subtask of generation is solvable in linear time, since

the consumption of computational resources is proportional to the number of such substitutions.

When it comes to the task of identifying the correct target word forms in type 2 cases, we have seen that information about TL-specific function words is accessible along with the processing of syntactic information for the purpose of generating TL-specific subtrees. Otherwise, identifying the target word forms requires, as explained in 3.3.3.2, the same information sources as in translation tasks of type 1, which means that the necessary information is directly accessible as a result of the analysis of the source string, and that the complexity of the task is comparable to that of table lookup (cf. 3.3.2.3).

3.3.3.4 Summary of type 2 correspondences

Type 2 correspondences represent the second lowest degree of translational complexity on the scale ranging from type 1 to 4. As in the case of type 1 correspondences, relations of equivalence hold between source and target string on the levels of syntax, semantics, and pragmatics, but with respect to syntactic equivalence, certain divergences are allowed: source and target string may differ with respect to the sequence of constituents, and/or with respect to the occurrence of language-specific function words. These divergences cannot violate the requirements that source and target string must be equivalent with respect to the assignment of syntactic functions to constituents, and that all lexical words in the source string must have a target correspondent of the same category and with the same function. Thus, source and target string are equivalent with respect to linguistic properties that influence the meanings of the two strings and which are obligatorily expressed in respectively source and target language. As in the case of type 1, there exist implicational relations between the equivalence relations: in a type 2 correspondence syntactic near-equivalence between source and target string implies semantic equivalence, which again is taken to imply pragmatic equivalence in the given texts.

We assume that translation tasks conforming to the characteristics of type 2 correspondences are computable since they fall within the domain of the linguistically predictable translation tasks, as tasks of type 1 do. Solving translation tasks of

type 2 requires the following information sources: firstly, sufficient information about the source language to parse the source string; secondly, sufficient information about the interrelations between source and target language to find out that each lexical word in the source string has a syntactically and semantically matching TL correspondent, and that the source string structure likewise has a target language match conforming with the type 2 requirements described above; thirdly, information about the constituent structure of the source string and information about how the constituent structure of the target string must be different, and, finally, sufficient information about morphological restrictions in cases where the lexical interrelations between SL and TL are not enough to identify the correct word forms in the target string.

With respect to processing effort, translation tasks of type 2 require, in comparison to type 1, the added effort involved in computing TL-specific subtrees in the constituent structure. Otherwise, tasks of type 2 are quite similar to those of type 1: firstly, the various kinds of information required to solve the translation task can be represented in a finite way, and, secondly, source string analysis is potentially the most resource-intensive part of the translation task since it involves syntactic parsing, whereas the computational complexity of type identification and generation of target word forms (other than TL-specific function words) is assumed to be comparable to that of table lookup. Thus, we again consider type identification and generation to be solvable in linear time, while analysis will require a larger number of calculations, determined by the demands of the parsing stage. With respect to the computing of target-specific subtrees in the constituent structure, we have explained it to be of modest complexity, assuming it to be solvable in linear time (cf. 3.3.3.3).

3.3.4 Type 3 correspondences

Type 3 correspondences constitute the second most complex class of translational correspondences. They represent translation tasks where linguistic divergences between source and target violate the restrictions on types 1 and 2, but where source and target text express the same meaning. The linguistic relation between source and target string involves greater structural differences in type 3 correspondences than in

string pairs of the two lower types. With respect to the language pair English-Norwegian, type 3 cases are more frequent than instances of both types 1 and 2.[34] In 1.3.1 we gave the following string pair as an example of a type 3 correspondence:

(4a) Hildegun himlet lidende mot taket og svarte med uforskammet høflighet: (BV)
'Hildegun rolled-eyes suffering towards ceiling.DEF and answered with brazen politeness'

(4b) Hildegun rolled her eyes in suffering towards the ceiling and answered with brazen politeness.

3.3.4.1 Linguistic characteristics of type 3

The defining characteristic of type 3 correspondences is that they violate one of the restrictions on type 2 correspondences in the following way: in a string pair of type 3 it is the case that for at least one lexical word in one of the strings there is no translational correspondent in the other string of the same category and/or with the same syntactic function as that lexical word. Source-target divergences of this kind will cause greater differences in constituent structure between source and target string than the differences allowed within type 2 correspondences, but they must not violate the requirement of semantic equivalence between original and translation. I.e., source and target string have to be equivalent with regard to the sets of expressed predicates and arguments, and the relations between the predicates and their arguments.[35] On the other hand, predicates and arguments in respectively source and target need not be contributed by translationally corresponding constituents between which there must exist relations of syntactic functional equivalence, which is a requirement of type 2 correspondences. The characteristic of semantic equivalence is shared by string pairs of types 1, 2, and 3, and, in general, this means that in correspondences of these three

[34] The proportion of type 3 correspondences within the recorded data is given in table 5.1 in 5.2.1.

[35] According to Alsina (1996: 4), a predicate "expresses a relation (or relations) among participants; these participants are called the *arguments* of the predicate." Thus, the argument structure of a predicate is the specification of how a set of arguments is involved in the relation expressed by the predicate. The predicate-argument structure specifies the number and internal ordering of the arguments of the predicate.

types, the same informational content is *linguistically* encoded in the source string, as well as in the target string.[36] This is a central principle of our analytical framework.

As previously described in 1.3.1, example (4) above contains two instances of divergences which violate type 2 requirements while being allowed within type 3. Firstly, the intransitive verb phrase *himlet* in the Norwegian source string (4a) corresponds with the expression *rolled her eyes* in (4b), which consists of a transitive verb phrase and a noun phrase functioning as direct object.[37] Thus, between these translationally corresponding expressions there is a considerable difference with respect to constituent structure although they are equivalent in terms of predicate-argument structure: the Norwegian verb *himle* ('roll one's eyes') describes the activity of rolling the eyes of the agent, and this is lexically encoded information, so that the existence of the referent of the English object NP *her eyes* is implied by the Norwegian expression. Secondly, the adverb phrase *lidende* ('suffering') in the Norwegian sentence corresponds translationally with the preposition phrase *in suffering* in the English sentence. These phrases share the function of verb-phrase modification, but belong to different syntactic categories. They are semantically equivalent with respect to the way in which they describe *how* the activity of eye-rolling is performed. We may observe that the remaining parts of the string pair (4), i.e. *og svarte med uforskammet høflighet — and answered with brazen politeness*, is a word-by-word correspondence of the lowest degree of translationally complexity, but, as previously noted in 3.3.3.2, the classification of a given string pair is determined by the most complex subpart(s) (cf. 4.3.6.1).

To sum up, in translational correspondences of type 3 we do not find, as in types 1 and 2, syntactic functional equivalence between source and target string. But in order to fall within type 3, source and target string must be equivalent with respect to compositionally derived semantic properties, and in the given texts they must be pragmatically equivalent in the sense of being used to perform corresponding

---

[36] In chapter 6 we will discuss some minor exceptions to this. These are cases of predictable differences between source and target string in the amount of grammatically expressed information; cf. 6.3.1.1–2.

[37] Cf. 1.4.2.3, where this example is mentioned in connection with mapping problems for automatic translation. The correspondence between *himlet* and *rolled her eyes* can be described as an instance of conflational divergence between SL and TL.

pragmatic functions, or speech acts. The structural divergences between source and target text in type 3 correspondences show that the degree to which there exist implicational relations between equivalence relations on different linguistic levels is smaller in translational correspondences of type 3 than in those of lower types. The information that there is a type 3 correspondence between a source string and a given target string entails that the strings are structurally different, but semantically equivalent, which in turn implies, within the domain of type 3, that they are also pragmatically equivalent in the given texts. Due to the requirement of semantic equivalence, type 3 correspondences are, like those of types 1 and 2, included among the linguistically predictable translations.

3.3.4.2 The structure of the translation task in type 3: information sources

Since type 3 correspondences are linguistically predictable, a translation task of type 3 is solvable within the pre-structured domain of linguistic information, as it is not necessary to process extra-linguistic information or information from the textual context of the given translation task in order to generate a semantically and pragmatically equivalent target text. Because type 3 correspondences exhibit more complex structural differences between source and target string than correspondences of types 1 and 2 do, solving the translation task in type 3 cases is also more complex than in cases of the lower types. We shall see that in comparison to types 1 and 2, the need for information in type 3 grows with respect to the subtasks of analysis and generation, but not with respect to type identification.

As pointed out in 3.3.1.3, the initial subtask of analysing the source string involves the same kind of parsing task in any type of correspondence. In cases of type 3 the syntactic analysis of the source string thus requires the same types of information as it does in cases of types 1 and 2, i.e. sufficient morphological and syntactic information to identify all lexemes in the source string and to derive its constituent structure.

With respect to the subtask of type identification, the same facts apply in type 3 correspondences as in those of the lower types: the solution to the subtask of type identification is implicit in the result of the analysis task (cf. 3.3.1.4 and 3.3.2.2), and

the information needed to solve type identification is the amount of bilingual information associated with the constituent structure derived for the source string.

Identification of a type 3 case may be illustrated by string pair (4), shown in 3.3.4. After parsing, the translational properties associated with the lexemes and constituents identified in the Norwegian sentence (4a) will reveal that the English translation (4b) corresponds with the source string according to the requirements of type 3, but not according to those of types 1 and 2, and this is due to two facts. Firstly, the Norwegian intransitive verb *himle* corresponds translationally with the syntactically complex expression *roll one's eyes*, which means that in the translation there will be (at least) two lexical words, *roll* and *eye*, which do not have correspondents in the source text. Still, there is semantic equivalence between the translationally corresponding expressions since the predicate-argument structure encoded in *roll one's eyes*, including the relation of possession expressed by *one's*, is incorporated in the lexical content of the verb *himle* (cf. 3.3.4.1). Secondly, the Norwegian present participle *lidende*, functioning as an adverbial modifying the verb phrase, corresponds translationally with the preposition phrase *in suffering* which is semantically equivalent to *lidende* in the target text. In this case the corresponding expressions carry the same syntactic function, but as the target expression contains a lexical word, the preposition *in*, which has no match in the source expression, the requirements of types 1 and 2 are nevertheless violated.

When a translation task has been identified as a type 3 case, it is necessary to derive information about the semantic structure of the source text in order to compute the target string. I.e., a semantic representation of the source string must be produced, and this is derived compositionally from the syntactic representation together with semantic information associated with the lexemes identified in the source string. The derivation of a semantic representation requires information about the constituent structure of the source string, about the assignment of syntactic functions to constituents, and about any components of meaning encoded linguistically in the source text (e.g. predicate-argument relations, spatial and temporal relations). It is rather a question of definition whether the semantic analysis of the source string should be regarded as part of the analysis step, or as the first step in the subtask of

generation since the output of semantic analysis will serve as the input to generation. What is more important is that the solution of translation tasks of type 3 — in contrast to those of the lower types — requires a semantic analysis of the source string because of structural divergences between source and target string. Since the derivation of a semantic representation of the source string is a kind of linguistic analysis, we prefer to regard this step as part of the subtask of analysis.

We assume that the generation of the target string in cases of type 3 must be based on information about the semantic structure of the source string because type 3 correspondences involve structural source-target divergences of a kind that is qualitatively different from those found in type 2.[38] When generation is based on a semantic representation of the source string, the meaning components identified in that representation provide the information required to retrieve from the target language description the specific lexical units and grammatical structures needed in order to generate the given target string.

It is too simple to view the generation task as parsing in reverse. According to Vander Linden (2000: 765), we may regard the *goal* of natural language generation (NLG) "as the inverse of that of natural language understanding (NLU) in that NLG maps from meaning to text, while NLU maps from text to meaning." However, with respect to *methods*, Vander Linden (2000: 765–766) points out important differences between the two types of processes. Firstly, while there is great variation among generation systems with respect to the nature of the input, systems for NLU (including parsers) receive linguistic input only, which is "governed by relatively common grammatical rules" (2000: 765). Secondly, since NLU aims at interpreting natural language input, "[its] dominant concerns include ambiguity, under-specification, and ill-formed input", matters which are not so relevant in NLG because "[the] non-linguistic representations input to an NLG system tend to be relatively unambiguous, well-specified, and well-formed" (2000: 766). Hence, Vander Linden concludes that "the dominant concern of NLG is *choice*" (2000: 766),

[38] Notably, there is no relation of syntactic functional equivalence between the entire source and target strings.

i.e. choosing the best way of expressing the meaning which is input to a generation system for a given natural language.

Thus, the task of generating a target string from a semantic representation of the source string is, above all, a task of making choices, and the search space for these choices is the entire lexicon and grammar representing the target language. Lexical units and grammatical structures are not selected independently of each other in natural language generation, as there are always close interconnections between meaning and structure in linguistic expressions: “In practice (and perhaps in theory too), it is not possible to separate cleanly selection of lexical items and commitments to particular grammatical organizations” (Bateman and Zock 2003: 289). Anyway, the purpose behind the selection is to extract elements of the TL lexicon and grammar in order to cover all of, but no more than, the components of meaning contained in the semantic representation of the source text. In this manner, the semantic representation, i.e. the input to generation, provides the information needed to carry out the target text generation.

In order to make the choices required to cover the input as precisely as possible, it is a precondition that the input is sufficiently specific. The latter is a challenge to be faced by the design of the framework used for semantic representation, as the framework must facilitate the required level of specificity in the representation of the meaning expressed by the source string. In the case of string pair (4), the semantic representation must provide the information needed to select the lexemes behind the word forms in the target string (4b). Moreover, the semantic representation must provide information specific enough to contribute to the identification of correct target word forms. In contrast to the identification of target word forms in types 1 and 2, information about morphological features identified in the source string word forms is not exploited when word forms are selected for the generation of target word forms in translation tasks of type 3, since the selection must be done by combining the information contained in the semantic representation with TL-specific lexical and grammatical information. As far as other details in the generation task are concerned (e.g. the determination of surface word order), they will be dependent on the design

of the generation algorithm, and we do not want to discuss further such issues of implementation.[39]

3.3.4.3 The weight of the translation task in type 3: processing effort

As before, processing effort is considered in terms of the extent to which necessary information sources can be represented in a finite way, and the amount of effort required in order to access and process them. Like translation tasks of types 1 and 2, we assume that those of type 3 are computable on the basis of the linguistic information encoded in the source string together with the general linguistic information sources given prior to translation (cf. 2.4.2.2). As previously argued, these information sources are available in a finite domain, and can be represented in a finite way.

With respect to the amount of effort required in order to access and process the required information sources, there are marked differences between, on the one hand, translation tasks of type 3, and, on the other hand, those of types 1 and 2, and the differences pertain to the subtasks of analysis and generation. As explained in 3.3.4.2, in cases of type 3 the need for information in the subtask of type identification is the same as in cases of the lower types, and the required information, i.e. about SL-TL interrelations, is, as in the case of types 1 and 2, directly accessible through information associated with the syntactic representation produced by parsing the source string. As previously argued, we thus regard this subtask to be solvable in linear time.

Then, with respect to the subtask of analysis, it is noticeably more demanding to access the required information in type 3 cases than in translation tasks of the lower types. Firstly, this is due to the fact that analysis in type 3 involves not only syntactic parsing, but also a semantic analysis of the source string, which makes it necessary to process a larger amount of the source language information available prior to translation, as all semantic information given about the lexemes of the source string must be analysed. Performing a full linguistic analysis of the source string in type 3 cases demands a far greater number of calculations in order to retrieve all necessary

[39] For further information see Vander Linden (2000), Bateman and Zock (2003).

information than the number of calculations required to perform the more shallow syntactic analyses which are sufficient in cases of types 1 and 2.

Also with respect to the subtask of generation, it is far more demanding computationally to access the required information in type 3 cases than in translation tasks of the lower types. In 3.3.4.2 we observed that generating the target string from the semantic representation of the source string involves a number of choices for which the search space is the entire TL language description. Firstly, choices such as selecting the appropriate lexemes, grammatical constructions, and inflectional word forms must be settled by processing information contained in the TL language description together with the information available in the semantic representation. Secondly, a given semantic representation may correspond with not one, but a set of linguistic expressions. A natural language will normally offer more than one way of encoding the same informational content. E.g., we assume that the passive sentence *The boy was given a book by the girl* is semantically equivalent with each of the two active sentences *The girl gave the boy a book* and *The girl gave a book to the boy*. With respect to example (4), we suggest that a semantically equivalent alternative to target string (4b) could be *In a suffering manner Hildegun rolled her eyes towards the ceiling and answered with impudent politeness*. Hence, generating a specific target string from a semantic representation of the source string will frequently involve the problem of choosing the most appropriate target string among a set of alternatives which all correspond with the semantic representation, and this selection task can add to the overall complexity of the translation task, unless a random choice is made.

Altogether, the generation task in type 3 correspondences is clearly a resource-intensive problem. It falls outside the present project to describe in detail the computational complexity of natural language generation from semantic representations, but a few general aspects may be noted. If we assume semantic equivalence between a semantic representation and a corresponding linguistic expression, then the generation task may be regarded as a special case of what Shieber (1993) calls the *problem of logical-form equivalence*. Shieber uses the term *logical form* to refer to any kind of non-linguistic representation of meaning serving as input to a natural

language generator (1993: 181). Different logical forms may correspond with one and the same syntactic form (i.e. surface expression) in a language, and hence Shieber argues that a language generator must be able to decide whether two logical forms "mean the same" (1993: 180), i.e. whether one of them may be translated into the other, and this is the problem of logical-form equivalence. It is the view of Shieber (1993) that the problem of logical-form equivalence is computable, but probably intractable due to the lack of an efficient solution algorithm.[40] Since we assume semantic equivalence between a semantic representation and a corresponding linguistic expression, we regard generation from a semantic representation as a special case of translating one logical form into another, semantically equivalent, logical form. For this reason we will suggest that the computational complexity of the generation task in type 3 correspondences is in the same class as that of logical-form equivalence. That is, the generation of the target text probably belongs to the set of intractable problems, and will in general be a more computationally demanding task than parsing is.[41] We will not pursue the topic of natural language generation further; it is also still a field with many unanswered research questions.

3.3.4.4 Summary of type 3 correspondences

Type 3 correspondences represent the second highest degree of translational complexity on the scale ranging from type 1 to 4. Relations of equivalence hold between source and target string on the levels of semantics and pragmatics. Implicational relations between such equivalence relations exist to a lesser degree than in the cases of types 1 and 2: in type 3 correspondences there is not syntactic equivalence between the entire source and target strings, but there is semantic equivalence, which implies pragmatic equivalence in the given texts. Source and target string are structurally different in the sense that for at least one lexical word in one of the strings, there is no translational correspondent in the other string of the same category and/or with the same syntactic function as that lexical word. Correspondences of

[40] The notion of an 'efficient solution algorithm' is described in 3.2.1.

[41] Given a sufficiently high degree of syntactic complexity, the task of parsing a natural language expression may also be intractable.

types 1, 2, and 3 have in common that source and target string are semantically equivalent in the sense that the same informational content is linguistically encoded in both of them.

We assume that translation tasks conforming to the characteristics of type 3 correspondences are computable since they fall within the domain of the linguistically predictable translation tasks, as tasks of types 1 and 2 do. The requirement of compositional semantic equivalence between source and target string in type 3 correspondences will be refined in chapter 6, where we will discuss certain cases of predictable semantic differences between translationally corresponding strings.

Solving translation tasks of type 3 requires the following information sources: firstly, sufficient information about the source language to identify all lexemes in the source string, to derive its constituent structure, and to derive a semantic representation containing all relevant components of meaning expressed by the source string; secondly, sufficient information about the interrelations between source and target language to find out that the target string is structurally different in the sense described above; thirdly, sufficient lexical, morphological, syntactic, and semantic information about the target language in order to generate a target string on the basis of the semantic representation of the source string.

With respect to processing effort, we assume that type 3 has in common with types 1 and 2 that all kinds of information required to solve the translation task can be represented in a finite way. Further, we assume for type 3, as for types 1 and 2, that the subtask of type identification is solvable in linear time.

Concerning the subtask of analysis, we have previously observed in 3.3.1.4 that the complexity of syntactic parsing is in principle the same for all types of translational correspondences. The added processing requirement of type 3, compared with types 1 and 2, is caused by the need for a semantic analysis of the source string.

Finally, with respect to the subtask of generation, we assume that translation tasks of type 3 differ sharply from those of the lower types in the sense that whereas a modest processing effort is required by target string generation in cases of types 1 and 2, generation from semantic representations in type 3 is very resource-intensive, in computational terms probably belonging to the set of intractable problems.

3.3.5 Type 4 correspondences

Type 4 correspondences constitute the most complex class of translational correspondences in our hierarchy of correspondence types. They represent translation tasks where linguistic divergences between source and target violate the restrictions on types 1, 2, and 3: in type 4 correspondences there are not only structural, but also semantic, differences between source and target string. With respect to the language pair English-Norwegian, type 4 cases represent the most frequent class of translational correspondences.[42] In 1.3.1 we gave string pair (5) as an example of a type 4 correspondence, pointing out the semantic difference between the translationally corresponding expressions *for å gå i melkebutikken eller til bakeren* ('to go to the milk shop or to the baker') and *to go and buy milk or bread*.

(5a) Her kunne de snakke sammen uten å bli ropt inn for å gå i melkebutikken eller til bakeren. (BV)
'Here could they talk together without to be called in for to go in milk-shop.DEF or to baker.DEF'

(5b) They could talk here without being called in to go and buy milk or bread.

As explained in 1.3.1, and further discussed in 2.4.2.1, the italicised expressions denote different activities, but it may be inferred from background information about the world that both activities can have the same result, and hence (5b) may be chosen as an appropriate translation of (5a).

3.3.5.1 Linguistic characteristics: type 4 correspondences are different

In contrast to the correspondence types of lower translational complexity it is not possible to describe specific linguistic characteristics of type 4 cases. Rather, the class is negatively defined: in correspondences of type 4 source and target string are not equivalent with respect to constituent structure as in type 1 cases, or they are not equivalent with respect to the assignment of syntactic functions to constituents as in

[42] The proportion of type 4 correspondences within the recorded data is given in table 5.1 in 5.2.1.

type 2 cases, or they are not equivalent with respect to compositional semantic properties. As pointed out in Thunes (1998: 29–30), pragmatic equivalence may hold between source and target string in a type 4 correspondence, but not necessarily. Example (6) may illustrate absence of pragmatic equivalence between original and translation:

(6a) …'har du nå vært på et av disse foredragene igjen.' (EFH)
'Have you now been at one of these lectures.DEF again?'
(6b) 'Have you been to one of those lectures again?'

String pair (6) is almost a minimal case of type 4: it is nearly a word-by-word correspondence, but in the pragmatic particle *nå* in (6a) has no translational match in (6b).[43] The Norwegian adverb *nå* ('now') functions in (6a) as a pragmatic particle creating the impression that the speaker probably disapproves of something the addressee has done. The lack of a corresponding expression in (6b) means, firstly, that a certain semantic component present in the source text is not contained in the target text, and, secondly, that the pragmatic effect created by the particle *nå* is not present in the target text. Because of this semantic difference the pair of sentences (6a) and (6b) is categorised as a type 4 correspondence. This example also illustrates the point that even if two corresponding strings are, by and large, structurally equivalent, the correspondence between them is nevertheless of type 4 if it exhibits some semantic divergence.

Moreover, (6) is an instance of pragmatic non-equivalence where we assume that pragmatic equivalence could have been achieved by choosing a different target text. There may be still other cases where original and translation are pragmatically non-equivalent due to cultural divergences between SL and TL. In such cases cultural differences between the two language communities may have the consequence that some semantic content encoded in the source language cannot be matched by a target

[43] There is also a semantic deviation between the translationally corresponding demonstrative determiners *disse* and *those*. The former expresses proximity, and the latter distance.

language expression with a corresponding communicative effect, and it may even be impossible to find a TL expression with matching semantic content.

Here we shall not pursue the factors governing pragmatic equivalence, since our focus is on the fact that type 4 correspondences distinguish themselves through non-predictable semantic deviation between source and target text, which means that the target string does not belong to the LPT set of the source string. The most notable difference between, on the one hand, correspondence type 4 and, on the other hand, types 1, 2, and 3 is that the semantic representation of the source expression is shared by the target expression in correspondences of types 1–3, but not in those of type 4. Also, within the domain of type 4 correspondences there do not exist, as in string pairs of the lower types, any implicational relations between equivalence relations on different linguistic levels.

3.3.5.2 The structure of the translation task in type 4: information sources

We assume that translation tasks of type 4 differ principally from those of types 1–3 in the sense that tasks of type 4 are not solvable within the pre-structured domain of linguistic information, and the need for information required to translate is larger in type 4 correspondences than in any of the other types. The growth in required information, in comparison to the lower types 1–3, concerns notably the subtasks of analysis and generation.

By definition, translation tasks of type 4 are non-computable since they require information types not included in the pre-structured domain of linguistic information. In 2.3.2 we have argued that there is no principle for delimiting a representation of the information sources lying outside the pre-structured domain, granted that our scope is the translation of general language, and not translation within a restricted semantic area. Thus, there is no principled limit on the amount and types of information that could be needed to solve a task of type 4. We regard type 4 cases as those that demand human translation: the human translator is normally capable of collecting as much information as the task requires, for instance by considering a wider textual context, by looking up more background information of various kinds, or by asking other translators for help, so that a target text can eventually be produced. In

this manner translation tasks of type 4 are seen as translatable, although they are not computable (cf. 3.2.5).

With respect to the lower types of translational correspondence, we have discussed the structure of the translation task in relation to the assumption that the task is computable. If we consider type 4 tasks to be non-computable, but solvable by humans, it may seem odd to describe the translation task in terms similar to those applied to the computable tasks. On the other hand, since we do not study the human translation process, the descriptive approach is not altered in relation to the structure of type 4 tasks, and the discussion will focus on the aspects that make type 4 cases fall outside the computable domain.

In order to solve the initial subtask of analysis, a type 4 case requires the same kinds of linguistic information as are required in type 3 (cf. 3.3.4.2) in order to derive a constituent structure as well as a semantic representation of the source string. But in type 4 cases, solving the translation task demands an understanding of the source string which goes beyond a syntactic and semantic analysis, and which requires sources of information included neither in the pre-structured domain of linguistic information nor in the information that is explicitly encoded in the linguistic form of the source string.

It is an important aspect of type 4 that on the basis of the source string alone, it cannot be predicted exactly which additional information sources that are necessary. To mention some possibilities, the required additional sources may include general background information about the world, domain-specific technical information, task-specific linguistic information about reference relations, as well as task-specific extra-linguistic information about the utterance situation of the source text, and about the described situation of the source text. With reference to example (5) above, we have previously pointed out in 2.4.2.1 that general background information about the world is required to fully interpret the Norwegian expression *for å gå i melkebutikken eller til bakeren*. At this point we do not want to illustrate further the types of additional information sources required in type 4 cases, as examples of them will be discussed in chapter 6.

As regards the subtask of type identification, we again want to assume that its solution is implicit in the result of the analysis task. In cases of types 1–3 the analysis yields information about the translational properties, with respect to the target language, of the linguistic items identified in the source string. In translation tasks of type 4, either the translator has chosen a target string deviating semantically from the source string although a literal translation (cf. 2.3.3) could have been produced, or the analysis will reveal that for at least some subpart of the source string there is no linguistically predictable correspondence in the specific target string.

Concerning the subtask of target text generation, we assume that required information sources are the semantic structure of the source string, together with information about the semantic deviation between source and target, as well as information about the grammar and lexicon of the target language. In addition, the generation task requires one or more of the information sources mentioned above in connection with the subtask of analysis. As in type 3 cases, the generation task is first and foremost a question of selecting the appropriate lexemes and structures for the target string. In type 3 this is done by choosing elements of the TL lexicon and grammar in order to cover all of, but no more than, the components of meaning contained in the semantic representation of the source string. In type 4 additional information must contribute to deciding which of those semantic components of the source string that are expressed in the target string, and which are not — as well as which components, if any, that are expressed instead. With reference to example (5) again, we assume that inferencing about the situation described by the Norwegian expression *for å gå i melkebutikken eller til bakeren*, together with access to background information about the world, makes the translator choose the semantically non-equivalent English target expression *to go and buy milk or bread*. As in the analysis subtask, it cannot be predicted, on the basis of the source string alone, which additional information sources are required in order to generate the target string. In general, that question is determined by what we will regard as *parole*-related factors, i.e. factors existing either in a wider textual context, or in the extra-linguistic context, or in both.

3.3.5.3 The weight of the translation task in type 4: processing effort

Since translation tasks of type 4 are non-computable, such tasks are, in computational terms, even harder problems than the intractable ones.[44] Like in the case of the information structure of type 4 tasks, the most relevant topic concerning the necessary processing effort is our assumption that the subtasks of analysis and generation require access to additional information not available in the finite domain of linguistic information sources. Type identification demands no more effort than in the lower correspondence types.

With respect to the subtasks of analysis and generation, we may observe that since some of the information needed to translate in a type 4 correspondence is not included in the finite domain of pre-structured linguistic information, there is in principle no limit on the size of the search job involved in compiling the necessary information. Moreover, within the present approach to translational complexity we have no framework for describing the amount of computational resources needed to access and process such additional information. As stated in 3.3.5.2, we consider type 4 tasks to be cases where human translation is needed, and how hard or easy it is for a human to solve a given translation task will be dependent on that individual's translator competence. That topic falls outside the scope of the present study.

3.3.5.4 Summary of type 4 correspondences

In contrast to translation tasks of the lower types 1–3, type 4 tasks are non-computable as they do not belong to the domain of linguistically predictable translation tasks. Type 4 correspondences represent the highest degree of translational complexity on the scale ranging from type 1 to 4. There is not semantic equivalence between the entire source and target strings; pragmatic equivalence may exist, but not necessarily. Hence, there do not exist, as in string pairs of the lower types, any implicational relations between equivalence relations on different linguistic levels. Type 4 correspondences typically exhibit structural divergences between original and translation, although in certain cases these may be of a minimal kind, as illustrated in

[44] Intractable problems may be computable; cf. 3.2.1.

3.3.5.1. The defining characteristic of type 4 is non-predictable semantic deviation between source and target text, which means that the target string does not belong to the LPT set of the source string. It also means that there will be certain semantic components which are not shared between the semantic representations of each of the two strings.

Solving a translation task of type 4 requires, like tasks of the lower types, access to the information linguistically encoded in the source text, as well as to general, given information about SL and TL, and their interrelations. But producing the semantically non-equivalent target text requires access to additional information sources in order to understand the source string beyond the levels of syntactic and semantic structure. As explained in 3.3.5.2, these additional sources are not included in the finite domain of linguistic information.

Since it is necessary, in order to solve translation tasks of type 4, to access information sources falling outside the finite, pre-structured domain, there is in principle no limit on the processing effort required to search for the needed information.

3.4 Summary

The present chapter is divided into two main parts. The first part (3.2 with subsections) provides a theoretical background for our approach to measuring translational complexity, and the second part (3.3 with subsections) contains a detailed description of the correspondence type hierarchy.

The main purpose of 3.2 with subsections is to view the notion of complexity from different angles, i.e. those of information-theory, linguistics, and translation, respectively, in order to explicate the approach taken to translational complexity in the present approach. In 3.2.1 we start by defining ‘computability’ as a property of tasks: a task that can be solved by a specifiable procedure is a computable task. Further, ‘complexity’ is a mathematical property describing the amount of time and space needed to solve a computable task, and computational complexity theory offers tools for analysing the information processing structure of computable tasks (or problems), as well as for sorting such problems into classes according to complexity measurements. Section 3.2.2 describes the relevance of complexity theory to studies

of natural language: firstly, applying complexity analysis to natural language problems has provided new knowledge about the structure of such problems, and, secondly, complexity analysis has proved to be a useful tool in the development of grammar formalisms. However, as pointed out in 3.2.3, applying complexity theory in linguistics requires that natural language problems are studied as problems of computation. Hence, there are several researchers, e.g. Dahl (2004) and Miestamo (2006), who study complexity in natural languages without using the tools of complexity theory. A common denominator of these two contributions is that complexity in natural languages is measured in terms of the length of the description of a given linguistic phenomenon.

Section 3.2.4 discusses the present approach to translational complexity. Like Dahl (2004) and Miestamo (2006), we have chosen a quantifiable, objective commodity as a basis for the analysis of complexity, but whereas their measurement is tied to the length of the description of linguistic phenomena, our measurement concerns the information needed in translation. Moreover, our approach resembles computational complexity analysis in several ways. For one thing, as computability is a precondition for complexity analysis, so translatability is a prerequisite for describing translational complexity. Further, in our analysis translational correspondences are viewed as tasks to be solved, and the complexity of these tasks is described in terms of the structure and weight of a given task. Also, our sorting of translational correspondences into types according to the degree of translational complexity resembles the sorting of computational problems into complexity classes. Section 3.2.5 stresses the point that when analysing translational complexity our focus is on isolating the computable, i.e. the linguistically predictable part of the translational relation, within the translatable.

The presentation of the correspondence type hierarchy in 3.3 with subsections is intended as a way of describing the information situation of the translation task: which information sources are available, and how much information is needed in order to solve a specific translation task? The description is thus an attempt at abstracting away from the human translator in order to investigate to what extent

specific bodies of parallel texts could have been translated automatically. Chapter 4 presents the empirical method applied in this investigation.

The four correspondence types differ in several respects. They are described in terms of the linguistic characteristics of the relation between source and target string, and with respect to the structure and weight of the translation task. The structure of a translation task pertains to the amounts and types of information required to solve it, and its weight concerns the effort needed to process those information sources. We keep the subtask of source text disambiguation apart from the translation task, which means that the complexity analysis of given translational correspondences applies to source strings, given their relevant interpretation, in relation to specific target strings.

The correspondence types are organised in a hierarchy reflecting a gradual increase in the degree of translational complexity: type 1 is the least complex, and type 4 is the most complex. Along this hierarchy we may observe, firstly, an increase with respect to linguistic divergence between source and target string; secondly, an increase in the need for information and in the amount of effort required to translate, and, thirdly, a decrease in the extent to which there exist implications between relations of source-target equivalence at different linguistic levels.

Different levels of interpretation of the source string are required in order to solve the translation task. With respect to types 1 and 2, it is not necessary to analyse the source string further than to the levels of constituent and functional structure, as there is a high degree of structural similarity between source and target string. Thus, the target string may be generated on the basis of information about the structure of the source string, and about lexical and structural correspondences between SL and TL. With respect to type 3, it is necessary to derive also a semantic representation of the source string, since source and target string exhibit structural divergences, although they are semantically equivalent. The target string may be generated on the basis of the information contained in the semantic representation of the source string, and the information available in the TL grammar and lexicon.

With respect to processing effort, the workload required by syntactic parsing of the source expression is shared by all correspondence types. In types 1 and 2, since the target expression can be generated mainly on the basis of the constituent structure

of the source expression, the major effort of the translation task is involved in the syntactic analysis of the source text. Moreover, in the language pair English-Norwegian, types 1 and 2 tend to occur in correspondences between relatively short and syntactically simple expressions, so that, altogether, a modest processing effort is required by translation tasks of these types. Then, in type 3, the translation task is far heavier than in the two lower types, firstly, since the subtask of analysis requires a full semantic analysis of the source expression, and secondly, because it is a resource-intensive computationally problem to generate the target expression on the basis of a semantic representation of the source string. In the case of type 4, analysing the source string to the level of semantic structure is no longer sufficient, as there is semantic divergence between original and translation, and it is necessary to exploit additional information not explicitly encoded in the source string, nor available through given, general linguistic information, in order to interpret the source string and generate the given target string. Hence, we assume that translation tasks of type 4 are non-computable as they fall outside the linguistically predictable part of the translational relation, and, given our framework, there is in principle no limit on the processing effort that may be required to solve them. On the other hand, tasks of types 1–3 are computable, and their solutions are predictable from the linguistic information contained in the source expression and in the finite domain of information about the two language systems.

According to the view taken in 2.3.3, the notion of literal translation covers correspondences of types 1–3. The most important distinction drawn in the present framework is the division between, on the one hand, the computable correspondences of types 1, 2, and 3, and, on the other hand, the non-computable correspondences of type 4. The computable, or linguistically predictable correspondences have in common that source and target expression are semantically equivalent in the sense that the same informational content is linguistically encoded in both of them. The importance of the analytical distinction between the computable and the non-computable will become clear through the discussion of empirical results in chapter 5.

Non-literal translation, represented by type 4 correspondences, can be seen as the topic of studies of human translation, and in a sense falls outside of the present

project. However, in order to clarify the division between the computable and the non-computable, we find it useful to discuss both literal and non-literal translation. For that reason we will in chapter 6 present certain linguistic phenomena which seem impossible, or at least very difficult, to include in literal translation.

PART III
METHOD

4 Empirical investigation

4.1 Overview

This chapter is divided into three main parts. The first part, 4.2 with subsections, presents an overview of the English-Norwegian parallel texts from which the empirical data in our study have been extracted, and discusses the concerns underlying the selection of texts, as well as certain features characteristic of the analysed texts.

The second part, 4.3 with subsections, presents the principles behind our empirical method. Here the main topics are, first, the syntactic criteria which determine how units of translation are identified; second, challenges encountered when applying those criteria, and, third, the principles governing the classification of string pairs with respect to the degree of translational complexity.

The third part, 4.4 with subsections, describes practical aspects of the recording of translational correspondences, such as the linguistic analysis carried out in order to identify correspondences, and the software used for storing and organising the recorded data.

4.2 Text material

The centrepiece of the present project is an empirical investigation of selected parallel texts of English and Norwegian. The collected data constitute a manually analysed and annotated corpus of about 68 000 words. The corpus covers both directions of translation, and it includes two text types, fiction and law texts. The texts from which data are compiled were produced in the period 1979–1996; further bibliographical information is given in the list of primary sources.

Table 4.1. An overview of the analysed text pairs with respect to text type, direction of translation, and numbers of running words.

Authors and texts	Text type	SL	TL	No. of words
Agreement on the European Economic Area, Articles 1–99 (*AEEA*)	law text	English		9 202
Avtale om Det europeiske økonomiske samarbeids-område, artiklene 1–99			Norwegian	8 015
Lov om petroleums-virksomhet, §§1–65 (*Petro*)	law text	Norwegian		7 929
Act relating to petroleum activities, Sections 1–65			English	9 647
André Brink (AB)	fiction			
The Wall of the Plague		English		4 021
Pestens mur			Norwegian	4 230
Doris Lessing (DL)	fiction			
The Good Terrorist		English		4 008
Den gode terroristen			Norwegian	4 652
Erik Fosnes Hansen (EFH)	fiction			
Salme ved reisens slutt		Norwegian		4 022
Psalm at Journey's End			English	4 395
Bjørg Vik (BV)	fiction			
En håndfull lengsel		Norwegian		4 010
Out of Season and Other Stories			English	4 550
Total				68 681

From pairs of source and target texts we have extracted translationally corresponding text units, using the finite clause as the basic level of analysis. How the empirical data have been analysed and compiled is described in 4.3 and 4.4 with subsections, and table 4.1 gives an overview of text type, direction of translation, and numbers of running words for each of the text pairs that have been investigated.

The total number of words in the compiled data (about 68 000) testifies that the scope of our empirical investigation is quite limited. In comparison, the English-Norwegian Parallel Corpus (ENPC), contains about 2,6 million words (cf. 1.4.3.2). Moreover, the total number of text pairs is as low as 6. Hence, the amount of data, as well as the number of text samples, do not satisfy the requirements of representativity in corpus building (cf. 1.4.3.1). Further, among the empirical data it is not possible to detect statistically significant differences between the various text pairs with respect to dimensions such as author preferences, gender, text type, and direction of translation.[1] However, it has been an aim that if these dimensions could not be sufficiently controlled for, they should at least be present among the collected data, in an attempt to avoid accidental overrepresentation of one or more of the mentioned aspects. We have aimed at compiling comparable amounts of data for each of the text types and directions of translation included among the text pairs.

4.2.1 Concerns underlying the selection of texts

The texts have been chosen with regard to certain criteria. These may be summed up as direction of translation, text type, variation between individual authors, and lawful access, and are discussed in 4.2.1.1–4.

4.2.1.1 Direction of translation

Another dimension that may have consequences for translational complexity is the direction of translation, and hence English and Norwegian appear as both source and target language for both text types.

[1] Cf. further comments in 5.2.2 on the limited scope of the empirical material.

That the direction of translation is important is for instance shown by Cathrine Fabricius Hansen's research on informational density in a cross-linguistic perspective.[2] Her notion of 'informational density' relates to discourse structure and may be understood as the amount of information expressed per linguistic unit: the measurement applied is "the relative frequency of new discourse referents and nonredundant conditions per sentence", and if two discourses are compared, "the discourse with the highest average concentration of information per sentence will be judged more loaded with information than the other" (Fabricius-Hansen 1996: 526). Fabricius-Hansen (1999) demonstrates differences in informational density between translationally parallel text sequences of English, German, and Norwegian, respectively, and she relates this to structural differences between the three languages. Among them, German shows the greatest tendency of using strongly hierarchical syntactic structures that allow a larger amount of information to be expressed per linguistic unit than what may be encoded in structures of a less hierarchical nature. Thus, translating from German into English or Norwegian involves unpacking elements of information and recoding them in informationally less dense structures of the target language. This is due to the tendency that non-clausal elements in German texts are frequently translated as clausal elements in English and Norwegian.

This phenomenon appears either as *clausal expansion*, where a non-clausal structure in the source text is converted into a subclause in the target text, or as *sentence splitting*, where one independent sentence in the source text is converted into a sequence of independent sentences in the target text. With respect to clausal expansion, the tendency is that while the German source sentence is informationally more dense than English or Norwegian translations, the English or Norwegian target sentences are more explicit than the German original because the translations contain a larger amount of overt linguistic material expressing the message of the source sentence (cf. Fabricius-Hansen 1999: 178–179). In cases of sentence splitting there is also a higher degree of informational density in the German source text than in the translations, and as the information conveyed by one independent sentence in the

[2] Cf. Fabricius-Hansen (1996), (1999).

original is distributed over a sequence of sentences in the translations, the discourse structures of the target texts are more incremental than that of the source text (cf. Fabricius-Hansen 1999: 183).[3]

Granted that such cross-linguistic differences with respect to informational density, explicitness, and incremental structure are reflections of structural differences between these languages, it becomes evident, in the light of Fabricius Hansen's studies, that the task of decoding the information expressed by a given text is influenced by the structural properties of the language in which the text is created. Further, as translation relies (at least) on the successful extraction of the information encoded in the source text, it is clear that the challenges involved in translating from English or Norwegian into German are different from those connected with translating German into English or Norwegian. From the observations reported in Fabricius-Hansen (1999) regarding English and Norwegian as target languages in relation to German, we do not want to predict anything with respect to how aspects like informational density and explicitness have consequences for translation within the language pair English-Norwegian, and our data from English-Norwegian parallel text have not been analysed in terms of Fabricius Hansen's notion of informational density. However, on the background of her studies, it is highly likely that the direction of translation is a factor that may influence the degree of translational complexity in specific translation tasks, and this motivates including both directions of translation in our empirical investigation.

4.2.1.2 Text type

As previously indicated in 1.1, it is an objective in our project to investigate how the dimension of text type may influence translational complexity. We apply a quite general understanding of the concept of 'text type'; it is a class of texts with a set of properties defining what the members of the class have in common and how they can be distinguished from instances of other text types. Defining the notion of text type calls for relating it to the concept of 'genre'.

---

[3] Clausal expansion and sentence splitting are also relevant to the translational phenomenon of *explicitation*, presented in 5.2.2.

Trosborg (1997b: 6) defines ‘genre’ as a category of texts, both written and spoken, used in particular situations for a particular purpose; Swales (1990: 45–47) explains ‘genre’ as “a class of communicative events” with “a shared set of communicative purposes.” Further, Trosborg (1997b: 12) shows how genres and text types may cut across each other: “Texts within particular genres can differ greatly in their linguistic characteristics; ... On the other hand, different genres can be quite similar linguistically; ...”[4] Thus, the two notions intersect, but they do not overlap fully: “Whereas the notion of genre refers to completed texts, communicative function and text type, being properties of a text, cut across genres” (Trosborg 1997b: 12). Hence, the two notions should be kept apart, and citing Biber (1989: 6), Trosborg concludes that “[g]enres and text types are clearly to be distinguished, as linguistically distinct texts within a genre may represent different text types, while linguistically similar texts from different genres may represent a single text type” (1997b: 12).

Within various fields of linguistic research there exist different approaches to the notion of text type, and Trosborg (1997b: 13–17) presents several of these. In one approach, represented by Hatim and Mason (1990), text types are sorted on the basis of “communicative intentions serving an overall rhetorical purpose” (1990: 140), and within this tradition text types are seen as constituting a limited set: description, narration, exposition, argumentation, and instruction (cf. Trosborg 1997b: 15). This is in contrast to the approach of Biber (1988, 1989), in which text types are identified on the basis of purely linguistic criteria, and Biber (1988: 13) explains the contrast clearly: “Most analyses begin with a situational or functional distinction and identify linguistic features associated with that distinction as a second step. ... The opposite approach is used here: quantitative techniques are used to identify the groups of features that actually co-occur in texts, and afterwards these groupings are interpreted in functional terms.” In Biber’s study the term *genre* refers to “categorizations assigned on the basis of external criteria”, and the term *text type* refers to “groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories” (1988: 70). In our view, it is an attractive feature of Biber’s approach that

[4] What is omitted in this quotation are Trosborg’s illustrations of the points she has made.

it is based on directly observable, empirical facts about language use. With respect to the opposite approach, we find its slightly essentialistic character problematic, and we are skeptical of viewing text types as a closed set since analysing new instances of text in relation to a fixed set of categories may easily lead to a revision of those categories.

However, both approaches are useful since they contribute in various ways to a better understanding of textual features, and the importance of the dimension of text type for translation is nicely summed up by the concluding remarks of Trosborg (1997b: 18): "Text typology with genre conventions and knowledge of how communicative functions and text types are realized in different languages within and across genres are useful knowledge in translator training and in translation itself."[5]

It could be argued that the notion of genre may safely be applied to the two text types investigated in our study, i.e. the literary genre of fiction, and the legal genre of law text. Also, both types are, according to linguistic criteria, sufficiently distinct from each other to qualify as different text types in Biber's sense (cf. 4.2.1.2). In order to prevent any confusion, we will avoid using the term *genre* and concentrate on the notion of text type. This is primarily a choice of term, and several textual aspects treated in our investigation are relevant to genres as well as to text types.

On the basis of the preceding discussion, we regard it as uncontroversial that text types will differ with respect to characteristic linguistic features, and with respect to the properties of the optimal translation. For instance, in the case of an informative, non-fictional text, it is important that the informational content of the original is preserved in the translation, whereas in the case of a literary text such as a poem, the preservation of informational content may be secondary to the aim of creating a translation with aesthetic and pragmatic qualities as similar as possible to those of the original. Such types of variation may create different challenges to the translator, and may influence the complexity of the translation task. Text type is thus relevant to the present investigation. As previously mentioned in 1.1, the analysed fiction texts are chosen as instances of unrestricted text types, whereas the law texts instantiate

[5] The language-specificity of textual conventions has previously been commented on in 2.4.2.1.

restricted text types. The difference in restrictedness between the two types will be discussed in 5.4.2 with subsections.

4.2.1.3 Variation between individual authors

A third concern underlying the selection of texts is that there may be variation in the language of individual authors with respect to linguistic characteristics caused by stylistic preferences. Leech and Short (2007: 9) point out that the concept of 'style' belongs to the level of *parole* as it is created by the choices that speakers or writers make within the repertoire offered by a language system. Further, they argue that style pertains to "characteristic uses of language" within a certain domain, or "corpus of writings" (2007: 10), and their preferred definition of 'style' is "the linguistic characteristics of a particular text" (2007: 11).

Since stylistic preferences on the part of individual writers may influence the linguistic properties of texts, our empirical material covers texts produced by more than one author for each direction of translation, and we have used texts written by both males and females, with respect to the translations as well as the originals (cf. table 4.2 in 4.2.2.2). Among the fiction texts, empirical data are compiled from two different text pairs for each direction of translation, with an even distribution of male and female source text writers. Regarding the law texts, we have no information about the specific persons who have produced them, but such texts are, typically, written by more than one legislator and also translated by teams, which made it reasonable, in this perspective, to choose only one text pair for each direction of translation.

4.2.1.4 Lawful access

A different issue that has been quite decisive for the selection of texts is that of gaining lawful access to text material. As pointed out in 1.4.3.1, a requirement for working efficiently with corpus data is that such resources are machine-readable, and when text is stored electronically, it may easily be copied, distributed, or otherwise manipulated in ways that are illegal in relation to copyright provisions. Hence, specific permission must be obtained from the copyright holder if copyrighted text is

to be stored electronically and exploited for research purposes. Thus, in the present project the selection of texts has been limited to text material for which lawful access would exist. We have used some publicly available texts, as well as texts which have been made available specifically for research purposes by the copyright owners. As regards the law texts, they are in principle publicly available. *Agreement on the European Economic Area* is distributed in Norway by the public foundation Lovdata, through which the English version of the agreement text was made available for our purposes, with no restrictions. We were granted similar permission by the Norwegian Ministry of Foreign Affairs for the Norwegian translation, which was created by translators of the Ministry. Concerning *Lov om petroleumsvirksomhet* (*Act relating to petroleum activities*) we were given free access by the Norwegian Petroleum Directorate to exploit the original as well as its translation. With respect to the fiction texts we have gained access through cooperation with the English-Norwegian Parallel Corpus (ENPC) Project of the University of Oslo.[6] For the ENPC permissions have been granted to store electronically extracts of maximum 15000 words from each text and to exploit them for research purposes only (cf. Johansson 2007: 13).

4.2.2 Textual features

Sections 4.2.2.1–2 present certain features of the analysed texts. We will focus on properties that distinguish the two chosen text types from each other, and we will mention some characteristics of the various texts, seen as instances of parallel text. The presentation is intended merely as a brief description, and not as an exhaustive discussion of the linguistic characteristics of the two text types. In chapter 5 we will discuss further the text-typological contrast between law text and fiction, as well as particular aspects of each text pair.

4.2.2.1 The law texts

The investigated law texts contain sets of sequentially numbered sections, or articles. We will prefer to use the term *law text* about the text type at hand. For our purposes

[6] The ENPC is documented in Johansson (1998, 2007), and Johansson et al. (1999/2002). Cf. 1.4.3.2.

law may, depending on the context, be used as a shorthand for *law text*, although there is of course a clear distinction between the law itself, i.e. its legal content, and the text which expresses it. Synonymous terms for *law* are *act* and *statute*. The authors of law texts may be referred to as *legislators* or *drafters*. Written laws are a kind of legal text, but since *legal text* also refers to other text types related to the legal domain, that expression will not be applied to the investigated law texts. According to Cao (2010: 78), "[l]egal texts refer to the texts produced or used for legal purposes in legal settings." Cao presents one possible way of dividing legal texts into subtypes, as she distinguishes between four "major variants or sub-varieties of written legal texts": (i) "legislative texts", among which law texts are included, (ii) "judicial texts produced in the judicial process", (iii) "legal scholarly texts", and (iv) "private legal texts" (2010: 79). It is only the first kind of legal text that is treated in this study. The writing of law texts is frequently referred to as *legal drafting*, and we may also use the term *(legal) drafting* to refer to the writing of a law text, although that term can also refer to the production of legal text types other than law texts.

In the present project one pair of translationally corresponding law texts has been analysed for each direction of translation (cf. table 4.1 in 4.2). They are written, respectively, in British English and in the Norwegian standard of writing referred to as *bokmål*. Both text pairs are extracts, running from the first section onwards, and excluding tables of contents. The law texts are written in a formal, impersonal style, with frequent use of long, complex sentences.[7] The texts are repetitive in the sense that specific expressions are recurrent.[8] Other characteristics are heavy constituents, enumerative listing, complex coordination, no occurrences of first and second person pronouns, and numerous instances of nonfinite constructions, especially in the English texts. Another salient feature is the high frequency of headings, normally realised as noun phrases. The latter is clearly a text type-specific feature, a result of the convention of introducing each numbered article with a heading, like *Article 1*. In addition, the documents contain a number of chapter and subchapter headings.

[7] Mattila (2006: 98), citing Laurén (1993: 74), observes that "[s]entences in legal language are longer than those of other languages for special purposes and they contain more subordinate clauses."

[8] E.g., some recurrent expressions found in the *AEEA* are: *within the framework of*, *with a view to*, *without prejudice to*.

As already stated, we regard law text as a restricted text type, and this point will be developed further in 5.4.2 with subsections. For discussions of the linguistic aspects of law texts, see Bowers (1989), Tiersma (1999), Mattila (2006), Cao (2007), Hutton (2009), Coulthard and Johnson (2010); on English statutory language, see Bowers (1989), Tiersma (1999), Mattila (2006), and on the language of Norwegian laws, see Vinje (1990b), (1995). Figure 4.1 shows a sample of translationally parallel law texts, namely the English and Norwegian versions of Article 96 of the *Agreement on the European Economic Area*:

Article 96.

1. Members of the Economic and Social Committee and other bodies representing the social partners in the Community and the corresponding bodies in the EFTA States shall work to strengthen contacts between them and to cooperate in an organized and regular manner in order to enhance the awareness of the economic and social aspects of the growing interdependence of the economies of the Contracting Parties and of their interests within the context of the EEA.

2. To this end, an EEA Consultative Committee is hereby established. It shall be composed of equal numbers of, on the one hand, members of the Economic and Social Committee of the Community and, on the other, members of the EFTA Consultative Committee. The EEA Consultative Committee may express its views in the form of reports or resolutions, as appropriate.

3. The EEA Consultative Committee shall adopt its rules of procedure.

Artikkel 96

1. Medlemmer i Den økonomiske og sosiale komité og andre organer som representerer arbeidslivets parter i Fellesskapet, og de tilsvarende organer i EFTA-statene skal bestrebe seg på å styrke kontakten seg imellom og å samarbeide på en organisert og regelmessig måte for å styrke bevisstheten om de økonomiske og sosiale sider ved den økende samhørighet mellom avtalepartenes økonomi og deres interesser i EØS-sammenheng.

2. For dette formål skal det opprettes en Rådgivende komité for EØS. Den skal være sammensatt av et likt antall medlemmer fra Den økonomiske og sosiale komité i Fellesskapet på den ene side og fra EFTAs Rådgivende komité på den annen side. Den rådgivende komité for EØS kan gi uttrykk for sine synspunkter i form av rapporter eller resolusjoner.

3. Den rådgivende komité for EØS skal vedta sin forretningsorden.

Figure 4.1. A sample of translationally parallel law texts.

The *Agreement on the European Economic Area* is an instance of an international legal instrument, more precisely an example of supranational legislation. The *EEA Agreement*, together with its translations, is a parallel text in the special sense that the status of each version is equal in each of the languages in which it exists. This means

that although some of its versions are *de facto* translations, all versions have the same legal status in their respective language communities.[9] The Norwegian version of the *EEA Agreement* has been translated from English, but, according to the Norwegian Ministry of Foreign Affairs, the translation is based also on the French version. That is, on several occasions choices made by the translators are determined not only by the English source text, but also by expressions used in the French version. This illustrates the supranational aspect. Certain textual features shared by both of the analysed pairs of law texts are mentioned above; for further details on the textual features of international legal instruments, see Cao (2007: 143–7), (2010: 89–90).

In the case of the second pair of law texts, original and translation do not have equal status. The Norwegian version *Lov om petroleumsvirksomhet* contains a law, regulating petroleum activities in areas belonging to Norway, whereas the English *Act relating to petroleum activities* is an unofficial translation of the Norwegian text. Thus, the English translation does not have the status of a law text, and to the target language reader it functions as a source of information about the content of the Norwegian law. *Lov om petroleumsvirksomhet* is an example of domestic legislation. The English translation can be seen as a clear instance of overt translation, whereas the Norwegian version of the *AEEA* is an example of covert translation, as defined by House (1997) (cf. 1.4.1.2). More information on domestic and supranational law, as well as on the two pairs of law texts, is provided by the discussion in 5.5.1.2.

4.2.2.2 The fiction texts

The analysed fiction texts are extracts of novels, except for the text by Bjørg Vik, which is taken from a short story (cf. the list of primary sources). Each extract runs from the beginning of the narrative, and none of them is a complete text. The selected fiction texts are stories evolving around a certain protagonist and other characters, and passages of dialogue are found in all of them.

In contrast to law text, a literary text type like narrative fiction can be described as unrestricted, at least in terms of the inventory of syntactic constructions that may

[9] Cf. the discussion of the legal principle of equality of authentic texts in 5.5.1.2.

occur. Narrative fiction texts may comprise all kinds of sentence types: simple as well as complex, declarative, interrogative, and imperative sentences. Also, direct speech may occur.[10] Furthermore, literary texts can include passages of other text types, which may add to the structural diversity. E.g., in a novel, poetry, songs, or even passages of non-fiction may be included. For discussions of the linguistic aspects of English fiction texts, see Leech and Short (2007), as well as Leech (2008), and with respect to Norwegian fiction texts, see Dahl (1995). Figure 4.2 presents a sample of translationally parallel fiction texts, i.e. the first two paragraphs of Doris Lessing's novel *The Good Terrorist*, shown together with its Norwegian translation.

THE house was set back from the noisy main road in what seemed to be a rubbish tip. A large house. Solid. Black tiles stood at angles along the gutter, and into a gap near the base of a fat chimney a bird flew, trailing a piece of grass several times its length.

"I should think, 1910," said Alice, "look how thick the walls are." This could be seen through the broken window just above them on the first floor. She got no response, but nevertheless shrugged off her backpack, letting it tumble on to a living rug of young nettles that was trying to digest rusting tins and plastic cups. She took a step back to get a better view of the roof. This brought Jasper into vision. His face, as she had expected it would be, was critical and meant to be noticed. For her part she did not have to be told that she was wearing *her look*, described by him as silly. "Stop it," he ordered. His hand shot out, and her wrist was encircled by hard bone. It hurt. She faced him, undefiant but confident, and said, "I wonder if they will accept us?" And, as she had known he would, he said, "It is a question of whether we will accept them."

Huset lå litt tilbaketrukket fra hovedveien, midt i noe som minnet om en søppelfylling. Et stort hus. Massivt. Svarte takstein hadde kilt seg fast i uryddige vinkler langsmed takrennene, og oppe ved skorsteinen gapte et mørkt hull; en fugl smatt inn i hullet med et strå i nebbet, strået var flere ganger lengre enn den vesle fuglekroppen.

"1910, ville jeg tro," sa Alice, "se hvor tykke veggene er." Hun kunne se dette gjennom den istykkerslåtte glassruten i vinduet rett over dem, i annen etasje. Noe svar fikk hun ikke, men likevel slapp hun av seg ryggsekken og lot den falle blant de spede brenneslene hun sto i, et levende teppe på bakken, de prøvde å ernære seg av rustne hermetikkbokser og engangsglass av plast. Hun tok et skritt bakover for å få overblikk over taket. Jasper kom inn i synsfeltet hennes. Ansiktet hans var som hun hadde ventet, bevisst kritisk. Ingen behøvde å fortelle henne at hun, for sin del, hadde *fjeset sitt* på, et ansiktsuttrykk han pleide å karakterisere som tåpelig. "Kutt ut," befalte han. Hånden hans skjøt fram, og håndleddet hennes var fanget i et hardt, benete grep. Det gjorde vondt. Hun så på ham, trygt og uten opprør. "Tror du de kommer til å godta oss?" sa hun. Og han sa det hun hadde visst han ville si: "Det spørs vel heller om vi kommer til å godta dem."

Figure 4.2. A sample of translationally parallel fiction texts.

[10] Cf. Ochs (1997: 185–189) on the diversity of narratives. On the narrative in general, see Abbott (2002), and Toolan (2001).

The English fiction texts are created in different varieties of English, i.e. American English, British English and South African English, whereas the Norwegian texts are all written in *bokmål*. Table 4.2 displays the distribution of varieties among the fiction text pairs. It has not been a concern behind the selection of texts to have different language varities represented among our data. This is rather an accidental feature of the fiction texts that were available, but should nevertheless be mentioned.

Table 4.2. Overview of the language varieties used in the fiction text pairs.

Author, source text	**SL variety**	**TL variety**	**Translator(s), target text**
André Brink (AB) *The Wall of the Plague*	South African English	*bokmål*	Per Malde *Pestens mur*
Doris Lessing (DL) *The Good Terrorist*	British English	*bokmål*	Kia Halling *Den gode terroristen*
Erik Fosnes Hansen (EFH) *Salme ved reisens slutt*	*bokmål*	American English	Joan Tate *Psalm at Journey's End*
Bjørg Vik (BV) *En håndfull lengsel*	*bokmål*	British English	David McDuff, Patrick Browne *Out of Season and Other Stories*

In the case of the investigated fiction texts, the TL versions are all examples of overt translation. They are presented to the target language readers as translated literature; they are not translations of the kind where cultural-specific features of the source text are adapted to the target language community. The fiction texts are translated by professional translators, and there is an interesting degree of variation within these four text pairs concerning the traditional opposition in translation studies between freeness and literalness in relation to the original text. This topic is discussed further in 5.5.2.2.

4.3 Methodological principles

In the empirical method applied in the present project, translationally corresponding text units, or string pairs, are extracted from parallel texts and classified according to the measure of translational complexity described in chapter 3. The linguistic analysis underlying the identification and classification of string pairs is done "manually" by a bilingually competent human annotator (i.e. the present author); cf. 4.4.1. A computer program, to be described in 4.4.2, is used for storing and organising the analysed data.

In the present project the analysis of translational correspondences is applied to running text. In principle, however, the analysis method does not require continuous text, since the classification of correspondences pertains to individual string pairs only. Translational correspondences are extracted from the parallel texts according to syntactic selection criteria, which in turn define what is regarded as units of translation for the purposes of our analysis. When all collected string pairs have been assigned one of the four correspondence types, we may calculate the distribution of types within the recorded data. It has been a process of numerous revisions both to determine what types of syntactic units to extract, and to define operational criteria for drawing the borders between the four correspondence types. Thus, data collected initially were reanalysed until the categories and criteria had reached a form that seemed appropriate to the purposes of our project.

In 4.3.1–6 with subsections we present the notion of 'translational correspondence'; we discuss the syntactic criteria for string pair extraction, as well as practical problems involved in applying the criteria, will be discussed, and we present the principles governing the assignment of correspondence type to individual string pairs.

4.3.1 The notion 'translational correspondence'

In the present work the notion of 'a translational correspondence' covers a pair of translationally related linguistic units of two different languages. In 1.3 the alternate terms *correspondence* and *string pair* are introduced as referring to translationally related pairs of word strings. It is a prerequisite for the extraction of string pairs from

parallel texts that it is possible, for each unit of translation identified, to find its translational match in the parallel text.[11]

In our analysis translational units are identified, in a given text, by means of the syntactic criteria to be discussed in 4.3.2 with subsections. However, with respect to identifying the translational correspondent of a certain unit, the meaning and interpretation of the parallel string are more important than its syntactic properties.[12] When identifying translational correspondences we apply a rather heuristic method in which we look for the closest possible match, given the texts. In the majority of cases it is a straightforward task to see what part of the target text is the translational correspondent of a given source string. In other cases some piece of meaning expressed in a certain string may not have any match in the parallel text, as it happens that meaning can be added or deleted during translation. There may even be cases where two particular strings, although they do not correspond with respect to what they express, constitute a string pair simply because other possible correspondence relations are excluded, and since neighbouring strings clearly belong to other string pairs. The present notion of 'translational correspondence' is a parallel to Toury's (1995: 77) idea of "coupled pairs" of source and target text segments, which are seen as units of analysis that mutually determine each other (cf. 3.3.1.1).

In principle, if some unit of meaning α, expressed in $text_1$, corresponds with a unit of meaning β, expressed in $text_2$, then α and β must be contained in the same string pair (unless technical limitations on part of the software makes it impossible; cf. 4.3.5.1). For the lower correspondence types 1, 2, and 3 this criterion is necessarily satisfied when a pair of translationally matching strings otherwise fulfils the requirements given by the definitions of those correspondence types. For string pairs of type 4, the criterion implies that although there are differences in meaning between the translationally related strings, it should not be the case that translationally corresponding units of meaning are contained in different string pairs.

---

[11] Cf. 4.3.5.3 for a discussion of problem cases where a given string does not have any translational correspondent in the parallel text.

[12] This point is illustrated by the approach chosen for handling partial translational correspondences between units of extraction; cf. 4.3.5.2.

4.3.2 Syntactic criteria for string pair extraction

As pointed out in 4.3.1, string pair extraction is based on assumptions about what may constitute units of translation. For our purposes we have chosen a limited set of syntactic units, and the selection of units is influenced by the wish to make our study of translational complexity relevant to the field of machine translation (MT). We have tried to envisage a way of segmenting text material that would be suitable for automatic translation regardless of specific algorithms for implementation.[13] If a fairly broad generalisation is allowed with respect to the linguistic approaches to automatic translation, we may claim that MT systems typically operate sentence by sentence, and hence we have chosen *the finite clause* as the basic unit of translation in this investigation of English-Norwegian parallel texts. This is not to say that rule-based MT systems exclusively process finite sentences; there are applications also operating on sub-sentential units. This topic will not be pursued, since we do not study algorithms for automatic translation. Clearly, in order to be of any use, an MT system must handle syntactic units at least as complex as those of the sentence level.

As a starting point, we apply a very simple understanding of 'finite clause': it is a syntactic unit containing a finite verb not embedded in any other unit. There is a variety of finite constructions in these two languages, and (1) below is a summary of the types of syntactic categories that occur as translational units in our analysis. It has been an aim to define search criteria allowing us to delimit translational units on the basis of surface syntactic structure.[14]

With respect to the language pair in question, it is quite common to find translational links between, respectively, finite and nonfinite constructions.[15] In this context nonfinite constructions cover non-clausal as well as clausal nonfinite constructions. A nominal subclause may for instance be matched by a noun phrase, and *category crossing* is a frequent phenomenon among the string pairs collected. That is, many correspondences hold between strings of different syntactic categories.[16] In the

---

[13] As stated in 1.3.2 and 1.4.2.3, the present investigation is not related to any particular MT architecture.
[14] This point is commented on towards the end of section 4.3.2.3.
[15] This topic will be developed further in 5.2.2 and 6.3.1.3.
[16] Such category crossing occurs only in correspondences of types 3 and 4, as it violates the syntactic restrictions that apply to types 1 and 2.

illustrations of extraction criteria in 4.3.2.1–4, examples will be presented where the top nodes of respectively source and target string are of the same syntactic category. The correspondences have been chosen to illustrate the extraction criteria, and the high degree of structural parallelism between source and target strings must not be seen as representative for the entire set of recorded data.

Syntactic analysis of source and target text is a prerequisite for compiling string pairs. In the present project syntactic analysis and description is based on rudimentary X'-analysis. Our syntactic approach is primarily to apply a classificatory system serving the practical purposes of the empirical investigation. What we have aimed at is to identify the syntactic type of each extracted string without being too detailed with respect to underlying assumptions pertaining to syntactic theory. In general terms, we apply an X'-analysis in line with the framework of Lexical-Functional Grammar (LFG).[17] In syntactic representations we assume no transformations, and we do not assume the existence of empty nodes. Further, we adhere to the principle of lexical integrity, according to which the terminal nodes of constituent structures are assumed to be morphologically complete words; cf. Bresnan (2001: 92). Moreover, we assume the noun to be the head in phrases with determiners; such phrases are thus described as noun phrases (NPs), not as determiner phrases (DPs).

A rule of thumb in our empirical method is that the occurrence of a finite verb, in either of the two parallel texts, will trigger the extraction of a string pair.[18] (1) is a summary of the extraction criteria applied in our investigation:

(1) A pair of translationally corresponding strings is extracted if at least one of the two strings is:
(a) a matrix sentence, including both simple and multiple sentences,
(b) or a finite subclause, if functioning as a sentence element,
(c) or an XP, where X is a lexical category, and the XP contains at least one finite subclause as a complement of X:[19] [$_{XP}$ ... X ... [$_{CP}$ ...] ...],

[17] For an introduction to LFG, see Dalrymple (2001). On the topic of X'-syntax in LFG, see Bresnan 2001).
[18] Some exceptional cases where a finite clause does not cause the extraction of a string pair are discussed in 4.3.5.3.

(d) or an expression with no finite verb which is marked by punctuation as a textual sentence.

Note that according to the requirements of (1b) and (1c), subclauses which are complements of lexical phrases are not extracted separately as long as the entire phrase is (cf. 4.3.2.3). The criteria summed up by (1) are further discussed and illustrated in 4.3.2.1–4.

As English and Norwegian are configurational languages exhibiting recursivity in syntactic structures, it is frequently the case that a syntactic unit satisfying one of the extraction criteria is embedded in another unit also satisfying an extraction criterion. The embedded unit and its translational match will then constitute a subcorrespondence embedded in a supercorrespondence, which is made up of the superordinate syntactic unit and its correspondent. This can be described as *nesting* of string pairs, and appears in several examples to be discussed in 4.3.2.1–4. When one string pair is embedded in another, they count as separate items among our data, but the subcorrespondence is not included in its full length in the supercorrespondence, as only the category labels of the embedded strings are included in the supercorrespondence.[20] The reason is that a sequence of words should not, in the quality of a word string, belong to more than one extracted correspondence in order to avoid duplicates among the compiled data. In 4.3.3 we will return to the topic of nested string pairs.

4.3.2.1 Matrix sentence

According to extraction criterion (1a), a pair of translationally corresponding strings is extracted if at least one of the two strings is a *matrix sentence*, which may be of several kinds.

Firstly, *simple* matrix sentences contain a single independent clause; cf. Quirk et al. (1985: 719). In our analysis, all instances of this category are treated as trans-

[19] *At least* is here included in order to cover cases where conjoined finite clauses appear as the syntactic complement of X; cf. example (19) in 4.3.3. The syntactic label *CP* (finite subclause) is commented on in 4.3.2.2, and an overview of syntactic labels used in the present study is given in 4.4.3 in tables 4.3–4.

[20] See examples (8), (10), and (14) in the following sections.

lational units. String pair (2) is a pair of simple matrix sentences, found among our empirical data.[21]

(2a) IP: She took a step back to get a better view of the roof. (DL)
(2b) IP: Hun tok et skritt bakover for å få overblikk over taket.

In (2) the labels *IP* stand for independent, or matrix, sentences. In X'-syntax *I* (shorthand for INFL, or *inflection*) is the label associated with the syntactic position allocated to the head of a sentence. According to this, a sentence is an IP, or inflectional phrase. In relation to X'-syntax it is an oversimplification to categorise all independent sentences as IPs. Still, in the present investigation we have chosen *IP* as the label for matrix sentences, in English as well as in Norwegian, primarily because it suits the purposes of our syntactic analysis to apply one label to all units included in this general category.

Secondly, there are *multiple* matrix sentences, which contain more than one finite clause, and these clauses are either coordinated or subordinated; cf. Quirk et al. (1985: 719). In their terminology coordinated clauses constitute *compound sentences*, whereas a multiple sentence containing one or more subordinated clauses is a *complex sentence*. These observations hold for Norwegian as well as for English.

In the case of compound, or conjoined, matrix sentences, each conjunct forms an independent clause and will, accordingly, be extracted as a separate unit. Normally in such cases, the coordinating conjunction is extracted together with its immediately following conjunct. The entire compound sentence is not extracted as a unit in addition to its subparts, as will be illustrated by (3)–(5).[22]

[21] In this chapter information about the syntactic category of extracted translational units will be indicated by category labels immediately following the example numbers. Norwegian examples will be glossed only where it is necessary to bring across properties which are relevant to the discussions of the correspondences.

[22] The reason is, as in the case of nested string pairs (cf. 4.3.2), that any string of words should not belong to more than one string pair, in order to avoid duplicates among the compiled data.

(3a) I'd skipped lunch and felt hungry; but that other hunger was more demanding. (AB)
(3b) Jeg hadde hoppet over lunsjen og var sulten; men den andre sulten var mer påtrengende.

(3) is a pair of coordinated matrix sentences; (4) and (5) show the string pairs that are entered in our set of data. (4) is a pair of simple sentences, and (5) is also a pair of simple sentences, each introduced by a coordinating conjunction.

(4a) IP: I'd skipped lunch and felt hungry; (AB)
(4b) IP: Jeg hadde hoppet over lunsjen og var sulten;

(5a) IP: but that other hunger was more demanding. (AB)
(5b) IP: men den andre sulten var mer påtrengende.

As stated above, complex sentences contain one or more subordinated clauses. These may be finite subclauses (cf. 4.3.2.2), or they may be independent sentences. Within the recorded data there are several occurrences of the latter kind, typically found in passages of direct speech, where an independent sentence may be the direct object of the verb referring to the act of speaking. This is the case in (6a), as well as in (6b):

(6a) Jeg går i vaskekjelleren, sa moren spisst. (BV)
(6b) "I'm going down into the laundry room," said the mother in a sharp voice;

However, example (6) does not illustrate a recorded string pair, because the compiled correspondences are not simply pieces of running parallel text.[23] For one thing, they include information about syntactic embedding that is relevant to the identification of translational units. Thus, whereas (6) shows a piece of authentic parallel text, (7) and (8) represent the string pairs extracted from (6):

[23] For more information, cf. 4.4 with subsections.

(7a) IP: Jeg går i vaskekjelleren, (BV)
(7b) IP: "I'm going down into the laundry room,"

(8a) IP: [IP:4] sa moren spisst. (BV)
(8b) IP: [IP:4] said the mother in a sharp voice;

String pairs (7)–(8) illustrate nesting of correspondences, introduced in 4.3.2. (8) is the string pair of top-level matrix sentences, the supercorrespondence, and (7) is the pair of embedded independent sentences, the subcorrespondence. In (8) the embedded independent sentences are represented by syntactic labels given inside square brackets. Category labels of embedded translational units will be given in brackets, and will be followed by a colon and a number, either 1, 2, 3, or 4. The number shows the correspondence type of the embedded string pair.

4.3.2.2 Finite subclause

Complex sentences may contain one or more finite subclauses. As stated by extraction criterion (1b) in 4.3.2, a string pair is extracted if at least one of the two strings is a finite subclause functioning as a sentence element. The subclause and its match in the parallel text then form a subcorrespondence embedded in the pair of superordinated sentences.[24] (9) is a pair of matrix sentences, each containing a finite subclause:

(9a) I'll tell you when I come back. (AB)
(9b) Jeg skal si fra når jeg kommer tilbake.
'I shall say from when I come back.'

Depending on the interpretation assigned to the sentences when situated in contexts of utterance, the subclauses in (9) function either as adverbials or as direct objects,

[24] When a subcorrespondence consists of a subclause and its match, the supercorrespondence is not necessarily a pair of independent sentences, since subclauses functioning as sentence elements may be recursively embedded in other subclauses.

but in either case the segmentation into translational units will be the same.[25] (10) and (11) represent the string pairs extracted from (9): (10) is the supercorrespondence containing the matrix sentences, and (11) is the embedded correspondence of finite subclauses.

(10a) IP: I'll tell you [CP:1] (AB)
(10b) IP: Jeg skal si fra [CP:1]

(11a) CP: when I come back. (AB)
(11b) CP: når jeg kommer tilbake.

In (10) the subclauses are represented by the category labels *CP*. In X'-syntax *C* (shorthand for *complementiser*) is the label associated with the syntactic position allocated to the function word introducing a subclause, the complementiser. This is the head of the subclause, and according to this, a finite subclause is categorised as a CP, or complementiser phrase. In our analysis, the label *CP* is generally used for finite subclauses in English, as well as in Norwegian.

4.3.2.3 Lexical phrase with finite subclause as complement

As mentioned in 4.3.2, the finite clause is regarded as the basic unit of translation in our study, and hence we want to single out all finite clauses in the texts investigated. However, when occurring as a phrase-internal constituent, a finite subclause by itself does not seem a natural unit of translation.[26] Also, as stated in 4.3.2, it has been an aim to define search criteria that allow translational units to be delimited on the basis of surface syntactic structure, a point we will return to below. With respect to phrase-internal finite subclauses, syntactic complementation is chosen as the criterion by which translational units are identified. The notion of 'complement' is here used in a purely syntactic sense. In the basic X'-schema of phrase structure the complement

[25] Due to semantic differences between (9a) and (9b), the object reading of the finite subclause *når jeg kommer tilbake* is not as likely in (9b) as the adverbial reading is.
[26] This point is also discussed in connection with the opacity principle in 4.3.6.5.

position is the sister node of the head of the phrase, X, which means that the complement is immediately dominated by the X'-node; cf. figure 4.3.[27]

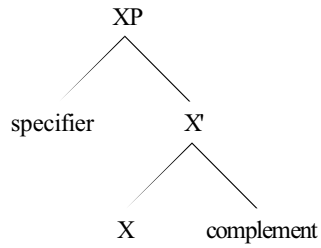

Figure 4.3. Basic X'-schema of phrase structure.

Extraction criterion (1c) in 4.3.2 states that a pair of translationally corresponding strings is extracted if at least one of the two strings is an XP, where X is a lexical category, and the XP contains at least one finite subclause as a complement of X. In other words, a lexical phrase in which a finite subclause is the complement of the head is treated as a translational unit, and is extracted together with its correspondent in the parallel text. Since the subclause is not, in such cases, identified as a translational unit of its own, this is a deviation from the main principle of treating any finite clause as a unit of analysis. The motivation for this is discussed below, and criterion (1c) is illustrated by examples (12)–(14):

(12a) From the instant I turned right at the fork in the road to put Orléans out of reach the urge to be back in the cathedral grew inside me like hunger. (AB)

(12b) Fra det øyeblikket da jeg svingte til høyre i krysset for å legge Orléans utenfor rekkevidde, vokste trangen til å være i katedralen igjen i meg lik en sult.

[27] As pointed out in 4.3.2, we apply certain concepts of X'-syntax, but our analysis is not based on all assumptions underlying that framework; for instance, we do not assume strict binary phrase structure.

(12) is a pair of matrix sentences, each containing a phrase-internal finite subclause. In (12a) the subclause *I turned right at the fork in the road to put Orléans out of reach* is embedded in a noun phrase where the subclause is the complement of the head noun *instant*. The structure of (12b) is parallel to that of (12a); in (12b) the matching subclause is *da jeg svingte til høyre i krysset for å legge Orléans utenfor rekkevidde*. The two translationally corresponding subclauses trigger string pair extraction, i.e. extraction of the entire noun phrases in which the subclauses are embedded. This pair of noun phrases is shown in (13), and the pair of superordinate matrix sentences in (14). Thus, (13) and (14) represent the data extracted from (12).

(13a) NP: the instant I turned right at the fork in the road to put Orléans out of reach (AB)
(13b) NP: det øyeblikket da jeg svingte til høyre i krysset for å legge Orléans utenfor rekkevidde,

(14a) IP: From [NP:3] the urge to be back in the cathedral grew inside me like hunger. (AB)
(14b) IP: Fra [NP:3] vokste trangen til å være i katedralen igjen i meg lik en sult.

That the finite subclause is not a natural unit of translation when occurring as a phrase-internal complement is supported by an observed tendency that there will more likely be a correspondent in the parallel text for an entire phrase than for its subparts. Within the recorded data there are several translational correspondences where each of the two strings contains subparts not matched by corresponding subparts in the other string.[28] This may be illustrated by (15):

(15a) NP: Paul's favourite (AB)
(15b) NP: den Paul likte best
'the-one Paul liked best'

[28] Compositional non-equivalence between translationally corresponding units is discussed in 6.2.4.1.

(15) is a pair of noun phrases, extracted separately because the Norwegian string contains a finite subclause (*Paul likte best*). (15a) is a noun phrase with a genitive determiner (*Paul's*), whereas (15b) is a noun phrase with a postmodifying, restrictive relative clause. There is a direct translational match between the entire NPs, and the two NPs, seen as units, are both referentially and denotationally equivalent in the given texts. On the other hand, if we were to extract the Norwegian subclause *Paul likte best* as a separate unit, it would be difficult to delimit a subpart of the parallel English NP as a translational correspondent to the Norwegian relative clause.

Another point motivating extraction criterion (1c) is that NPs containing relative clauses, like (15b), illustrate how a phrase-internal finite subclause may be closely linked, semantically as well as syntactically, to the rest of the phrase in which it occurs. In the case of (15b) there is a relation of coreference between the antecedent and the gap in the relative clause, and the antecedent forms a unit together with the relative clause. Due to the close link between the relative clause and the rest of the NP it seems unnatural to identify the clause as a translational unit of its own.

NPs with relative clauses is only one kind of lexical phrase with finite subclause as complement. If we consider the general class of lexical phrases with embedded subclauses, there is some heterogeneity with respect to the nature of the relation between the embedded subclause and the rest of the phrase. In some cases the subclause may express information needed in order to determine the interpretation of the entire phrase (as in the case of restrictive relative clauses); in other cases it may simply convey additional information, and there may be instances where it is not quite clear whether the finite subclause has a restrictive or a non-restrictive function. These various types of cases have been encountered in connection with the extraction and analysis of translational correspondences. We have not tried to account for this heterogeneity among the data, since we do not think it would contribute substantially to the analysis of translational complexity. Rather, for the purpose of string pair extraction, we have aimed at treating the whole class of lexical phrases with embedded subclauses in a uniform way: if it is possible, on the basis of surface syntactic structure, to identify a finite subclause as a syntactic complement to the head of a lexical phrase, then the entire phrase is extracted as a translational unit,

regardless of the kind of relation that may hold between the subclause and the rest of the phrase.

4.3.2.4 Punctuation

Finally, according to extraction criterion (1d) in 4.3.2, an expression with no finite verb will be extracted as a translational unit if it is marked by punctuation as a textual sentence. That is, these are expressions which are separated by punctuation in ways that normally delimit sentences, although they do not constitute syntactic sentences. They may, or may not, be syntactically complete phrase-level categories. (1d) specifies the residual category within the extraction criteria: it is meant to cover linguistic units that will not match any of the other criteria because these expressions do not contain finite verbs. Moreover, since such strings are syntactically independent, there are no superordinated translational units in which they may be included. The motivation behind criterion (1d) is that we want to apply our analysis to running text without omitting any parts of the material.

To illustrate, the parallel text samples in (16) consist of sequences of expressions separated by periods, and a pair of noun phrases satisfying criterion (1d) are italicised. Since these noun phrases start with a capital letter and end with a period, they are treated as translational units in our analysis.

(16a) "It's not all that far to Avignon."
"*Seven hundred kilometres*. I don't want you to get there after dark." (AB)
(16b) "Det er ikke så langt til Avignon."
"*Syv hundre kilometer*. Jeg vil ikke du skal komme frem i mørke."

Also in accordance with criterion (1d), phrases functioning as headings, normally noun phrases, are treated as translational units in our analysis. This is particularly relevant in the case of the investigated law texts; cf. 4.2.2.1, where this is illustrated.

4.3.3 Embedded string pairs

Examples (8), (10), and (14) have shown that in the extracted data embedded translational units appear as *opaque* items inside the superstrings: the embedded substrings are represented by syntactic category symbols, and the words contained in the substrings are not visible. A string where all words are visible may then be described as *non-opaque*. The property of opacity is relevant to the classification of embedded string pairs, and this topic is discussed further in 4.3.6.5.

There are mainly two reasons why embedded correspondences are treated as opaque units. Firstly, it is an aim to analyse the degree of translational complexity in superordinate correspondences independently of the correspondence type assigned to embedded string pairs.[29] Secondly, we want to avoid duplicating strings within the recorded data, because that could corrupt the calculation of the amount of parallel text covered by each correspondence type.[30] In principle, no word string should, in the quality of non-opaque, belong to more than one string pair.

The empirical data include cases of *multiple embedding*, or *multiple nesting*, i.e. cases where an embedded correspondence itself contains at least one pair of substrings. This is an effect of syntactic recursivity. Since embedded translational units are treated as opaque strings in our analysis, no information about the internal structure of subcorrespondences is displayed in a supercorrespondence. Examples (17)–(20) illustrate a case of multiple nesting:

(17a) Og med uvante hender, som allikevel har lært hva de skal gjøre og som snart skal greie dette helt på egenhånd; med hvite, litt kalde barnehender sikter han mot solen, dreier på skruene og plasserer teleskopet i riktig stilling. (EFH)

(17b) With awkward hands, which have nevertheless learned what they have to do and will soon be able to do this on their own, with chilly white childhands, he focuses on the sun, turning the knobs and adjusting the telescope into the right position.

[29] Cf. the opacity principle, presented in 4.3.6.5.

[30] Section 4.3.4 explains how string length is calculated in the present approach.

The parallel text presented in (17) is a pair of complex matrix sentences. (18)–(20) show the correspondences extracted from (17). The top-level string pair of matrix sentences is given in (18):

(18a) IP: Og med [NP:3] med hvite, litt kalde barnehender sikter han mot solen, dreier på skruene og plasserer teleskopet i riktig stilling. (EFH)
(18b) IP: With [NP:3] with chilly white childhands, he focuses on the sun, turning the knobs and adjusting the telescope into the right position.

In (18) an opaque subcorrespondence is represented by the category symbols *NP*. This subcorrespondence is shown as a pair of complex noun phrases in (19):

(19a) NP: uvante hender, som allikevel har lært [CPwh:3] og som snart skal greie dette helt på egenhånd; (EFH)
(19b) NP: awkward hands, which have nevertheless learned [CPwh:3] and will soon be able to do this on their own,

The pair of NPs in (19) is extracted because they contain finite relative clauses functioning as complements of the head nouns.[31] (19) is here an intermediate-level string pair containing another embedded string pair represented by the category symbols *CPwh*. The bottom-level correspondence is shown in (20), which is a pair of finite interrogative subclauses (CPwh) functioning as direct objects of, respectively, the verbs *lære* and *learn*.

(20a) CPwh: hva de skal gjøre (EFH)
(20b) CPwh: what they have to do

To sum up, (17)–(20) illustrate that in cases of multiple nesting, it is only immediately embedded string pairs, represented by category labels, which are visible in a

[31] (19a) is a special instance of extraction criterion (1c) in 4.3.2, since the complement is a pair of conjoined relative clauses, thus illustrating the necessity of the expression *at least* in (1c). The conjoined relative clauses are, respectively: *som allikevel har lært hva de skal gjøre og som snart skal greie dette helt på egenhånd*, and *which have nevertheless learned what they have to do and will soon be able to do this on their own.*

superordinate string pair, and correspondences embedded at levels deeper than the immediately subordinate one are not displayed in the superordinate string pair.[32]

4.3.4 String length

In order to quantify how much parallel text that is covered by each type of translational correspondences, string length is measured by counting the number of word forms included in individual strings. This may be simply illustrated by repeating examples (10)–(11), where (11) is the pair of finite subclauses embedded in (10):

(10a) IP: I'll tell you [CP:1] (AB)
(10b) IP: Jeg skal si fra [CP:1]

(11a) CP: when I come back. (AB)
(11b) CP: når jeg kommer tilbake.

If a string contains no embedded translational unit, then its length equals the number of word forms it contains. Thus, in the subcorrespondence (11) the non-opaque strings (11a) and (11b) each has a string length of 4. If a string contains one embedded translational unit, then the opaque substring will add only 1, for the category symbol representing it, to the length of the superstring. This means that the string length of (10a) is 4, and that of (10b) is 5. A contracted form like *I'll* in (10a) counts as 1 word form, since no morphological analysis is involved in the automatic routine for string length calculation.

4.3.5 Extraction problems

Although we have tried to establish criteria by which translational units can be easily identified, extraction problems have arisen in given cases. For instance, it has proved very difficult to avoid completely situations where a non-opaque string is included in more than one translational correspondence. Also, there are certain string pairs

[32] The software used for recording string pairs can display information about multiple embedding in a comment field associated with each recorded correspondence; cf. 4.4.4.

among the compiled data where none of the extracted strings is a syntactic unit satisfying any of the extraction criteria.[33] A rather diverse set of extraction problems have been encountered during data compilation. A subset of these problems represent recurring phenomena, and may be grouped into three types of cases. Firstly, in cases where a translational unit is discontinuous we encounter a technical difficulty, and, secondly, the problem of choice occurs in cases where the criteria motivate the extraction of strings between which the translational relation is only partial. Thirdly, there are cases where a unit to be extracted does not have a direct match at all in the corresponding text. These three types of cases are discussed in 4.3.5.1–3.

4.3.5.1 Discontinuous translation units

Example (21) may illustrate discontinuous translation units among our data:

(21a) Just like three years before, when I'd got the news of Ma's death, the name made *something* take shape in my mind *which I hadn't been able to grasp before*. (AB)

(21b) Akkurat som for tre år siden da jeg fikk høre nyheten om Mammas død, hadde navnet fremkalt *noe* i tankene mine, *som jeg tidligere ikke hadde klart å fatte*.
'Just as for three years since when I got hear news.DEF about mummy's death, had name.DEF evoked something in thoughts.DEF mine, which I earlier not had managed to understand.'

(21) is a pair of complex matrix sentences, each containing two syntactically embedded finite clauses; the focus of interest here lies with the discontinuous noun phrases given in italics in (21): *something … which I hadn't been able to grasp before – noe … som jeg tidligere ikke hadde klart å fatte*. There is a translational correspondence between these NPs, and as they both contain finite relative clauses as syntactic complements, they qualify as translational units according to extraction criterion (1c) in 4.3.2.[34] At matrix level in the English source sentence (21a), the causative verb *make*

[33] See for instance examples (23)–(24) in 4.3.5.1, where relative clauses detached from their antecedents are selected as translational units.

[34] Some readers will perhaps object to this analysis and might claim that the sentences in (21) are examples of extraposition. On that reading, the embedded clauses in question are not relative clauses dominated by the

takes as object the infinitival clause *something take shape in my mind which I hadn't been able to grasp before*. The discontinuous NP already referred to is the subject of this infinitival clause, and due to "the principle of end-weight" (cf. Quirk *et al.* 1985: 1397–1398), the subject NP is cleft in order to avoid the stylistically heavy expression shown in (22):

(22) ? ... the name made something which I hadn't been able to grasp before take shape in my mind.

Consequently, the relative clause *which I hadn't been able to grasp before* is placed at the end of the infinitival clause in (21a), dislocated from its antecedent *something*. In the Norwegian translation the corresponding relative clause is also given sentence-final position, due to similar stylistic preferences. Thus, we have a pair of discontinuous phrases which both should be extracted as translational units according to our extraction criteria.

Then, a practical problem arises because the software used for storing and organising string pairs (cf. 4.4.2) does not allow the extraction of strings whose parts are not strictly contiguous. As it is an overall principle in our method that the occurrence of a finite verb should always give rise to a string pair, we have, in this case, chosen to solve the problem by extracting the translationally corresponding relative clauses as a string pair of its own. We do not extract the antecedents, *something* and *noe*, as a separate string pair, since they do not meet any of the extraction criteria, and they are otherwise included in the string pair shown in (24). Accordingly, the subcorrespondence of relative clauses (23) is entered among our data:

(23a) CPrel: which I hadn't been able to grasp before (AB)
(23b) CPrel: som jeg tidligere ikke hadde klart å fatte

nominal expressions *something* and *noe*, but, rather, they are, as finite subclauses, immediate constituents of the matrix sentences. This would yield a subcorrespondence of finite subclauses, but would not in other ways alter the way in which (21) is divided into respectively superordinate and embedded correspondences.

The top-level correspondence extracted from the piece of parallel text shown in (21) is given in (24), where the subcorrespondence (23) is represented by the category labels *CPrel* (relative clause; cf. the lists of syntactic categories in tables 4.3 and 4.4, in section 4.4.).[35]

(24a) IP: Just like [AdvP:4] the name made something take shape in my mind [CPrel:3]. (AB)
(24b) IP: Akkurat som [PP:4] hadde navnet fremkalt noe i tankene mine, [CPrel:3].

Thus, technical limitations on the part of the record-keeping software makes it necessary to treat correspondences involving discontinuous translation units in a somewhat *ad hoc* fashion. While (23)–(24) have illustrated a case of discontinuous phrases, string pair (25) contains discontinuous sentences:

(25a) "Det røde øyet," sier faren rolig, "er kanskje en stor ø som flyter omkring på overflaten." (EFH)
(25b) "That red eye," his father says calmly, "could be a large island floating around on the surface."

In (25) both source and target string contain direct speech. The verbs referring to the act of speaking take independent sentences as direct objects, and these sentences are discontinuous. The two matrix sentences are recorded as a string pair, and due to technical limitations the embedded sentences are extracted in chunks; cf. (26)–(28). The pair of matrix sentences is shown in (26), and the embedded sentences consists of the pair of noun phrases in (27), together with the pair of finite verb phrases (VPfin) in (28).

(26a) IP: [NP:1] sier faren rolig, [VPfin:3] (EFH)
(26b) IP: [NP:1] his father says calmly, [VPfin:3]

[35] In (24) the category symbols *AdvP* (adverb phrase) and *PP* (preposition phrase) indicate another embedded string pair, which is not relevant to the present discussion.

(27a) NP: “Det røde øyet,” (EFH)
(27b) NP: “That red eye,”

(28a) VPfin: “er kanskje en stor ø som flyter omkring på overflaten.” (EFH)
(28b) VPfin: “could be a large island floating around on the surface.”

Thus, both string pairs (27) and (28) consist of units which do not satisfy any of the extraction criteria, but they are nevertheless included among the recorded data. Firstly, the correspondence of verb phrases in (28) should be extracted since it involves finite constructions. Secondly, the extraction of the pair of noun phrases in (27) is a consequence of the extraction of the finite verb phrases, as it is only as part of a unit formed together with the relevant NP that each VPfin is embedded in the matrix sentence, and in order to capture this, the pair of noun phrases is extracted in addition to the pair of finite verb phrases. Thirdly, in order to analyse the translational complexity of the superordinate correspondence (26), it is necessary to represent the syntactic categories of the embedded “chunks”.[36]

4.3.5.2 Partial translational correspondence

Cases where there is only a partial correspondence between strings to be extracted from source and target text constitute a second class of problems arising from the application of the syntactic extraction criteria. As laid down in 4.3.2, an extraction criterion need not be satisfied in more than one of the two parallel texts when a string pair is recorded. Hence, the recorded data contain many correspondences where only one of the strings conforms with an extraction criterion, and the search for translational units in a given text, be it an original or a translation, is in principle independent of the structure of the given parallel text. As a result, there may be cases where the extraction criteria pick out as translational units a string *a* in $text_1$ and a string *b* in $text_2$, where the translational correspondence between *a* and *b* is only partial. If the extraction criteria are to be followed strictly, this situation would give rise to two separate string pairs: (i) string *a* matched by a subpart of *b* plus possibly a contiguous

[36] By the expression *chunk* we understand a word string that may, or may not, constitute a syntactic unit.

part of $text_2$, *x*, not contained in *b*, and (ii) string *b* matched by a subpart of *a* plus possibly a contiguous part of $text_1$, *y*, not contained in *a*. This is illustrated in figure 4.4:[37]

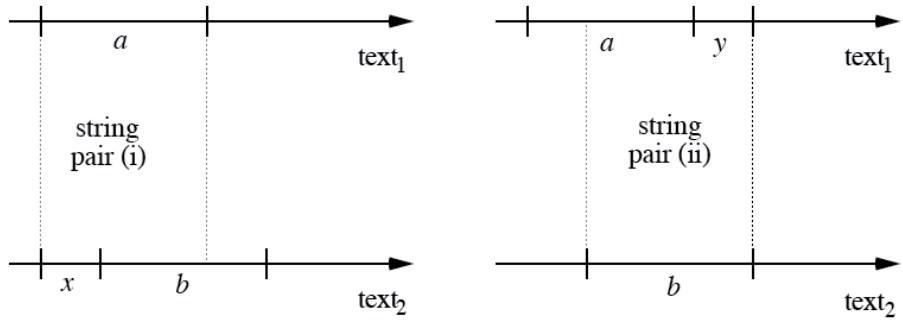

Figure 4.4. Partial translational correspondence between strings *a* and *b*, which are translational units according to the syntactic extraction criteria.

The problematic aspect of situations of this kind is that if both string pairs (i) and (ii) are extracted, then certain pieces of text, i.e. subparts of *a* and *b*, will be included as non-opaque strings in more than one string pair. Thus, these pieces of text would be counted twice in connection with the calculation of string length. Such problem cases are not very frequent in our investigation, but the duplication of non-opaque strings among the data should be avoided as far as possible, since it may disturb the measurement of translational complexity (cf. 4.3.3). In situations of the kind illustrated here, we avoid such duplication by extracting only one string pair, and this is done by choosing translation units which comprise both strings *a* and *b*. To achieve this either one, or both, of the two translational units are expanded in order to extract a string pair where each of the two competing units will be fully included. These expanded units should constitute complete units syntactically as well as semantically. Moreover, the translational correspondence between the two strings should be full,

[37] Figure 4.4 is meant to display the most general picture of partial translational matches. In cases of partial matches, either one or the other of the text sequences *x* and *y* may be the empty string. However, if both *x* and *y* are empty strings, then there is a full translational match between *a* and *b*.

not partial, and the strings will, if possible, be of the same syntactic category, so that the complexity of the supercorrespondence will be minimised.

Relevant here is Toury's (1995: 88–89) discussion of the task of identifying coupled pairs of source and target text segments in order to analyse translational phenomena (cf. 3.3.1.1). He observes that "[t]he pairing is subject to a *heuristic* principle … that beyond the boundaries of a target textual segment no leftovers of the 'solution' to a certain 'problem', posed by a corresponding segment of the source text, will be present" (1995: 89). The strategy we have chosen for handling partial correspondences between translational units is a parallel to the observation made by Toury. Our approach is illustrated by (29)–(32):

(29a) More slowly they went up generously wide stairs, and confronted a stench *which made Jasper briefly retch*. (DL)

(29b) De gikk langsommere nå, opp de gavmildt brede trappene, og møtte en stank så ille *at Jasper brakk seg*.
'The walked more-slowly now, up the generously wide stairs.DEF, and met a stench so bad that Jasper retched himself.'

(29) is a pair of corresponding matrix sentences, each containing a phrase-internal finite clause (given in italics). According to criterion (1c) in 4.3.2, the English noun phrase [$_{NP}$ *a stench* [$_{CPrel}$ *which made Jasper briefly retch*]] is a translational unit to be extracted from the source text, and the Norwegian adjective phrase [$_{AdjP}$ *så ille* [$_{CP}$ *at Jasper brakk seg*]] ('so bad that Jasper retched himself') is likewise a translational unit to be extracted from the target text. However, the Norwegian adjective phrase is only a partial translation of the English noun phrase. In relation to the picture in figure 4.4, the situation in example (29) can be seen as a special case where string *a* is matched by *x* plus *b*, and *b* is matched by a subpart of *a* (*y* is empty). According to the strategy outlined above, we handle this case by expanding the translational unit in the Norwegian text (*b*), so that the following pair of NPs are extracted:

(30a) NP: a stench which made Jasper briefly retch (DL)

(30b) NP: en stank så ille at Jasper brakk seg

String pair (30) includes both of the translational units indicated by italics in (29); (30a) and (30b) are full translations of each other, and, as syntactic units, source and target string belong to the same syntactic category, and carry the same function (object) within the supercorrespondence (31).[38]

(31a) IP: More slowly they went up generously wide stairs, and confronted [NP:4] (DL)
(31b) IP: De gikk langsommere nå, opp de gavmildt brede trappene, og møtte [NP:4]

If the translational unit *så ille at Jasper brakk seg* had not been expanded into the string (30b), the supercorrespondence would have turned out as (32):

(32a) IP: More slowly they went up generously wide stairs, and confronted [NP:4] (DL)
(32b) IP: De gikk langsommere nå, opp de gavmildt brede trappene, og møtte en stank [AdjP:4]

Thus, (31), compared with (32), shows that selecting the pair of NPs given in (30) as the subcorrespondence will reduce translational complexity in the supercorrespondence. There is greater divergence, structurally, functionally, as well as semantically, in the pair of chunks *confronted NP – møtte en stank AdjP* ('met a stench AdjP'), than in *confronted NP – møtte NP* ('met NP').

In conclusion, we may observe that our way of handling the problem of partial correspondences illustrates the point made in 4.3.1 that although syntactic criteria serve to identify translational units in our analysis, the meaning and interpretation of units in the parallel text are more important than their syntactic properties when it comes to identifying the correspondents of given translational units.

[38] In the sense described by Toury (1995: 77), (30a) and (30b) mutually determine each other as units of analysis; cf. 3.3.1.1.

4.3.5.3 Absence of translational correspondent

A third type of problems observed in connection with the application of the syntactic extraction principles, is the situation that arises when a certain translational unit has no match in the corresponding text. A string pair necessarily consists of two strings, so that when a translational unit has no correspondent, not even a partial one, in the parallel text, it is impossible to extract a string *pair* where the unit in question is contained. Among our data there are, broadly, two groups of such cases.

In the first group of cases the unmatched string is syntactically incorporated in a larger linguistic unit, as in example (33):

(33a) og navnet hennes hadde nok stått på lerretet *før filmen begynte*. (BV)
'and name.DEF her had surely stood on screen.DEF before film.DEF started'
(33b) and her name would surely have appeared in the film credits.

String pair (33) is a pair of matrix sentences, where the source text (33a) contains a finite subclause *før filmen begynte* ('before the film started') functioning as an adverbial in the matrix sentence. The subclause satisfies extraction criterion (1b) in 4.3.2, but in the English translation (33b) there is no linguistically expressed correspondent to the Norwegian string. However, the information that may be inferred from general world knowledge through the English preposition phrase *in the film credits* is matched by the information that is linguistically expressed through the Norwegian word sequence *på lerretet før filmen begynte* ('on the screen before the film started').[39] However, no element within the English PP *in the film credits* is a direct match to the Norwegian translational unit *før filmen begynte*. The problem is solved by leaving the Norwegian subclause as it is, embedded in the string pair (33), but when recording the correspondence (33) we supply the information that the source string contains a finite subclause. This information is entered in a comment, stored together with the string pair (cf. 4.4.4). The finite subclause *før filmen begynte*

[39] The latter is a sequence of two adverbials, realised as a preposition phrase and a finite subclause, respectively. The subclause is not embedded in the PP. Had that been the case, this substring would have been extracted as a translational unit, matched by the English PP *in the film credits*.

will thus not contribute to the number of compiled string pairs, but that seems appropriate since it is an unmatched translational unit.

String pairs like (33), where a certain translational unit lacks a correspondent in the parallel text, normally exhibit a difference in the amount of information expressed in the corresponding strings.[40] Such facts are also indicated in the comments that may be attached to the compiled string pairs, and the comments will state whether information has been deleted or added during translation.

In the second group of cases where a translational unit has no correspondent, the unmatched string is syntactically independent, and not part of a larger linguistic constituent. Typically, such cases involve independent matrix sentences, like the italicised sentence in (34):

(34a) Alice began to cry. *It was from pure rage*. "The bastards," she cursed. (DL)
(34b) Alice begynte å gråte."De svina," svor hun.
'Alice started to cry. These bastards.DEF, cursed she.'

In cases like (34) we do not embed the unmatched unit in a supercorrespondence since it is not incorporated in any larger syntactic constituent. On the other hand, the unit cannot be omitted as our measurement of translational complexity is based on running text. We have chosen to match units of this kind with a blank character in the parallel text, so that a string pair may after all be extracted. Such pairs will consist of the unit in question mapped onto the empty string, represented by the symbol NIL, as in (35).[41]

(35a) IP: It was from pure rage. (DL)
(35b) NIL

[40] It is necessary to say "normally" since there are certain cases where a syntactic unit satisfying our extraction principles has no direct match, but the information it contains is nevertheless present in the parallel text, only not in a single syntactic unit which can readily be extracted. These are instances of compositional non-equivalence in translation, presented in 6.2.4.1.

[41] The symbol "NIL" represents the empty list in Lisp, the programming language used for creating the software applied to the recording of string pairs (cf. 4.4.2).

To string pairs like (35), where either the source or the target text is the empty string, we also attach comments stating whether information has been deleted or added by the translator.

The ways in which we have handled unmatched translational units may seem somewhat *ad hoc*, and have been determined by the possibilities offered by the record-keeping software. Nevertheless, we have found it necessary to incorporate such cases in the material. In general, the phenomena of deletion and addition in translation receive considerable attention in various fields. They have been widely studied by translation researchers as types of changes that occur in translation; cf. e.g. Chesterman (1997: 109–110), who describes deletion and addition as types of *information change* in translation.[42] Toury (1995: 78–79) regards deletion and addition as cases where one of the members of a coupled pair is zero. Further, the absence of translational correspondents is investigated in contrastive language studies; see, e.g., Aijmer and Altenberg (2002). Johansson (2007: 23, 26) describes the phenomenon as *zero correspondence*. Within computational linguistics, the problem is sometimes referred to as *null link*, and it is addressed in various approaches to parallel text processing; cf. e.g. Merkel (1999: 184, 185), Merkel et al. (2002: 157), and Macken (2010: 36).

In the texts we have analysed there are not many cases where entire units of extraction have no match in the parallel text. By far the most frequent instances of zero correspondences are found in string pairs where it is only a subpart of an identified translational unit that is unmatched. We will describe such cases as specification and despecification, respectively, and they are discussed in 6.3.1 with subsections.

4.3.6 Assignment of correspondence types

In chapter 3 the presentation of the correspondence type hierarchy specifies the characteristics of each type, as well as criteria for distinguishing between the types,

[42] Changes in translation, so-called *shifts*, are discussed in 6.2.1.

understood as complexity classes.[43] The characteristics of each correspondence type is the basis for measuring the degree of translational complexity in specific string pairs, seen as translation tasks, and provide the criteria for deciding what type each extracted string pair belongs to. As explained in 3.3.1.1, the purpose of the empirical analysis is to measure the degree of complexity in the task of generating automatically the specific translations that have been produced by human translators for selected texts. This measure is implemented as the assignment of one of the four correspondence types to each extracted string pair. Sections 4.3.6.1–6 describe the principles governing the classification of individual string pairs.

4.3.6.1 An elimination procedure

After a translational unit has been identified and its correspondent found in the parallel text, the string pair is classified by the human annotator. The classification, or type identification, is based on a linguistic analysis of the translational correspondence between the two strings. The translational complexity of an extracted string pair is identified by analysing the degree to which linguistic matching relations (as explained in 3.3.1.2) hold between source and target string, this is the task of deciding whether the translational correspondence satisfies the requirements of either type 1, 2, 3, or 4.

The first step in this task is to compare the syntactic properties of the two strings. If there is full agreement between source and target string with respect to the sequence of word forms, and there is also syntactic functional equivalence, semantic compositional equivalence, as well as pragmatic equivalence, between the two strings, then the correspondence is of type 1 (cf. 3.3.2.1). If these requirements hold, except for at least one deviation between the two strings with respect to word order and/or the use of grammatical form words, then type 2 is assigned (cf. 3.3.3.1). However, if for at least one lexical word in one of the strings there is no translational correspondent in the other string of the same category, and/or with the same syntactic

[43] The correspondence types resemble complexity classes insofar as they can be distinguished in terms of a lower bound on how easy it can be to solve a translation task and an upper bound on how hard it can be; cf. 3.2.1 and 3.2.4.

function as of that lexical word, then the correspondence is of type 3, provided that source and target string are still equivalent with respect to the sets of expressed predicates and arguments, and with respect to the relations between the predicates and their arguments. Also, pragmatic equivalence must hold (cf. 3.3.4.1). Finally, if there is any deviation between source and target string with respect to the requirements of semantic equivalence, then type 4 is assigned, and pragmatic equivalence may, or may not, hold (cf. 3.3.5.1).

Thus, type assignment works like an elimination procedure where we start by testing for the lowest correspondence type and then move upwards in the hierarchy if the test fails. This may seem a fairly straightforward task, but not in any case. In particular, it can be difficult to distinguish between instances of types 3 and 4, since that may involve fine-grained semantic analyses.

A string pair is assigned the correspondence type of its most complex subpart.[44] If there is only one violation, in the correspondence between two translational units, of the restrictions defined for a lower type, then a higher type is assigned to the entire string pair. This point has already been observed in connection with linguistic examples illustrating types 2 and 3.[45]

The following four principles, to be presented in 4.3.6.2–5, are involved when correspondence type is assigned to a given string pair:

1. After extraction, the strings are treated as items on the level of *langue*, as units belonging to the language systems.
2. Type assignment is done solely on the basis of the information expressed by the linguistic material present in the extracted strings.
3. Assignment of type to an embedded correspondence is done with no regard to information contained in the supercorrespondence.
4. Likewise, an embedded correspondence is an opaque unit inside the supercorrespondence, identified only by its top node categories.

[44] Embedded correspondences represent an exception to this general principle: a matrix string pair is classified independently of the degree of complexity in embedded correspondences (cf. 4.3.6.5–6).
[45] Cf. 3.3.3.2 and 3.3.4.1.

4.3.6.2 System-level units

The first principle for type assignment states that the classification applies to items on the level of *langue*, i.e. to system-level units, or linguistic types, and not to tokens of language use, or items on the level of *parole*. The parallel texts that the string pairs are taken from are, on the other hand, situated texts, made up by sets of utterances. Thus, every translational unit, when placed in its textual context, can be seen as a token, but after string extraction is done, the pairs of corresponding strings are classified as linguistic types.

The type-token distinction can be illustrated by the phenomenon of ambiguity. A linguistic expression, a *langue* item, may be syntactically or semantically ambiguous in the sense that it is possible to derive more than one interpretation of it. However, when the same expression is placed in a linguistic context, it becomes a situated utterance, and the ambiguity may be resolved if the context makes it clear that only one of the interpretations is intended by the speaker, or sender.

As stated in 3.3.1.1, the problem of source text disambiguation is kept apart from the measurement of translational complexity. Thus, when analysing the complexity of extracted string pairs (i.e. correspondences between strings a_{L1} and b_{L2}), we assume that the task of translating a_{L1} into b_{L2} is solved on the basis of one relevant interpretation of a_{L1}, and that is the interpretation which lies behind the chosen translation b_{L2}.

That the classification of translational correspondences applies to *langue*-items means that a_{L1} and b_{L2} are regarded as system units, and, in the case of a_{L1}, as a disambiguated system unit. From this it follows that when we analyse the correspondence, we do not regard the interpretations of a_{L1} and b_{L2} in relation to the contexts from which they are extracted. In order to classify the correspondence between a_{L1} and b_{L2}, we consider whether it is possible to derive, on the basis of the pre-structured domain of linguistic information (cf. 2.3.2), at least one interpretation of b_{L2} that matches the given, relevant interpretation of a_{L1}. If there is such a match, then b_{L2} is a linguistically predictable translation of a_{L1}. Given the definition of LPT in 2.3.2, this means that b_{L2} shares a maximum of the meaning properties of a_{L1}, taking into account differences between the two language systems. According to the

present analytical framework, the relation between a_{L1} and b_{L2} will then satisfy the requirements of semantic equivalence defined for the correspondence types 1, 2, and 3. On the other hand, if the derivation of an interpretation of b_{L2} that matches the given interpretation of a_{L1} requires access to information sources other than the source string and the pre-structured domain of linguistic information, then b_{L2} is not included in the LPT set of a_{L1}, and the string pair is a type 4 correspondence.

The type-token distinction is relevant to the issue of computability in the translational relation, as presented in 3.2.5. The language descriptions incorporated in a system for machine translation will be descriptions that represent source and target languages on the level of *langue*, the language systems. Thus, the output computed by the MT system for a specific translation task is determined by what may be predicted from the representations of the language systems, and aspects of the level of *parole* cannot be drawn on, since we assume that those are not encoded in the language descriptions. Consequently, the translational relation computed by the system is a relation between linguistic types, not between tokens of language use (cf. 2.3.1–2). This provides some motivation for the principle of classifying translational correspondences as system-level units, since that conforms with the aim of making our investigation of translational complexity relevant to the field of machine translation.

4.3.6.3 Available information

The second principle for type assignment follows from the first one: if string pairs are classified as expressions on type level, i.e. analysed as items out of context, then the classification must be done solely on the basis of the information available in the extracted strings. The evaluation of correspondence type is for instance not influenced by any kind of inference made possible by information contained in the linguistic contexts preceding the two strings. Such inference might serve to delimit the set of possible interpretations of a given string. However, any inference that is

predictable from the linguistic material present *within* the strings is naturally relevant to the analysis.[46]

The principle may be illustrated by a case where a referential relation holds between an anaphor inside an extracted string and its antecedent in the preceding context. Consider the piece of parallel text given in (36), where relevant expressions are given in italics:

(36a) Det fantes *en bok* oppe i stuen også, som Jason så meget i — men *den* var annerledes. (EFH)
'It existed a book up in sitting-room.DEF also, which Jason looked much into — but it/that was different.'
(36b) There was *another book* up in their living room which Jason often looked at — but *that book* was different.

In (36) both source and target text are sequences of two independent sentences, connected by the coordinating conjunctions *men* and *but*, respectively. The pair of final sentences in these sequences is shown in (37):

(37a) IP: men den var annerledes. (EFH)
(37b) IP: but that book was different.

Apart from the correspondence between the anaphor *den* and the referential NP *that book*, (37b) is a word-by-word translation of (37a). By accessing the information provided by the preceding linguistic contexts (cf. (36a) and (36b)), it is possible to see that *den* and *that book* are coreferential, but the information that these two expressions refer to the same extra-linguistic entity lies outside the domain of the string pair (37). On the basis of the information contained in the extracted string, (37) must be classified as a type 4 correspondence, since the expressions *den* ('it/den') and *that book* are not denotationally equivalent. In this case denotational non-equivalence between *den* and *that book* means that (37b) cannot be a linguistically predictable translation of (37a). Although it does not influence type assignment here, we may

[46] This point is illustrated in 6.3.1.2 by the discussion of non-prototypical cases of what we have described as the inalienability pattern in English-Norwegian translation.

observe that information about the relation of coreference makes it possible to derive an interpretation of (37b) matching the given interpretation of (37a).

4.3.6.4 Self-contained embedded correspondences

The third principle for type assignment follows from the second one: if a string pair is classified solely on the basis of the linguistic material contained in the extracted strings, then correspondence type is assigned to an embedded string pair with no regard to information that may be derived either from the supercorrespondence or from other, possibly adjacent, embedded correspondences. Thus, embedded string pairs are analysed as self-contained units.

The principle may be illustrated by a case where an anaphoric relation holds between a personal pronoun in an embedded string and a referring expression in the superstring. Consider the parallel text given in (38), which is a pair of complex matrix sentences:

(38a) Brita løftet det rødmende ansiktet og så inn i de undersøkende øynene og så at fru Bendixen ikke smilte. (BV)
'Brita raised the blushing face.DEF and looked in into the searching eyes.DEF and saw that Mrs Bendixen not smiled.'

(38b) Brita raised her blushing face and looked into Mrs Bendixen's searching eyes and saw that she was not smiling.

Both (38a) and (38b) contain a finite subclause functioning as a syntactic object. Thus, according to the extraction criteria in 4.3.2, we identify as string pairs both the superordinate correspondence of matrix sentences shown in (39), and the embedded correspondence of subclauses given in (40):

(39a) IP: Brita løftet det rødmende ansiktet og så inn i de undersøkende øynene og så [CP:4]. (BV)

(39b) IP: Brita raised her blushing face and looked into Mrs Bendixen's searching eyes and saw [CP:4].

(40a) CP: at fru Bendixen ikke smilte (BV)
(40b) CP: that she was not smiling

In the embedded string pair (40) the correspondence between source and target string does not fulfil the requirements on syntactic equivalence as specified for types 1 and 2 in chapter 3, and whether (40) is of type 3 or 4, depends on the translational relation between the referential noun phrase *fru Bendixen* in (40a) and the anaphor *she* in (40b).[47] As in the case of (37) in 4.3.6.3, the conclusion is that (40) is of type 4 since the two expressions are not denotationally equivalent, and thus the personal pronoun *she* cannot be a linguistically predictable translation of *fru Bendixen*. Although it does not alter the type assignment, access to the information contained in the supercorrespondence (39) makes it possible to see that *fru Bendixen* and *she* are coreferential.

The principle of analysing embedded correspondences as self-contained units strengthens the motivation behind extraction criterion (1c) in 4.3.2, which defines lexical phrases with finite clauses as units of analysis.[48] A good illustration is provided by the case of noun phrases with embedded relative clauses, such as the expression *a stench which made Jasper briefly retch*, given in (30a) in 4.3.5.2. In such phrases the antecedent of the relative clause is the lexical head of the phrase (i.e. *stench* in (30a)). If we had chosen to identify phrase-embedded subclauses as separate translational units, then relative clauses would be extracted in isolation from their antecedents (i.e. *which made Jasper briefly retch* would be a unit of analysis). Given the third principle for type assignment, information about the antecedent would then not be accessible from inside the extracted relative clause. The principle of analysing embedded correspondences as self-contained units thus makes it rational to extract phrase-internal finite clauses together with the rest of the phrases in which they occur.

[47] Regarding the translational relation between the Norwegian simple verb form *smilte* and the English progressive construction *was ... smiling*, it is argued in 6.3.1.1 that such correspondences are linguistically predictable.
[48] Cf. also the discussion of that criterion in 4.3.2.3.

4.3.6.5 The opacity principle

The fourth principle for type assignment is a direct reflection of the third one: seen from the inside, an embedded string pair is analysed independently of any super-correspondences, and seen from the outside it appears as an opaque unit (cf. 4.3.3). Previously given examples of nested string pairs have shown that embedded units are represented by their top node categories inside a supercorrespondence, and that the words contained in an embedded, opaque substring are not visible at the level of the superordinate, or matrix, correspondence. That embedded units are opaque means that nothing of the information expressed "inside" the embedded strings is available when correspondence type is assigned to a superordinate string pair. We will refer to this as *the opacity principle*.

However, two elements of information *about* embedded translational units are available at the matrix level when a supercorrespondence is classified. These elements are, respectively, the syntactic category and the syntactic function of the embedded unit as a constituent within the superordinate unit.

Information about the category and function of embedded translational units follows from the linguistic analysis performed in order to identify translational units according to the extraction criteria of 4.3.2. This refers to the linguistic analysis carried out by the human annotator for the purpose of string pair extraction, to be described in 4.4.1. In chapter 3 we have discussed linguistic analysis as a subtask of translation, but the linguistic analysis involved in solving a translation task is, at least on a conceptual level, distinct from the analysis involved in string pair extraction.[49] The former kind applies to translational units of the source text only, whereas the linguistic analysis involved in data compilation is performed, in parallel, on strings extracted from *both* source and target text.

Thus, information about the syntactic category of an embedded unit is a result of the syntactic analysis by which the unit is identified as a constituent of the super-

[49] Cf. the discussions of the subtask of analysis in 3.3.1.4, as well as in later sections presenting each of the four correspondence types.

ordinate string.[50] The syntactic function of a constituent does not exist inside it, but belongs to the matrix level, and once the category of an embedded, opaque string is identified, information about its syntactic function is derivable from the matrix sentence by combining information about its constituent structure, its predicate-argument structure, and about the linking between arguments and syntactic constituents. The two latter pieces of information are given through the lexical specification of the main verb.

Information about the syntactic category and function of a substring does not pertain to the internal structure of the string, but is part of the syntactic and grammatical structure of the superordinate string. Hence, these are pieces of information needed for the classification of the matrix correspondence.

In accordance with the opacity principle, a superordinate string pair is classified independently of the correspondence types that are assigned to embedded string pairs.[51] In particular, if translational complexity is higher in a subcorrespondence than in the supercorrespondence, then the degree of complexity in the matrix string pair is not influenced by that of the subcorrespondence.[52] This can be illustrated by example (41):

(41a) but I could see *he was finding it difficult* (AB)
(41b) men jeg kunne se *det var vanskelig*,
'but I could see it was difficult,'

(41) is a pair of complex sentences, each containing a finite subclause functioning as direct object to the main verb. The direct objects are given in italics, and the pair of subclauses is extracted separately as the correspondence shown in (42):

[50] Naturally, to recognise the category of an embedded string requires that the internal structure of the opaque unit has been analysed. As will be explained in 4.4.1, embedded correspondences are analysed before superordinate string pairs are processed.
[51] This point has previously been made in 4.3.3.
[52] This is an amendment to the analysis method described in Thunes (1998), where a supercorrespondence cannot be assigned a type lower than the highest type found in any embedded string pair; cf. Thunes (1998: 33).

(42a) CP: he was finding it difficult (AB)
(42b) CP: det var vanskelig

The embedded string pair (42) is classified as a type 4 correspondence since there are denotational differences between source and target expression, and because the amount of linguistically expressed information is smaller in the translation (42b) than the original (42a).[53] On the other hand, type 1 is assigned to the supercorrespondence (43):

(43a) IP: but I could see [CP:4] (AB)
(43b) IP: men jeg kunne se [CP:4]

In (43) the relation between source and target string conforms with the requirements specified for type 1 correspondences in 3.3.2.1. Thus, the low degree of complexity in the supercorrespondence is not influenced by the high degree of complexity in the subcorrespondence.

4.3.6.6 Classification of nested correspondences

According to the opacity principle, the classification of a supercorrespondence is independent of the translational complexity of embedded string pairs, and at the matrix level the only information available about embedded correspondences is the specification of their syntactic category and function as constituents of the supercorrespondence. Thus, the classification of matrix string pairs involves evaluating to what extent translational links between embedded constituents fulfils the various linguistic requirements defining the four correspondence types, and in this regard certain principles are followed. In order to satisfy the requirements of types 1 and 2, translationally corresponding substrings, as constituents of the superstrings, must be identical with respect to top node categories, as well as syntactic functions. If the substrings have different categories, but still identical functions, then the require-

[53] Denotational differences, and differences in the amount linguistically expressed information between translationally corresponding expressions are discussed in chapter 6.

ments of type 3 are met. Finally, if the substrings are associated with different syntactic functions, the correspondence between them is of type 4. This can be illustrated by example (44):

(44a) Der han stod kunne han se kuppelen på St. Paul's; (EFH)
'There he stood could he see cupola.DEF on St. Paul's;'
(44b) From where he was standing, he could see the cupola on St. Paul's,

(44) is a pair of complex matrix sentences. The data extracted from (44) are, respectively, the matrix string pair shown in (45), and the subordinate correspondence shown in (46). Both (46a) and (46b) are lexical phrases with finite subclauses as complement, thus instantiating extraction criterion (1c) in 4.3.2. The top node categories of the embedded strings are entered in parentheses in (45):

(45a) IP: [AdvP:4] kunne han se kuppelen på St. Paul's;
(45b) IP: [PP:4] he could see the cupola on St. Paul's,

(46a) AdvP: Der han stod
(46b) PP: From where he was standing,

In order to classify the matrix string pair (45), we may first consider the non-opaque substrings: there is a type 2 correspondence between the expression *kunne han se kuppelen på St. Paul's* in (45a), and the expression *he could see the cupola on St. Paul's* in (45b).[54] But since the matrix string pair also involves a correspondence between different, sense-carrying categories (i.e. an adverb phrase and a preposition phrase), it cannot be assigned type 2 according to the classification criteria. Then, since the adverb phrase in (45a) and the preposition phrase in (45b) both fill the function of locative adverbial, the conclusion is that (45) is a type 3 correspondence because the embedded, opaque units provide the same kind of underspecified semantic information in the semantic structures of the sentences in which they are em-

[54] Between the two word sequences there are linguistically predictable differences with respect to word order and the use of the grammatical form word *the*; cf. 3.3.3.1.

bedded. Thus, the translational link between the two adverbials conforms with the requirement that in type 3 correspondences source and target expression are equivalent with respect to the sets of expressed predicates and arguments, and with respect to the relations between the predicates and their arguments (cf. 3.3.4.1).

In (45) the correspondence between an adverb phrase and a preposition phrase is an example of category crossing (cf. 4.3.2). In relation to the classification of matrix correspondences, it could be questioned whether category crossing between embedded units necessarily means that the supercorrespondence cannot fulfil the requirements of syntactic functional equivalence which apply to correspondences of types 1 and 2, because there are certain syntactic types that typically realise the same kinds of syntactic functions. E.g., noun phrases and nominal subclauses both function as nominals; adverb phrases and preposition phrases can realise various kinds of adverbial functions, and adjective phrases and relative clauses can both function as modifiers of nominals. These observations are true for English, as well as for Norwegian. On this background it could be argued that the correspondence between the adverb phrase in (45a) and the preposition phrase in (45b) is a predictable correspondence, derivable from information about the syntactic interrelations between the two language systems. In line with this, we could modify the requirement that in correspondences of types 1 and 2 embedded strings must agree with respect to both category and function, so that category crossing would be allowed in cases where the substrings have the same syntactic function within the respective superstrings. In such cases translational links between non-identical, but functionally equivalent, categories could be compared to correspondences between translationally related lexical items. Such instances of category crossing would hence be regarded as formal deviations not affecting syntactic functional equivalence between the strings, i.e. on a par with differences in constituent sequence, and in the use of grammatical form words. Still, we have chosen to keep the requirement of category match in types 1 and 2, although it is rather strict. Adopting the alternative approach would in our opinion demand a prior study of systematic English-Norwegian correspondences between various syntactic types, which has not been part of the present investigation.

4.4 Implementation of method

This part of the present chapter presents the practical implementation of the methodology described in 4.3 with subsections. As previously pointed out, the linguistic analysis needed in order to identify and classify string pairs is done "manually" by a bilingually competent human annotator. The parallel texts have been processed by the computer program Text Pair Mapper (cf. 4.4.2), which is specially designed for annotating, storing and organising translational correspondences.

4.4.1 Parsing "by brain"

In the compilation of data each pair of translationally parallel texts has been processed from the beginning to the end of the text extracts. The work follows a certain procedure where the annotator starts at the top, analysing source and target text in parallel, in order to find the first translational unit satisfying the search criteria discussed in 4.3.2. Once the unit, and its translational correspondent, are identified, the string pair is recorded and a correspondence type assigned to it. The annotator then identifies the next translational unit, and proceeds through source and target text until the end of the extracts, parsing the texts in parallel, identifying corresponding text strings, and classifying the translational correspondence in each string pair.

With respect to simple string pairs with no embedded correspondences, the analysis and classification of a string pair is completed before the immediately succeeding correspondence is identified. String pairs are thus processed sequentially, one after the other.

In the case of nested string pairs, we have chosen a bottom-up strategy: all subcorrespondences within a supercorrespondence are fully analysed before the matrix string pair is processed. As discussed in 4.3.6.5–6, the syntactic categories and functions of embedded correspondences may influence the degree of complexity in a superordinate string pair, and hence it is practicable to finish the processing of embedded correspondences before the matrix string pair is analysed. However, the first thing to do with a superordinate correspondence is to identify the beginning and end of each matrix string. The following tasks are to find all finite verbs occurring within

the two parallel strings, and, applying the syntactic extraction criteria, to identify the unit of translation that each finite verb belongs to. Thus, all embedded correspondences are recorded and classified, one by one. In the case of multiple embedding, where a subcorrespondence itself contains one or more embedded string pairs (cf. 4.3.3), we start by analysing the most deeply embedded correspondences, and then move upwards in the syntactic hierarchy. Finally, the supercorrespondence is evaluated and assigned a correspondence type, as described in 4.3.6.6.

With respect to data compilation, the most important search criterion is the occurrence of finite verbs. The question could be raised whether it would be helpful to apply automatic syntactic tagging to the text material before string pair extraction, since that would ease the identification of finite verbs. This issue has been considered, but disregarded. Automatic taggers, as well as the human parser, are not infallible, so that in either way error checking would be required after extraction. Since the error rate of a competent human annotator is in general lower than that of an automatic tagger, we expected that the use of automatic tagging could increase the amount of correction to be done. Moreover, it is fairly straightforward to identify finite verbs in English and Norwegian, and for these reasons we have chosen to rely on the human parser, also because we then avoided the insertion of syntactic tags in the texts.

4.4.2 The software: Text Pair Mapper

The compiled data have been recorded by means of the computer program Text Pair Mapper, created by Helge Dyvik.[55] This is software specially designed for compiling string pairs from parallel texts, as it facilitates the storing and organisation of string pairs, and it offers efficient tools for performing search and sort operations on the compiled data. The program is written in Lisp, and runs in the Medley Lisp environment.

Text Pair Mapper is designed for processing one text pair at a time. Figure 4.5 is a snapshot of its user interface. The upper left window contains the source text, and the

[55] Cf. Dyvik (1993).

target text is displayed in the upper right window. A recorded string pair (cf. example (18) in 4.3.3) is marked by inversion in the source and target text windows, and it is shown in the “String Correspondence Display Window” in the lower right part of the picture.

Firstly, the program offers facilities for data compilation. The user, i.e. the annotator, selects a translational unit in the source text window, and the translational correspondent in the target text window, and subsequently picks, from the “Add pointers” menu, the appropriate option for the type of correspondence holding between the two strings. Next, the program prompts the user to select a syntactic category for each string. After that is done, the program prints the selected string pair in the “String Correspondence Display Window”, displaying the two strings together with their syntactic categories and the correspondence type of the string pair. Also, the user may add a comment to the string pair by typing it into the “Comment Edit Window”.

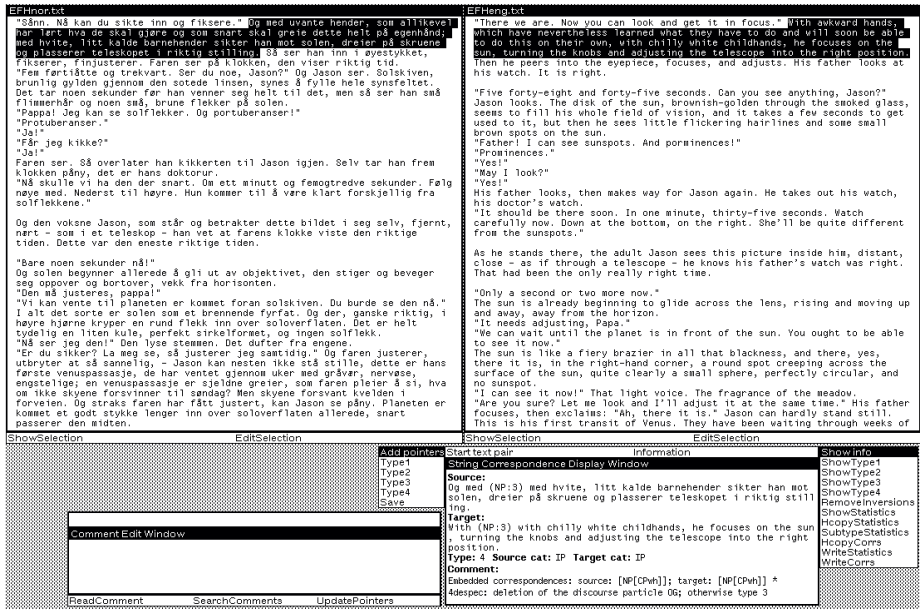

Figure 4.5. A picture of the user interface in Text Pair Mapper.

Secondly, Text Pair Mapper provides tools for editing, inspecting and exporting the recorded string pairs. If some text is selected with the mouse either in the source or target text window, both menu options "Show Correspondence" and "Edit Correspondence" will display the string pair in which the selected text is included, together with the information entered about that correspondence. If the selected text is included in more than one string pair, which is the case in nested correspondences, the user may choose among the relevant string pairs via a pop-up menu. When nested correspondences are recorded, the software automatically keeps track of the string pair(s) embedded in any superordinate correspondence. If a superordinate string pair is selected, either for inspection or editing, information about the embedded units is given as part of the comment accompanying the string pair. The embedded units are represented by their syntactic categories, as shown in the "String Correspondence Display Window" in figure 4.5.[56] If no correspondence includes the selected text, the program responds by stating this in the prompt window.

The editing options give access to the correspondence type, syntactic categories, and comment field of individual string pairs, so that these may be changed according to the wish of the user. There is also an editing option for the removal of entire string pairs. The recorded correspondences can be output to a printer, and they may be exported from the program since the string pairs extracted from a text pair may be written to a text file. The correspondences of a given text pair may be output in the sequence in which they occur in the parallel texts. They may also be output according to options for sorting and searching as the program offers facilities for performing sort and search operations on compiled string pairs. These are particularly useful for the inspection of recorded data. The tools for sorting and searching are applied to correspondences of one text pair at a time.

With respect to *sorting*, string pairs may be sorted by correspondence type, i.e. either by main type (1, 2, 3, 4), or by subtype, and the result of a sort operation can be output to a printer or written to a file. Subcategories within the main correspondence types 3 and 4 are presented in chapter 6. During the compilation of string pairs,

[56] Figure 4.5 illustrates a case of multiple embedding, previously discussed in 4.3.3.

occurrences of subtypes have been recorded by manual insertion of tags in the comment fields of individual string pairs, and each subtype has been associated with a unique tag, or label (cf. 4.4.4).

The *search* operations that may be done in Text Pair Mapper yield sets of string pairs satisfying a search parameter as specified by the user, and this kind of output, too, can be printed or written to a file. The program allows two kinds of search parameters: syntactic categories and strings entered in comment fields. The search operations are, however, not designed for combinations of more than one parameter at a time. In a search according to syntactic categories, the user is prompted to specify a source string category and a target string category, of which one may be a wild card, and the program then compiles all correspondences between strings of the specified categories. The other kind of search operation pertains to the comments that may be entered together with the correspondences: the user enters a string of characters, and the program collects all correspondences whose comment field contains the search string.

Another important function in Text Pair Mapper is the calculation of the quantitative distribution of the four main correspondence types within a given text pair. For each type the program calculates both the number of string pairs and the lengths covered in respectively source and target text. Text length is here measured in the same manner as described for the calculation of string length in 4.3.4. Also, for each correspondence type the program calculates its percentage of the total number of string pairs and its percentages of the total lengths of the strings extracted from respectively source and target text. As regards the subtypes of main types 3 and 4, the program can, for a given text pair, provide a count of the number of string pairs instantiating each subtype.

4.4.3 Syntactic labels for empirical data

We have used a certain set of syntactic category labels for the strings extracted from the analysed texts. It is has previously been mentioned in 4.3.2 that the syntactic analysis is based on a rudimentary X'-framework, and, by and large, in line with the LFG framework, although it has not been an aim in this project to perform syntactic

analysis in strict accordance with any specific theory of syntax, and the set of syntactic labels applied does not conform in detail to any particular variety of X'-syntax. Moreover, we have not tried to identify exhaustive sets of categories for English and Norwegian. The set of categories that has been used reflects the application of the syntactic extraction criteria. It also reflects certain challenges involved in the task of analysing running text: when we need to find a category for every extracted unit, there are cases where the classification is difficult because running text may include string fragments that are not complete syntactic units. As pointed out in 4.3.2, the goal of the analysis is to identify the syntactic type of each translational unit, and to do so without applying very detailed theoretical assumptions.

The analysis has required a varied set of syntactic categories, and we have tried to be consistent in the use of category labels. The overall majority of the category labels are neutral with respect to the two languages involved, i.e. neutral in the sense that the categories they represent are found in both languages. The language-neutral syntactic labels are listed in table 4.3. Then there are a few category labels which are language-specific, shown in table 4.4.

Some of the categories listed are extracted obligatorily (e.g. IP and CP) because they always conform with an extraction criterion. Other categories, like NP and AdjP, qualify as translational units only if they carry certain syntactic properties, and are hence not extracted unless such properties are present. Further, we have identified certain categories that do not conform with any of the extraction criteria, e.g. the relative clause and various nonfinite constructions. Such categories are extracted only in cases where they correspond with a translational unit in the parallel text.

Several of the syntactic labels listed in tables 4.3–4 do not require further comments as they refer to types of syntactic categories which are generally known and accepted, such as PP (preposition phrase). Other labels may appear less transparent, but will still not be commented on as the examples make it reasonably clear what classes of constructions are covered by the given categories, e.g. VPinf (infinitival verb phrase). Finally, certain categories referred to in tables 4.3–4 are entities that are not normally identified as types of syntactic constituents, such as IPinc (incomplete matrix sentence). Since the criteria governing their application may not be self-

evident, the motivation behind categories of the latter kind is commented on in 4.4.3.1–4. These may appear as rather *ad hoc* categories, but they have arisen from the need to categorise strings which are not units of translation according to our extraction criteria, but, as pointed out above, must be identified because they correspond translationally with a string matching one of the extraction criteria.

In tables 4.3–4, tokens of each category are given in italics. For the purpose of illustration, a few of the tokens are provided with some context, and in those cases the context surrounding the relevant token is *not* italicised. The examples are taken from the recorded string pairs, and the italicised expressions all appear as units of translation within the data. With respect to the language-neutral syntactic categories, only English examples are given, to avoid the need for glossing.

Table 4.3 (continues overleaf). Syntactic categories found in texts of both languages.

Label	**Category**	**Token**
IP	matrix sentence	*This could be seen through the broken window just above them on the first floor.* (DL)
IP-seq	sequence of matrix sentences	*It's all right, it's O.K., don't worry!"* (DL)
IPwh	independent *wh*-interrogative	*Who can they be?* (DL)
IPpot	a potential IP, i.e. a substring of an IP, which on type level can be parsed as an IP	*She faced him, undefiant but confident*, and said, "I wonder if they will accept us?" (DL)
IPinc	incomplete matrix sentence	*They didn't even...* (DL)
CP	finite subclause	*as she could see* (DL)
CP-seq	sequence of finite clauses	*if it meets the needs of coordination of transport or if it represents reimbursement for the discharge of certain obligations inherent in the concept of a public service* (*AEEA*)
CPrel	relative clause	a living rug of young nettles *that was trying to digest rusting tins and plastic cups* (DL)

Table 4.3 (continued). Syntactic categories found in texts of both languages.

CPinf	nonfinite clause with infinitival verb phrase[57]	Alice saw *Bert's body stiffen*, ... (DL)
CPwh	dependent *wh*-interrogative	"You know *what the question was*," ... (DL)
CPpot	a potential CP, i.e. a substring of a CP, which on type level can be parsed as a CP	Bert said, after a pause, "*That this group should make approaches to the I.R.A. leadership, offering our services* as an England-based entity." (DL)
CPinc	incomplete finite subclause	*Where compliance with the provisions of Articles 10 and 12 leads to:* (*AEEA*)
CPnovrb	verbless clause	*her heart full of pain because of the capacious, beautiful and unloved house* (DL)
NP	noun phrase	*the bell, which did not ring* (DL)
NP-seq	sequence of noun phrases	*A brochure on the cathedral, photographs of its windows.* (AB)
NPpot	a potential NP, i.e. a substring of an NP, which on type level can be parsed as an NP	*a living rug of young nettles* that was trying to digest rusting tins and plastic cups (DL)
WhP	headless relative clause	That's *what the police said.* (DL)
QP	quantifier phrase	Ahead of you lies *everything that you do not know*. (EFH, in translation)
PP	preposition phrase	*by the time I saw him again, two or three weeks from now* (AB)
AdjP	adjective phrase	*as near to normal as any they had seen* (DL)
AdvP	adverb phrase	*just about the farthest I could hope to get away from District Six* (AB)
VPfin	finite verb phrase	Only the burning colours of those tall windows, I knew in my guts, *would lessen it* (AB)
VPinf	infinitival verb phrase	*Marry Paul,* and I could no longer choose. (AB)
VP'	infinitival verb phrase with infinitive marker	*to accept offers of employment actually made* (*AEEA*)
seq	sequence of constituents that cannot be parsed as one constituent	She offered Alice a cigarette, *which was refused, and smoked hers needfully, greedily*. (DL)

[57] Traditionally called *accusative with infinitive*. Among the strings extracted from the Norwegian texts there is only one occurrence of this category, here given in italics: "Han hørte *henne dundre på dør etter dør, og rive dem opp ettersom ingen svarte*" (DL, in translation). Gloss: 'He heard her pound on door after door, and tear them up as nobody answered.'

Table 4.4. Language-specific syntactic categories.

English:		
Label	**Category**	**Token**
VPed	verb phrase with *-ed* participle	*described by him as silly* (DL)
VPing	verb phrase with *-ing* participle	*stepping from a bath* (AB)
Sing[58]	nonfinite clause with *-ing* participle	*his cheeks and teeth shining in candlelight* (DL)
Norwegian:		
Label	**Category**	**Token**
VPptptc	verb phrase with past participle	*Sist endret ved lov av 27. november 1992 nr 119.* (*Petro*) Gloss: 'Last changed by act of 27th November 1992 number 119'
VPprptc	verb phrase with present participle	*eller noen ganger pilende gjennom gatene midt på lyse dagen* (EFH) Gloss: 'or some times running through streets.DEF middle on bright day.DEF'

4.4.3.1 Sequences of the same category

In the labels *IP-seq*, *CP-seq*, and *NP-seq* the segment *seq* is short for *sequence*, and all these labels refer to conjoined constituents, respectively conjoined matrix sentences, conjoined finite subclauses, and conjoined noun phrases. The conjuncts may be joined together either by means of coordinating conjunctions, or by means of punctuation only. That constructions of the category IP-seq have been extracted as translational units will seem contradictory to what has previously been stated with respect to conjoined matrix sentences in 4.3.2.1: when (at least) two matrix sentences are conjoined into a compound matrix sentence, then each conjunct is identified as a translational unit, but the entire compound sentence is not extracted as a unit in addition to its subparts. This is the main rule, but there are exceptional cases where a single matrix sentence in one text is matched in the parallel text by a string consisting of more than one matrix sentence; cf. example (47):

[58] This is a type of nonfinite clausal construction that cannot have a complementiser; hence, the segment *CP* is not used in its label.

(47a) Otherwise, just as consciously, I would have to resign myself to the prospect of a lasting emptiness, the very idea of which threatened and offended me. (AB)

(47b) Eller jeg måtte like bevisst resignere for utsikten til en varig tomhet, bare tanken på det truet og krenket meg.
'Or I had-to just consciously resign for prospect.DEF to a lasting emptiness, only thought.DEF on that threatened and offended me.'

(47) can be seen as an example of sentence splitting (cf. 4.2.1.1): whereas the translation (47b) consists of two matrix sentences, conjoined by a comma, the original (47a) is one independent sentence with an embedded sentential relative clause, whose antecedent is the propositional content of the sentence preceding the relative clause.[59] Thus, (47b) contains two translational units, and hence the two string pairs shown in respectively (48) and (49) are recorded among the data:[60]

(48a) IPpot: Otherwise, just as consciously, I would have to resign myself to the prospect of a lasting emptiness, (AB)

(48b) IP: Eller jeg måtte like bevisst resignere for utsikten til en varig tomhet,

(49a) CPrel: the very idea of which threatened and offended me. (AB)

(49b) IP: bare tanken på det truet og krenket meg.

In (48) the English sentence (47a), with its sentential relative clause omitted, corresponds with the first matrix sentence of (47b), and in (49) the relative clause of (47a) is paired with the second matrix sentence of (47b). Since the entire matrix sentence (47a) is a translational unit as well, we need to compile a third string pair consisting of the entire strings given in (47), and in this correspondence the source string is labelled *IP*, and the target string *IP-seq*, since it is a sequence of conjoined matrix sentences; cf. string pair (50) in 4.4.3.2. The compound matrix sentence (47b) is not a translational unit according to the extraction criteria, but its extraction is forced by the unit (47a), and hence it has been necessary to find a syntactic label for (47b).

---

[59] On English sentential relative clauses, see Quirk et al. (1985: 1118–1120).

[60] The category label given for (48a) is commented on in 4.4.3.2.

In the present study the need for syntactic categories denoting sequences of constructions of the same type has been seen most frequently in relation to matrix sentences, in cases similar to that of (47). In the analysed texts there are also sequences of finite subclauses corresponding translationally with a single CP, and sequences of noun phrases corresponding with a single NP.

4.4.3.2 "Potential" constituents

In the category labels *IPpot*, *CPpot*, and *NPpot* the segment *pot* is short for *potential*, and these labels are attached to expressions which on type level can be parsed as, respectively, a matrix sentence, a finite subclause, and a noun phrase. Thus, such linguistic units have the potential of occurring as tokens of IP, CP, and NP. Seen as linguistic types, such strings may be analysed as complete IPs, CP, or NPs, but as linguistic tokens they are incomplete in the sense that there is at least one other adjacent constituent together with which they form the given category. These special category labels are always assigned to strings which are embedded in superordinate correspondences, and are extracted because they correspond translationally with expressions conforming with one of the syntactic extraction criteria.

This can be illustrated by the potential matrix sentence, IPpot, shown in (48a) in 4.4.3.1. (48a) is recorded as a translational unit because it corresponds with the Norwegian matrix sentence (48b). On type level, (48a) can be analysed as an IP, but in the linguistic context from which it is taken it forms an independent sentence together with the sentential relative clause in (49a). As pointed out in 4.4.3.1, a third string pair is also recorded, since the entire matrix sentence shown in (49) is also a unit of extraction. String pairs (48) and (49) are both embedded in this third string pair, shown in (50):

(50a) IP: [IPpot:3] [CPrel:3] (AB)
(50b) IP-seq: [IP:3] [IP:3]

The category of (50a) is IP, and that of (50b) is IP-seq (cf. 4.4.3.1). String pair (50) includes no words, only placeholders for embedded units, and this is a consequence

of the opacity principle (cf. 4.3.6.5), as well as of the principle of avoiding the duplication of word strings among the compiled data (cf. 4.3.3).

Since strings of the categories IPpot, CPpot, and NPpot always appear as embedded units, it is relevant to consider how translational complexity may be affected by correspondences between categories of the types Xpot and X, where *X* refers to the same syntactic type. Since an Xpot can be analysed as an X on type level, we regard X and Xpot to be of the same category when seen as linguistic types. Then, the question whether an X and an Xpot have the same syntactic function cannot be answered generally but must be decided for each given case. In the light of the discussion of type assignment in nested correspondences (cf. 4.3.6.6), the conclusion is that correspondences between an X and an Xpot are allowed within types 1, 2 and 3 if the units have the same syntactic function at the level of the matrix string pair. If there is disagreement with respect to function, the correspondence is of type 4.

Another question is whether it is necessary to categorise strings like (48a) as *potential* matrix sentences, since they seem to be treated exactly as other matrix sentences in the analysis. However, since units of type Xpot are embedded translational units, the categorisation of a string as an Xpot, carries the information that there is some constituent adjacent to it, together with which it forms a constituent of the category X. Thus, information about syntactic composition is tied to the use of the categories IPpot, CPpot, and NPpot. For instance, we know that a potential noun phrase followed by a relative sentence forms a noun phrase.

4.4.3.3 Verbless clauses

Quirk et al. (1985: 996) describe English verbless clauses as a type of syntactic compression, and state that in such constructions "... we can usually postulate a missing BE and to recover the subject, when omitted, from the context." This is a suitable characterisation of the set of English, as well as Norwegian, strings which have been labelled *CPnovrb* within the recorded data. With respect to Norwegian, Faarlund et al. (1997: 958–972) have classified this group of constructions as sentence fragments (*setningsfragment*), which indicates a certain heterogeneity among its members, and heterogeneity is indeed a feature of the set of recorded

verbless clauses. This may be illustrated by examples (51)–(53), in which the verbless clauses are indicated by italics.[61] The examples present the verbless clauses with either some preceding or succeeding context.

(51a) Moren nikket mot etasjen over, munnen var stram. (BV)
'Mother.DEF nodded towards floor.DEF above, mouth.DEF was tight.'
(51b) She nodded towards the ceiling, *her lips pressed tight*.

In (51b) the expression *her lips pressed tight* consists of a subject NP, a past participle and a manner adverbial, and, as a syntactic unit, it functions as an adverbial in the matrix sentence.[62] Since *her lips pressed tight* is the translational correspondent of the Norwegian matrix sentence *munnen var stram* in (51b), the English sentence fragment has been extracted as a unit of translation. We have chosen to classify it as a verbless clause, as we may assume that the copula verb is missing: the sentence *her lips were pressed tight* would have been contextually appropriate.

(52a) "Jeg mener han ville ha godt av det."
"Helt sikkert, John." (Gloss: 'Quite surely, John.') (EFH)
(52b) "I think it will be good for him."
"I'm sure it will, John."

The verbless clause *Helt sikkert, John* in (52a) is a unit of extraction, firstly, because it corresponds translationally with the English matrix sentence *I'm sure it will, John*, and, secondly, because it functions as a textual sentence (cf. 4.3.2.4). In this case, the verbless clause contains a sentence adverbial (*helt sikkert*) and a proper name functioning as a vocative. The text sequences shown in respectively (52a) and (52b) are passages of dialogue, and the verbless clause in (52a) is a comment to the

[61] Note that the examples (51)–(53) do not represent extracted string pairs. The italicised verbless clauses are extracted as translational units because they correspond translationally with strings matching one of the syntactic extraction criteria. The examples are given to provide some linguistic context for the verbless clauses being discussed.
[62] The expression *her lips pressed tight* is an example of an English construction described by Hasselgård (forthcoming) as possessive absolutes, and by Quirk et al. (1985: 1120–1121) as absolute adverbial clauses, which may be nonfinite or verbless.

preceding statement. The expression could be expanded into a matrix sentence by adding a referring subject (*det*, 'it/that') and a copula verb (*er*, 'is'): *Det er helt sikkert, John*.

(53a) Når særlige grunner tilsier det, kan departementet forlenge tillatelsen for ett år om gangen i inntil 4 år. (*Petro*)
(53b) *When justified by special reasons*, the Ministry may extend the licence for periods of one year each up to a total of 4 years.

(53b) illustrates another kind of sentence fragment: the expression *when justified by special reasons* is extracted because it corresponds translationally with the Norwegian finite subclause *når særlige grunner tilsier det* ('when special reasons justify it'). This construction may be described as a nonfinite clause introduced by a complementiser. In analysed texts examples of this type are normally subjectless. The example given in (53b) functions as a conditional adverbial in the matrix sentence. We have described it as a verbless clause since an expletive subject and a copula can be seen as missing: the finite subclause *when it is justified by special reasons* would also have been felicitous in the given context.

The presence of past participle verb forms (*pressed*, *justified*) in two of the examples indicates that syntactic units belonging to this group could perhaps more appropriately be described as *nonfinite sentence fragment*, or *nonfinite construction* than as *verbless clause*. In the examples recorded from the analysed texts, the common denominator for this class is principally the absence of a finite verb, which is normally a copula. That the category CPnovrb is attached to the whole group primarily reflects the need for a shorthand label for a rather heterogeneous group, whose members are identified as a practical consequence of the syntactic extraction criteria. In the present project it is of greater interest to study how these linguistic units are translationally related to their correspondents in the parallel texts than to discuss linguistic properties of the various members of this group. It should be emphasised that in our analysis verbless clauses, or nonfinite sentence fragments, are extracted only when at least one of the following conditions applies: either the non-

finite construction occurs as the translational correspondent of a string which qualifies as a translational unit, or it is marked by punctuation as a textual sentence.

4.4.3.4 Incomplete constituents

Finally, we should comment on the category labels *IPinc*, *CPinc*, and *seq*. The segment *inc* stands for *incomplete*, and *seq* is an abbreviation of *sequence*. Strings recorded with these labels have been extracted either because they have been marked by punctuation as a textual sentence (cf. criterion (1d) in 4.3.2), or because they correspond with a translational unit in the parallel text.

Among the recorded units of analysis there are certain strings which are marked by punctuation as textual sentences, but cannot be analysed as complete syntactic units, whether at phrase or sentence level. Some of these strings may be recognised as broken off sentences, and these are labelled *IPinc* or *CPinc*, depending on whether they would be matrix sentences or finite subclauses if necessary constituents were added to make them syntactically complete. (54) is an example of a pair of incomplete matrix sentences:

(54a) IPinc: alle ting tyder på at — (EFH)
'all things indicate on that —'
(54b) IPinc: everything indicates that...

The wider textual contexts provided in (55) show in what sense the strings (54a–b) are textual sentences:

(55)

Ute i anretningen kan han høre foreldrenes stemmer i brokker og bølger inne fra stuen. Han skjønner de er kommet til det kritiske punktet; det er avgjørelsen som tas nå. "— ikke på tale — huset — oppussing — vitenskapen — vitenskapen! — men meget? — alle ting tyder på at — guttens utvikling — og leksene? — vitenskapens økende betydning i årene som — frisk luft — betingelse —"	Outside in the pantry he can hear his parents' voices piecemeal and in waves coming from inside the living room. He realizes they have come to the critical moment and a decision is now being made. "... not for anything... the house... decorating... science... science!... but a lot?... everything indicates that... the boy's development... and his homework?... the increasing importance of science in the years to... fresh air... conditions..."

(EFH)

Then, we use the label *seq* in cases where it seems difficult to identify a word sequence as something that would be a sentence structure if completed, and where we cannot find any other suitable description of the string. This is truly a residual category in our analysis: word strings which cannot be parsed as any kind of syntactic unit.[63] An example of a sequence is given in italics in (56a):

(56a) Og faren justerer, utbryter *at så sannelig, —* (EFH)
'And father.DEF adjusts, exclaims that so truly, —'
(56b) His father focuses, then exclaims: "Ah, there it is."

In (56a) the transitive verb *utbryte* ('exclaim') takes an object that cannot be parsed as one constituent, namely the sequence *at så sannelig*, which consists of the complementiser *at* followed by the adverbial phrase *så sannelig*. Because the sequence corresponds with a translational unit in the parallel text, i.e. the matrix sentence *Ah, there it is*, it is extracted and recorded in the embedded string pair (57):

(57a) seq: at så sannelig, — (EFH)
(57b) IP: "Ah, there it is."

4.4.4 Other notational conventions

When a string pair is recorded in Text Pair Mapper, it is output to the screen in a fixed manner. This will be illustrated by the case of multiple embedding previously discussed in 4.3.3 (example (17)), here repeated in (58), where syntactic bracketing indicates how the strings are structured into super- and subordinate units:

(58a) [$_{IP}$ Og med [$_{NP}$ uvante hender, som allikevel har lært [$_{CPwh}$ hva de skal gjøre] og som snart skal greie dette helt på egenhånd;] med hvite, litt kalde barnehender sikter han mot solen, dreier på skruene og plasserer teleskopet i riktig stilling.] (EFH)

[63] The notion of 'sequence' applied here is similar to, but not quite the same as, the understanding of 'chunk' used in 4.3.5.1, since chunks may, but need not, constitute syntactic units.

(58b) [$_{IP}$ With [$_{NP}$ awkward hands, which have nevertheless learned [$_{CPwh}$ what they have to do] and will soon be able to do this on their own,] with chilly white childhands, he focuses on the sun, turning the knobs and adjusting the telescope into the right position.]

From the parallel text in (58) three string pairs are extracted: the superordinate correspondence, the intermediate correspondence, and the most deeply embedded subcorrespondence; cf. (17)–(20) in 4.3.3. After recording, the supercorrespondence will be presented by the program as illustrated in (59):

(59) *Source:* Og med (NP:3) med hvite, litt kalde barnehender sikter han mot solen, dreier på skruene og plasserer teleskopet i riktig stilling.
Target: With (NP:3) with chilly white childhands, he focuses on the sun, turning the knobs and adjusting the telescope into the right position.
Type: 4 *Source cat:* IP *Target cat:* IP
Comment:
Embedded correspondences: source: [NP[CPwh]]; target: [NP[CPwh]]
4despec: deletion of the discourse particle *og*; otherwise type 3

(59) illustrates the principle that when a supercorrespondence is displayed, embedded correspondences are treated as opaque units. In the screen output, source and target strings are followed by information about the correspondence type, the syntactic category of each string, and, finally, the string pair comment, which may be empty.

In the comment field, information is presented in a fixed order. The first element is information about embedded correspondences, if any, and this is given automatically by the software. Then follows any information entered by the annotator, and such comments have been written in accord with certain user-defined conventions in order to facilitate searching for instances of specific phenomena among the collected data. These conventions are independent of the design of the software, and have been chosen specifically for the present study. Since Text Pair Mapper allows searching for specific character strings contained in the comment field (cf. 4.4.2), the use of standardised notation makes it possible to compile instances of annotated phenomena within the correspondences recorded for a given text pair. In particular, we have taken advantage of the comment field in order to mark recurring observations among

the collected data, and most important among these are the subtypes of correspondence types 3 and 4, which have been tagged by a set of fixed labels. Examples of such labels are shown in (59) and (61), respectively: *4despec* signifies non-predictable despecification, and *4spec* is shorthand for non-predictable specification; cf. 6.3.1.1.

In the comment field, another prominent type of user-entered information pertains to phrase-internal clauses. When lexical phrases containing a finite clause as syntactic complement are recorded as translational units (cf. extraction criterion (1c) in 4.3.2), then information about the embedded clause is entered in the comment field. Since the comment pertains to the *pair* of strings, it is necessary to indicate whether it is the source or target string, or both, that contain an embedded clause. In example (60) both strings contain a phrasal subclause:

(60) *Source:* så blått som det bare er i april (EFH)
Target: blue as it can be only in April
Type: 3 *Source cat:* AdjP *Target cat:* AdjP
Comment:
>CP - >CP

In (60) the right angle brackets (>) signify syntactic embedding, more precisely, that the syntactic categories following the angle brackets are embedded in respectively source and target string, and a hyphen (-) is used to indicate the division between source and target string. The category entered to the left of the hyphen is embedded in the source string and the category to the right of the hyphen is embedded in the target string. By contrast, in (61) there is an embedded finite clause only in the target string:

(61) *Source:* described by him as silly (DL)
Target: et ansiktsuttrykk han pleide å karakterisere som tåpelig
Type: 4 *Source cat:* VPed *Target cat:* NP
Comment:
- >CPrel
4spec

In the comment field of (61) the absence of syntactic information to the left of the hyphen signifies that the source string contains no embedded finite clause, whereas the character sequence *>CPrel* to the right of the hyphen indicates that the target string contains an embedded relative clause. In a case where only the source string contains an embedded clause (e.g. a CP) this is indicated by *>CP* -, where the positioning of the hyphen to the right of the syntactic information shows that there is no clausal embedding in the target string. Thus, the hyphen is obligatory when we mark embedded constituents in this manner. By convention, information about syntactic complementation in phrase-level units precedes other types of user-entered information about the string pair. This is shown in example (61), where the second line in the comment field contains the label *4spec* (cf. above). By observing the notational conventions, it is possible to search among the recorded string pairs for correspondences involving specific patterns of phrase-internal subclauses.

In addition to the marking of subtypes and phrase-internal clauses, the comment field has been used for entering brief notes on various properties of individual string pairs. Such notes are not necessarily standardised, and they include remarks on interesting translational and linguistic phenomena, on linguistic and/or translational quality, or on the correspondence type assigned. Remarks of the latter kind mainly deal with cases of dubious, or problematic, classification.

4.5 Summary

The empirical data collected for the present study constitute a manually analysed and annotated corpus of about 68 000 words. The data are taken from English-Norwegian parallel texts of two types, respectively fiction and law texts, and both directions of translation are represented. An overview of the analysed text pairs is given in 4.2, and some characteristic features of the two text types are presented in 4.2.2.1–2.

Different concerns that have governed the selection of texts are discussed in 4.2.1 with subsections. Firstly, because structural differences between two language systems may have consequences for the amount of information that is normally encoded per linguistic unit, the degree of complexity in translation tasks may be influenced by the direction of translation in a given language pair, and for this reason

Norwegian and English appear as both source and target language in the compiled data. Secondly, it is an aim in the present project to investigate whether samples of two different text types exhibit variation with respect to the degree of translational complexity, and for this purpose narrative fiction is chosen as an example of an unrestricted text type, and law text as an instance of a restricted type. Thirdly, since the linguistic features of texts may (at least in the unrestricted case) be influenced by variation in the stylistic preferences of individual authors, texts produced by more than one author are included for each direction of translation. Finally, the selection of texts for analysis has been constrained by the issue of lawful access.

The basic approach of the empirical method is to extract translationally corresponding strings from parallel texts, and to classify each string pair according to the scale of translational complexity defined by the correspondence type hierarchy presented in chapter 3. Since the present study aims at investigating how far it would be possible to automatise the translation of the selected texts, the finite clause is chosen as the primary unit of translation because MT systems typically operate sentence by sentence. Also, a central principle behind the analysis is to delimit translational units on the basis of surface syntactic structure. The main syntactic types identified as units of analysis are matrix sentences, finite subclauses, and lexical phrases with finite clause as syntactic complement (cf. 4.3.2).

Certain challenges encountered when applying the syntactic extraction criteria are discussed in 4.3.5 with subsections. Firstly, discontinuous units of translation present practical problems in relation to the software used for storing and organising the recorded data. Secondly, problems of choice will occur in cases where the syntactic criteria identify translational units between which there is only a partial correspondence. Such challenges may be solved by expanding the translational units. Thirdly, there are cases where certain strings identified as translational units have no correspondent in the parallel text.

In 4.3.6 with subsections the assignment of correspondence type to string pairs is described as an elimination procedure where we start by testing each correspondence for the lowest type and then move upwards in the hierarchy if the test fails. In practice, this is an analysis of the degree to which linguistic matching relations hold

in a pair of translationally corresponding strings. When type assignment is carried out, strings identified as translational units are seen as items on the level of the language system. I.e., type assignment is done solely on the basis of the information expressed by the linguistic material present within the two translationally corresponding strings. Furthermore, in the case of nested string pairs, embedded correspondences are opaque units in relation to the superordinate string pair. Thus, the classification of a matrix correspondence is done independently of the degree of complexity in embedded string pairs. The only information available, at the matrix level, about subordinate strings is information about their syntactic categories and functions.

The identification of translational units, as well as the categorisation of each extracted correspondence, is done manually. When string pairs are extracted and classified, source and target text are analysed in parallel by the human annotator. The syntactic analysis involved in this is based on rudimentary X'-analysis. The primary aim of the analysis is to identify the syntactic type of each recorded string, and to do so without applying very detailed assumptions of syntactic theory. The inventory of syntactic categories used for describing the compiled translational units is presented in 4.4.3 with subsections.

While the analysis of each string pair is done manually, specially designed software, the Text Pair Mapper (Dyvik 1993), is used for storing and organising the recorded data. The program calculates the distribution of the four types of translational correspondence within an analysed text pair, and results of this kind are the basis for the measurements of translational complexity to be discussed in chapter 5. Further, the software offers a range of options for sorting, and searching within, the correspondences extracted from a given text pair. This has enabled us to extract from the data information about certain recurrent linguistic phenomena, which will be presented in chapter 6.

PART IV
RESULTS AND DISCUSSION

5 Complexity measurement

5.1 Overview

In the presentation of the correspondence type hierarchy in chapter 3, as well as in the discussion of units of extraction in chapter 4, the focus has been on individual translation tasks and on individual string pairs. In this chapter attention is given to the *pairs of texts* that the empirical data are collected from, and we will present the results of applying the method described in chapter 4 to the selected English-Norwegian parallel texts. The collection of data is a set of type-sorted string pairs, and the distribution of the four types of translational correspondence within a set of data provides a measurement of the degree of translational complexity in the parallel texts that the data set is extracted from.

The chapter is divided into four main parts. The first part, 5.2 with subsections, presents the complexity measurement across the entire collection of data. In the second part, 5.3 with subsections, where we discuss complexity measurements within each of the two directions of translation. The third part, 5.4 with subsections, presents the results for the investigated text types, and includes a discussion of various text-typological aspects that may have influenced the differences in translational complexity found between, respectively, fiction and law texts. Complexity measurements for individual text pairs are discussed in the fourth part, 5.5 with subsections.

5.2 Translational complexity across all data

The characteristics of the four types of translational correspondence may be summed up as follows: Within type 1, the translationally corresponding strings are pragmatically, semantically, and syntactically equivalent, even down to the level of word forms. Also in type 2 correspondences relations of pragmatic, semantic, and syntactic

functional equivalence hold between source and target string, but the strings exhibit at least one structural mismatch with respect to the sequence of constituents or the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, and, in contrast to types 1 and 2, there is at least one structural difference violating syntactic functional equivalence between the strings. Finally, in type 4 correspondences there is at least one semantic difference between source and target string, so that they are not semantically equivalent. Pragmatic equivalence may, or may not, hold in type 4 correspondences. Types 1, 2, and 3 together cover the correspondences where we assume that the translation task is computable, i.e. where the target expression can be predicted on the basis of the source expression together with given, linguistic information sources. Type 4 correspondences cover the non-computable translation tasks where additional information is needed in order to produce the target expression.

Correspondence types 3 and 4 can be further divided into a set of subtypes identified through semantic criteria, and these subcategories will be discussed in chapter 6. Hence, we may refer to types 1, 2, 3, and 4 as the *main* correspondence types. The subtypes constitute a characterisation of patterns of semantic divergences observed within the data, whereas the distribution of the main correspondence types is a measure of the degree of translational complexity in the compiled data. As explained in 3.2.4, the classification of translational correspondences is a way of sorting the compiled string pairs into classes according to the types and amounts of information needed to produce the target string, and the amount of effort required in order to access and process the necessary information.

5.2.1 Global measurement of translational complexity

As presented in 4.2, the compiled data constitute an annotated English-Norwegian parallel corpus of about 68 000 words, including two text types (fiction and law text), and both directions of translation. Table 5.1 presents the "global" results of our investigation, i.e. the distribution of main correspondence types within the total collection of data.

Table 5.1. The global distribution of correspondence types in the investigated texts.

Total results, all text pairs	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	601	272	1 347	2 219	4 439
Percentage of string pairs	*13,5*	*6,1*	*30,4*	*50,0*	*100,0*
Length of source text	1 906	1 642	12 179	19 263	34 990
Percentage of source text	*5,4*	*4,7*	*34,8*	*55,1*	*100,0*
Length of target text	1 926	1 741	12 940	20 547	37 154
Percentage of target text	*5,2*	*4,7*	*34,8*	*55,3*	*100,0*

When discussing the results displayed in table 5.1, we want to focus on the *proportions* of text covered by each correspondence type, rather than on the absolute numbers of occurrences of each type. The reason for this is that, in the given language pair, the two least complex types (1–2) normally occur in pairs of short and syntactically simple strings of words, whereas pairs of longer and more complex strings tend to be of the two higher types (3–4). Thus, types 1 and 2 would appear as covering an unproportionally large amount of the analysed texts if the distribution of the main correspondence types would be presented merely on the basis of the numbers of string pairs.

The proportions of text covered by each correspondence type are given as the lengths of, respectively, source and target text. More precisely, text length is here measured through the notion of string length which is explained in 4.3.4. That is, the length of a recorded translational unit equals its number of word forms, and if the unit contains any embedded strings, then each embedded unit adds 1 to the length of the superordinate string.[1] Thus, when table 5.1 shows, e.g., that the whole set of type 1 correspondences amounts to a string length of 1 906 word forms in the source texts, then this figure is the result of adding together the lengths of all source strings contained in the recorded type 1 correspondences. The corresponding percentage (5,4) is the proportion of type 1, given in string length, in relation to the total length

[1] The motivation behind this is to avoid duplicate strings among the recorded data; cf. 4.3.3.

of all recorded source strings. The discussion will focus on proportions of text length, and not on numbers of string pairs, also in the later presentations of complexity results relative to directions of translation, text types, and individual text pairs.

That the total amount of analysed data comprises about 68 000 words calls for a comment on the information given in table 5.1 for, respectively, source and target text length. According to this, source and target text together have a total string length of 72 144. As explained in 4.3.4, the category symbol of an embedded unit adds 1 to the length of superordinate string, and this accounts for the difference between the figures given for string length in table 5.1 above and the figures given for numbers of words in table 4.1 in 4.2.

We will observe four points that we regard as the most striking results displayed in table 5.1. Note that the percentages to be given here are calculated as the average values of the proportions of, respectively, source and target text, given in per cent in table 5.1.

Firstly, table 5.1 shows that more than the half of all recorded data are classified as non-computable translational correspondences, as 55,2% of the analysed parallel texts are covered by type 4 correspondences. That is, granted the analytical framework described in chapter 3, we assume that 55,2% of the compiled data are not included in linguistically predictable translations; they are correspondences where the translation task is not computable.

Secondly, table 5.1 shows that the remainder of the data, 44,8%, is covered by correspondences of types 1–3, which is to say that in only 44,8% of the recorded translational correspondences the target string is assumed to be linguistically predictable.

Thirdly, the sets of correspondences classified as types 1 and 2, respectively, together cover 10,0% of the analysed parallel texts. On the basis of the assumptions given in chapter 3 regarding the expected amount of processing effort in these correspondence types (cf. 3.3.2.3 and 3.3.3.3), we thus assume that in only 10,0% of the compiled data the task of generating the target string is both computable and easily solvable, as it would require a modest amount of processing effort.

Fourthly, type 3 correspondences cover 34,8% of the analysed texts. These are string pairs where the translation task is assumed to be computable, but it may be expected that solving it is a highly resource-intensive task (cf. 3.3.4.3). Within the domain of linguistically predictable correspondences there is a marked division between, on the one hand, correspondence types 1 and 2, and, on the other hand, correspondence type 3, as we assume that types 1 and 2 represent translation tasks solvable by using a moderate amount of processing effort, whereas tasks of type 3 can be highly resource-intensive.[2] From table 5.1 it follows that within the subset of linguistically predictable correspondences, i.e. types 1–3, the subset of type 3 correspondences constitute a large majority. I.e., in terms of string length, type 3 covers 77,7% of the amount of text included in the linguistically predictable string pairs.

Points (i)–(iv) are a summary of these four observations concerning the distribution of main correspondence types across all collected data:

(i) 55,2% of all recorded data are classified as non-computable translational correspondences.

(ii) 44,8% of the data fall within the domain of linguistically predictable translations.

(iii) In only 10,0% of the compiled data, we assume that the target string can be generated automatically using a modest amount of processing effort.

(iv) In 34,8% of the data we assume that the translation task is computable, but resource-intensive.

5.2.2 Discussion of complexity across all data

Since correspondence types 1 and 2 cover a very modest proportion of the analysed texts (on average 10,0% across all data), we regard the distinction between the computable correspondences (types 1–3) and the non-computable correspondences (type 4) as the most informative indicator of translational complexity. This holds not only for the complexity measurement across all recorded string pairs, but also for the

[2] As explained in chapter 3, in types 1 and 2 required processing effort is mainly determined by the complexity of the parsing task, whereas in type 3, translation also demands semantic analysis of the source string, and the resource-intensive task of generating the target string from the semantic representation of the original (cf. 3.3.2.4, 3.3.3.4, and 3.3.4.4).

measurements within different subsets of the data, to be discussed in 5.3, 5.4, and 5.5, with subsections.

We may briefly note that the large proportion of non-computable translational correspondences found in this study is in agreement with the results presented in Thunes (1998), where the correspondence type hierarchy is applied in a study of translational complexity in four pairs of English-Norwegian parallel texts, covering about 33 000 words. The analysed material there includes fiction, law text, and technical documentation. However, the findings of that study are not directly relatable to the present investigation, firstly, because the quantitative results are given in terms of numbers of string pairs, not string lengths, and, secondly, because the opacity principle (cf. 4.3.6.5) is not incorporated in the applied method.[3] Thus, we will not compare the results of the present study in any detail with those of Thunes (1998), but merely observe that both studies show that there is a majority of non-computable correspondences in the sets of analysed parallel texts.

As we now evaluate the results presented in table 5.1 in 5.2.1, our focus is on the following: in 55,2% of the analysed texts the translation task is assumed to be non-computable; in 10,0% of the material we expect that the target strings can be produced using a modest processing effort, and in 77,7% of the data classified as computable correspondences, the generation of the translation is assumed to be a highly resource-intensive task. These facts provide a basis for answering the question of how far it is possible to automatise translation in the analysed texts. If this is understood as producing, with no manual intervention, translations which will exhibit the same properties as those found in the human-produced target texts, then MT does not appear very helpful in relation to the investigated material: according to our analysis, human translation can be simulated by fully automatic translation in only 44,8% of the analysed texts. Notably, this conclusion is drawn on the basis of a framework for complexity analysis which assumes a linguistic approach to automatic

[3] Hence, in that study the complexity of matrix string pairs is determined by the correspondence type assigned to embedded string pairs in cases where subcorrespondences are more complex than the superordinate ones; cf. Thunes (1998: 33–34).

translation; we have no basis for making any claims about the suitability of statistical methods in automatic translation.[4]

Granted that the majority of the recorded translational correspondences are of type 4, it is natural to ask to what extent linguistically predictable translations could be possible target strings for the source strings identified in the compiled type 4 correspondences. This question brings attention to the distinction drawn in 3.3.1.1 between the general task of translating a source string a_{L1} given some target language L_2, and the specific task of translating a_{L1} into the target string b_{L2}. The solution to the general task is predictable from information about the interrelations between source and target language systems, and the output of that task is the set of literal translations of a_{L1} with respect to L_2.[5] The complexity of the general task is the minimal complexity of translating a_{L1} into L_2. The complexity of the specific task, that of translating a_{L1} into the target string b_{L2}, is the same as the complexity of the general task if b_{L2} is a member of the LPT set of a_{L1} with respect to L_2. However, if b_{L2} is not included in the LPT set, then the complexity of the specific translation task is higher because generating the target string b_{L2} requires more information than what is available in the source string and in the descriptions of source and target languages and their interrelations. Given our analytical framework, the latter point applies to the correspondences classified as type 4 within the empirical data. Moreover, in type 4 correspondences linguistically predictable translations are in principle *possible* provided that the LPT set of each source string is not empty. As we do not aim at opening a discussion of criteria for translation quality, we will at this point merely observe that although literal translations belong to the set of possible translations of a given source expression, it is a different matter whether a literal translation would be felicitous, or even acceptable, in relation to the context of each given source string.

Without doing a systematic empirical analysis, it is impossible to estimate to what extent literal translations are possible target strings for the source strings identified in

---

[4] It is, however, not plausible that statistical methods could do much better, since the identified source strings do not contain those types of required information which are the reason why type 4 cases fall outside the domain of computable translation tasks. Cf. 1.4.2.5 on the dichotomy between linguistic and non-linguistic approaches to MT.

[5] That is, the LPT set of a_{L1}; cf. 2.3.2–3.

the compiled type 4 correspondences. However, since English and Norwegian both belong to the Germanic language family, and are used in language communities which are, in cultural terms, not very far apart, we assume that among the recorded data there are only few source expressions with an empty LPT set.[6] Thus, if we assume that it is possible to provide a rule-based system for automatic translation with a full description of the two languages and their interrelations, then we suppose that the system would be able to generate literal translations for most of the source texts covered by our investigation.

The global distribution of main correspondence types, given in table 5.1 in 5.2.1, can be seen as an average measurement of the degree of translational complexity in the entire set of collected data. The global average will be referred to in the later presentations of complexity measurements within various subsets of data. The results for each subset of recorded string pairs will be related to the global average according to points (i)–(iv) in 5.2.1, i.e. in terms of the proportions of (i) non-computable translational correspondences (type 4), (ii) of linguistically predictable translations (types 1–3), (iii) of "easily" computable correspondences (types 1–2), and (iv) of resource-intensive, computable correspondences (type 3). The figures representing the complexity measurement of the entire set of data will be calculated as the average values of the percentages given in table 5.1 of, respectively, source and target text lengths.[7]

We assume that in a very large, representative parallel corpus for a certain language pair, the distribution of the four correspondence types would reflect the degree of translational complexity in the *parole* relation between texts in the given two languages.[8] But on the basis of the modest size of our empirical material, it will remain a mere speculation whether the distribution of correspondence types across the entire set of data may reflect the general degree of complexity in the translational relation between English and Norwegian, as instantiated on the level of *parole*. We have previously mentioned in 4.2 that the limited size of the recorded data prevents

[6] Cf. example (1) in 2.3.2.
[7] Cf., e.g., the right-most column in table 5.4 in 5.3.1.
[8] Such a corpus would include both directions of translations, as well as a large variety of text types and authors; cf. 1.4.3.1.

the detection of statistically significant results, and only tendencies may be observed within the recorded material. As will be shown in 5.5 with subsections, the complexity measurements for individual text pairs reveal a considerable degree of variation among them. Since only six text pairs have been analysed, this means that the standard deviation within the group of text pairs is considerable, too.[9] Because of this, it is difficult to generalise from the average measurement of translational complexity across the entire collection of data. Had there been only small variations among the text pairs, it would have been more likely that the global average could indicate the degree of complexity of this language pair in general. Still, since it is a measurement across the entire data set, we will use the global average as a basis for comparisons in the later discussions of complexity within various subsets of data.

Within the analysed texts, the distribution of correspondence types may be influenced by certain principles of the chosen empirical method. In particular, a combination of the criteria for identifying translational units and the principles for assignment of correspondence type may have an effect on the proportion of type 4 within the recorded data. Most notably, this pertains to string pairs exhibiting only a minimal semantic difference between source and target unit. In the analysed texts, the probably most frequent kind of minimal type 4 correspondences includes string pairs where the only semantic deviation between the two translational units is the presence or absence of linguistically expressed temporal information; these are normally cases where only one of the two strings includes a finite verb with a tense marker.[10]

Such correspondences reflect differences between English and Norwegian concerning the use of, respectively, finite and nonfinite constructions. Although the two languages have rather similar inventories of finite and nonfinite verb forms, there are also important divergences. In English, the use of nonfinite constructions, such as *-ing*-clauses and *-ed*-clauses, is far more frequent than the use of syntactically congruent structures in Norwegian. E.g., the various kinds of adverbial functions that may be realised by English *-ing*-clauses tend to be associated with finite subclauses in

[9] 'Standard deviation' is here understood intuitively as deviation from the standard, i.e. from the average value of a data set.

[10] Cases of this kind are also discussed in 6.3.1.3, where the phenomenon is illustrated by string pair (20), and discussed further in connection with occurrences of specification and despecification.

Norwegian.[11] In our view, cases where English nonfinite constructions correspond translationally with Norwegian finite clauses reflect a certain regularity which is included among the interrelations between the two language systems. With respect to English-Norwegian parallel texts, we will refer to this as the *nonfinite-finite pattern*. This is not an absolute regularity that excludes the speaker, or writer, from making choices between alternative expressions. In translation, an English nonfinite construction will not always be matched by a Norwegian finite subclause, or vice versa. We regard the nonfinite-finite pattern as created by an interplay between *langue* and *parole*. That is, the language systems determine what syntactic functions that may be associated with the various kinds of finite and nonfinite constructions, as well as the semantic contribution of those functions, but whether a finite or nonfinite construction is chosen in a specific context may also be influenced by factors pertaining to language use.

It falls outside the scope of the present project to study this topic in detail, but instances of the nonfinite-finite pattern are found in two classes of recorded translational correspondences. These are, firstly, string pairs where one of the units is a finite subclause, and the other is a nonfinite construction, and, secondly, correspondences between complex lexical phrases where only one of the extracted units contains a finite subclause, and where the syntactic complement in the parallel unit is some kind of nonfinite construction (cf. 4.3.2). Normally in such cases, the tensed expression is included in the Norwegian string, but the opposite situation may occur as well.[12]

Correspondences of these kinds are categorised as type 4 due to the absence of temporal information in the nonfinite expression, even if this is the only semantic

[11] To our knowledge, there exists no comprehensive study of how the various finite and nonfinite constructions of, respectively, English and Norwegian are translationally interrelated, but within constrastive language studies there are several contributions which deal with parts of this large topic, and some examples may be mentioned. Hasselgård (forthcoming) discusses English possessive absolutes, i.e. nonfinite and verbless adverbial clauses introduced by a possessive determiner (e.g. *her lips pressed tight*), and their Norwegian correspondences. Nordrum (2007) studies English nominalisations and translations of them in Norwegian and Swedish. Smith (2004) investigates sentence-initial *-ing* participle constructions in English and their translation into Norwegian. Behrens (1998, 1999) analyses free *-ing*-participial adjuncts in English and how they are translated into Norwegian. Thunes (1998: 31–32, 37–38) includes a few observations regarding correspondences between English nonfinite constructions and Norwegian finite subclauses.

[12] For more information on the latter point, cf. the discussion of occurrences of (de)specification in 6.3.1.3.

difference between source and target expression. As minimal cases of type 4, such string pairs can be said to be on the verge of semantic equivalence between source and target unit, and hence may be seen as concealing relations of cross-linguistic semantic equivalence in the compiled data.[13] E.g., if, in a given string pair, the only semantic difference between source and target expression is the absence or presence of grammatically expressed temporal information, then a non-linguist would most likely regard the two strings as expressing the same meaning when interpreting the expressions in relation to given contexts. The piece of temporal information missing in the nonfinite construction is normally available within the matrix sentence in which the tenseless unit is embedded. In 6.3.1.3 we will argue that correspondences instantiating the nonfinite-finite pattern constitute the most important factor that has created minimal cases of type 4.

The longer and the more frequent minimal type 4 correspondences are within the analysed texts, the larger is the amount of text included in them, and if they are sufficiently long and/or frequent, then the category of type 4 will cover a disproportionally large part of the recorded string pairs.[14] This illustrates how the results of our investigation are influenced by the way in which the parallel texts have been segmented into units of analysis. If we had chosen matrix sentences as the primary unit of extraction, then translational links between nonfinite constructions and finite subclauses would not have contributed to complexity in the same way, as the relevant piece of temporal information would be available in both units of translation. On the other hand, this would have increased the average string length within the data, which could have created other effects influencing the complexity measurements. We return to this point in 7.3.

[13] Cf. 3.2.2, where we cite van de Koot (1995: 39), who observes that natural language computations are "on the verge of tractability".

[14] In 5.4.2.6 this point will be discussed in relation to differences between the investigated text types.

5.3 Complexity relative to directions of translation

As discussed in 4.2.1.1, the direction of translation is a factor that may have consequences for the degree of translational complexity. Hence, we will present complexity measurements for each direction of translation in the recorded data.

5.3.1 Complexity measurements for the two directions

Tables 5.2–3 show the distribution of the main correspondence types relative to directions of translation. Table 5.2 presents the total results for three text pairs translated from English into Norwegian, and table 5.3 gives the total results for three text pairs translated in the opposite direction. For both directions, we have investigated one pair of law text and two pairs of fiction, and, as explained in 4.2, for each direction we have analysed comparable amounts of parallel text.[15]

Tables 5.2–3 show interesting differences between the two directions of translation, differences which can be related to points (i)–(iv) previously presented in 5.2.1, i.e. the proportion of non-computable correspondences, the proportion of linguistically predictable correspondences, the proportion of "easily" computable correspondences, and the proportion of resource-intensive, computable correspondences. As explained in 5.2.2, the global distribution of main correspondence types, given in table 5.1 in 5.2.1, can be seen as an average measurement of the degree of translational complexity in the entire set of collected data, and this average is displayed in the right-most column in table 5.4. The results presented in tables 5.2–3 for each direction of translation may be compared both with the global average, and with each other. The outcome of these comparisons is summed up in table 5.4, where the figures are calculated as the average values of the percentages of, respectively, source and target text lengths.

[15] Cf. table 4.1 in 4.2 for an overview of the analysed text pairs with respect to text type, direction of translation, and numbers of running words.

Table 5.2. The distribution of correspondence types within data recorded from English-to-Norwegian translation.

Total results, E → N	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	303	96	565	1 140	2 104
Percentage of string pairs	*14,4*	*4,5*	*26,9*	*54,2*	*100,0*
Length of source text	889	633	5 637	10 835	17 994
Percentage of source text	*5,0*	*3,5*	*31,3*	*60,2*	*100,0*
Length of target text	891	597	5 224	10 865	17 577
Percentage of target text	*5,1*	*3,4*	*29,7*	*61,8*	*100,0*

Table 5.3. The distribution of correspondence types within data recorded from Norwegian-to-English translation.

Total results, N → E	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	298	176	782	1 079	2 335
Percentage of string pairs	*12,8*	*7,5*	*33,5*	*46,2*	*100,0*
Length of source text	1 017	1 009	6 542	8 428	16 996
Percentage of source text	*6,0*	*5,9*	*38,5*	*49,6*	*100,0*
Length of target text	1 035	1 144	7 716	9 682	19 577
Percentage of target text	*5,3*	*5,8*	*39,4*	*49,5*	*100,0*

Table 5.4 shows that the degree of translational complexity is higher than the global average in the string pairs compiled from English-to-Norwegian parallel texts, and lower than the average in the Norwegian-to-English correspondences. The difference in complexity is particularly evident from rows (i) and (ii) in table 5.4, which highlight the division between computable and non-computable correspondences. Thus, while the global average of text included in the set of computable, or linguistically predictable, correspondences is 44,8%, the average across English-to-Norwegian data is 39,0%, and 50,5% across Norwegian-to-English. Rows (iii) and (iv) in table 5.4 provide further information on to what extent the three least complex corre-

spondence types are more frequent in English-to-Norwegian than in Norwegian-to-English.

Table 5.4. Differences in complexity between the two directions of translation.

Proportions of...	**E → N**	**N → E**	**in all data**
(i) non-computable translational correspondences (type 4)	61,0%	49,5%	55,2%
(ii) linguistically predictable correspondences (types 1–3)	39,0%	50,5%	44,8%
(iii) "easily" computable correspondences (types 1–2)	8,5%	11,5%	10,0%
(iv)resource-intensive, computable correspondences (type 3)	30,5%	39,0%	34,8%

In the presentation of the global results in 5.2.1, we commented on the proportion of the computationally resource-intensive type 3 within the subset of linguistically predictable correspondences. With respect to English-to-Norwegian, it follows from table 5.2 that, in terms of string length, type 3 correspondences cover 78,3% of the amount of text included in the computable correspondences. Likewise in the case of Norwegian-to-English, it follows from table 5.3 that type 3 correspondences cover 77,2% of the amount of text included in the computable correspondences. Both these percentages are fairly close to the global average of 77,7%, which was presented in 5.2.1.

5.3.2 Discussion of differences between the directions

As explained in 5.2.2, since the present study includes only a small number of text pairs, the complexity measurement across the entire set of data is not a reliable indicator for the general degree of translational complexity of the investigated language pair. Likewise, since only three text pairs have been analysed for each direction of translation, the results presented in 5.3.1 cannot be seen as representative of the general degree of complexity in the two directions, and they are no firm basis for claiming that the degree of complexity is higher in English-to-Norwegian than in

Norwegian-to-English translation. Tables 5.2–3 merely present the average results of the three different text pairs investigated for each different direction. Hence, the difference in complexity found between the two directions is basically a function of variation among the individual text pairs, a topic to be dealt with in 5.5 with subsections.

Across the two pairs of law texts, type 4 covers 61,2% of the English-to-Norwegian data (cf. table 5.11 in 5.5.1.1), and 39,1% of the Norwegian-to-English data (cf. table 5.13 in 5.5.1.1). Then, across the four pairs of fiction texts, type 4 covers about 60,1% of the recorded data in both directions of translation (cf. tables 5.15, 5.17, 5.19, and 5.21 in 5.5.2.1). Hence, the variation found in the degree of complexity between the two directions is mainly caused by the quite modest proportion of type 4 in the Norwegian-to-English law data (i.e. the *Petro* text pair, cf. 5.5.1.1), which reduces the average proportion of non-computable correspondences in that direction of translation.

In 4.2.1.1 the notion of informational density was discussed in connection with the methodological issue of direction of translation, and it was explained that if two given languages are different with respect to informational density, then the challenges that the translator will encounter may vary according to the direction of translation.[16] Thus, if there are systematic differences between English and Norwegian texts with respect to informational density, there will be asymmetry between the two directions of translation with respect to the information needed in order to solve translation tasks. As pointed out in 4.2.1.1, the notion of discourse information is in Fabricius Hansen's work associated with frequencies of new referents and non-redundant semantic conditions. This is different from the quantitative notion of information on which our analytical framework is based (cf. 2.4.1.1). Moreover, empirical analyses in terms of informational density rely on a fairly fine-grained semantic analysis that is not incorporated in the correspondence type hierarchy. Hence, although they would be highly relevant, research findings on informational density in English and Norwegian could not be applied directly to the present investigation.

[16] On informational density, cf. Fabricius-Hansen (1996), (1999).

Moreover, contrastive studies of Fabricius Hansen's notion of informational density have not been done for this language pair.[17]

When comparing tables 5.2 and 5.3, it is noteworthy that in the set of English-to-Norwegian correspondences the total length of target strings is somewhat shorter than that of source strings. The difference can be attributed to the text pair containing Articles 1–99 of the *Agreement on the European Economic Area* (*AEEA*), and the parallel Norwegian version. From the information provided by table 4.1 in 4.2 about the analysed texts extracts, it may be calculated that in the *AEEA* text pair the target text is as much as 12,9% shorter than the source text, in terms of numbers of word forms. Table 5.5 illustrates how the shortness of the Norwegian *AEEA* version influences the average measurements of target string length, as well as the differences between the average lengths of source and target strings. In table 5.5 string length is measured as explained in 5.2.1, and source-to-target length difference is calculated by subtracting the average source string length from the average target string length. The difference is given as a percentage of the average source string length. The *AEEA* is the only text pair where recorded target strings are, on average, shorter than recorded source strings (cf. table 5.22 in 5.5.2.2).

Table 5.5. Average string lengths, given for the entire data set, and for the data representing each direction of translation.

	Average source string length	Average target string length	Source-to-target length difference
Across **all** string pairs	7,9	8,4	+6,3%
Across string pairs **E → N**	8,6	8,4	–2,3%
Across string pairs **N → E**	7,3	8,4	+15,1%

The shortness of the Norwegian version of the *AEEA* is quite atypical of translated text. It is now generally agreed among translation researchers that translations are more explicit than non-translations, and this typically causes translated text to

[17] This information is provided by personal communication with Cathrine Fabricius Hansen, March 2011.

contain a larger number of words than the corresponding original text. Pym (2005: 30) observes that "[e]xplicitation is now bound to the study of the norms of translational behavior; it is a candidate for status as a universal or even law of translation." Within translation studies, the notion of explicitation is usually ascribed to Vinay and Darbelnet (1995), and Blum-Kulka (1986). According to Pym (2005: 30), it has been described by Vinay and Darbelnet as making explicit in the translation information which is only implicit in the original.[18] The so-called *explicitation hypothesis* is formulated in Blum-Kulka's (1986) article, and, according to Pym (2005: 31), it primarily links explicitation to redundancy in the target text. Pym (2005: 31–32) further mentions a range of later studies investigating various phenomena included in the notion of explicitation. He cites from Klaudy and Károly (2003) several examples of such phenomena, all of which are of relevance to the present investigation: (i) the replacement of an SL unit of meaning with a semantically more specific unit of meaning in the TL; (ii) the distribution of the meaning components of one semantically complex SL word over a set of TL words; (iii) the addition of new elements of meaning in the translation; (iv) the dividing of one SL sentence into two or more TL sentences, and (v) the expansion of an SL phrasal construction into a TL clause structure.[19] Among these five examples, all but the first one will typically increase the number of word forms of the target in relation to the source text.[20] This very straightforward effect of explicitation is evident even in the short extract of translationally parallel fiction texts shown in 4.2.2.2. Likewise, in 4.2.2.1 it is quite noticeable in the *AEEA* text pair sample that the target text is shorter than the original. Among the six text pairs investigated, the *AEEA* is the only one where this is the case. Possible explanations for it are presented in 5.5.1.2, where we will argue that in the case of the *AEEA* certain translation norms specific to law texts have worked against explicitation.

[18] Cf. Vinay and Darbelnet (1995); the French original version of that work appeared in 1958.

[19] (i) and (iii) have parallels in our notion of specification, presented in 6.3.1; (ii) is relevant to compositional non-equivalence, explained in 6.2.4.1.; (iv) and (v) are described by Fabricius-Hansen (1999) as, respectively, sentence splitting and clausal expansion (cf. 4.2.1.2).

[20] (i) may also cause an increase in word forms, but not necessarily.

As pointed out above, the higher degree of translational complexity in the English-to-Norwegian data than in those of the opposite direction is primarily a function of differences between individual text pairs. A relevant topic in this respect is the phenomenon of *specification* (cf. 6.3.1), which belongs to the set of subtypes identified within the main correspondence types 3 and 4. We shall see in chapter 6 that specification has been identified as the most frequent among various recurring phenomena that contribute to semantic non-equivalence between translationally corresponding units (cf. 6.3.1.3). Moreover, within the English-to-Norwegian data, the occurrence of specification is particularly large in the law text pair (the *AEEA*), and in one of the fiction pairs (DL). String pairs compiled from these two text pairs together cover about 75% of the parallel texts analysed for this direction of translation. Hence, it is reasonable to assume that the high frequencies of semantic specification in these two text pairs have contributed to a greater degree of complexity across the English-to-Norwegian data.

5.4 Translational complexity and text type

As previously commented on, an important aim in the present project is to study how differences between text types may have consequences for the degree of translational complexity, and hence fiction and law text are chosen as representatives of, respectively, unrestricted and restricted text types.[21]

5.4.1 Complexity measurements for the two text types

Tables 5.6 and 5.7 present the distribution of main correspondence types in relation to the dimension of text type. For both text types comparable amounts of data are collected from each direction of translation.

We will here present the results for the two text types following the line of the presentation in 5.3.1. Thus, the complexity measurements given for each text type will be compared with each other, and with the global average discussed in 5.2.2. The

[21] Cf. 1.1, 1.4.2.3, and 4.2.1.2.

outcome of these comparisons is summed up in table 5.8, which gives the average values of the percentages of source and target text lengths.

Table 5.6. The distribution of correspondence types within the data recorded from law texts.

Total results, law text	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	304	71	598	740	1 713
Percentage of string pairs	*17,8*	*4,1*	*34,9*	*43,2*	*100,0*
Length of source text	813	485	7 455	8 899	17 652
Percentage of source text	*4,6*	*2,8*	*42,2*	*50,4*	*100,0*
Length of target text	827	541	7 839	8 897	18 104
Percentage of target text	*4,6*	*3,0*	*43,3*	*49,1*	*100,0*

Table 5.7. The distribution of correspondence types within the data recorded from fiction.

Total results, fiction	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	297	201	749	1479	2726
Percentage of string pairs	*10,9*	*7,4*	*27,5*	*54,2*	*100,0*
Length of source text	1 093	1 157	4 724	10 364	17 338
Percentage of source text	*6,3*	*6,7*	*27,2*	*59,8*	*100,0*
Length of target text	1 099	1 200	5 101	11 650	19 050
Percentage of target text	*5,8*	*6,3*	*26,8*	*61,1*	*100,0*

Table 5.8 shows that there is a lower degree of translational complexity in the data recorded from law texts than in those compiled from fiction. This is evident from the division between computable and non-computable correspondences, displayed by rows (i) and (ii). While the proportion of text covered by computable string pairs is 44,8% across all recorded data, it is 50,2% in the law texts, and merely 39,6% in the fiction texts.

Table 5.8. Differences in translational complexity between the two text types.

Proportions of...	**in law text**	**in fiction**	**in all data**
(i) non-computable translational correspondences (type 4)	49,8%	60,4%	55,2%
(ii) linguistically predictable correspondences (types 1–3)	50,2%	39,6%	44,8%
(iii) "easily" computable correspondences (types 1–2)	7,5%	12,6%	10,0%
(iv)resource-intensive, computable correspondences (type 3)	42,7%	27,0%	34,8%

It is a pertinent question whether we should disregard, within the law data, all string pairs of the form *Article n – Artikkel n* since they are relatively frequent within the pairs of law text, and to some extent causes an overrepresentation of type 1 correspondences, indicated by the relatively high number of type 1 string pairs given in table 5.6.[22] But as explained in 4.3.2.4, we want to apply our investigation to running text, and hence such string pairs are not disregarded, since they are an integral part of the language used in the law texts. In terms of string length, the proportion of type 1 is anyway small within the law data.

Given the overall lower degree of translational complexity within the pairs of law text, it is surprising to see from row (iii) in table 5.8 that the proportion of the two least complex correspondence types is *smaller* in the law texts than in the fiction texts: while an average of 10,0% of all analysed parallel texts are covered by correspondences of types 1 and 2 together, only 7,5% of the law texts, but 12,6% of the fiction texts, are included in string pairs of the two lowest types. In combination with the marked difference between the text types with respect to the proportions of type 4, this results in a quite sharp difference concerning the proportion of type 3, which is, according to row (iv) in table 5.8, as large as 42,7% in the law data, and as modest as 27,0% in the fiction data.

With respect to the amount of text included in computable correspondences, it follows from table 5.6 that in the case of the law data, type 3 covers as much as

[22] Expressions of the form *Article n* in the law texts are commented on in 4.2.2.1.

85,2% of it, and for the fiction text pairs, it follows from table 5.7 that 68,4% of it is covered by type 3. Thus, within the subset of computable correspondences, the law data exhibit a larger proportion of presumably resource-intensive translation tasks than the fiction data do. This can be seen as correlated with the relatively small proportions of types 1 and 2 within the law data, which is further discussed in 5.4.2.7.

5.4.2 Discussion of text-typological differences

As previously explained in 5.2.2 and 5.3.2, on the basis of only six investigated text pairs, we cannot generalise from the complexity measurements for the entire data set, and for each direction of translation. Likewise, the text-type specific results presented in 5.4.1 cannot be seen as indicative of the general complexity of translating, respectively, law text and fiction between English and Norwegian. The results show that the degree of complexity is, on average, lower in the selected pairs of law texts than in those of fiction, but this is primarily due to the relatively low complexity measured in one of the two law text pairs (*Petro*; cf. 5.5.1–2). In the other pair of law texts, the *AEEA*, the degree of complexity is higher, and, in fact, quite similar to the average found across the four pairs of fictions texts.[23]

The discussion of the notion of text type in 4.2.1.2 made it clear that aspects of text typology have consequences for translation, and given that law text and fiction are instances of, respectively, restricted and unrestricted text types, we had anticipated a lower degree of translational complexity in the pairs of law texts than in the fiction text pairs. Still, the results do not indicate that while the analysed fiction texts appear as clearly unsuitable for automatic translation, the law texts appear as suitable. Also, the text type differences presented in 5.4.1 in tables 5.6, 5.7, and 5.8 are no more striking than deviations that may be observed between the individual text pairs (cf. 5.5 with subsections). As far as the investigated material is concerned, the degree of translational complexity is found to be so high that fully automatic translation does not seem to be a fruitful option for any of the analysed text types, at least if human-

[23] Cf. table 5.11 in 5.5.1.1 in comparison to table 5.8 in 5.4.1.

quality output is aimed for. This is in line with the evaluation of the global complexity measurement, presented in 5.2.2.

Then, we may consider the fact that for non-literary text types the use of automatic translation tools has become fairly widespread because it reduces the manual workload, and can be helpful even if post-editing of the output is required.[24] Chapter 6 will provide more information on various linguistic phenomena involved in the non-computable correspondences. Some of these factors are relevant to the text type dimension, and may indicate differences between the analysed fiction and law text concerning how demanding it would be to edit machine output in order to achieve the quality of human translation. This topic is to be developed further in 7.4.

The following discussion will focus on how the two text types deviate in terms of the degree of restrictedness, and we will relate differences in translational complexity found in the analysed samples of the two text types to differences in restrictedness. Concerning the law data, the complexity measurements show a noticeable diversity between the two analysed text pairs (cf. 5.5.1.1). Those results will be discussed in 5.5.1.2, where we explain that the two pairs of law texts instantiate quite different kinds of legal translation, and also present factors that may have contributed to the amount of non-computable translational correspondences found among the law data. Then, 5.5.2 with subsections will discuss further the results extracted from fiction, focussing on the observations that there is a sharp difference between the two text pairs translated from English into Norwegian, whereas there are striking similarities in the pairs translated in the opposite direction.

5.4.2.1 Norms and differences in restrictedness

The difference in restrictedness between the two text types is the direct reflection of a basic opposition between the language of the law and that of fiction: the former is strictly norm-governed in ways that the latter is not. In law-regulated societies the law is nothing less than the highest power, and this gives law texts their authority. Because of the optimally authoritative status of a law text, its production as well as its

[24] On post-editing, cf. 1.4.2.2.

interpretation are strongly governed by the intersubjective norms of the legal domain of society (cf. e.g. Bowers 1989: 53–54, Cao 2007: 13–14). The principal constraint on law text is perhaps linguistic clarity, described by Mattila (2006: 65) as "an absolute norm of legislation". The authority of a law text necessitates that its content is expressed with high precision, and without ambiguity. Hutton (2009: 80) points out that the authoritative nature of the language of law requires internal coherence, and that it should be "like an ideal language of science", consisting "of terms which correspond to precisely defined concepts." Moreover, legal interpretation is largely controlled by institutionalised rules.[25] Bowers (1989: 9) holds that legal drafting is still more strongly governed by norms than legal interpretation is. According to Bhatia (2010: 46–47), the primary concern of drafters is "loyalty to legislative intentions", and he describes four different norms of law writing (2010: 38–39): clarity of expression (i.e. avoiding vagueness), precision (by using as few words as possible), unambiguity, and all-inclusiveness (i.e. specifying adequately the scope of application of the law text).

Fiction texts are, like any kind of language use, subject to the linguistic norms of the language community, and there are norms of literary language use that shape the characteristics of various kinds of styles and genres.[26] Still, fiction texts are in no way as norm-governed as law texts are, and although literary norms, too, can have intersubjective existence, they are not institutionalised like legal norms. Moreover, it would be wrong to claim that, unlike law texts, fiction texts are without authority, since particularly successful fiction texts may become highly regarded. In such cases, the authoritative status of the fiction text is determined, firstly, by the creative ability of the author to express a story, and, secondly, since literature without receivers has little effect, by the capacity of that story to create great experiences in the minds of the readers. The subjective factors attributed to the sender and recipient of a fiction

[25] Among these the "literal" rule of interpretation is the most important. According to Bowers (1989: 113–118), the common understanding of this rule is that the literal meaning of the words in a law text is their meaning as given by ordinary dictionaries, and, basically, the literal rule says that "words should be taken literally unless they are anomalous in the context of the Act" (1989: 118). On literal interpretation, see also Hutton (2009: 71–77).

[26] The kind of norms that shapes the linguistic characteristics of literary styles is described by Leech and Short (2007: 41–44) as *relative norms*.

text are a true opposite to the institutionalised norms controlling the drafting and interpretation of law texts. This is illustrated by Leech (2008: 193) in his general claim about literary language that its "interpretation is particularly multivalent and open-ended", and he attributes the open-endedness of literary interpretation to deviations from the different kinds of norms governing non-literary language use.[27] Thus, the production of a fiction text is governed by the individual choices of the author, which may include norm violations, and its reception is determined by the subjective experiences of the readers. This is in sharp contrast to the norms of law texts, which are determined by the collective purpose of regulating society.[28]

In terms of Popper's division of reality into three domains (cf. 2.2.1), the creation and reception of a fiction text can to a large extent be ascribed to the second world of subjective mental states, whereas the production and interpretation of a law text is mainly governed by objective knowledge of the third world. The norms controlling legal drafting and interpretation constitute a set of conventions which is shared knowledge among the participants of the legal domain.[29] This Popper-inspired description is probably too simple, as it is most likely that elements of the second as well as of the third world are involved in both text types. Leech (2008: 190–194) warns against overestimating the contribution of the subjective experience of the reader in the interpretation of a literary text, since a fair amount of the reception of a text relies on what he describes as "common ground" within the audience. In this common ground Leech includes linguistic as well as extra-linguistic knowledge, and in relation to the typology of information sources given in chapter 2, it corresponds with general linguistic information and background information about the world.

---

[27] Leech explains the open-endedness of literary interpretation by referring to *foregrounding*, which can be explained as "motivated deviation from linguistic, or other socially accepted norms, ... a basic principle of aesthetic communication" (2008: 30), and the effect of foregrounding in literary language is that it "invites interpretations over and above the commonplace meanings which word strings have in typical non-literary texts" (2008: 193).

[28] This contrast is also expressed by Bowers (1989: 53–54), who points out that "[t]here is a chain of accountability in the drafting and interpretation of a statute that doesn't exist, for example, in the writing and reading of a novel: the novelist may write whatever and however he wishes, while the reader may interpret with the utmost of subjectivity, but the draftsman has the obligation of rendering as faithfully as possible what the government instructs him to do, and the courts must just as faithfully search for the basic intention of parliament."

[29] The objective character of the rules of legal interpretation is explained by Hutton (2009: 64), citing Wendel (2005: 1190–1191).

The point made by Leech elucidates one way in which fiction and law text can be described, respectively, as unrestricted and restricted text types. In the case of law text, drafting and interpretation is done on the basis of objective, common ground knowledge, the linguistic part of which is controlled by the norms of the language community, and the extra-linguistic part of which is controlled by the norms of the field of law. At least in the ideal case, drafting and interpretation is fully governed by this body of objective knowledge, although the work of an individual writer or reader of a law text will necessarily be influenced by previous experience. In the case of fiction, creation as well as interpretation involves an interplay between, on the one hand, linguistic and extra-linguistic common ground knowledge, and, on the other hand, individual mental states and subjective knowledge. The larger the contribution of the latter group of factors, the greater the possibilities of an open-ended set of interpretations. The opposite circumstances in the writing and reading of law text are characterised by Bowers (1989: 3) as an "explicit" process of "intention-to-expression-to-interpretation that is not to be found in other writing activities". Ideally, a piece of law text should thus be written so that there is only one available interpretation of it, whereas it is generally seen as an attractive quality of literary language that it can invoke what Leech (2008: 194) describes as "the superabundance of interpretations available".

5.4.2.2 Linguistic effects of differences in restrictedness

The difference in restrictedness between the two text types has already been described in 4.2.2.1–2 with reference to the inventories of linguistic constructions found in each type, and in this respect it is clear that the structural diversity showed by fiction texts is not present in law texts. Several other linguistic differences between the text types can be related to the issue of restrictedness.

At the level of macrostructure the two text types exhibit striking differences. The notion of macrostructure can be explained as "the general line of thought of a text and the sequence of passages typical for the text type" (Kussmaul 1997: 71). In this regard it is difficult to identify a canonical structure of narrative fiction texts, since

the individual author decides how to shape the telling of a given story.[30] In law texts the situation is quite the opposite: legal drafters must follow strict macrostructural conventions, which are an example of the text type-specific norms controlling the writing of law texts.[31] As pointed out in 4.2.2, one of the analysed pairs of law texts (the *AEEA*) is an example of an international legal instrument, and, as described by Cao (2007: 143–7), (2010: 89–90), this is a text type with a fairly fixed sequence of elements: title, preamble (normally describing the parties involved and the purpose behind the instrument), main text (typically starting with definitions), final clauses, attestation clause with signatures, and annex(es). The other analysed pair of law texts (*Petro*) also exhibits a rigid macrostructure quite similar to the former, but, being an instance of domestic law, it does not contain all the elements characteristic of international legal instruments.[32] At this point macrostructural differences between the two text types will not be discussed further; principally because the finite clause is the basic unit of analysis in our investigation of translational complexity, and hence macrostructure, as a notion related to entire texts, is of limited relevance. Still, the observed macrostructural properties reflect the difference in restrictedness between the two text types, and the macrostructural conventions followed in law texts contribute substantially to determining the formal characteristics of law texts.

Further, while parallel law texts exhibit strict one-to-one correspondences between translationally matching orthographic sentences, there is a fair amount of deviation from this pattern in parallel fiction texts. As described in 4.4.3.1, there are cases among the recorded data where a single matrix sentence in one text is matched in the parallel text by a string consisting of more than one matrix sentence.[33] Such correspondences are found in the pairs of fiction texts; they do not occur in the pairs of law texts. In the latter type, the translationally parallel versions are perfectly matched with respect to how the texts are divided into articles, numbers, matrix

[30] As discussed by Abbott (2002: 14–16), the structure of a narrative can be quite different from the structure of the story (i.e. an event, or series of events) that is represented by that narrative.

[31] Cf. the discussion in 2.4.2.1 of textual norms in the legal discipline.

[32] Cf. Cao (2007: 104–106) on the structure of domestic law texts.

[33] If the single sentence is contained in the source text, and the translationally corresponding set of sentences in the target text, then this is an example of sentence splitting, and probably an effect of explicitation, described in 5.3.2.

sentences, and enumerative lists. This is another consequence of the text type-specific norms which constrain the writing and, in the cross-linguistic setting, also the translation of law texts. To a certain degree, this is an effect of the macrostructural conventions described above, and it is primarily a consequence of the so-called "one-sentence rule", which is described by Šarčević (1997: 130) as "the practice of formulating each section or subsection [of an act] as a single sentence". Related to this is the convention of expressing only one idea per sentence. E.g., Šarčević (2007: 46) reports that drafting guidelines for EU legislation recommend that "each sentence should express one idea only and that there must be a logical link between the ideas expressed in an article". Thus, legal significance is attributed to the way in which the content of a law text is split into matrix sentence units, and the sequential ordering of the elements in a law text is of legal importance.[34] Due to these constraints, it is obligatory that the sequential ordering of the text is the same in different language versions. The translation of fiction texts are not subject to norms of this kind, although there may of course be cases where translators try to create a sentence structure in the target text as close as possible to that of the source text. This may be desirable in cases where the original sentence structure has a significant function, and the translator's challenge is then to achieve what Koller (1992: 116) describes as formal-aesthetic equivalence (cf. 1.4.1.1).

Another manifestation of the difference in restrictedness is that the individual factors involved in the creation of a fiction text invite linguistic creativity, whereas the authority of a law text requires strict standardisation in the language in which the law is expressed. The use of standard language secures the authority of the law text, which is further strengthened by keeping the linguistic form of a law text stable over time (Bowers 1989: 10), and by using a language variety accepted by the entire society (Bowers 1989: 69). All these aspects are supported by the norms of legal language. In contrast, fiction texts are generally expected to reflect the personal linguistic style of the author who is at liberty to break any linguistic, social, or cultural norms as long as it contributes to the expression of the story. Unlike the lan-

[34] See also Mattila (2006: 81–82) on the structure of law texts.

guage of law texts, the language of fiction will also naturally reflect the linguistic changes taking place in the given language community, and the use of non-standard linguistic elements is fully accepted.

5.4.2.3 Special-purpose texts

As previously mentioned, we have chosen texts of fiction and law as representatives of, respectively, unrestricted and restricted text types. The language of law is normally seen as an example of LSP, language for special purposes, which is in contrast to LGP, language for general purposes. It is commonly agreed that the distinction between LSP and LGP is not unproblematic; e.g. Laurén (1993: 14) points out that it is just as difficult to provide precise definitions of notions such as general language, literary language, spoken language, and others, as it is to define LSP. We will understand 'LSP' as the use of language for communication within a field where specialist knowledge is required, and this is in agreement with the definition of 'special languages' provided by Sager et al. (1980: 69), a definition that is still valid: "Special languages are semi-autonomous, complex semiotic systems based on and derived from general language; their use presupposes special education and is restricted to communication among specialists in the same or closely related fields."

Cao (2007: 8) prefers to describe legal language (in which the language of law texts is included, cf. 4.2.2.1) as LLP, "language for legal purpose", since it involves "language for special purpose (LSP) in the context of law". Concerning the language of fiction, it is problematic to regard it as representative of LGP if we assume that general language is both non-technical and non-literary. Since literary language involves elements of deviation from the norms applying to non-literary usage (cf. Leech 2008: 193), it is probably more appropriate to describe it, for instance, as LFP, language for fictional purposes.

Still, we regard it as uncontroversial to classify law text as a restricted text type, and fiction as an unrestricted text type. Bowers (1989: 354) sums up why the language of law texts, as an LSP, is syntactically restricted in order to suit its purposes: "... statutory language ... selects from the universal inventory those items and structures which most effectively carry out its special purpose. The selection of

structures is not a limitation of expressiveness but a functional application of effective structures and a rejection of unfunctional — interrogative sentences, cleft sentences, first- and second-person pronouns, and so on have no function in the categoric declaration of rights, privileges, and obligations." This also throws some light on why a broader inventory of structures is found in fiction: it is required for the multifaceted purpose of storytelling.

5.4.2.4 Pragmatic functions

In our view, the difference between fiction and law text concerning the inventories of linguistic constructions is related to what kinds of pragmatic functions that are found in the two text types. Topics such as communicative function, illocutionary force, and speech acts are frequently discussed in studies of the language of law texts.[35] E.g., Šarčević (1997: 121) observes that "regulatory speech acts" are found in the prescriptive parts of legal rules, and she lists the following types of such speech acts: "commands, prohibitions, permissions, and authorizations."[36] In a discussion of illocutionary force, Bowers (1989: 27–48) divides the law text into three parts: title and preamble, the enacting formula, and the legal provisions following the enacting formula.[37] He attributes representative illocutionary force to the first part; declarative illocutionary force to the enacting formula, and, with respect to legal provisions, he identifies three different kinds of illocutionary force: "facultative" (conferring rights, permissions, or power), "imperative" (imposing obligations), and "prohibitory" (stating prohibitions).[38] Moreover, in English law texts, certain expressions occur frequently as markers of some of these types of speech acts: the modal auxiliary *may* signals permissions and authorisations, and *shall* and *shall not*, encode, respectively,

---

[35] The notions of 'speech act' and 'illocutionary force' are briefly commented on in 2.4.2.1. For further information, see chapter 4 in Huang (2007).

[36] According to Šarčević (1997: 121), legal rules contain a descriptive part expressing "the conditions under which a rule becomes operative", and a prescriptive part "expressing legal actions".

[37] Examples of enacting formulae are, in English: *Be it enacted that...*; and in Norwegian: *Det gjøres herved vitterlig...* The use of such formulae is rather archaic in statutory language, and it is not found in the investigated law texts.

[38] Bower's description of illocutionary force in legal provisions is supported by Cao (2007: 115).

commands and prohibitions.[39] The types of illocutionary force identified by Bowers (1989) in legal provisions thus corresponds well with the types of speech acts mentioned by Šarčević (1997: 121). In the analysed pairs of law texts, definitions (e.g. of the form *for the purposes of this Agreement X means Y*) are also an important type of textual element, and in accord with the neo-Austinian speech act typology, these can be seen as typical assertives, or representatives (cf. 2.4.2.1).

The preceding observations may be summed up in relation to the mentioned typology. Thus, characteristic types of speech acts in law texts are, firstly, within the group of assertives: definition, and stating of purposes; secondly, within the group of declaratives: enactment, and, thirdly, within the group of directives: permission, authorisation, command, and prohibition. This is a tentative list, and must not be seen as exhaustive. For present purposes two points are particularly relevant. First, our tentative list of characteristic types of speech acts in law texts is noticeably limited in relation to the multitude of possible speech acts in language use. Second, it is striking that two groups of speech acts are not included in our list, and these are, respectively, commissives, of which typical examples are offers, promises, refusals, and threats, and expressives, such as apologies, accusations, congratulations, and thanking. Clearly, all these types are speech acts that are likely to be found in narrative fiction texts, especially when passages of dialogue are included, which is the case in the analysed pairs of fiction texts. A greater variety of pragmatic functions requires a larger inventory of linguistic constructions than the limited set that is used in statutory language. In law texts, the dominating sentence type is declarative, with no first or second person nominals, and this fits the characteristic types of speech acts mentioned above. Thus, the text-typological difference in the use of syntactic constructions can be seen as an effect of a contrast concerning the set of pragmatic functions, which is far more restricted in law texts than in fiction.

[39] Thus, the meaning of these modal expressions is narrower within the domain of law text, than within general language use; cf. e.g. Bowers (1989: 32–33), and Witczak-Plisiecka (2007).

5.4.2.5 The role of extra-linguistic information sources

Although the analysed law texts exhibit a lower degree of translational complexity than the samples of fiction texts do, the average proportion of non-computable, and hence semantically non-equivalent, translational correspondences is fairly large also within the law data.[40] This is a bit surprising since we would expect the strict norms of legal translation to work against semantic non-equivalence. Given our analytical framework, this result indicates that in both text types extra-linguistic sources of information have contributed substantially to the production of the target texts, since the amounts of literal, predictable translation are not large.

In our view, the important text-typological difference in this respect is that in the case of the law texts, objective knowledge about the domain of law writing, as well as about the domains to which the given laws apply, constitutes an essential extra-linguistic information source, whereas in the case of the fiction texts, the subjective knowledge of individual writers, translators, and readers is a significant source of extra-linguistic information.[41] With respect to the law texts, the norms governing drafting and translation are a vital part of extra-linguistic, objective knowledge. Examples are the four norms described by Bhatia (2010) (cf. 5.4.2.1), as well as the one-sentence rule (cf. 5.4.2.2), and the related convention of expressing only one idea per sentence. Such norms can either exist in the form of unwritten conventions shared by skilled legislators and legal translators,[42] or they can exist as formalised, written instructions to drafters and translators, described by Šarčević (1997: 122) as "institutional drafting guidelines", which will always be available for the production of multilingual law texts.[43]

---

[40] The proportion of type 4 is 49,8% in law text, and 60,4% in fiction; cf. table 5.8 in 5.4.1.

[41] As discussed in chapter 2, other kinds of extra-linguistic information, e.g. general background information about the world, are also necessary in both text types, but they do not in the same way highlight the text-typological differences.

[42] These are described by Bhatia (1997) as conventions specific to the genres of the legal discipline; cf. 2.4.2.1.

[43] Šarčević (2007) discusses guidelines applying to the laws of the European Union; cf. 5.5.1.2.

5.4.2.6 Semantic equivalence and non-equivalence

The identified text-typological difference in translational complexity reflects a difference between the law data and the fiction data in the extent to which there holds semantic equivalence between source and target strings (cf. row (ii) in table 5.8 in 5.4.1). Considering the difference in restrictedness between the two text types, this seems a fairly predictable result. Given that fiction texts are not norm-governed in the same way as law texts are, it is to be expected that there is a larger proportion of semantic equivalence between source and target string in the law data than in the fiction data. To preserve the authority of the target version of a law text it is necessary that it contains, as far as possible, the same meaning as the source text does. In the case of fiction, there is a greater tendency than in law text that semantic non-equivalence between translationally corresponding word strings is caused by various extra-linguistic elements influencing the choice of, respectively, source and target expression.

However, as noted in 5.4.2.5, the average proportion of non-computable, and hence semantically non-equivalent string pairs, is fairly large also within the law data. One factor that has contributed to this is the average length of the extracted string pairs: the translational units extracted from the law texts are on average longer than those compiled from the fiction texts, and the figures given in table 5.9 indicate that in this respect there is a considerable difference between the two text types. It is likely that the relatively large average string lengths of the analysed law texts have had an impact on the proportion of type 4 correspondences across the law data, since pairs of long and syntactically complex strings tend to be of the higher correspondence types (cf. 5.2.1). Another factor influencing the proportion of semantically non-equivalent string pairs is created by our analysis method: it was discussed in 5.2.2 that, independently of text type, the proportion of semantic non-equivalence in the analysed text pairs may be influenced by certain principles involved in the empirical method. More specifically, minimal type 4 correspondences function as a cover for relations of cross-linguistic semantic equivalence in the compiled data. This effect seems to be stronger in the pairs of law texts than in the fiction text pairs, firstly, because the translational units are on average longer among the law data, and,

secondly, since correspondences showing only one semantic difference between source and target string (such as the presence or absence of grammatically expressed temporal information, discussed in 5.2.2) are considerably more frequent within the law data than within the fiction data.[44]

Due to the factors described here, and given the norms that have shaped the law texts, we will claim that in the investigated pairs of law texts there is a higher degree of semantic equivalence between translationally corresponding expressions than what is attested by the quantitative results. Other aspects that may have influenced the proportion of type 4 correspondences within the law data are discussed in 5.5.1.2.

Table 5.9. Average string lengths, given for the entire data set, and for the data representing each text type.[45]

	Average source string length	Average target string length	Source-to-target length difference
Across **all** string pairs	7,9	8,4	+6,3%
Across **legal** string pairs	10,3	10,6	+2,9%
Across **fiction** string pairs	6,4	7,0	+9,4%

5.4.2.7 The proportions of types 1 and 2

As shown by table 5.8 in 5.4.1, the two least complex correspondence types (1–2) together cover 7,5% of the law texts, and 12,6% of the fiction texts. The main reason why both text types exhibit only small proportions of types 1 and 2 is the nature of the structural interrelations between the English and Norwegian language systems. There are important structural differences between the languages, so that also in a representative English-Norwegian parallel corpus we would expect only a small proportion of easily computable correspondences.[46]

---

[44] The frequency of minimal type 4 correspondences is commented on towards the end of 6.2.4.2.

[45] String length is here measured as explained in 5.2.1. Otherwise, cf. the presentation of table 5.5 in 5.3.2.

[46] On the basis of this limited study, our expectation is somewhere between 5 and 15 per cent, across text types.

As previously observed in 5.2.1, within the given language pair correspondence types 1 and 2 normally occur in pairs of short and syntactically simple strings of words. It is then our opinion that two different factors may explain why the proportion of types 1 and 2, together, is even smaller in the law texts than in the fiction texts. Firstly, the recorded translational units are, on average, longer within the law data than within the fiction data (cf. table 5.9 in 5.4.2.6). Secondly, the text-typological difference in restrictedness also contributes to the low proportion of the least complex types in the law texts. We have argued that statutory language contains a limited inventory of linguistic constructions, in comparison to the language of narrative fiction, and, as pointed out in 4.2.2.1, the analysed law texts are characterised by long sentences with syntactically heavy constituents. Hence, the kinds of syntactic constructions where correspondence types 1 and 2 can occur in translation between English and Norwegian seem to be disfavoured by the textual norms that apply to the domain of law writing. Within the law data the very small proportions of types 1 and 2 reflect to what extent there is structural equivalence between source and target strings in the set of constructions occurring in those texts. Thus, we expect that if a body of parallel text exhibits a wider selection of constructions, and a higher frequency of short translational units, then there will be larger proportions of types 1 and 2, as in the case of the fiction data.[47]

5.5 Translational complexity in individual text pairs

The complexity measurements for each of the six investigated text pairs show a considerable degree of variation in terms of how the main correspondence types are distributed within the data compiled from each text pair. As already explained, we cannot generalise from the average complexity measurements for, respectively, the entire data set, and the sets representing each text type and direction of translation, because only a very small number of text pairs has been studied. Still, in the following presentations of results for individual text pairs the average measures will be used as a basis for comparisons within the recorded data. The comparisons will be

[47] In relation to this, the text pair DL is an exception: it has the smallest average string lengths, and relatively small proportions of types 1 and 2. Cf. 5.5.2.1–2.

done in line with the pattern of the previous presentations of results in 5.2.1, 5.3.1, and 5.4.1, and the discussions will focus on one, or a few, aspects which may distinguish a given text pair from other parts of the investigated data.

5.5.1 The pairs of law texts

As presented in 4.2, the data recorded from law texts comprise string pairs extracted from two parallel texts: firstly, the *Agreement on the European Economic Area*, Articles 1–99 (*AEEA*), together with its Norwegian version, and, secondly, *Lov om petroleumsvirksomhet*, §§1–65 (*Petro*), together with its English translation.

5.5.1.1 Complexity measurements for the law texts

Table 5.10 presents the distribution of correspondence types within the data recorded from the *AEEA* text pair. As pointed out in 5.3.2, this body of parallel text is anomalous in the sense that the translated text is shorter than the original, and it is the only one among the investigated text pairs where this is the case. In table 5.10 the difference is evident from the measurements of, respectively, source and target text length.

Table 5.10. The distribution of correspondence types within the data recorded from the *Agreement on the European Economic Area* (*AEEA*), Articles 1–99, and its translation into Norwegian.

AEEA	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	163	15	208	405	791
Percentage of string pairs	*20,6*	*1,9*	*26,3*	*51,2*	*100,0*
Length of source text	384	119	3 170	5 644	9 317
Percentage of source text	*4,1*	*1,3*	*34,0*	*60,6*	*100,0*
Length of target text	377	98	2 593	4 961	8 029
Percentage of target text	*4,7*	*1,2*	*32,3*	*61,8*	*100,0*

Table 5.11 presents the outcome of comparing the results for the *AEEA* with the corresponding average measurements of the entire set of data, of the set of English-to-Norwegian data, and of the set of law data. With respect to the dividing line between the computable and the non-computable (cf. rows (i) and (ii) in table 5.11), we may observe that the degree of translational complexity in the *AEEA* is not only noticeably higher than the average of the two pairs of law text, but also higher than the average of the entire collection of string pairs. The results found for the *AEEA* are, however, fairly close to the average of the whole set of English-to-Norwegian data. With respect to the division between the computable and the non-computable, the results for the *AEEA* are strikingly similar to the average figures for the set of fiction data, which can be seen by a comparison with table 5.8 in 5.4.1. It is surprising that in the *AEEA* the proportion of text included in correspondences where source and target string are semantically equivalent (38,8%, cf. row (ii) in table 5.11) is very close to the average found across the fiction data (39,6%, see row (ii) in table 5.8).

Table 5.11. The results for the text pair *AEEA* compared with the average measurements for all data, for English-to-Norwegian, and for the law data.

Proportions of...	**in *AEEA***	**in all data**	**in E → N**	**in law text**
(i) non-computable translational correspondences (type 4)	61,2%	55,2%	61,0%	49,8%
(ii) linguistically predictable correspondences (types 1–3)	38,8%	44,8%	39,0%	50,2%
(iii) "easily" computable correspondences (types 1–2)	5,6%	10,0%	8,5%	7,5%
(iv) resource-intensive, computable correspondences (type 3)	33,2%	34,8%	30,5%	42,7%

Table 5.12 presents the distribution of correspondence types within the data recorded from the *Petro* text pair. In contrast to the *AEEA*, this is a body of parallel text following the normal pattern where the translation contains a larger number of words than the original does. Based on counting of word forms, the *Petro* target text is as

much as 21,7% longer than the source text, and among the five text pairs where the translation is longer than the original, *Petro* exhibits the greatest increase in word length.[48]

Table 5.12. The distribution of correspondence types within the data recorded from *Lov om petroleumsvirksomhet* (*Petro*), §§1–65, and its translation into English.

Petro	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	141	56	390	335	922
Percentage of string pairs	*15,3*	*6,1*	*42,3*	*36,3*	*100,0*
Length of source text	429	366	4 285	3 255	8 335
Percentage of source text	*5,1*	*4,4*	*51,4*	*39,1*	*100,0*
Length of target text	450	443	5 246	3 936	10 075
Percentage of target text	*4,5*	*4,4*	*52,1*	*39,0*	*100,0*

The probably most striking feature of the results given in table 5.12 is that, if we focus on the computability issue, *Petro* exhibits the lowest degree of translational complexity among all text pairs analysed in this study. This is also emphasised by table 13, which presents the comparison of the *Petro* results with the average measurements of, respectively, the entire set of data, the set of English-to-Norwegian data, and the set of law data. *Petro* is the text pair exhibiting the largest proportion of texts covered by computable correspondences: according to row (ii) in table 5.13, the proportion is as large as 60,9% of the analysed texts in *Petro*. This is about 10 percentage points higher than the average of law data, as well as of all Norwegian-to-English data, and it is 16,1 percentage points higher than the average of the entire data set.

[48] In comparison, the following percentages represent the length increase found in the other text pairs: AB +5,2%, DL +16,0%, EFH +9,3%, BV +13,5%, *AEEA* –12,9%. These percentages are based on the numbers given for word forms in table 4.1 in 4.2.

Table 5.13. The results for the text pair *Petro* compared with the average measurements for all data, for English-to-Norwegian, and for the law data.

Proportions of...	**in *Petro***	**in all data**	**in N → E**	**in law text**
(i) non-computable translational correspondences (type 4)	39,1%	55,2%	49,5%	49,8%
(ii) linguistically predictable correspondences (types 1–3)	60,9%	44,8%	50,5%	50,2%
(iii) "easily" computable correspondences (types 1–2)	9,2%	10,0%	11,5%	7,5%
(iv) resource-intensive, computable correspondences (type 3)	51,7%	34,8%	39,0%	42,7%

5.5.1.2 Discussion of the pairs of law texts

The following discussion will evolve mainly around two observations made on the basis of the quantitative results for the two pairs of law texts. Firstly, in the *AEEA* the proportion of non-computable, and hence semantically non-equivalent, correspondences is surprisingly large (61,2%), given the institutionalised norms of legal translation which should secure the preservation of the meaning of the original text. Secondly, in *Petro* the proportion of computable correspondences is so large (60,9%) that for this text pair machine translation might be useful, depending on the workload involved in correcting the output for the non-computable part of the translation task.[49] We will present several aspects illustrating that these two parallel texts instantiate quite diverse types of legal translation, and we will try to relate this opposition to the marked difference between the *AEEA* and *Petro* text pairs with respect to the extent to which corresponding source and target strings have been categorised as semantically equivalent.

TYPES OF LEGISLATION. Within studies of legal language it is common to distinguish between domestic and supranational law (cf. e.g. Cao 2007: 134, 2010: 80). The *Agreement on the European Economic Area* is an example of the second cate-

[49] This point will be discussed in 7.4.

gory; its legal force has authority over all the members states included in the European Economic Area (EEA).[50] The legislation regulating the European Union (EU) is another, and frequently cited, example of supranational law. *Lov om petroleumsvirksomhet* is an instance of domestic law; its legal force applies only within the jurisdiction of the Norwegian state. Within the field of domestic law, Cao (2007: 101–103, 2010: 84–86) draws the distinction between, on the one hand, monolingual jurisdictions, and, on the other hand, bi- or multilingual jurisdictions, and the reason is that the purpose behind the translation of law texts varies according to this division.

With respect to the present example of domestic law, we observed in 4.2.2 that the Norwegian original version has the authority of a law text, whereas the English version is without legal force and functions as a source of information on the Norwegian legislation regulating petroleum activities. This description matches the characterisation given by Cao (2007: 103, 2010: 85) of the translation of domestic law in monolingual jurisdictions.[51] The demand for an English version of the Norwegian act is obvious since many non-Norwegian agents are involved in petroleum activities on the Norwegian continental shelf. Although English has played a dominant role in the Norwegian oil industry, the field can be described as a multilingual setting within a monolingual jurisdiction, and its numerous non-Norwegian agents need access to information about the Norwegian legislation.

When domestic law is translated in bi- or multilingual jurisdictions, the purpose of the translation is normative, and not just informative (cf. Cao 2010: 85). In such cases the law text may be drafted in one language, and translated subsequently, or drafting and translation may take place in parallel. The outcome of this process is a bi- or multilingual body of law text where each version has the same legal force and the same authority within the jurisdiction. In legal terms, each version of the law text is regarded as equally *authentic*, and the formal process through which each version acquires legal status is referred to as *authentication* (cf. Cao 2007: 102). In principle,

[50] At present the EEA member states include the 27 members of the European Union, together with Iceland, Liechtenstein, and Norway.

[51] Strictly speaking, Norway is a bilingual nation: Norwegian and Sami are its two official languages, but until recently petroleum activities have not been carried out in the regions where Sami is spoken, so that Norwegian has in effect been the only legal language of the Norwegian oil industry. Some examples of bi- or multilingual jurisdictions are Belgium, Canada, Finland, Hong Kong, and Switzerland.

the different versions are to be drafted in parallel, or co-drafted, but, in practice, translation is most often involved in the process of creating equally authentic language versions of the law text (cf. Cao 2010: 85).

The characteristics presented here with respect to domestic law in bi- or multilingual jurisdictions are also found in supranational legislation, including the *Agreement on the European Economic Area*, and this highlights the fact that in the *AEEA* text pair both versions of the law text have the same legal status. This is a basic principle of supranational legislation, and within the EU it is referred to as the *principle of equality of authentic texts* (abbreviated as PEAT; cf. Doczekalska 2007: 60, Šarčević 2007: 36–37). PEAT is derived from the "principle of legal universality", which demands that all citizens of the EU are governed by the same laws (cf. Correia 2003: 39), and these principles apply also within the EEA. The legal effect of these principles on the different language versions of the *EEA Agreement* is that "one version cannot be regarded as a translation of another", so that "all authentic texts are originals regardless of the way of production" (Doczekalska 2007: 60).[52] Thus, whereas the *Petro* text pair is the result of a regular instance of translation, and is carried out for information purposes, the Norwegian version of the *EEA Agreement* is created for the purpose of legal enactment, and it is the product of what Kjær (2007: 87) describes as "interlingual text reproduction". These two cases of translation have been governed by quite different factors, which may to a certain extent elucidate the difference in the degree of translational complexity found between the two pairs of law texts.

Probably the most significant contrasts between the two pairs of law texts are, firstly, that while both language versions have the authority of a law text in the *AEEA*, this is true only of the source text in the case of *Petro*, and, secondly, that the *EEA Agreement* is translated for normative purposes, whereas *Lov om petroleumsvirksomhet* is translated for informative purposes. These factors may have contributed to differences in linguistic quality between these two translations. In general, the Norwegian version of the *Agreement* is a text of high linguistic quality, which is not

[52] This is an example of a *legal fiction*. Schane (2006: 7) characterises legal fictions in the following way: "Acknowledged not to be literally true, nonetheless fictions are treated as though they were."

surprising, firstly, because it has been produced by professional translators employed by the Norwegian Ministry of Foreign Affairs, and, secondly, because poor linguistic quality would not be acceptable given the high authority of the text. In contrast, there is a tendency of lower linguistic quality in the unofficial translation of *Lov om petroleumsvirksomhet*. Information about the translator(s) of this text has not been available, but occasional language errors indicate that English may not be the first language of the translator(s).[53]

EXPLICITATION. There is a striking difference between the two pairs of law texts with respect to target text length in relation to source text length. In our view, the numerical differences in word forms indicate to a certain extent that in the *Petro* text pair the translation is, as usual, more explicit than the original, whereas in the *AEEA* the translators drafting the Norwegian version have tried to avoid explicitation as far as possible.[54] That is, we assume that the legal norms controlling the creation of this text pair will work against the tendency of explicitation in translations. One typical surface indicator of explicitation, sentence splitting (cf. 5.3.2), is not present in the *AEEA* text pair, as it is excluded by the legal translation norm demanding one-to-one correspondences between original and translation at matrix sentence level (cf. 5.4.2.2). However, since this norm applies also in the case of the *Petro* text pair, in which we claim that explicitation does occur, the absence of sentence splitting is an insufficient basis for our assumption, and it is probably of greater significance that the principle of legal universality demands that all versions of the *Agreement* must express the same law. Moreover, granted the authority of the Norwegian version of the *EEA Agreement*, and the fact that it is produced by expert translators observing the legal requirements of equality between the texts, we assume the Norwegian text to be reliable in the sense of conveying the same informational content as the English version does.

---

[53] For the present project, we gained access to a version of *Lov om petroleumsvirksomhet*, with English translation, published in 1994. There is reason to believe that linguistic errors in the translation have been corrected in later versions.

[54] Even if the English and Norwegian versions of the *EEA Agreement* are, in legal terms, texts of equal authenticity, it is a fact that they were not co-drafted: the Norwegian version was created as a *de facto* translation from English, although the French version was also consulted (cf. 4.2.2.1).

Avoiding explicitation is among the challenges to be dealt with by translators and co-drafters working with supranational legislation: if a target expression is semantically more specific than the original expression, then the informational content of the law is altered. Ideally, since the writing of law texts is controlled in the source language by the demands of linguistic clarity and non-ambiguity (cf. 5.4.2.1), it should be unproblematic for a legal translator to identify the intended interpretation of a given SL expression. However, as pointed out by several, it is a characteristic of supranational laws that they are expressed through *negotiated texts*.[55] Supranational legislation is the result of political negotiations between two or more states, and, normally, compromises have to be found in order to reach agreement on the content of international legal instruments. Such compromises can result in intended imprecisions in the law texts, either in the form of ambiguity or vagueness. The principle is that the text should be as linguistically clear as possible, but its content must not be more specific than what is acceptable to each of the parties involved — it must not say more than the compromise allows. If negotiations over legal content have created intended vagueness or ambiguity in the language version(s) in which the law is originally drafted, or co-drafted, then it is a challenge when new language versions are produced to find target language expressions which have equally vague meanings, or are ambiguous in the same way. This is especially difficult in cases where semantic distinctions, either lexically or grammatically encoded, are drawn in different ways in source and target language.[56] Translators who have not participated in the negotiations over the original drafting of the supranational law may not be aware of intended imprecisions in the source version of the text, and hence their instruction is to refrain from increasing the degree of semantic specificity in new language versions, since explicating intended imprecisions will change the legal content.[57] This

[55] See e.g. Šarčević (1997: 204, 2007: 44), Cao (2007: 153, 2010: 88).
[56] Cf. the discussion of generation problems for automatic translation in 1.4.2.3.
[57] This point is made in several contributions; cf. e.g. Doczekalska (2007: 62), Correia (2003: 42), Šarčević (1997: 204, 2007: 44), Cao (2007: 153, 2010: 88).

constraint has applied to the writing of the Norwegian version of the *EEA Agreement*, and is our basis for claiming that explicitation has been avoided in this text pair.[58]

Even if the normal pattern of explicitation has not been followed, it is surprising that the Norwegian version of the *EEA Agreement* contains 1 187 fewer word forms than the English version does.[59] However, this probably does not indicate that translation has shortened the content of the original. Firstly, the applied measure is the simple kind of word counting available in word processing tools. Hence, parsing is not involved, so that lexical units encoded as multi-word expressions are not recognised as such. For this reason, the word count is not a reliable indication that, e.g., the number of lexemes has been reduced in the Norwegian text. These two parallel texts are written in specialised language with a high frequency of technical terms, which are often realised as multi-word expressions, in particular compound nouns. It is a structural difference between English and Norwegian that while compound nouns in Norwegian are written as single-word expressions, they are in English typically multi-word expressions where the components are separated by a space. Thus, the English phrase *the Contracting Parties* counts as 3 word forms, whereas the Norwegian translation *avtalepartene* is 1 word form. Within the analysed material there are 73 occurrences of this lexical correspondence. Further, there are 34 instances of *the EEA Joint Committee*, counting as 4 word forms, and corresponding with the single word form *EØS-komitéen*. Also, correspondences between non-compound nouns will, whenever definite form is expressed, involve a reduction by 1 word form in English-to-Norwegian translation since the English definite article *the* corresponds with a noun suffix in Norwegian. Even if this has not been investigated systematically within the compiled data, it appears reasonable to attribute the difference in number of word forms between the English and Norwegian versions of the *EEA Agreement* to surface phenomena of these kinds.

Moreover, the Norwegian *Lov om petroleumsvirksomhet* is not an instance of negotiated text to the extent that the *EEA Agreement* is: as a case of domestic law it

[58] This claim is also compatible with an observation presented in 6.3.1.3, in the discussion of occurrences of non-predictable specification and despecification: a large majority of the cases of specification identified in the text pair *AEEA* can be attributed to the nonfinite-finite pattern.
[59] Cf. table 4.1 in 4.2.

expresses the legislative intent of only one state, represented by its parliament. Although the possibility cannot be excluded that the text contains instances of intended imprecisions caused by compromises reached after political disagreement over the legislative content, such phenomena will be far less frequent than in supranational legislation, and we assume that the Norwegian text has been drafted in accord with the legal norms of linguistic clarity and non-ambiguity.[60] We have no basis for judging whether the work of the translator(s) has been influenced by the legal norm of avoiding a higher degree of semantic specificity in the target text, but, considering the informative purpose of the translation (cf. above), we assume it has been produced with the primary aim of rendering the original content as accurately as possible. This is compatible with the fact that the *Petro* text pair exhibits the largest proportion of semantic equivalence between source and target strings among the six text pairs studied. It is likely that the translation has been produced by one or more experts on the specific domain of this act. If the norm working against explicitation in the translation of law texts has not been operative, and the primary aim of the translation is to convey the informational content of the original, then it seems reasonable to assume that target expressions which are more precise than a given source expression will be preferred over possible target expressions which are less precise. Such choices will result in explicitation in the translation.[61]

Among the investigated text pairs, the translation of the Norwegian *Lov om petroleumsvirksomhet* shows the largest increase (+21,7%) with respect to the number of word forms in relation to its corresponding source text. However, this does not necessarily indicate that explicitation occurs to a larger extent in *Petro* than in any of the other text pairs. Given that the primary aim of the translation has been to render the original content as accurately as possible, it is plausible that there is a more modest amount of explicitation in *Petro* than in the four pairs of fiction texts, and it is likely that a substantial part of the increase in word forms is caused by the same kinds

[60] In relation to EU legislation, Šarčević (2007:44) observes that "Community law is negotiated to a much greater degree than national law".

[61] Relevant in this connection is the phenomenon of specification. In 6.3.1.3 we present the frequency of identified cases of non-predictable specification, and it follows from the discussion there that if we disregard the instances which are caused by the nonfinite-finite pattern, then specification is about twice as frequent in the text pair *Petro* than in the *AEEA*.

of factors which may explain why the English version of the *EEA Agreement* contains a greater number of word forms than the Norwegian version does.[62]

SEMANTIC EQUIVALENCE. Thus, various factors indicate an important difference between the *AEEA* and *Petro* text pairs with respect to the extent to which the norms of legal translation have influenced the target text production. Since the Norwegian version of the *EEA Agreement* appears to have been shaped by the aims of maintaining linguistic clarity and avoiding semantic specification in a stronger degree than the English translation of *Lov om petroleumsvirksomhet*, it is surprising that the extent to which semantic equivalence has been identified between translationally corresponding units is larger in the *Petro* text pair than in the *AEEA*. That is, on the one hand, it appears reasonable that the informative purpose behind the translation of *Lov om petroleumsvirksomhet* results in a relatively high proportion (60,9%) of semantic equivalence within the compiled data.[63] But, on the other hand, it is contrary to our expectations to find a considerably lower proportion (38,8%) within the *AEEA* data, because it seems plausible that the strict norms applying to the production of the Norwegian version of the *EEA Agreement*, in combination with the principles of equal authenticity and legal universality, would result in a high degree of semantic equivalence between the translationally parallel texts.

In relation to the large proportion of semantically non-equivalent type 4 correspondences (61,2%) within the *AEEA* data, it is relevant that, depending on contexts, translationally parallel pieces of text may contain the same informational content even if certain string pairs included in them are categorised as type 4. It is a central principle of our analytical framework that if two translationally corresponding expressions are semantically equivalent, then the same informational content is *linguistically* encoded in both of them (cf. 3.3.4.1). However, when interpreted relative to the contexts, linguistic as well as extra-linguistic, in which they are situated, two trans-

[62] Firstly, in translational correspondences between definite nouns, the word count is increased by 1 in the English text due to the definite article *the*. Secondly, in *Petro* there is a very high frequency of correspondences between technical terms realised as single word forms in Norwegian and as multi-word expressions in English. E.g., only within the first period of Section 1 there are five examples of this: *petroleumsvirksomhet – petroleum activities*, *leteboring – exploration drilling*, *rørledningstransport – pipeline transportation*, *sjøterritorium – sea territory*, and *kontinentalsokkel – continental shelf*.

[63] It is noteworthy that certain linguistic flaws do not prevent the translator from attaining the goal of preserving the information contained in the source text.

lational units which are semantically *non*-equivalent may be perceived as carrying the same message if the recipients of the parallel texts are able to understand the expressions in a uniform way by accessing information sources other than the content which is linguistically encoded in the strings. As explained in 3.3.5.2, this additional information may involve linguistically encoded information in the contexts of the two strings, or various kinds of extra-linguistic information. Thus, by merging linguistic and extra-linguistic information, the readers of two different language versions of a law text can receive the same message, even if type 4 correspondences have been identified among string pairs compiled from the translationally parallel versions. Due to this, we will not regard the high degree of semantic non-equivalence identified among the *AEEA* data as a symptom that the translation is inaccurate, i.e. that it fails to express the same informational content as the source text does. That would simply not be allowed according to the principle of legal universality.

In order to explain why there is a noticeably larger proportion of type 4 correspondences within the *AEEA* data than within those compiled from *Petro*, it is hard to discover facts without first-hand knowledge of the work of the translators, but we will indicate three different factors that may have contributed to this result.

Firstly, as indicated above, it is likely that the extent to which legislative negotiations have preceded the drafting of the original texts is greater in the case of the *EEA Agreement* than in the case of *Lov om petroleumsvirksomhet*. If this is correct, there may be more elements of intended imprecisions in the English agreement text than in the Norwegian act dealing with petroleum activities, and textual imprecisions may in turn cause semantic non-equivalence between translationally corresponding expressions if the target text is not likewise imprecise. Such translational challenges could yield cases where a target expression is in some way denotationally non-equivalent with its source text correspondent, and in the present study we have indeed found a higher frequency of such deviations within the *AEEA* data than within the *Petro* data (cf. table 6.17 in 6.3.2.1).

Secondly, it may be relevant that there is a stronger element of domain-uniqueness in *Lov om petroleumsvirksomhet* than in the *EEA Agreement*. In this respect we will focus on terminology since technical fields and their respective terminologies

seem to be mutually dependent on each other (cf. Laurén et al. 1997: 14), and since the presence of technical terms is regarded as the most basic characteristic of language for special purposes (cf. Laurén 1993: 10). According to Mattila (2006: 5), "legal language can be divided into sub-genres on the basis of branches of law", and "[t]he main distinguishing criterion then becomes the specialist terminology of each branch." He also observes that whereas "a large part of the legal terminology of the various branches of law is universal" (2006: 5), other terms will occur only in texts belonging to a particular legal sub-field.[64] Thus, both of the investigated pairs of law texts contain terms belonging to the general domain of law. The *EEA Agreement* deals with a range of technical areas, among which the following can be mentioned as an illustration of its diversity: trade, customs, workers' rights, consumer protection, social policy, the environment, state aid, and policy cooperation in economic and monetary matters. *Lov om petroleumsvirksomhet* regulates the exploration, production, utilisation, and pipeline transportation of petroleum in Norwegian waters. Although these are different topics, they may be said to belong to a single field, that of petroleum activities. Thus, in addition to the general legal terms, the *AEEA* text pair contains terms from a large set of technical areas, whereas the *Petro* text pair contains terms belonging to a single field.

In this connection it is relevant that term frequency varies between different technolects; cf. Laurén (1993: 74, 99–101).[65] Since technical terms are semantically precise lexical units, we may expect a high degree of semantic equivalence between translationally parallel texts with a large element of technical terms. If the domain variation between the *AEEA* and *Petro* text pairs could be correlated with a higher frequency of technical terms (relative to the number of lexical units) in *Lov om petroleumsvirksomhet* than in the *EEA Agreement*, then that might account for the larger proportion of semantic equivalence among the *Petro* data. A systematic comparison of the two text pairs with respect to such term ratio unfortunately falls outside the scope of the present project. Considering that Laurén (1993: 74, 99–101)

[64] An example provided by Mattila (2006: 5) is that terms specific to criminal law "are almost never used in texts on the law of property or constitutional law."
[65] A technolect is understood as the specialist language associated with a technical domain.

reports that within a variety of technolects the term frequency of legal language is relatively high, it is perhaps not likely that the *EEA Agreement* and *Lov om petroleumsvirksomhet* will show any marked difference in this respect, since both are law texts. However, to a reader the former text may appear as less specialised than the latter, since the *EEA Agreement* covers a wide range of subject matters, while *Lov om petroleumsvirksomhet* deals with a considerably more restricted domain. Possibly, this has contributed to a higher level of semantic precision in the Norwegian act, and it is compatible with the lower frequency of denotational non-equivalence between corresponding strings in *Petro*, as noted above.

Thirdly, that the amount of semantically equivalent string pairs is greater among the *Petro* data than among the *AEEA* data may be related to the fact that whereas the *EEA Agreement* involves many independent nations, each with its own legal tradition, *Lov om petroleumsvirksomhet* belongs to the jurisdiction of one state only. As explained by Mattila (2006: 105), the legal concepts of a society are shaped by its history and culture, so that the concepts of different legal orders may not correspond with each other. This may create problems for the translation of legal texts, and it can be a challenge especially to find adequate target language correspondents for legal terms. With respect to legal traditions, it is normal to distinguish between common law (also referred to as *case law*) and civil law. According to Mattila (2006: 106), the common law tradition generally belongs to the various English-speaking countries, whereas civil law systems are found in continental Europe, in Latin America, and in the Nordic countries. He describes the supranational legislation of Europe as a "hybrid" legal system, in which the traditions of common and civil law have been mixed.[66] In addition, Mattila (2006: 106) points out that due to "interaction between Community organs and national legal orders", EU law has developed its own characteristics, and represents, in his view, a new kind of legal system, different from common law as well as from civil law. This is the legal tradition within which the *EEA Agreement* has been created.

[66] On this point Mattila (2006: 106) cites de Cruz (1995: 158–163, 180). The hybrid character of EU legislation is also commented on by Šarčević (2007: 44).

A symptom of the individual character of the EU legal system is that there are legal concepts which are original to EU legislation, and, according to Mattila (2006: 118), attempts are made to find new terms for expressing such concepts, of which the term *principle of subsidiarity* is an example. In a discussion of guidelines for the creation of mulitilingual EU laws, Šarčević (2007: 49) observes that when new legal terms are needed, drafters and translators are instructed to avoid expressions which already exist within national legal systems, so that new terms are, as far as possible, neutral to the legislation of each member state. Another drafting principle, which has become influential in recent years, is that when a new language version of an EU law text is drafted, its linguistic form should be as true to the target language as possible, since fidelity to the original version may create the unwanted result of a law text showing a foreign appearance to the target language recipients (cf. Šarčević 2007: 50). Balancing these two principles demands highly skilled translators.

At the time of the creation of the Norwegian version of the *EEA Agreement*, the principle of faithfulness to the target language was not yet formally in operation, but the mentioned guidelines are relevant to the *AEEA* text pair since they reflect challenges that most likely had to be handled by the Norwegian translators of the *EEA Agreement*, and they throw more light on differences between the translation tasks which lie behind the *AEEA* and *Petro* text pairs, respectively. The basis for drafting the Norwegian version of the *EEA Agreement* was a law text created within the hybrid, supranational legal system of Europe, and the content of this agreement was to be expressed in a legal language developed within the civil law system of Norway. Whether this involved solving problems of conceptual mismatches, we are unable to tell without first-hand knowledge of the writing of the Norwegian version, but it appears likely that this factor may have influenced the proportion of semantic non-equivalence within the data compiled from the *AEEA*. This may be in agreement with a higher frequency of denotational non-equivalence found in the *AEEA* than in *Petro*; at least, it appears even more reasonable that there is a larger proportion of semantically equivalent string pairs within the *Petro* data than in the other pair of law texts. Since *Lov om petroleumsvirksomhet* deals with internal Norwegian matters, and the recipients of the English target text will, broadly, be agents operating within the

jurisdiction of Norway, we regard this as a case where a translator should primarily aim at fidelity to the source text, especially given the informative purpose of the translation. Further, there is no need to shape the target text in accord with concepts or textual norms specific to the legal systems of any English-speaking countries. Due to faithfulness to the content of the Norwegian law text, it is not surprising, within the present study, that the *Petro* text pair shows the largest proportion of linguistically predictable translational correspondences.[67]

5.5.2 The pairs of fiction texts

The data recorded from fiction comprise correspondences compiled from four different text pairs, two pairs for each direction of translation; cf. table 4.1 in 4.2. All four text pairs exhibit the general characteristic of translation where target texts prove to be longer than the corresponding source texts. The results of the analysis show quite different complexity measurements for the fiction pairs translated from English into Norwegian, whereas the fiction pairs representing the opposite direction of translation exhibit fairly similar measurements. Hence, the presentations of results in 5.5.2.1 will be organised according to the dimension of direction of translation.

5.5.2.1 Complexity measurements for the fiction texts

ENGLISH-TO-NORWEGIAN. Table 5.14 presents the distribution of correspondence types within the data recorded from the text pair AB, and table 5.15 gives the outcome of comparing the results for this text pair with average measurements for the entire data set, for the English-to-Norwegian data, and for the fiction data, respectively.

The most striking aspect revealed by the data recorded from AB is that among the pairs representing fiction this is the one that exhibits the largest proportion of computable correspondences (56,1% of the texts). Also, in this regard AB clearly stands out from the other three pairs of fiction text, since they all exhibit considerably small-

[67] Cf. the discussion of faithfulness to the original in 5.5.2.2.

er proportions of computable correspondences; the figures are, respectively, 23,6% (DL), 40,7% (EFH), and 39,0% (BV); cf. tables 5.17, 5.19, and 5.21.

Table 5.14. The distribution of correspondence types within the data recorded from an extract of 4000 words of *The Wall of the Plague*, by André Brink (AB), and its translation into Norwegian.

AB	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	70	50	193	208	521
Percentage of string pairs	*13,4*	*9,6*	*37,1*	*39,9*	*100,0*
Length of source text	282	363	1 758	1 846	4 249
Percentage of source text	*6,6*	*8,5*	*41,4*	*43,5*	*100,0*
Length of target text	286	350	1 857	1 972	4 465
Percentage of target text	*6,4*	*7,8*	*41,6*	*44,2*	*100,0*

Table 5.15. The results for the text pair AB compared with the average measurements for all data, for English-to-Norwegian, and for the fiction data.

Proportions of...	**in AB**	**in all data**	**in E → N**	**in fiction**
(i) non-computable translational correspondences (type 4)	43,9%	55,2%	61,0%	60,4%
(ii) linguistically predictable correspondences (types 1–3)	56,1%	44,8%	39,0%	39,6%
(iii) "easily" computable correspondences (types 1–2)	14,6%	10,0%	8,5%	12,6%
(iv) resource-intensive, computable correspondences (type 3)	41,5%	34,8%	30,5%	27,0%

Furthermore, table 5.15 shows that with respect to the proportions of the least complex correspondence types, 1 and 2, the results for AB are fairly similar to the average measurements for the fiction data. This emphasises that it is particularly the division between computable and non-computable correspondences that distinguishes this text pair from the other three fiction pairs.

Table 5.16 presents the distribution of correspondence types within the data recorded from the text pair DL, and table 5.17 displays the results for this text pair compared with average measurements for the entire data set, for the English-to-Norwegian data, and for the fiction data, respectively.

Table 5.16. The distribution of correspondence types within the data recorded from an extract of 4000 words of *The Good Terrorist*, by Doris Lessing (DL), and its translation into Norwegian.

DL	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	70	31	164	527	792
Percentage of string pairs	*8,9*	*3,9*	*20,7*	*66,5*	*100,0*
Length of source text	223	151	709	3 345	4 428
Percentage of source text	*5,1*	*3,4*	*16,0*	*75,5*	*100,0*
Length of target text	228	149	774	3 932	5 083
Percentage of target text	*4,5*	*2,9*	*15,2*	*77,4*	*100,0*

Table 5.17. The results for the text pair DL compared with the average measurements for all data, for English-to-Norwegian, and for the fiction data.

Proportions of...	**in DL**	**in all data**	**in E → N**	**in fiction**
(i) non-computable translational correspondences (type 4)	76,4%	55,2%	61,0%	60,4%
(ii) linguistically predictable correspondences (types 1–3)	23,6%	44,8%	39,0%	39,6%
(iii) "easily" computable correspondences (types 1–2)	8,0%	10,0%	8,5%	12,6%
(iv) resource-intensive, computable correspondences (type 3)	15,6%	34,8%	30,5%	27,0%

The most noticeable aspect of the results found for DL is that in comparison to all the other analysed texts, regardless of text type, this text pair exhibits the smallest proportion of computable correspondences, merely 23,6% of the analysed texts. This

is 21,2 percentage points lower than the global average, and 16,0 percentage points lower than the average within the subset of fiction data. Furthermore, within the group of fiction texts, DL is a case where type 4 correspondences constitute a large majority (76,4%; cf. table 5.17) and all the other correspondence types are less frequent than in any of the other three pairs of fiction texts. It is interesting that the proportions of the two least complex correspondence types are quite similar to the results found within the law data: on average, types 1 and 2 together cover 7,5% of the analysed law texts (cf. e.g. table 5.13 in 5.5.1.1), whereas the corresponding result for the text pair DL is 8,0% (cf. table 5.17). This will be commented on in 5.5.2.2.

NORWEGIAN-TO-ENGLISH. For the fiction text pairs EFH and BV the results are so similar that they can be presented together. Tables 5.18 and 5.20 give the distribution of correspondence types within the data recorded, respectively, from EFH and BV, and tables 5.19 and 5.21 present the results for the text pairs compared with average measurements for all data, for all Norwegian-to-English data, and for all fiction data.

Table 5.18. The distribution of correspondence types within the data recorded from an extract of 4000 words of *Salme ved reisens slutt*, by Erik Fosnes Hansen (EFH), and its translation into English.

EFH	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	79	64	195	365	703
Percentage of string pairs	*11,2*	*9,1*	*27,8*	*51,9*	*100,0*
Length of source text	285	326	1 145	2 581	4 337
Percentage of source text	*6,6*	*7,5*	*26,4*	*59,5*	*100,0*
Length of target text	280	358	1 276	2 753	4 667
Percentage of target text	*6,0*	*7,7*	*27,3*	*59,0*	*100,0*

Table 5.19. The results for the text pair EFH compared with the average measurements for all data, for Norwegian-to-English data, and for the fiction data.

Proportions of...	**in EFH**	**in all data**	**in N → E**	**in fiction**
(i) non-computable translational correspondences (type 4)	59,3%	55,2%	49,5%	60,4%
(ii) linguistically predictable correspondences (types 1–3)	40,7%	44,8%	50,5%	39,6%
(iii) "easily" computable correspondences (types 1–2)	13,9%	10,0%	11,5%	12,6%
(iv) resource-intensive, computable correspondences (type 3)	26,8%	34,8%	39,0%	27,0%

The most striking aspect of the results presented for EFH and BV is that the complexity measurements for both text pairs are very close to the average measurements for the whole set of analysed fiction texts. In particular, tables 5.19 and 5.21 show that with respect to the division between computable and non-computable correspondences, the results for EFH and BV are almost exact matches of the average measurements across the fiction data (cf. rows (i) and (ii) in the tables).

Table 5.20. The distribution of correspondence types within the data recorded from an extract of 4000 words of *En håndfull lengsel*, by Bjørg Vik (BV), and its translation into English.

BV	**Type 1**	**Type 2**	**Type 3**	**Type 4**	**All types**
Number of string pairs	78	56	197	379	710
Percentage of string pairs	*11,0*	*7,9*	*27,7*	*53,4*	*100,0*
Length of source text	303	317	1 112	2 592	4 324
Percentage of source text	*7,0*	*7,3*	*25,7*	*60,0*	*100,0*
Length of target text	305	343	1 194	2 993	4 835
Percentage of target text	*6,3*	*7,1*	*24,7*	*61,9*	*100,0*

Table 5.21. The results for the text pair BV compared with the average measurements for all data, for Norwegian-to-English data, and for the fiction data.

Proportions of...	**in BV**	**in all data**	**in N → E**	**in fiction**
(i) non-computable translational correspondences (type 4)	61,0%	55,2%	49,5%	60,4%
(ii) linguistically predictable correspondences (types 1–3)	39,0%	44,8%	50,5%	39,6%
(iii) "easily" computable correspondences (types 1–2)	13,8%	10,0%	11,5%	12,6%
(iv) resource-intensive, computable correspondences (type 3)	25,2%	34,8%	39,0%	27,0%

5.5.2.2 Discussion of the pairs of fiction texts

As previously explained, we anticipated a relatively high degree of translational complexity in the pairs of fiction texts (cf. 5.4.2). In general, we expect that the output produced by applying automatic translation to fiction would be of such a low linguistic quality that the workload involved in post-editing might approximate the effort of fully manual translation. It could even make things worse, if errors in the machine output would disturb the translator's attention towards the original text. This would probably be the case for three of the fiction text pairs (DL, EFH, and BV), where less than half of the analysed texts are covered by computable correspondences. Then, in the case of AB, which proves to have a lower degree of complexity, as much as 56,1% of the analysed texts represent computable translation tasks, and the possible usefulness of automatic translation would depend on how much the machine output deviates from human-quality translation. Probably in this case, too, post-editing would require a considerable workload.[68]

The complexity measurements for the individual pairs of fiction texts primarily indicate differences on the part of the translators concerning the extent to which the informational content of the target text matches that of the original. This may reflect

[68] We return to this topic in 7.4.

variation along a continuum between two approaches to literary translation: at one end of the continuum the translator is mostly oriented towards the SL author and the source text, at the other end mostly towards the TL readers (cf. e.g. Landers 2001: 50–51). In our view, the complexity measurements primarily reflect differences between the individual translators in terms of the degree of faithfulness to the original text. Such diversity is reasonable within the group of analysed fiction texts, given that they are examples of an unrestricted text type (cf. 5.4.2.1), and it is also to be expected that this kind of variation will influence the degree of translational complexity in different bodies of parallel texts.

By *faithfulness* to the original we refer to the classic translatological distinction between *free* and *literal* translation, described by Palumbo (2009: 49) as "the binary opposition that has dominated the debate on translation over the centuries." He defines the two latter notions as follows (2009: 49): "Free translation is usually taken to concentrate on conveying the meaning of the ST disregarding the formal or structural aspects of the ST. Literal translation is normally taken to be a mode of translation that remains close to the form of the original."[69] Referring to Robinson (1991, 1998), Palumbo (2009: 49–50) points out that it is problematic to define the concept of 'free translation', since it has been understood in various ways, "depending on the exact nature of the type (or types) of translation it is opposed to." On the basis of (Robinson 1998: 88–89), Palumbo (2009: 50) concludes that the probably most useful way of defining 'free translation' is "to see it as translation that deviates from the 'hegemonic norms' that establish, in a given period or community, what faithful translation is."

Likewise, we adopt a relativised understanding of 'free translation', although we do not want to describe it in terms of translation norms. In the first case, we prefer to avoid speaking of an opposition between *free* and *literal* translation, since we apply a special definition of 'literal translation', defined in 2.3.2. In our view, this is a distinction between *free* and *faithful* translation, and we regard it not simply as a dichotomy, but as a scale ranging from free translations where the informational

[69] *ST* stands for *source text*.

content of the target text deviates considerably from that of the source text, to translations which are faithful, or *true*, to the original in the sense of preserving, as far as possible, its informational content. Thus, our understanding of faithfulness to the original does not pay attention to structural or formal correspondences between source and target text; it is a semantic notion in the same way as 'informational content' was defined as a semantic notion in 2.4.1.2.

The following discussion will comment on how the complexity measurements for the pairs of fiction texts reflect different degrees of faithfulness to the original. In the case of the two pairs of law texts we assumed that the strict norms of legal translation ensured faithfulness, even in the *AEEA*, where a relatively large proportion of the texts are included in semantically non-equivalent correspondences (cf. 5.5.1.2). We made use of background information on the types of legislation represented by the two pairs of law texts in order to account for the difference in translational complexity observed between them. With respect to the analysed fiction texts, far less information is available on aspects that may have influenced the creation of the source texts, and the choices made by the translators.[70] Thus, on the basis of the complexity measurements presented in 5.5.2.1, it is difficult to explain the variation in faithfulness among the pairs of fiction texts. Subsequent to comments on the individual text pairs, we will try to relate the results for the fiction texts as a group to a distinction between dominating and dominated languages, to be explained below. Then, additional information on individual text pairs will be provided in chapter 6 by the presentations of certain recurrent semantic phenomena within the recorded data.

As already noted, with its relatively modest degree of translational complexity, the text pair AB is a special case, because the extent to which there is semantic equivalence between source and target strings (56,1%) is distinctly larger than in the other pairs of fiction texts.[71] Interestingly, on this point it is the pair of law texts *Petro* which shows a result most similar to that of AB. Moreover, while there are certain linguistic flaws in the target text of the *Petro* text pair (cf. 5.5.1.2), we find no reason

[70] Relevant factors could be, e.g., the author's intention of writing the story, the translator's previous experience with translation, or experience with the cultures of the source and target language communities, respectively.

[71] As shown in table 5.15 in 5.5.2.1, correspondence types 1–3 cover 56,1% of the AB texts.

to criticise the quality of any of the translations of fiction, and this shows that the level of linguistic perfection in the translation can be independent of the proportion of semantically non-equivalent correspondences in a given text pair. That is, the relatively large elements of semantic deviation found in the fiction pairs DL, EFH, and BV, are certainly compatible with good quality in the translation.

In our view, the complexity measurement for the text pair AB primarily shows that in this case the translator seems to be more faithful to the original than the translators of the three other fiction texts.[72] This also illustrates that in order to create idiomatic target expressions, it is not always required to avoid literal translations, since occurrences of idiomatic language use will naturally be included in the domain of translational correspondences delimited by the linguistically predictable interrelations between SL and TL. Thus, relatively large proportions of literal translation, as in AB and *Petro*, do not necessarily lead to a target text with linguistic imperfections, as in the case of *Petro*.

The high degree of translational complexity in the text pair DL indicates that in this case the translator has chosen semantically equivalent target expressions in a relatively low degree. More information on this will be provided in chapter 6, where the discussions of subcategories within the main correspondence types 3 and 4 will show, e.g., that two of the most frequent types of semantic deviations between translationally corresponding units, specification and denotational non-equivalence, are noticeably more frequent in DL than in the other pairs of fiction texts (cf. 6.3.1.3 and 6.3.2.2).

Concerning the two least complex correspondence types, there are, as observed in 5.5.2.1, strikingly small proportions of these in DL, i.e. proportions similar to those found across the data recorded from the pairs of law texts. Previously, in 5.4.2.7, we have attributed the low frequency of types 1 and 2 in the law data to two factors. Firstly, extracted translational units are, on average, longer in the pairs of law texts than in those of fiction. Secondly, due to the difference between the two text types concerning the degree of restrictedness, the law texts contain a more limited set of

[72] This view is supported by observations made in 6.3.1.3 regarding the frequency of specification and despecification in the text pair AB.

syntactic constructions than the fiction texts do, and types 1 and 2 are infrequent within the inventory of relatively complex constructions found in the law texts. These factors are influential because correspondence types 1 and 2 tend to occur in short and syntactically simple string pairs (cf. 5.2.1). On this background the small proportions of types 1 and 2 in the fiction text pair DL are surprising, because the average string lengths in DL are the shortest among all investigated text pairs; cf. table 5.22.

Table 5.22. Average string lengths, given for the entire data set, for each direction of translation, for each text type, and for individual text pairs.[73]

	Average source string length	Average target string length	Source-to-target length difference
Across **all** string pairs	7,9	8,4	+6,3%
Across string pairs **E → N**	8,6	8,4	–2,3%
Across string pairs **N → E**	7,3	8,4	+15,1%
Across **legal** string pairs	10,3	10,6	+2,9%
Across **fiction** string pairs	6,4	7,0	+9,4%
Across string pairs in ***AEEA***	11,8	10,2	–13,6%
Across string pairs in ***Petro***	9,0	10,9	+21,1%
Across string pairs in **AB**	8,2	8,6	+4,9%
Across string pairs in **DL**	5,6	6,4	+14,3%
Across string pairs in **EFH**	6,2	6,6	+6,5%
Across string pairs in **BV**	6,1	6,8	+11,5%

Since the writing of fiction is not norm-governed to the extent that law writing is, textual constraints cannot account for the low proportion of the least complex corre-

[73] String length is here measured as explained in 5.2.1. Otherwise, cf. the presentation of table 5.5 in 5.3.2.

spondence types in the text pair DL.[74] We regard this, like the high frequency of semantically non-equivalent correspondences, as an indication that the translation is less faithful to the original than in the other pairs of fiction texts. The considerable increase in average string length from source to target text in DL is also compatible with this: the relatively high frequency of semantic specification in DL (noted above) indicates that elements of explicitation are involved in the fairly free translation.

Concerning the complexity measurements for the two text pairs EFH and BV, it is, as pointed out in 5.5.2.1, striking that they are so similar, even in relation to all four correspondence types (cf. tables 5.18 and 5.20 in 5.5.2.1). Also, as shown above in table 5.22, these two text pairs are quite similar with respect to the average lengths of source and target strings. These facts make it tempting to ask whether the results for EFH and BV are indicative of how the correspondence types would in general be distributed across a representative parallel corpus for the translation of fiction from Norwegian into English. In both text pairs originals as well as translations are texts produced by different persons, but it is impossible to generalise on the basis of only two text pairs. In comparison to the other pairs of fiction texts, we can say that in the cases of EFH and BV, the translations show moderate freedom in relation to the originals.

In our view, perhaps the most interesting result concerning the entire set of fiction data is that the complexity measurements show considerable differences between the English-to-Norwegian text pairs, whereas those representing Norwegian-to-English exhibit a very high degree of convergence. Since only two text pairs have been analysed for each direction of translation, these findings may be fully accidental.

Still, it is tempting to relate these results to the distinction between *dominating* and *dominated* languages, explicated by Casanova (2010) in a description of dominance relations across the literary field. She presents an approach to literary translation where translation practice is placed "in the universe of international literary exchanges" (2010: 287). The world literary field is seen as organised in "literary and

[74] Since DL is the only fiction pair exhibiting such small proportions of types 1 and 2, the average of the other three fiction pairs is probably closer to the proportions of types 1 and 2 which would be found in a representative English-Norwegian parallel corpus.

linguistic inequalities and hierarchies", and hence the translation of literature is an exchange between languages that may have unequal status (2010: 288). Under this view, Casanova (2010: 288) describes translation as "one of the specific forms that the relationship of domination assumes in the literary field." Then, she argues that when different languages have unequal status, they are unequal in terms of the volume of their *linguistic capital*, and, likewise, the amount of prestige associated with individual works of literature determines the volume of their *literary capital* (2010: 288–289). On this basis Casanova (2010: 289) observes that "[t]he unequal distribution of this capital organizes the linguistic-literary field according to the opposition between dominated literary languages and dominating literary languages. Dominated languages have been recently nationalized (that is, have become national languages relatively recently), are relatively deprived of literary capital, have little international recognition, a small number of national or international translators, or are little known and have remained invisible for a long time in the great literary centres. Dominating languages are endowed with a relatively large volume of literary capital due to their specific prestige, their age, and the number of texts which are considered universal and which are written in these languages."

Thus, English is a clear example of a dominating literary language. Casanova (2010: 289–290) further divides the class of dominated languages into four groups, and one of these comprises "languages of ancient cultures and traditions used in 'small' countries." Such languages "have quite an important history and prestige, but few speakers; they are used by few polyglots, and are little recognized outside national borders, that is, they are accorded little value in the world literary market." Norwegian is a fairly clear example of this type of dominated language.

According to Casanova (2010: 290) translation between languages of unequal status involves a "power struggle", and the nature of this struggle depends on whether the source and target languages are, respectively, dominating or dominated. She further claims that in analyses of literary translation it is also necessary to consider, firstly, the status of the source text author within his or her national literary field, and, secondly, the position of the national field within the entire world of literature (2010: 290).

In relation to the language pair of the present study, the relevant scenarios are translation from a dominating language into a dominated language, instantiated by the text pairs AB and DL, and translation from a dominated language into a dominating language, represented by the text pairs EFH and BV. All source text writers are esteemed authors within their respective national fields, i.e. South African, British, and Norwegian literature. However, since both South African and British literature is produced in a language which is one of the world's most important (within, as well as outside, the literary domain), the position of those national fields clearly outranks the position of Norwegian literature within the international literary field. Hence, in line with the view of Casanova (2010), the prestige of André Brink and Doris Lessing surpasses that of Erik Fosnes Hansen and Bjørg Vik within the international field of literature.

Then, translating the texts by Brink and Lessing into the dominated literary language of Norwegian can be described, following Casanova (2010: 291), as a "diversion" of literary capital. That is, it increases the literary capital of the dominated language by making available works created in a dominating language. Concerning the opposite direction of translation, several effects can be seen of translating the texts by Fosnes Hansen and Vik into the dominating literary language of English. Firstly, it strengthens the prestige of these authors: according to Casanova (2010: 294–295) it *consecrates* the Norwegian authors in the sense of increasing their recognition within the world literary field. Secondly, it enhances the autonomy of the international literary field by adding to its literary capital (cf. Casanova 2010: 295). Thirdly, on the translated texts there is an effect of annexation by the target language. In this respect, Casanova (2010: 301) regards translation from a dominated language into a dominating one as "a kind of universalization", where the translators are described as "[m]ediators from the centre [who] reduce foreign literary works to their own categories of perception, which are set up as universal norms." She further claims that those categories function as translation norms which diminish the richness of the original texts (2010: 301–302).

If this line of argument is correct, then it would be reasonable to find differences between, respectively, the products of translation from a dominating language into a

dominated one, and the products of translation in the opposite direction for the same pair of languages. Apart from the point about reduced richness, Casanova (2010) unfortunately does not say much about the effects of the annexation, or universalisation, caused by translation from a dominated language into a more central one. It is then natural to ask: what kinds of consequences of this universalisation can be detected by studying target texts in comparison to their originals? And further, are there observable effects of the converse dominance relation found in the opposite direction of translation?

In our view, it is possible that the universalisation involved in translation from Norwegian into English could cause a relatively larger degree of convergence among individual translators concerning the extent to which the informational content of the original is preserved in the target text. Notably, this is *not* to expect that within this direction translators tend to be relatively faithful to the original. On the contrary, since Norwegian has the relatively low status of a "small" language, deviations from the original are likely to occur if the translators do not have first language competence in Norwegian. Rather, this is an expectation that target texts will exhibit some freedom in relation to their originals, but within limits. This fits the moderate level of faithfulness to the source texts in EFH and BV, as well as the high degree of convergence found between the measurements of translational complexity in these two text pairs. Moreover, such convergence is compatible with Casanova's (2010: 301) claim that in translation from a dominated language into a dominating one the richness of the source text is reduced through the operation of the norms of the target literary field.

Concerning the opposite direction of translation, it is possible that the diversion of literary capital caused by translation from English into Norwegian could be correlated with more variation between the level of faithfulness in different translations. This would be compatible with the differences in complexity found between the two text pairs AB and DL, which indicate that whereas the translator of Brink's novel has been fairly true to the original, the translator of Lessing's text has created a relatively free target text. Given the dominance relations observed between these two languages, it is not surprising that the informational content of the original can be pre-

served to a large extent in translation from English into Norwegian, as shown by the text pair AB: when the source language is dominating in relation to the target language, the translator is most likely highly competent in the SL. Moreover, the large element of semantic non-equivalence found in the text pair DL could be compatible with a lack of the norms of universalisation which apply, according to Casanova (2010), in Norwegian-to-English, but not in the opposite direction.

Since the present investigation covers only two text pairs for each direction of translation, it is impossible to generalise. But the distinction between dominating and dominated languages is clearly relevant to the language pair English-Norwegian, and would be interesting to study in a larger collection of parallel literary texts.

5.6 Summary

In the present chapter the discussion has been focussed, not on individual translational correspondences, but on the pairs of texts that we have analysed. The distribution of correspondence types within the string pairs extracted from a body of parallel texts provides a measurement of the degree of translational complexity in that text pair.

The scope of our empirical investigation is limited: only six text pairs have been investigated. Hence, the results are no more than complexity measurements of the analysed texts, and they cannot be seen as reflecting the general complexity of translation between English and Norwegian, nor as measurements of the general complexity of translating the two chosen text types for this language pair.

The complexity measurement across the entire collection of data is discussed in 5.2 with subsections. Calculated in terms of string lengths, the results show that as little as 44,8% of all recorded string pairs are classified as computable translational correspondences, i.e. as type 1, 2, or 3. Thus, non-computable string pairs of type 4 constitute a majority (55,2%) of the compiled data. Types 1 and 2 together cover 10,0% of the analysed texts, which means that in only 10,0% of the compiled data we assume that the target string can be generated automatically using a modest amount of processing effort. Finally, in 34,8% of the data (type 3) we assume that the translation task is computable, but resource-intensive. Thus, following the assumptions of

our analytical framework, human translation can be simulated by fully automatic translation in only 44,8% of the analysed texts.

Complexity measurements relative to each direction of translation are discussed in 5.3 with subsections. The results show that within the analysed text pairs, there is a higher level of complexity within the English-to-Norwegian data than within those recorded from the opposite direction of translation: within the former the proportion of computable correspondences is 39,0%, and within the latter it is 50,5%. This difference is largely due to a relatively low degree of complexity in the pair of law texts translated from Norwegian into English (*Petro*), which provides about 50% of the data for that direction of translation. Moreover, the phenomenon of non-predictable specification, has been found to be more frequent in the English-to-Norwegian data than in the string pairs extracted from the opposite direction (cf. 6.3.1.3).

5.4 with subsections deals with complexity measurements for each of the investigated text types. The analysis has detected a lower average degree of complexity across the pairs of law text than across the pairs of fiction texts. The proportion of computable correspondences is 50,2% in the law data, and 39,6% in fiction. Again it is the relatively low degree of complexity of the *Petro* text pair which has given this result, since the complexity of the other law text pair (the *AEEA*) is quite similar to the average measured across the fiction text pairs.

A range of text-typological differences between law and fiction can be explained by viewing law texts as a highly restricted text type, and fiction as a relatively unrestricted one. Whereas the writing, the interpretation, and the translation of law texts are restricted by institutionalised norms belonging to the legal domain, fiction texts are in no way as norm-governed as the former. A fiction text will to some extent be constrained by linguistic and stylistic norms, but its creation is determined by the individual choices of the author, which may include norm violations, and its reception is determined by the subjective experiences of the readers. The difference in restrictedness between the two text types is evident in several ways. Law texts have a rigid macrostructure; fiction does not. Whereas law texts exhibit limited inventories of, respectively, pragmatic functions and types of syntactic constructions, fiction texts are far more varied in these respects. Moreover, our analysis has shown that the

extent to which there holds semantic equivalence between translationally corresponding units is larger in the law texts than in the fiction texts.

Complexity measurements for individual text pairs are discussed in 5.5 with subsections. These results reveal considerable variation within each text type. Given the difference in restrictedness between law and fiction, we had expected a lower degree of complexity across the law data, and, on this background, it is striking that there is a considerable difference between the two pairs of law texts concerning the proportion of semantically equivalent correspondences: merely 38,8% in the *AEEA*, and as much as 60,9% in *Petro*. In our view, translating the *EEA Agreement* into Norwegian, and *Lov om petroleumsvirksomhet* into English, are quite different tasks, and this is reflected by the complexity results. The former text is an instance of supranational legislation; its translation is carried out for a normative purpose, and the legal status of the target version is equal to that of the original. The latter case exemplifies domestic legislation; it is translated for informative purposes, and the target version does not have the status of a law text. Moreover, we have argued that a larger proportion of semantic equivalence in the *Petro* text pair is compatible with three factors. Firstly, the *AEEA* is a negotiated text to an extent that the Norwegian act is not; hence it is likely that there are larger elements of intended linguistic imprecisions in the former than in the latter. Secondly, there is a stronger degree of domain-uniqueness in *Lov om petroleumsvirksomhet* than in the *EEA Agreement*, and due to this there may be a higher level of semantic precision in the Norwegian act than in the supranational agreement. Thirdly, it is likely that whereas the translation of the *AEEA*, given its normative purpose, has been influenced by faithfulness to the target language, the informative purpose behind the translation of the Norwegian act has promoted fidelity to the source text.

As regards the four pairs of fiction texts, it is our opinion that the various complexity measurements indicate differences among the translations concerning the degree of faithfulness to the informational content of the source text. Since these cases represent an unrestricted text type, such diversity may be expected. It is perhaps more interesting that whereas we have found a sharp difference in complexity between the two text pairs translated into Norwegian, there is a high degree of con-

vergence between those of the opposite direction. Of relevance to this is the difference in status between the literary languages of, respectively, English and Norwegian, but it would require studying a large group of text pairs to establish whether these aspects are related.

Concerning the question of automatisation of translation, raised in chapter 1, the picture does not look promising for the analysed texts, given the results of the complexity measurements. In the cases of the text pairs *AEEA*, DL, EFH, and BV, where a large majority of the data are classified as non-computable correspondences, we expect that automatic translation would not be useful, since the workload of correcting machine output would be too heavy. Then, for *Petro* and AB we have found that, respectively, 60,9% and 56,1% of the text pairs are included in computable correspondences. In these cases the usability of MT depends on the effort involved in editing the errors that a translation system would produce for those parts of the texts that fall outside the linguistically predictable domain. We will return to this topic in chapter 7 after the discussion of semantic phenomena in chapter 6.

6 Semantic phenomena

6.1 Overview

Having discussed the analysed text pairs in chapter 5, our focus is again directed towards individual types of translational correspondences. This chapter will discuss certain phenomena which are recurrent among the recorded data, and which involve some kind of semantic deviation between translationally corresponding units. We have organised the phenomena into classes, which we regard as subtypes within the main correspondence types 3 and 4.

The chapter is divided into two main parts. The first part, 6.2 with subsections, starts by presenting our motivation for identifying semantic subtypes within the recorded data. The classification is then seen in relation to the notion of shifts in translation, before we explain the criteria by which semantic subcategories have been identified and sorted into groups of related types. After this follows a brief presentation of the various sets of categories.

The second part, 6.2 with subsections, presents three main groups of subtypes. These are, firstly, classes of correspondences involving differences between source and target string in the amount of linguistically expressed information; secondly, classes of denotational non-equivalence between translationally corresponding expressions, and, thirdly, classes of referential differences between correspondents. All subtypes in these three groups have in common that there is some kind of difference in informational content between translationally related expressions. Within each group, we distinguish between predictable and non-predictable classes of correspondences, and for each class that is presented, the discussion provides a description with examples, as well as information on its frequency of occurrence within the compiled data.

6.2 The identification of semantic subtypes

The four types of translational correspondences fall naturally into two groups: whereas instances of types 1 and 2 can be identified by the presence of certain properties of surface syntactic structure, it is rather the absence of such properties which indicates that a string pair belongs to type 3 or 4. Normally, it is a straightforward task to identify translational correspondences of the two lower types by means of surface syntax, but once it is clear that a given string pair is neither of type 1 nor of type 2, it may be more difficult to decide whether it belongs to type 3 or 4. That decision relies on whether there is a non-predictable semantic difference between original and translation, which is a question that cannot be answered by means of surface-evident criteria, and it may require a thorough analysis of the semantics of each string.

Hence, the compilation of empirical data has forced us to observe semantic phenomena, and during this process we have found that certain types of semantic deviations between translationally corresponding units are recurrent among the data. Through a set of subtypes within the major categories of types 3 and 4, we try to describe how some such recurrent phenomena manifest themselves in translational correspondences. An earlier version of the set of semantic subtypes is presented in Thunes (1998: 38–49).

There are several reasons why semantic subtypes have been identified in the translational correspondences. Above all, since correspondences of types 3 and 4 cover a large majority of the entire set of data, it is desirable to analyse that part of the collected string pairs more thoroughly than just marking these data with a category label of *3* or *4*.[1] Further, it is in itself interesting that certain types of linguistic phenomena are recurrent among the translational data, and that motivates a description of them, as they may reveal something about the relationship between English and Norwegian. However, a comprehensive contrastive analysis of the English and Norwegian language systems is far beyond the scope of the present study, and only a quite limited selection of phenomena will be discussed. Moreover, if we assume that correspondences between semantically equivalent expressions can be produced by

[1] In terms of string length, correspondences of types 3 and 4 together cover 90% of the analysed texts; cf. table 5.1 in 5.2.1.

automatic translation, then the subcategories of type 4 indicate what kinds of linguistic challenges a post-editor will meet: since these mismatches are not computable correspondences, they must be handled by the human translator. We will return to this topic in 7.4.

6.2.1 Shifts in translation

The descriptions given in the present study of the main correspondence types 3 and 4, and of the semantic subtypes, may be seen as a parallel to the topic of *shifts* in translation studies. The concept of a 'shift' in translation is defined by Palumbo (2009: 104) as "a linguistic deviation from the original text, a change introduced in translation with respect to either the syntactic form or the meaning of the ST." In translation studies the term *shift* was first introduced by Catford (1965), although his contribution was not the earliest study of the range of phenomena that the term may refer to.[2] Palumbo (2009: 104–106) provides a brief historical overview of various approaches to shifts in translation, and points out that a variety of labels in addition to *shifts* have been applied to these phenomena: e.g., in Vinay and Darbelnet (1995) shifts are described as *translation procedures*, and in Chesterman (1997) as *(local) strategies*.[3] Shifts in translation constitute a broad topic, and several researchers have presented taxonomies of the different phenomena that are involved.[4] Chesterman (2005) gives a critical survey of the various approaches, and, across the field, he calls for greater terminological consistency, as well as conceptual stringency, in relation to the description of the phenomena involved in shifts (cf. 1.4.1).

There are also similarities between the present approach to cross-linguistic semantic deviations and analyses of translation shifts found in certain works that are rooted not only in translation studies, but in other disciplines as well. One example is the model given in Merkel (1999) for the description of structural and semantic correspondences in parallel texts (cf. 1.3.3). Another is found in Cyrus (2006), who

---

[2] The description presented in Vinay and Darbelnet (1995) appeared originally in 1958.

[3] Chesterman (1997: 90–91) explains that whereas *global strategies* apply to the translation of entire texts or kinds of texts, *local strategies* apply to translation units below text level.

[4] Among these, the model by van Leuven-Zwart (1989, 1990) is frequently cited. Also, chapter 4 in Chesterman (1997) provides a comprehensive typology.

presents a framework for manual annotation of translationally interrelated predicate-argument structures in an English-German parallel corpus. The aim of the analysis is to detect grammatical and semantic shifts in translational correspondences, and the annotated corpus is intended as a resource for linguists, translatologists, and MT researchers.

The correspondence type hierarchy, together with our discussion of subcategories within types 3 and 4, is not meant to be a new attempt to describe shifts in translation. For one thing, we want to avoid the term *shift*, because there has been a tendency in translation studies to apply this notion to translation methods, and the perspective of the present approach is to study relations between source expressions and their existing translations. Thus, as far as the characterisation of subtypes is concerned, this has not been developed from available descriptions of translation shifts. It is of course interesting to find parallels to our subtypes in categories commonly found in the various approaches to translation shifts. The subtype description is a truly data-driven classification that emerged during the analysis of the data on which Thunes (1998) is based, and which has been developed further in the present study. The categories arose solely from phenomena observed in the texts that were analysed. Furthermore, the subtype sorting is based on data representing only one language pair, and some of the phenomena to be discussed are language-pair specific. As indicated in 6.2, the subtypes are not intended as an exhaustive description of semantic deviations between the two languages. Their empirical basis is only a small selection of texts and text types, and recurrent semantic deviations may have been overlooked in the investigated texts.

6.2.2 Subtype sorting in relation to complexity sorting

Since every collected string pair is assigned one of the four types of translational correspondence, the compiled data can be seen as a set of type-sorted string pairs. As explained in chapters 1 and 3, the four main types of translational correspondences are a way of sorting string pairs according to an increasing degree of translational complexity. In contrast, the set of semantic subtypes within types 3 and 4 are a way of sorting correspondences on the basis of linguistic criteria, and they should not be

seen, from the outset, as representing more fine-grained distinctions on the scale defined for measuring translational complexity. The semantic criteria identifying the subtypes are independent of the complexity hierarchy; they refer to cross-linguistic phenomena which do have consequences for the degree of translational complexity, but the phenomena are not selected for description because of assumptions concerning their complexity. They have been analysed because they are recurrent, and this invites us to consider, subsequently, what effects the various phenomena have on translational complexity.

Since the semantic criteria of subtype sorting are independent of the complexity scale, we may find that if a certain type of semantic phenomenon occurs in two different string pairs, then that does not necessarily mean that those two correspondences belong to the same class of translational complexity. The type of a given string pair is determined by the entire correspondence, not only by the specific semantic phenomenon. Thus, an instance of a certain subtype in a string pair may be only one among several factors determining the degree of complexity assigned to the string pair as a whole. In practice, when a subcategory of type 3 is found in a given correspondence, then the entire string pair will be assigned type 4 if any other part of the string pair shows a degree of complexity higher than type 3. In such cases we do, however, keep track of the instance of the less complex subtype.[5]

6.2.3 Overview of semantic subtypes

Above all, the present study is a sorting project. Through the set of correspondence types, string pairs are sorted into four different classes, reflecting an increase in the amount and kinds of information needed in order to solve the translation task. In addition to the categories given by the correspondence type hierarchy, the framework behind our analysis also provides other distinctions that may serve as criteria for sorting. Firstly, the notion of predictability in the translational relation enables us to sort the empirical data into respectively computable and non-computable translation tasks (cf. 2.3.2). Secondly, through the qualitative notion of 'informational content'

[5] Cf. the description in 4.4.4 of how the tagging of subtypes has been implemented during the recording of data.

we may sort out correspondences where source and target vary with respect to the content of the linguistically encoded message (cf. 2.4.1.2). Thirdly, the quantitative notion of 'information' provides a basis for identifying translational correspondences exhibiting differences with respect to the amount of linguistically expressed information (cf. 2.4.1.1). Finally, the semantic subtypes, identified through linguistic criteria, constitute yet another dimension of sorting within correspondence types 3 and 4.

Thus, we have a set of five different dimensions which apply to the sorting of translational correspondences: translational complexity, predictability, informational content, amount of information, and semantic phenomena. As pointed out in 6.2.2, the different dimensions pertain to criteria that are independent of each other, but there are also important interconnections between them. For instance, the limit of predictability is linked with the scale of translational complexity, and source-target differences with respect to informational content, or the amount of information expressed, will obviously influence the degree of complexity in translational correspondences.

We have made an attempt at grouping the semantic subtypes according to the principles of our framework. In most of the subtypes, there is some kind of non-correspondence pertaining to the linguistically encoded informational content of, respectively, source and target string.[6] Since this is a very general description, we have tried to identify certain ways in which informational content is seen to differ between corresponding units. In this regard, one group of subtypes consists of cases where translationally related expressions vary with respect to the amount of linguistically encoded information. Another group comprises a range of subtypes where the common denominator is some kind of denotational difference between original and translation. A third group covers subtypes exhibiting source-target differences with respect to reference.[7] There is also a set of subtypes characterised by compositional non-equivalence between original and translation. Finally, in the later

[6] An exception is given by certain cases which fall within the domain of linguistically predictable correspondences although there is not a compositional relation of semantic equivalence between source and target string; cf. the discussion in 6.2.4.1.

[7] We maintain a distinction between 'reference' and 'denotation' as explicated in 6.3.2.

presentation we will make a further distinction within each group of subtypes between linguistically predictable and non-predictable cases. The sorting of subtypes is illustrated by table 6.1 in 6.2.4.2.

The categorisation should be seen as tentative, as there may be more than one possible way of describing individual subtypes given our framework. This is in line with an observation made by Chesterman (2005: 24) regarding the description and classification of shifts in translation, in his context referred to as *strategies*: "... a given change may be evidence of several strategies all operating at the same time."

That we in this chapter speak of semantic deviations within type 3 correspondences may appear as a contradiction to the central assumption, made clear in chapters 2 and 3, that a linguistically predictable translation is semantically equivalent with the source expression. The latter is indeed a main principle in our approach, but it may be slightly refined. In the later discussions of specific linguistic phenomena we will argue that certain semantic differences between translationally corresponding units can be seen as included in the domain of linguistically predictable correspondences. Thus, semantic subtypes within type 3 involve systematic, and predictable, semantic differences between source and target language systems. This means that the most important criterion for distinguishing between, on the one hand, correspondence types 1–3, and, on the other hand, type 4, is whether the given target expression is a member of the LPT set of the source expression, or not. In general, linguistic predictability, or computability, means full semantic correspondence, but in our view there are certain semantic deviations which may be predicted from information about how source and target language systems are interrelated.

Two aspects are in particular noticeable in relation to the recorded instances of semantic subtypes. Firstly, as regards the range of identified phenomena, there is a larger set of categories within the non-predictable domain than within the predictable. Secondly, with respect to the number of instances of the various categories, the frequency of non-predictable subtypes is noticeably larger than that of predictable ones.[8] That is, in the case of type 4 correspondences, nearly all string pairs are marked with

[8] Both aspects can be seen from the overview given by table 6.1 in 6.2.4.2.

one or more subtypes. With respect to type 3 correspondences, one or more subtypes have been identified in 20,8% of the number of string pairs extracted from fiction texts, and in 35,1% of the number of string pairs extracted from law texts. Given the difference in predictability between types 3 and 4, it appears reasonable that there is only a limited set of semantic divergences within the domain of predictable correspondences between two language systems, while the set of non-predictable semantic divergences is possibly open-ended, and its occurrences more frequent.

6.2.4 Brief presentation of individual subtypes

Having introduced a tentative grouping of the semantic subtypes in 6.2.3, we will briefly present each subtype in 6.2.4.1, together with an overview of occurrences within the entire set of compiled string pairs in 6.2.4.2. During the recording of empirical data, instances of semantic subtypes have been marked by short subtype labels entered in the comment field associated with each string pair, and this labelling has been done in order to count the number of occurrences of each semantic subtype.[9] Otherwise, the labels are of no importance, and will not be discussed further.

Some of the subtypes have been selected for more detailed discussions, which will focus on subtypes that occur relatively frequently, and on certain types that may reveal differences between the two investigated text types, and, to some extent, between the two directions of translation. It is more likely that such subtypes will have measurable effects on translational complexity than what is the case for infrequent categories or categories that are evenly distributed across the recorded data. Attention will also be paid to subtypes that are of special relevance to the issue of linguistic predictability in the translational relation.

6.2.4.1 Descriptions of subtypes

As stated in 6.2.3, the very general category of non-correspondence between source and target string with respect to the linguistically encoded informational content is divided into three main groups.

[9] This information can be produced automatically by the software used for data recording; cf. 4.4.2 and 4.4.4.

AMOUNT OF INFORMATION. The first group covers a set of subtypes exhibiting differences in the amount of linguistically encoded information in, respectively, source and target string. As shown by table 6.1 in 6.2.4.2, this group exhibits the largest occurrence within the analysed material, and the majority of such cases fall outside the domain of linguistically predictable translations. The predictable subset within this group concerns differences in grammaticalisation between source and target language systems, i.e. cases where certain distinctions of meaning are obligatorily expressed by grammatical markers in one translational unit, but not in its correspondent, because the distinctions are grammaticalised in only one of the languages. Thus, there will be a larger amount of linguistically encoded information in translational units expressing grammaticalised distinctions than in parallel units where those distinctions are absent. We have analysed cases of this kind as a subtype of type 3; they are further discussed in 6.3.1.1. Then, there are certain systematic differences between English and Norwegian in the use of possessive determiners which, in our view, constitute a special case of predictable differences in the amount of grammaticalised information, and in 6.3.1.2 this is presented as a separate subtype. Further, translational correspondences exhibiting non-predictable differences between source and target string in the amount of linguistically encoded information, are sorted into, respectively, cases of *specification*, where the information expressed in the source string is a subpart of the information in the target string, and cases of *despecification*, where the information expressed in the target string is a subpart of the information in the source string. Table 6.1 in 6.2.4.2 shows that non-predictable specification and despecification are the two most frequent subtypes within the recorded data; these categories are presented in more detail in 6.3.1.3.

DENOTATIONAL NON-EQUIVALENCE. The second group of subtypes exhibiting differences in informational content is a fairly heterogeneous set: the common denominator for its members is some kind of denotational non-equivalence, and hence a difference in linguistically expressed informational content, between source and target string. The notion of denotational non-equivalence is the topic of 6.3.2. As in the case of the former group, the majority of the correspondences included in this one do not represent linguistically predictable translations. There is, however, a subset of cases

exhibiting differences in the category of number between translationally matched nouns, and we regard this as a systematic, and hence predictable, denotational difference. This subtype is presented in 6.3.2.1, and the fairly wide category of non-predictable denotational differences between source and target string is discussed in 6.3.2.2.

Further, in type 4 correspondences we have observed several classes of other kinds of non-predictable denotational deviations. One of these subtypes concerns the phenomenon where co-referential noun phrases in, respectively, source and target string are denotationally non-equivalent. This occurs relatively frequently and is described in 6.3.2.3. A related subtype is found in correspondences between translationally linked anaphoric expressions which are denotationally non-equivalent, as shown by the italicised pronouns in example (1):[10]

(1a) ... but *one* soon learned either to get rid of them or accommodate them. (AB)
(1b) ... men *jeg* lærte snart enten å kvitte meg med dem eller gi etter for dem.

In (1a) the impersonal pronoun *one* corresponds with the singular, first person pronoun *jeg* in (1b). Instances of this subtype are not frequent (31 cases, according to table 6.1 in 6.2.4.2); they are fairly evenly distributed across the investigated text pairs, and will hence not be discussed further.

Another subtype of non-predictable denotational differences, which will for the same reasons not be dealt with in greater detail, is a class of correspondences exhibiting deviations in argument structure, as shown in example (2):

(2a) I told you. (DL)
(2b) Jeg sa jo det.[11]
'I said that.'

[10] In this chapter, Norwegian examples will be glossed only where it is necessary to bring across properties which are relevant to the discussions of the correspondences.
[11] The Norwegian adverb *jo* is not glossed because it has no English counterpart. Semantically it corresponds roughly with the expression *after all*.

In string pair (2) the Norwegian verb form *sa* ('said') is a linguistically predictable translation of the English verb form *told*. The relation expressed by these two verb forms can be represented as the predicate 'tell', which takes 3 arguments: argument 1 is linked with the agent role, argument 2 with the patient, and argument 3 with the beneficient. In English as well as in Norwegian, all three arguments can, but need not, be linguistically expressed in syntactic realisations of this predicate-argument structure. This may be accounted for by the distinction made by Pustejovsky (1995: 63–65) between true and default arguments: true arguments must be expressed syntactically; if not, the sentence will be ungrammatical. Default arguments are not obligatorily expressed, but "[t]hey are necessary for the logical well-formedness of the sentence" (Pustejovsky 1995: 64). In (2a) arguments 1 (*I*) and 3 (*you*) are linguistically expressed and argument 2 is implied, whereas in (2b) arguments 1 (*jeg* 'I') and 2 (*det* 'that') are expressed and argument 3 is implied. At this point we will not go more deeply into the distinction between true and default arguments in possible English and Norwegian realisations of the predicate 'tell', but merely observe that example (2) is a characteristic case of this subcategory: typically, in string pairs exhibiting denotational non-equivalence between translationally corresponding argument structures, source and target text differ with respect to the set of linguistically expressed arguments, and the difference may pertain to the number of expressed arguments, as well as to the order of those arguments.[12] Notably in such cases, the deviation in the argument structure of the target string cannot be predicted from the linguistically encoded information in the source string together with information about the interrelations between the two language systems. Such instances of cross-linguistic variation between corresponding argument structures are highly interesting from the point of view of theoretical linguistics, but since they are neither frequent within the analysed data (cf. table 6.1 in 6.2.4.2), nor indicate any text-typological differences, a more detailed discussion of them is peripheral to the issues investigated in the present project.

[12] *Order* in this context does not, of course, mean sequential order in surface syntax; it refers to order in argument structures, conventionally reflecting a ranking of semantic roles.

In contrast, a type of non-predictable denotational deviation which has a larger number of occurrences, and which is more frequent in the law texts than in the fiction texts is a class of correspondences where source and target string differ with respect to modality, as illustrated by example (3):[13]

(3a)[14] Unless otherwise specified, Articles 10 to 15, 19, 20 and 25 to 27 shall apply only to products originating in the Contracting Parties. (*AEEA*)

(3b) Med mindre annet er særskilt angitt, får artikkel 10 til 15, 19, 20 og 25 til 27 anvendelse bare for produkter med opprinnelse i avtalepartene.
'Unless something else is particularly specified, gets article 10 to 15, 19, 20 and 25 to 27 application only for products with origin in contracting-parties.DEF.'

The English modal verb *shall* in (3a) has no correspondent in the translation (3b). As noted in 5.4.2.4, in English law texts the modal *shall* is typically a marker of directive speech acts, normally commands, or prohibitions, if negated. The pragmatic function of command which is expressed by the modal in (3a) is not linguistically encoded in the string (3b), but it follows from extra-linguistic background information about the general directive function of the *EEA Agreement*. On this basis, we regard string pair (3) not only as an example of non-equivalence with respect to modality, but also as a case where the translation is semantically less specific than the original. Example (3) is characteristic of the majority of the identified cases. This subtype seems to reflect a tendency in the analysed law texts, where English modal verbs, normally *shall* or *will*, sometimes *may*, appear relatively often with no translational match in the corresponding Norwegian text. In our view, this is caused by a textual norm specific to the domain of law texts, which is different in the legal languages of English and Norwegian, respectively. In the English law texts we have studied, it seems to be a convention that speech acts such as command, prohibition, permission, and authorisation are expressed by the modals *shall*, *will*, or *may* (cf. 5.4.2.4), whereas in the Norwegian law texts, no modal (but simple present) is used because the pragmatic

[13] To include modality among the denotational properties of linguistic expressions is to apply a fairly wide sense of 'denotation', which we argue for in 6.3.2.

[14] For present purposes we disregard the internal structuring of (3) into a matrix correspondence and an embedded string pair. In this chapter this holds also for other examples in cases where the nesting is not relevant to the given discussion.

functions are implicit in information about the functions of the law itself. Even if this subtype seems to reflect a text-typological contrast within the recorded data, it will not be discussed further. Pragmatic functions of law text have previously been commented on in 5.4.2.4; the topic of textual conventions specific to law writing was introduced in 2.4.2.1, and will be further illustrated in 6.3.1.3. Moreover, most of the identified cases of mismatches in modality can also be seen as cases of specification or despecification (cf. 6.3.1.3), and within the recorded data we have identified only a few correspondences between semantically non-equivalent modal verbs.

Finally among the subtypes involving non-predictable denotational deviations, there is one class characterised by aspectual differences, and another by differences in grammatically expressed tense. In both classes, the non-equivalence concerns translationally corresponding verb phrases which may, but need not, constitute a linguistically predictable lexical correspondence between source and target language.

The correspondence between the italicised verb phrases in (4) is an instance of aspectual non-equivalence:

(4a) I still *meant* to go to Provence. (AB)
(4b) Jeg *hadde* fortsatt *tenkt* å dra til Provence.
'I had still thought to go to Provence.'

In example (4) the past perfect verb form *hadde tenkt* in the Norwegian translation conveys that the described act was completed before the time of utterance, whereas completion is not expressed by the simple past *meant* in the English source sentence. The tendency within the recorded cases of aspectual non-equivalence is that translationally corresponding verb phrases differ in the way illustrated by (4): one of the expressions is a complex verb phrase signalling that the verbal action has been completed (cf. the past perfect in (4b)), whereas its correspondent is a simple verb phrase which does not express completion (cf. the simple past in (4a)). Moreover, the choice of verb form in the translation cannot be predicted from the information contained in the source string together with information about the interrelations between source and target language. Within the identified cases, it varies whether the complex verb

form is found in the English translational units, or in the Norwegian one.[15] It is however too simple to regard this phenomenon merely as correspondences between simple and complex verb phrases, since there are normally also aspectual contributions from the lexical meanings of the verbs that are involved. Denotational non-equivalence of this kind appears to be more frequent within the analysed fiction texts than within the law texts, which appears reasonable given the strong demands of precision which apply to the drafting and translation of legal acts (cf. 5.4.2.1). But, as table 6.1 in 6.2.4.2 shows, this is not a very frequent subtype, and further discussion is left aside.

In (5) there are two correspondences, between italicised verb forms, which illustrate the subtype characterised by non-equivalence in grammatically expressed tense:

(5a) Brita *forstår* straks at hun *mener* fru Bendixen. (BV)
'Brita understands immediately that she intends Mrs Bendixen.'
(5b) Brita *understood* at once that she *was referring* to Mrs Bendixen.

The present tense verb form *forstår* in (5a) corresponds translationally with the past tense verb form *understood* in (5b); the present tense verb form *mener* in (5a) is matched by the past progressive *was referring* in (5b). These changes in tense cannot be predicted on the basis of the linguistic information contained in the source string together with information about the interrelations between source and target language systems. *Tense* is here understood according to Comrie (1985: 9) as "grammaticalised expression of location in time." Accordingly, this subtype includes only cases of temporal non-equivalence involving grammatical properties associated with verb phrases; it disregards translational differences between lexicalised expressions of temporal location.[16] As shown by table 6.1 in 6.2.4.2, non-equivalence with respect to tense appears to be an infrequent subtype within the recorded data. The overall majority of identified cases occur in the fiction texts, and, like in the case of aspectual non-equivalence, it is to be expected, given the norms of statutory language, that non-

[15] We have not recorded cases involving differences only in lexically encoded aspectual properties. Such cases are classified as instances of the more general subtype of non-predictable denotational differences; cf. 6.3.2.2.
[16] Deviations of that kind fall within the category of non-predictable denotational differences.

equivalence with respect to tense is rare in the law texts. In the fiction texts this subtype appears to be even less frequent in English-to-Norwegian correspondences than in Norwegian-to-English data. However, due to the limited size of the empirical material, we cannot judge whether this quantitative difference is correlated with the dimension of direction of translation, or whether it is accidental and caused by diverging preferences of individual translators. Due to the generally low frequency of this subtype it will not be discussed further.

REFERENTIAL DIFFERENCES. The third group of semantic subtypes covers cases where noun phrases that are translationally interrelated do not correspond to each other with regard to referential properties, mainly due to differences in the marking of definiteness. Altogether, the members of this group are not very frequent, compared with the other groups, but their distribution may reflect an interesting difference between the analysed text types. The law data contain a noticeable set of translational links between definite and indefinite noun phrases, respectively. Such correspondences are very rare within the fiction data. They are further discussed in 6.3.3.1, where we argue that they can be seen as a linguistically predictable type of correspondence. However, the recorded data also include some cases which cannot be classified as predictable. These appear to be more frequent in fiction than in law text, and they are presented in 6.3.3.2.

COMPOSITIONAL NON-EQUIVALENCE. In addition to the three groups of subtypes involving differences in informational content, we have identified a fourth group characterised by absence of compositionality in the correspondences between source and target strings. Compositional equivalence in the translational relation is defined by Thunes (1998: 39) in the following way: "... if there holds compositional equivalence between linguistic signs, there is not only global equivalence between the entire signs, but also local equivalence between corresponding constituents of the two signs." Thus, according to the requirements of compositional semantic equivalence specified for correspondence types 1, 2 and 3 (cf. 3.3.2.1, 3.3.3.1, and 3.3.4.1), compositional equivalence between source and target string is typically fulfilled in string pairs of these types. As explained by Thunes (1998: 39), there are cases of type 3 correspondences where we regard the global meaning of respectively source and

target string as equivalent, even if there is semantic non-equivalence between certain corresponding constituents of the two strings; cf. (6):

(6a) They gave us four days to leave. (DL)
(6b) De ga oss fire dager å komme oss vekk på.
'They gave us four days to come us away on.'

As indicated by the glossing of (6b), there is compositional equivalence between subparts of source and target strings, but it is violated by the correspondence between the English verb *leave* and the Norwegian expression *å komme oss vekk*, because the Norwegian lexical units *oss* ('us') and *vekk* ('away') have no direct correspondents in the English expression. Still, we find that *å komme oss vekk* is a linguistically predictable translation of *leave*, given the relevant interpretation of the source sentence (6a). That is, the translational relation between *leave* and *å komme oss vekk* is predictable from the lexical meanings of these two expressions. Hence, string pair (6) is a type 3 correspondence.[17] Moreover, since (6a) and (6b), seen as units, are semantically equivalent, we regard the linguistically encoded informational content of the two sentences to be the same.

Compositional non-equivalence in type 4 correspondences is not uncommon within the recorded data; it may be illustrated by (7):

(7a) She cursed steadily, the tears streaming. (DL)
(7b) Hun bannet og gråt, en jevn strøm av ord og tårer.
'She cursed and cried, an even stream of words and tears.'

The glossing of (7b) indicates the non-compositionality in this correspondence. Most notably, the meaning expressed by the adverb *steadily* in (7a) is matched by the adjective *jevn* in (7b), and the verb form *streaming* in (7a) corresponds with the noun *strøm* in (7b). In this case we regard the translation as semantically more specific than the original: both sentences (7a) and (7b) describe the referent of *she* as cursing and

[17] Cf. the remarks on linguistically predictable translation in 6.2.3.

crying, but the notion of steadiness, which in (7a) is associated with cursing, is in (7b) attributed explicitly to the crying as well as to the cursing. Hence, there is a difference between the two sentences in linguistically expressed informational content, and (7b) is not a predictable translation of (7a). Alternatively, these semantic deviations can be ascribed to the category of non-predictable denotational differences (cf. 6.3.2.2), and as already noted, the correspondence is also a case of semantic specification (cf. 6.3.1). Example (7) illustrates that it is not always easy to isolate instances of compositional non-equivalence from occurrences of other subtypes.

Identified instances of compositional non-equivalence are fairly evenly distributed among the investigated text pairs, as well as across both the dimensions of text type and direction of translation. This may indicate that compositional non-equivalence is not correlated with variations in translational complexity along these dimensions. Hence, this category will not be discussed further, although it is highly interesting, in particular from the viewpoint of semantic analysis.

6.2.4.2 Occurrences of subtypes

Quantitative data on the semantic subtypes should be seen as highly tentative results, as the registration of subtypes is in several ways prone to errors, especially since the identification of the instances of subtypes relies on semantic interpretation carried out by an individual annotator, and not on the recognition of surface-evident criteria, which are applied when translational units are extracted. Hence, cases may easily be overlooked, and their categorisation may be debatable. Also, there are cases where it has been possible to assign more than one subtype to specific semantic deviations identified in the compiled data.[18] Thus, table 6.1 presents what has been identified, and not everything that can be found, in the analysed texts. The quantitative results are given in terms of the number of string pairs in which at least one occurrence of each subtype has been identified.

[18] Insofar as it involves individual, linguistic judgments, the identification of semantic phenomena among our data is similar to a special annotation task practised within the field of word sense disambiguation. Automatic WSD tools can be trained on a corpus where human annotators have marked what sense occurrences of semantically ambiguous words belong to, and for this kind of task human inter-annotator agreement has been reported to be of merely about 80% with respect to English (cf. Jurafsky and Martin 2009: 679). This indicates that a certain element of inconsistency is probably unavoidable in the semantic annotation of the compiled data.

Table 6.1. Tentative frequency of occurrence for each semantic subtype across the entire set of recorded data.

	Type 3	Type 4
Differences in the amount of linguistically expressed information:		
predictable differences w.r.t. grammatically coded information	49	
predictable differences in the use of possessive determiners	54	
non-predictable specification		918
non-predictable despecification		604
Denotational non-equivalence:		
predictable denotational differences	115	
non-predictable denotational differences		433
denotational non-equivalence between coreferential NPs		304
denotational non-equivalence between corresponding anaphors		31
non-equivalence in argument structure		23
non-equivalence w.r.t. modality		114
non-equivalence w.r.t. aspect		73
non-equivalence w.r.t. tense		48
Referential differences:		
predictable differences in the use of definiteness	137	
non-predictable referential differences		55
Compositional non-equivalence:		
in predictable correspondences	240	
in non-predictable correspondences		138

More detailed, but equally tentative, quantitative data will be given in the following presentations of selected semantic subtypes. For each of these subtypes, figures will be given to indicate how the occurrences are distributed across the entire set of data, and across the various sets of data representing each direction of translation, each text type, and each text pair. Within each of these sets of data, we will calculate the proportion of string pairs where the subtype is identified in relation to the total

number of string pairs in that set (n_T). With respect to the subtypes involving non-predictable correspondences, we will also calculate proportions in relation to the number of type 4 correspondences within each set of data (n_4), since string pairs of type 4 constitute the majority of the analysed texts. For these purposes, tables 6.2 and 6.3 provide reference data: n_T and n_4 are, respectively, the total number of string pairs, and the number of type 4 correspondences, relative to the different sets of data. These figures will serve as a basis for comparison in the presentations of occurrences of individual subtypes.

Table 6.2. The values of n_T and n_4 relative to all data, to each direction of translation, and to each text type.

	all data	**E → N**	**N → E**	**law text**	**fiction**
$\boldsymbol{n_T}$	4439	2104	2335	1713	2726
$\boldsymbol{n_4}$	2219	1140	1079	740	1479

Table 6.3. The values of n_T and n_4 for each text pair.

	AEEA	***Petro***	**AB**	**DL**	**EFH**	**BV**
$\boldsymbol{n_T}$	791	922	521	792	703	710
$\boldsymbol{n_4}$	405	335	208	527	365	379

For each subtype it is possible that more than one instance of it is found within a string pair, and this is the main reason why occurrences are counted as the number of string pairs containing at least one token. In relation to the distribution of the main correspondence types, we focussed on the proportions of text, given in terms of string length, that are covered by each type (cf. 5.2.1). Our notion of string length is, however, not so easily applicable in connection with the semantic subtypes, since the subtypes involve phenomena that are not necessarily associated with entire translational units, or that may not be readily attributed to identifiable subparts of such units. E.g., in connection with the categories of specification and despecification, we will argue

that differences in the amount of linguistically expressed information can in certain types of cases be measured by counting linguistic signs in translationally corresponding expressions (cf. 6.3.1). But this cannot straightforwardly be converted into a counting of word forms, or string length, principally because there is no one-to-one relation between signs and word forms. For instance, a single word form may express one (or more) grammatical sign(s) in addition to a lexical sign. This is only one example showing that string length measurement is difficult in relation to the identification of occurrences of semantic subtypes. Thus, the least problematic approach is to estimate the frequencies of the various categories by counting the numbers of string pairs where at least one instance of each phenomenon is found.

As we have seen in chapter 5, the high degree of restrictedness in law texts makes it reasonable that the extent to which extracted translational units are semantically equivalent is greater within the data recorded from law than within those compiled from fiction. This picture is also confirmed by the identification of semantic subtypes. In table 6.1 we have not provided information on how the occurrences of the various kinds of semantic subtypes are distributed among the individual text pairs.[19] But we have counted the total number of occurrences of *non-predictable* semantic deviations in each text pair, and from this we have found that within the law data there are 972 identified instances of non-predictable semantic deviations, whereas the corresponding figure for the fiction data is 1769 occurrences. Moreover, in chapter 5 we have discussed the importance of minimal type 4 correspondences, i.e. string pairs which are classified as non-computable because of only one semantic difference between the two correspondents (cf. 5.2.2 and 5.4.2.6). Along with the tagging of semantic subtypes, we tried to keep track of such cases, and of minimal type 4 correspondences we have tentatively identified 338 occurrences in the law texts and 155 in the fiction texts. Given that we have recorded altogether 740 type 4 correspondences from the law texts, and 1479 from fiction,[20] this means that among the law data, as much as 45,7% of the string pairs of type 4 are minimal cases, and

[19] We will do so only for the subtypes to be discussed in 6.3 with subsections.
[20] Cf. tables 5.6–7 in 5.4.1.

that among the fiction data, only 10,5% of the compiled type 4 correspondences are minimal ones.[21]

These facts confirm two general observations made during the recording of string pairs. Firstly, in type 4 correspondences extracted from the fiction texts, there tends to be several semantic differences between source and target units, whereas in type 4 correspondences recorded from the law texts, there are normally only one or two semantic deviations. Secondly, minimal cases of type 4 are markedly more frequent in the law data than in the fiction data. The importance of minimal type 4 correspondences has already been discussed in chapter 5, and the relevance of these two observations will be seen in the discussions of certain subtypes (cf. 6.3.1.3), and will be discussed further in chapter 7.

6.3 Differences in informational content

As regards the grouping of semantic subtypes, the most general classification criterion is divergence with respect to the linguistically encoded informational content of, respectively, source and target string (cf. 6.2.3). By differences in informational content we understand the following: when a pair of translationally corresponding linguistic expressions do not have the same meaning, they carry different messages, and they do not have the same informational content.[22] This applies, in fact, to the overall majority of the recorded instances of semantic deviations between source and target string. As pointed out in 6.2.3, this is a very general category, but we find it useful, because it does not include all cases of semantic deviations: in some cases there is a linguistically predictable, semantic difference between source and target text, a difference that does not change the message of the original text. As we have seen in 6.2.4.1, this pertains to predictable correspondences involving compositional non-equivalence.

---

[21] The software applied to the recording of string pairs does not facilitate calculating, in terms of string length, the proportions of texts covered by the minimal cases of type 4. Had that been possible, we would have seen an even sharper text-typological contrast, since, as presented in 5.4.2.6, the average lengths of the recorded string pairs are greater within the law data than within the fiction data.

[22] Cf. the explication of 'informational content' in 2.4.1.2.

Since the category defined by differences in linguistically encoded informational content is very wide, the sorting into various groups of subtypes serve to describe it in a more interesting way. These groups will be presented as follows: 6.3.1 with subsections discusses differences in the amount of linguistically encoded information; denotational differences are the topic of 6.3.2 with subsections, and referential differences are presented in 6.3.3 with subsections.

6.3.1 Differences in the amount of information

In 6.2.3 translational correspondences involving differences in the amount of linguistically encoded information are introduced as a subtype of differences in informational content. The reason is simple: a particular message, or informational content, is supported by a certain amount of information, so that if a signal s_1 contains a smaller amount of information than a signal s_2, then s_1 and s_2 cannot express the same message.[23] Thus, a translational correspondence where source and target units contain different amounts of information is also an example of a correspondence between expressions carrying different informational contents.

This distinction makes it possible to identify string pairs exhibiting differences in the amount of information encoded linguistically in translationally corresponding expressions. Hence, such correspondences can be separated from cases where a source-target difference in linguistic informational content is not a question of quantity.[24] The recorded data include many correspondences where source and target string differ in the sense that the amount of information expressed linguistically by one string, or by a segment of it, is a subpart of the amount of linguistic information contained in its correspondent in the parallel text. Granted that information is a commodity that can be measured, there are, in such cases, quantitative source-target differences in the amount of linguistically expressed information, and by identifying string pairs where the difference is in the amount of linguistic information, it is

[23] Cf. the point made in 2.4.1.2 that to convey a specific message requires that *all* the information behind that message is transmitted (Dretske 1981: 60).
[24] 'Informational content' itself is not a quantitative notion; cf. 2.4.1.2.

possible to distinguish between cases where the target expression is more, or less, specific than the corresponding source expression.

According to information theory, information can be measured in terms of the reduction of uncertainty.[25] Although we have adopted the quantitative notion of 'information', the present approach does not apply any mathematical tools for measuring amounts of information in terms of numerical values. For our purposes we want to correlate the reduction of uncertainty with reductions in the sets of possible interpretations of linguistic expressions, and in relation to the empirical data we may individuate linguistic signs in translational correspondences. Signs contain an expression as well as a component of meaning, and they may be sorted into lexical signs and grammatical signs.[26] If, in a pair of translationally corresponding expressions, one or more signs in one of them are not matched by any linguistic material in the correspondent, and the expressions otherwise contain signs which are pairwise related to each other in a translationally predictable way, then the set of unmatched signs represents the difference in amount of linguistic information between the corresponding expressions. The amount of linguistic information which is shared by the two expressions is contained in the sets of signs which are pairwise related in a translationally predictable way. The number of signs which have no translational match may serve as a very simple quantitative measure of the difference in amount of linguistic information. To be more precise: the quantitative difference may be estimated in terms of the number of *non-identical* and *non-coreferential* linguistic signs which have no translational match: i.e., an unmatched sign counts only as 1 in this quantitative measure even if there are more than one coreferential tokens of it in the given translational unit.[27]

---

[25] Cf. Dretske (1981: 4), cited in 2.4.1.1.

[26] In English as well as in Norwegian an example of a grammatical sign is past tense, which is expressed through verbal morphology in these two languages, and whose meaning is that the situation referred to by the verb took place before the time of utterance.

[27] E.g. in the sentence *They have arrived* there are two coreferential tokens of the grammatical sign 'plural number', one in the pronoun *they*, and another in the auxiliary verb *have*. This sentence corresponds word-by-word with the Norwegian translation *De har ankommet*, in which the plural is encoded only in the pronoun *de*, since the category of number is not expressed in Norwegian verbal morphology. Still, we do not regard the English sentence as containing a larger amount of information than its Norwegian correspondent, since the number of non-identical and non-coreferential linguistic signs is the same in both sentences.

Further, in a situation of this kind, we assume that the translationally parallel sets of signs which contain the shared amount of information constitute expressions with shared sets of possible interpretations. Since the set of unmatched signs adds to the amount of information contained in one of the two expressions, this set of signs reduces the set of possible interpretations of that expression. In this manner the difference in the amount of linguistically encoded information is correlated with a reduction in uncertainty: when an expression has fewer possible interpretations, uncertainty is reduced with respect to what its correct interpretation is.[28] The points made here will be illustrated by the later discussions of linguistic examples. In 6.3.1.3 we will also discuss cases where relations of hyponymy, or hyperonymy, hold between translationally corresponding lexical signs, so that there is a quantitative difference in information between source and target string even if they contain the same number of non-identical and non-coreferential linguistic signs.

It should be noted that when measuring differences in the amount of linguistically expressed information in translationally corresponding units, we consider the sets of possible interpretations for each unit. Previously we have argued that when the translational complexity of given string pairs is analysed, we consider the target expression in relation to the relevant interpretation of the source expression, since we keep source text disambiguation apart from the translation task (cf. 3.3.1.1 and 4.3.6.2). However, in order to quantify differences in the amount of linguistically encoded information, it is necessary to take into account the sets of possible interpretations of both units. To consider only the relevant interpretation of the source string would mean an increase in uncertainty in every case where more than one interpretation is possible for the target string.

As explained in 6.2.4.1, correspondences where a source expression contains a subpart of the linguistic information included in the target expression are regarded as cases of *specification*, while correspondences where the target expression carries a subpart of the linguistic information contained in the source string are seen as cases

[28] A parallel to this approach is found in Fabricius-Hansen's (1996, 1999) notion of informational density, understood as the amount of information expressed per linguistic unit. Her analysis is commented on in 4.2.1.1 and 5.3.2.

of *despecification*. Within each of these categories we will make a further distinction between *lexical* and *grammatical* (de)specification, which follows from the division between lexical and grammatical signs. We find it natural to treat specification and despecification as semantic differences between original and translation: different amounts of information lead to different messages, and different messages do not convey the same meaning.

Since we apply the notions of specification and despecification in an analysis of translational data, the question may be raised why these are not described, respectively, as *explicitation* and *implicitation*. The latter terms are avoided because we do not see that our notion of specification overlaps fully with the notion of explicitation, as applied in translation studies. To add information which is not expressed in the original can be, but is not necessarily, the same as making explicit information which is implicit, but not linguistically expressed, in the source text. If the added information is not even implied in the original, but is a piece of genuinely new information, then such instances of specification are not examples of explicitation, at least according to a certain definition of that phenomenon (cf. 5.3.2). Also, the term *explicitation* is avoided because it tends to be associated with translation method, which is not a topic in our product-oriented approach.[29]

Chesterman (1997: 109–110) describes a notion of 'information change', which is presented as a type of translation strategies, where the translator either adds or omits information. The result of information change, in Chesterman's sense, corresponds with the phenomena we classify as specification and despecification, respectively. According to Chesterman, information change is due to a deliberate choice, and this is what makes addition and omission distinct from the strategies described by Chesterman (1997: 108–109) as *explicitness changes*, which comprise explicitation and implicitation. In our approach the category of differences in the amount of expressed information cover the results of information changes, as well as of explicitness changes, as defined by Chesterman (1997). That he draws the line between deliberate and non-deliberate changes is clearly useful in translation studies. How-

[29] Cf. the similar remarks in 6.2.1 on why the term *shift* is avoided.

ever, he does not distinguish between the quantitative notion of information and the semantic concept of informational content.

In relation to the compiled data, we do not treat specification and despecification as phenomena that must be associated necessarily with entire pairs of translational units. That is, instances of specification and despecification can be found in subparts of extracted strings, which will be shown by the later discussions of examples. This also means that within one given string pair, both categories of specification and despecification may be instantiated, and there may be more than one instance of each category.[30]

Specification and despecification have consequences for the degree of translational complexity. In the majority of the cases identified, source-target differences with respect to the amount of expressed information fall outside the domain of the linguistically predictable correspondences. In some special cases we regard the differences as linguistically predictable; these are discussed in 6.3.1.1 and 6.3.1.2.

6.3.1.1 Predictable differences in the amount of grammatical information

DESCRIPTION. As stated in 6.2.4.1, we have identified a class of correspondences where certain distinctions of meaning are obligatorily expressed by grammatical markers in one translational unit, but not in its correspondent, because the distinctions are grammaticalised in only one of the languages. Applying the terms introduced in 6.3.1, this subtype could be described as, respectively, *grammatical specification* and *grammatical despecification*. In correspondences exhibiting grammatical specification, there is at least one grammatical sign in the target string which has no match in the source string, and in cases of grammatical despecification, at least one grammatical sign expressed in the source string has no match in the target.

The tokens identified of this subtype largely concern the use of progressive aspect in English, which is not grammaticalised in Norwegian. Three cases have been found which involve the use of subjunctive mood. This is grammatically expressed in English, but in Norwegian no longer part of the language system, and subjunctive forms

[30] See, for instance, example (21) in 6.3.1.3.

of Norwegian verbs occur only in archaic texts. Since only three instances of the English subjunctive have been found among the recorded data, this will not be further discussed.

Example (8) is a case of grammatical specification involving the use of progressive aspect in the English translation. Progressive aspect in English is also referred to as durative or continuous aspect (cf. Quirk et al. 1985: 197). In Norwegian, there is no grammatical marker of the aspectual feature of duration; if it is expressed, it is through lexical means, such as by the complementiser *mens* ('while'), or by the verb-particle construction *holde på å* ('be in the process of'). In the source sentence (8a) the Norwegian present perfect *har ventet* ('have waited') corresponds translationally with the English present perfect progressive *have been waiting* in (8b):

(8a) ... de har ventet gjennom uker med gråvær, ... (EFH)
'they have waited through weeks with grey-weather'
(8b) They have been waiting through weeks of cloudy weather, ...

On the syntactic level, source and target strings in (8) are not sufficiently similar to fulfil the constraints on correspondence types 1 and 2, but it is our view that this is a type 3 case: each lexically encoded unit of meaning in the source string has a lexical match in the target string, and the only semantic difference we have identified between the two strings is the grammatical meaning expressed by progressive aspect in the English translation. The semantic component of duration is important in both sentences: it is included in the lexical meaning of both of the translationally corresponding verbs *vente* and *wait*, as well as in each of the temporal adverbials *gjennom uker med gråvær* and *through weeks of cloudy weather*.

The expression *har ventet* in (8a) is, in line with Faarlund et al. (1997), a present perfect verb form. According to Faarlund et al. (1997: 566), the temporal meaning expressed by the Norwegian present perfect tense on a durative verb (e.g. *har ventet*) in the context of a durative adverbial (e.g. *gjennom uker med gråvær*) is that the described situation applies during a period which starts in the past and which includes the time of utterance. There are at least two alternative interpretations of the Norwegian sentence (8a): the waiting has lasted for weeks until the present, but now stops,

or the waiting, which has lasted for weeks until the present, may even continue into the future.

The expression *have been waiting* in (8b) is, according to Quirk et al. (1985), a present perfective progressive. For English Quirk et al. (1985) identify two aspectual features, the perfective and the progressive, and they argue that it is not easy to isolate the semantic contribution of the different English aspectual markers, since it is intertwined with the meaning of tense, and since the use of aspectual features is influenced by the semantic content of verbs, and by the meaning of temporal adverbials.[31] In particular, if the basic meaning of perfective aspect is that an event is completed, whereas progressive aspect signals that an event is on-going, this highlights the point that "the perfective progressive has a semantic range that is not entirely predictable from the meanings of its components" (Quirk et al. 1985: 210–211). Concerning the use of the perfective progressive with so-called *stance verbs* (e.g. *live*, *stand*, *sit*, *lie*), of which we regard *wait* in (8b) as an example, Quirk et al. (1985: 205–206) observe some variation in speakers' intuitions, because the perfective interferes with the component of duration inherent in the verbal meaning. Still, there is a tendency that the use of the progressive in (8b) signals that the described situation, which has lasted for a certain period up to the present, is not necessarily over and may continue beyond the time of utterance.[32] This narrows the possibilities of interpreting the translation in relation to the possible interpretations given above for the source sentence (8a), and hence the use of the progressive in (8b) is an example of grammatical specification.

Example (9) illustrates grammatical despecification, and as in the case of example (8), it is our view that the only semantic difference between source and target string is the grammatical meaning expressed by the progressive in (9a):

(9a) "Who are you trying to repeat in me?" (AB)
(9b) "Hvem forsøker du å gjenta i meg?"
'Who tries you to repeat in me?'

[31] Cf. Quirk et al. (1985: 188–189), (1985: 189–197) on perfective aspect, (1985: 197–210) on progressive aspect, and (1985: 210–213) on the perfective progressive.
[32] In contrast, the simple present perfective *have waited* would, at least to some speakers, imply that the waiting is over at the time of utterance.

The present progressive *are trying* in (9a) corresponds translationally with the present verb form *forsøker* in (9b), which is the only possible Norwegian translation of the English verb phrase.

With respect to *are trying* in (9a), the use of the present progressive can be seen as a marked choice in relation to the simple present verb form. If the English simple present had been used, it would have offered (at least) two readings: when verbs express events, such as an act of trying, the two most common meanings of present tense are, according to Quirk et al. (1985: 179–180), the habitual present, where the verb refers to a sequence of repeated events, and the instantaneous present, where the verb describes a single event occurring at the time of utterance. With a simple present verb (i.e. *Who do you try to repeat in me?*), both readings would be possible for sentence (9a) if considered out of context, as a linguistic type. But a habitual reading appears improbable in relation to the context in which the sentence occurs.[33] It is more likely that the sentence describes a single event, and the effect of the present progressive in (9a) is exactly to exclude the habitual reading, since a habitual reading of the progressive would require the presence of an adverbial referring to the period during which the repetition would take place (cf. Quirk et al. 1985: 199). The semantic contribution of the present progressive in (9a) is to emphasise that the described situation is a single event. To describe it as instantaneous may appear odd, since an act of trying will have some duration, although limited, since there is also a punctual element included in the verbal meaning of *try*.

Concerning the Norwegian verb form *forsøker* in (9b), the meaning expressed by its present tense form is that the utterance time is included in the time span of the described situation. In Norwegian, present tense verb forms may also have a habitual reading when referring to repeated actions.[34] Thus, if considered out of context, both the habitual reading and the single-event reading are logically possible interpretations of sentence (9b). This means that, as linguistic types, the translation (9b) has a wider

[33] The question in (9a) is addressed to her lover by the female protagonist in André Brink's *The Wall of the Plague*. By asking it, she alludes to his previous loves, and it appears odd that she should refer to a habit of his by this question. As explained in 4.3.6.3, such extra-linguistic contextual information is not considered when correspondence type is assigned to string pairs.

[34] See Faarlund et al. (1997: 562–563) on meanings expressed by present tense in Norwegian.

set of interpretations than the original (9a) has, and since this is due to the absence in (9b) of a piece of information that is grammatically expressed in (9a), this is a case of grammatical despecification.

As previously indicated, the phenomenon illustrated by examples (8) and (9) relates to the fact that different languages vary with respect to the inventories of grammaticalised semantic distinctions. The examples have illustrated variation between the sets of categories that are grammaticalised. Languages may also vary in terms of the sets of features that are grammaticalised within a category. E.g., within the category of mood, three features are grammatically expressed in English (indicative, imperative, and subjunctive), whereas only two are grammaticalised in Norwegian (indicative and imperative). Due to facts of this kind, there are translational correspondences where grammaticalised semantic distinctions are obligatorily expressed in one language, but absent in the other. We regard such correspondences as linguistically predictable because they are derivable from information about the two language systems and their interrelations. In particular, the translation of Norwegian verb forms into English progressive forms is predictable from information about the distinction between progressive and non-progressive aspect in English.

The examples discussed of, respectively, grammatical specification and despecification, illustrate a certain asymmetry between the two directions of translation. In the given example of despecification, only one Norwegian verb form (*forsøker*) is available in the translation of (9a), if a linguistically predictable translation is to be chosen.[35] In the case of grammatical specification in example (8), producing the target sentence (8b) involves making a choice between the progressive and the non-progressive. Thus, there is asymmetry in the degree of translational complexity between grammatical despecification and specification, respectively, since there is lower complexity in the task of discarding a piece of information that is not grammaticalised in the target language than in the task of identifying a piece of infor-

[35] (9b) is of course not the only possible translation of (9a). A semantically equivalent alternative could be the stylistic variant *Hvem er det du forsøker å gjenta i meg?*, where the interrogative *hvem* is focussed using a so-called fronted construction with the expletive *det*. Another possible translation could be *Hvem vil du forsøke å gjenta i meg?*, which is not semantically equivalent with the original, since the modal verb *vil* adds elements of volition and futurity in the translation.

mation that is obligatorily expressed through grammatical distinctions specific to the target language. This point will be revisited in the discussion of occurrences of grammatical (de)specification.

To say that the distinction between progressive and non-progressive is obligatorily expressed in English implies, in principle, that there is a difference in the amount of grammatically encoded information also in cases where English non-progressive verb forms correspond with Norwegian verbs. However, the progressive is a marked choice in relation to simple verb forms in English,[36] and for that reason we have not identified correspondences between English non-progressive verb forms and Norwegian verbs as instances of grammatical (de)specification.

It is typical of cases of predictable grammatical (de)specification that for a single expression in one language there is a set of linguistically predictable translational correspondents (*LPT*) in the other language, and that the *LPT* is a finite and relatively small set. This point offers a way of distinguishing predictable grammatical (de)specification from non-predictable cases of (de)specification (cf. 6.3.1.3): in the non-predictable correspondences, the semantically more specific expression is only one alternative within an in principle unlimited set of non-predictable specifications.

OCCURRENCE. Table 6.1 in 6.2.4.2 shows that recorded instances of predictable differences in grammatically coded information are not very frequent in comparison to certain other subtypes. Within the entire set of compiled data, we have tentatively identified 49 string pairs containing at least one instance of predictable grammatical (de)specification. Tables 6.4 and 6.5 present further details on how these occurrences are distributed along the dimensions of text type and direction of translation, and across the different text pairs.[37] The absolute numbers, as well as the percentages, indicate the low frequency of this subtype: on average, it has been identified in only 1,1% of all recorded string pairs. Since the majority of the cases involve the progressive aspect, this can most likely be attributed to the following facts reported by Quirk et al. (1985: 198): “The progressive aspect is infrequent compared with the nonprogressive. A count of a large number of verb constructions has indicated that

[36] Cf. the remarks on the frequency of the progressive, quoted below from Quirk et al. (1985: 198).
[37] The different values of n_T are presented in tables 6.2 and 6.3 in 6.2.4.2.

less than 5 per cent of verb phrases are progressive, whereas more than 95 per cent are nonprogressive."

Table 6.4. Occurrences of predictable grammatical (de)specification, counted within all recorded string pairs, within each direction of translation, and within each text type.

	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T
Across **all data**:	49	1,1
Across all data **E** → **N**:	19	0,9
Across all data **N** → **E**:	30	1,3
Across all **law** data:	3	0,2
Across all **fiction** data:	46	1,7

Table 6.5. Occurrences of predictable grammatical (de)specification in individual text pairs.

Legal texts			Fiction texts		
Text pairs	Frequency of string pairs where the subtype is found:		Text pairs	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T		in absolute numbers	in per cent of n_T
AEEA	2	0,3	**AB**	10	1,9
			DL	7	0,9
Petro	1	0,1	**EFH**	17	2,4
			BV	12	1,7

The results given in table 6.4 indicate that there is a certain difference between the two directions of translation in relation to the frequency of predictable grammatical (de)specification: while the subtype has been identified in 19 string pairs among the English-to-Norwegian data, it is found in 30 string pairs among the Norwegian-to-

English data. However, considering the high degree of uncertainty associated with subtype identification (cf. 6.2.4.2), and the very limited size of the data, it would require a larger empirical study to establish to what extent differences in grammaticalisation between these two languages have consequences for the degree of translational complexity that may be correlated with the dimension of direction.

Then, the results reveal a clearer contrast between the two investigated text types: while predictable grammatical (de)specification has been found in only 0,2% of all law data, it is identified in 1,7% of the string pairs extracted from fiction texts (cf. table 6.4). Another fact about the frequency of the progressive, reported by Quirk et al. (1985: 198), is relevant in this connection: "The same count shows that progressive forms are more frequent in conversation than in scientific discourse..." The investigated law texts are not the same text type as "scientific discourse", but they do not contain passages of dialogue, which occurs in all the extracts of fiction texts which are included in our empirical material. 9 of the cases identified where the progressive aspect is used in the English text are found in reported speech. Other cases occur in sequences of internal monologue, where the story is told "...through the *words* or *thoughts* of a particular person" (Leech and Short 2007: 140), i.e. either through a first person narrator, or through another character in a third person point of view. In our opinion, the difference in restrictedness between the analysed texts of law and fiction may account for the variation observed between the two text types with respect to the occurrence of progressive aspect. As discussed in chapters 4 and 5, the unrestricted fiction texts exhibit a larger inventory of syntactic constructions, as well as of pragmatic functions, than the restricted law texts do.[38] In relation to this, the observation by Quirk et al. (1985: 198) is relevant since there will be similar linguistic differences between the two types of language use that they have mentioned. We find it reasonable that the progressive is more frequent in the fiction texts than in the law texts due to the larger degree of linguistic variation within the former text type.

[38] Cf. 4.2.2.1 and 5.4.2.3–4.

As mentioned in 6.2.4.1, systematic differences between English and Norwegian in the use of possessive determiners constitute a special case of predictable grammatical (de)specification, and is presented as a separate subtype in 6.3.1.2. Adding the number of occurrences of that subtype to the figures displayed in tables 6.4–5, gives the results shown in tables 6.6–7.

Table 6.6. Occurrences of all types of predictable grammatical (de)specification, counted within all recorded string pairs, within each direction of translation, and within each text type.

	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T
Across **all data**:	103	2,3
Across all data **E** → **N**:	34	1,6
Across all data **N** → **E**:	69	3,0
Across all **law** data:	5	0,3
Across all **fiction** data:	98	3,6

Table 6.7. Occurrences in individual text pairs of all types of predictable grammatical (de)specification.

Legal texts			Fiction texts		
Text pairs	Frequency of string pairs where the subtype is found:		Text pairs	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T		in absolute numbers	in per cent of n_T
AEEA	4	0,5	**AB**	19	3,6
			DL	11	1,4
Petro	1	0,1	**EFH**	33	4,7
			BV	35	4,9

Adding the figures emphasises, firstly, the clear difference between the two text types, and, secondly, the tendency that linguistically predictable differences in grammaticalisation occur to a larger extent in Norwegian-to-English translation than in English-to-Norwegian. This illustrates the asymmetry commented on above between grammatical specification and despecification, and within the recorded data this asymmetry is mirrored by a certain difference in translational complexity correlated with the direction of translation, as shown in table 6.6. In general terms, if a larger number of semantic distinctions are grammatically expressed in a given target language L_2 than in the source language L_1, then there will be cases where translating string a_{L1} into L_2 must involve making a choice between a set of linguistically predictable correspondents (b_{L2}, c_{L2}, ...). This is a more complex task than translating one of the strings b_{L2}, c_{L2}, ... into L_1 when a_{L1} is the only predictable translation of each of them.

6.3.1.2 Predictable differences in the use of possessives

DESCRIPTION. Within the analysed data, a special case of predictable grammatical (de)specification is caused by a systematic difference between English and Norwegian in the use of determiners in definite noun phrases referring to what may loosely be called objects of possession. With respect to English, *possessives* refer to the closed class of words *my*, *your*, *his*, *her*, *its*, *our*, *their*, categorised by Quirk et al. (1985: 256, 361) as possessive pronouns with determinative function. In Norwegian the possessives are *min*, *din*, *hans*, *hennes*, *dens*, *dets*, *sin*, *vår*, *deres*. These are described as a subcategory of determiners by Faarlund et al. (1997: 23), who also state that the primary syntactic function of determiners is to premodify nouns, and that their semantic function is to specify the reference of the noun phrases in which they occur. These facts apply likewise to English possessives as determiners; cf. Quirk et al. (1985: 253–256).

The translational correspondences between the italicised NPs in (10) illustrate the phenomenon to be discussed:

(10a) She dragged *her backpack* by *its strap* after her ... (DL)
(10b) Hun dro *ryggsekken* med seg etter *skulderremmen* ...
'She dragged backpack.DEF with self after shoulder-strap.DEF.'

The four noun phrases *her backpack*, *its strap*, *ryggsekken* and *skulderremmen* in (10) have in common that they refer uniquely to specific entities.[39] In the cases of the English NPs, unique reference is expressed through the possessives *her* and *its*, and in the Norwegian phrases through the suffix *-en*, which is a marker of definite form.[40] Compared with the definite markers in the Norwegian NPs, the English possessives *her* and *its* provide a larger amount of information since they not only specify the reference of the NPs, but also encode relations of possession: in (10a) the possessive *her* signals that the backpack referred to belongs to the referent of the pronoun *she*, and the possessive *its* shows that the referent of *strap* belongs to the mentioned backpack in the sense of being one of its parts. *Her* and *its* are obligatory markers of these possessive relations: if the markers are not used, the relations are not asserted. E.g., in neither of the sentences *She dragged the backpack by the strap* or *She dragged a backpack by a strap* is it necessarily true that the backpack belongs to the subject referent and that the strap is a part of the backpack, although these are possible interpretations. To conclude, the non-correspondence between (10a) and (10b) in the use of possessives is a semantic difference caused by a deviation in the amount of grammatically expressed information.

In this section we will show that this semantic deviation between (10a) and (10b) is predictable from information about regularities in the two language systems. It is our view that (10) instantiates a pattern where the possessive relations which are explicitly encoded in (10a) are implied in (10b): the definite form of the Norwegian NP *ryggsekken* implies that this expression refers to an item that belongs to the subject referent, and the definite form of *skulderremmen* implies that its referent is a

[39] In the present discussion we disregard the difference in semantic granularity between the two lexemes *strap* and *skulderrem* ('shoulder strap'). This topic is treated in 6.3.1.3. Although (10) illustrates a predictable difference in the use of possessives, other factors make it impossible to assign type 3 to the entire correspondence, and that *strap* is a hyperonym of *skulderrem* is one of these factors.
[40] The Norwegian suffix *-en* also marks singular number and common/masculine gender, which are grammatical features of the nouns *ryggsekk* and *skulderrem*, respectively.

part of a known entity, i.e. the backpack. The examples in (10) indicate that the encoding of such relations of possession requires the use of possessives in English, and the use of definite form, but no possessive, in Norwegian. If, e.g., the possessive is omitted in English, or, indefinite form, or a possessive, is used in Norwegian, it will change the intended meaning or create non-idiomatic expressions. If this pattern reflects linguistic regularities, the challenge is to find precise criteria for identifying the classes of nouns that follow the regularities in each language.

This difference between English and Norwegian in the use of possessives illustrates the point made in 6.3.1.1 that there is an asymmetry in terms of translational complexity between cases of grammatical despecification and cases of grammatical specification. Reversing the direction of translation in example (10) would increase the translator's need for information: generating the appropriate English noun phrases *her backpack* and *its strap* from the Norwegian sentence would require the derivation of explicit information about the possessive relations which are only implicit in the Norwegian sentence (10b). Thus, information about the relation of possession is less easily accessible in the Norwegian string than in its English correspondent.

There are certain differences between the type of grammatical (de)specification described here and the class presented in 6.3.1.1. In the case of the latter, there are systematic correspondences between single expressions in one language and sets of expressions in another language because a certain semantic distinction is grammatically expressed in only one of the languages. In this category there is normally only one appropriate expression in each language, if the same possessive relation is to be conveyed by both strings. Moreover, in the present type it is not the case that some information that is grammaticalised in English is absent from Norwegian; it is just not encoded in an equally explicit way.

With respect to the issue of linguistic predictability, and hence computability, we need to identify characteristic linguistic properties of the correspondences that follow the pattern illustrated by (10). On the basis of the cases found within the recorded data, it is our view that a subset of the cases constitute a prototypical kernel exhibiting certain properties. These canonical cases can be regarded as manifestations of a

translational correspondence between rules of English and Norwegian grammar, respectively. Then, there are other, more peripheral, instances of the pattern where one or more of the characteristic properties are missing. Whether those correspondences conform with the pattern or not seems to a larger degree to be determined by *parole*-related factors than what is the case in the prototypical instances. The characteristic properties of the canonical cases pertain to surface form as well as to semantic content. Concerning the formal properties, we have already observed that in this pattern a possessive determiner is obligatory in the English noun phrase, whereas the Norwegian noun phrase is typically in the definite form.[41] With respect to the semantic properties, the core cases involve a human possessor and a so-called inalienable possessee (i.e. object of possession).

Possession. It is relevant to consider the general notion of 'possession'. Heine (1997: 1) describes the concept of 'possession' as "inherently vague or fuzzy", and the set of relations that can be expressed by possessive constructions is quite heterogeneous (cf. Heine 1997: 2). Various contributions have identified a range of types of possession, and through different descriptive approaches.[42] Heine (1997: 5), as well as Herslund and Baron (2001: 2), point out that most researchers have a prototypical view of this phenomenon, in that some types of possession are regarded as more central instances than others. Citing Heine (1997: 39–41), Herslund and Baron (2001: 2) list the properties of prototypical possession "... such as 'human possessor, concrete possessee, possessor having the right to use the possessee, spatial proximity between the two, no temporal limit on the possessive relation'." It is not easy to try to define what the different types of possession have in common, but, at least, possession involves a relation between two entities, and it is always clear which of them is the possessor and which is the possessee (cf. Heine 1997: 2). In this very general sense, possession can be seen as a kind of part-whole relation.

In the string pairs where we have identified predictable differences between English and Norwegian in the use of possessives, the possessor is a human in 50 out

[41] In 8 of the 54 recorded cases, the Norwegian NP is indefinite, generally because indefinite form is more idiomatic in the local linguistic context.

[42] Heine (1997: 2–6, 33–41) presents several approaches, and different types of possession. Also, Seiler (2001: 27) lists several studies on possession.

of 54 recorded instances (92,6%). In two of the deviating cases, which are found in the *AEEA* text pair, the possessor is a kind of institution (a surveillance authority, and the EEA Council, respectively), and in the other two, which are found among the fiction data, the possessor is an inanimate object (a cathedral and a backpack).[43] In three of the four deviating cases, the encoded relations of possession are kinds of part-whole relations, as illustrated by the noun phrase *its strap* in example (10). Altogether, human possessors are dominating and prototypical in this class of correspondences.

Inalienability. With respect to the semantic properties of the possessees, inalienability seems to be the canonical feature. It may, however, appear inadequate to regard inalienability simply as a semantic notion associated with nouns. In a general-language sense, 'inalienable' can be understood as the quality of being something that cannot be taken away from a person. The grammatical notion of 'inalienability' is ascribed to Lévy-Bruhl (1914), who discussed possessive constructions in Melanesian languages. According to Chappell and McGregor (1996b: 3), Lévy-Bruhl (1914: 97–98) observed that in these languages, nouns could be divided into two classes due to two different kinds of morphological possession marking. The inalienable class of nouns comprised terms for kinship, body parts, spatial relations, and certain important personal belongings, whereas all other nouns were included in the alienable class. Chappell and McGregor (1996b): 3) note that "... this dichotomy represents a basic semantic pattern that recurs across many languages, regardless of genetic affiliation or grammatical type."

However, there is considerable variation across languages with respect to which nouns that are classified, respectively, as alienable or inalienable, and culture-specific, as well as pragmatic, factors determine where the division is drawn in individual languages (cf. Chappell and McGregor 1996b: 9, Heine 1997: 11–12). Moreover, there are several languages where certain nouns may be used either as inalienable or as alienable (cf. Chappell and McGregor 1996b: 3).[44] Such facts indicate that

[43] Example (10) shows one of the non-typical instances.

[44] Examples (13) and (14) below include alienable nouns that may appear with inalienability marking (*tea cup*, *cigarette*), and the examples illustrate how this is influenced by, respectively, culture-specific and pragmatic factors.

it is not always straightforward to predict, on the basis of the meaning of a given noun, whether it is inalienable or not. Heine (1997: 17–18) argues that inalienability is not merely a lexical property; it is rather an aspect of the relation between the possessor and the possessee, and one that has structural consequences reaching outside the noun phrase itself. Some researchers regard the alienability distinction as a type of noun classification similar to grammatical gender marking, but there is not general agreement on this point (cf. Heine 1997: 15–16). In a typological perspective, inalienability is a grammatical notion with semantic content, and its effects are visible on the levels of morphology as well as those of phrasal and clausal syntax.

Kinship and body part terms. Although it can be questioned whether inalienability is a lexical property, it is our view that the canonical instances of the translation pattern discussed here can be associated with nouns denoting inalienable possessions, and among the cases identified within the recorded data, the prototypical nouns appear to be body part terms, as in (11) below, and kinships terms, as in (12). This is in line with what Dahl and Koptjevskaja-Tamm (2001: 208) have observed, applying a language-typological perspective: "It is a well-known fact that kin terms and body part terms ... are the two semantic classes that are most often treated as "inalienable" whenever alienability distinctions are made."

(11a) Hun løftet haken (BV)
'She raised chin.DEF.'
(11b) She raised her chin

(12a) Jason spør moren. (EFH)
'Jason asks mother.DEF.'
(12b) Jason asks his mother.

Seiler (2001: 28) makes the point that languages vary with respect to the marking of inalienability, and, among the strategies he mentions, *possessor suppression* and *obligatory possessor marking* are the two most relevant ones in relation to the translation pattern illustrated by (11) and (12). Norwegian follows the strategy of possessor suppression, in which noun phrases referring to inalienable possessions have

definite form and contain no possessive determiner, as shown by the NPs *haken* ('the chin'), and *moren* ('the mother') in (11a) and (12a), respectively.[45] This is quite the opposite of the English strategy, which is obligatory possessor marking, shown by the NPs *her chin* and *his mother* in respectively (11b) and (12b). For the purposes of the present discussion, we will refer to the English-Norwegian translational regularity illustrated by (11) and (12) as *the inalienability pattern*: its canonical instances involve nouns denoting kinship and body parts, typically with human possessors, and it is realised in English through obligatory possessor marking, and in Norwegian through possessor suppression.

Highly relevant to the English sentence (11b) is the observation made by Dahl and Koptjevskaja-Tamm (2001: 211) that possessive determiners are obligatory in English in the case of "subject-controlled body part terms." This captures the predictability of the instance of the inalienability pattern shown by string pair (11): since the subject NP refers to a human, and the object NP refers to a body part, definite form is obligatory in the Norwegian object NP, and the English object NP requires a possessive determiner encoding the relation of possession between the subject and the object. If we assume it to be linguistic information whether a noun or a pronoun can refer to a human, and whether a noun denotes an inalienable entity, then the translation (11b) is predictable on the basis of the information that is linguistically encoded in the original (11a) together with information about the interrelations between source and target language systems. Likewise, the NP correspondence *moren – his mother* in (12) can be described as a linguistic regularity. However, in sentence (11a) the presence of a human subject and an object referred to by a body part term does not imply that the given relation of possession is a logical necessity, and although a context where the described chin belongs to someone other than the subject referent may appear odd, it is not unthinkable. With respect to (12a), the likelihood that the subject referent "owns" the object referent is perhaps smaller than in the case of (11a):[46] the sentence *Jason spør moren* may occur in a context where

[45] Possessor suppression is also illustrated by the NP *ryggsekken* in the Norwegian sentence (10b).

[46] Alternatively, it can be said that the subject is the *anchor* of the kinship term in (12a); cf. Dahl and Koptjevskaja-Tamm (2001: 201).

the NP *moren* refers to a person who is the mother of somebody else than Jason, and then *his mother* is an appropriate English translation only if it is known that the kinship term *moren* is anchored to a male individual. Depending on who the anchor is in the given context, other possible translations could be *her mother*, *its mother*, *their mother*, or even *the mother*.

Considering kinship and body part terms in a typological perspective, Dahl and Koptjevskaja-Tamm (1998: 43–44) observe how the two categories differ with respect to the way in which the possessor (or the anchor) is identifiable, and they claim that kinship terms are "pragmatically anchored", while body part terms are "syntactically anchored". Possibly, the English-Norwegian inalienability pattern is more of a linguistic regularity in the case of body part terms than in the case of kinship terms, but on the basis of our highly limited empirical data it cannot be decided if there is any significant difference between the two categories with respect to the extent to which *parole*-related factors determine whether they occur in the pattern or not. Dahl and Koptjevskaja-Tamm (1998: 44) state that "[i]n both cases, we are dealing with highly predictable possessors", and in our opinion, both classes of nouns are included in the prototypical kernel of this translation pattern.

A prototype view. Our prototype view of this language-pair specific phenomenon is supported by typological research on inalienability. Based on empirical investigations, different linguists have tried to establish *alienability scales*, or hierarchies where the most prototypically inalienable class(es) of nouns are at the top of the scale, and gradually less typical classes follow below.[47] The tendency is that the central kinds of inalienable constructions are linked to kinship, body parts, part-whole and spatial relations, but, as Chappell and McGregor (1996b: 8–9) point out, languages vary so much with respect to the organisation of such hierarchies that a universal scale cannot be assumed.

In the string pairs where we have identified predictable differences between English and Norwegian in the use of possessives, prototypical kinds of inalienable possessions are involved in 34 out of 54 recorded instances (63,0%). Among the cano-

[47] Cf. Chappell and McGregor (1996b: 8–9).

nical cases, we have included nouns denoting concepts such as 'mind', 'life', and 'voice', which are, like kinship and body parts, non-transferrable properties of the possessor. Further, the nouns occurring in the less prototypical instances denote concepts that we have tentatively grouped into (i) clothes, parts of garments, and other things attached to the body (e.g. *jacket*, *inside pocket*, *make-up*), (ii) other objects used by humans (e.g. *backpack*, *book*, *cigarette*, *instrument*, *tea cup*), and (iii) human activities (e.g. *exercises*, *work*). The nouns observed within these three groups have in common that they may occur with inalienability marking when they are associated with only one possessor (i.e. user, or agent) at a time, because there is some kind of close connection, such as physical contact, between the unique anchor and the possessee.[48] Also, the connection in question is normally required for typical use of the possessed object. In contexts where these nouns are not marked as inalienable possessions, the criterion of a close connection tends to be absent.[49] Within the recorded data, these facts seem to apply in both of the investigated languages, but the empirical material is too limited to conclude that there is full correspondence between English and Norwegian on this point. The generalisation across the three groups of concepts is supported by typological studies of alienability splits in connection with clothing and related notions: in several languages inalienability marking occurs in possessive constructions referring to clothes and similar objects *that are worn*, whereas the same objects are treated as alienable possessions when they are not attached to a person's body (cf. Heine 1997: 17–18).

Non-prototypical cases. An example may illustrate that the computing of the target text can be more complex in the non-prototypical cases of the inalienability pattern than in the canonical ones. In the sentence pair (13) the pattern is instantiated by the italicised NPs:

(13a) Hun drakk stadig av *tekoppen*.
'She drank continually from tea-cup.DEF.'
(13b) She drank continually from *her tea cup*. (BV)

[48] Naturally, the possessor/anchor is not necessarily an individual; it may also be a group.
[49] That this is a tendency, and not a rule is shown by example (14b) where *cigarette* is used without inalienability marking in spite of physical contact with the possessor.

We regard strings (13a) and (13b) as semantically equivalent except for the occurrence of grammatical specification through the possessive determiner *her* in (13b). The translationally corresponding NPs *tekoppen* and *her tea cup* carry inalienability marking although they refer to an alienable concept. 'Tea cup' is a kind of object typically used by humans, and in the given context these NPs refer to an object that is used the subject referent, and by nobody else. Because of this close connection between the tea cup and its possessor, the cup is similar to clothing, and the use of inalienability marking (by means of, respectively, possessor suppression and obligatory possessor marking) is idiomatic in both sentences (13a) and (13b). The example also illustrates the influence of culture-bound factors on alienability splits: in cultures where people do not share items used for drinking and eating it is natural that expressions referring to such objects are marked in the same way as those referring to important personal belongings and to clothes being worn.

With respect to the translational correspondence (13), the predictability issue relies, in our view, on whether it is linguistically expressed in the source sentence (13a) that the referent of the subject (*hun*) is the only user of the referent of the object (*tekoppen*), and that the connection between them is of the kind required for the typical use of tea cups. This may illustrate the point discussed in 2.3.2 that the limit of predictability in translational relations is determined by where the division is drawn between linguistic and extra-linguistic information. In 2.4.2.1 this distinction is further linked with the task of defining the kinds of information that are included in formal representations of language systems, an issue that relies on chosen presuppositions concerning the design of language descriptions. If the pieces of information by means of which the possessive relation in sentence (13a) may be inferred are regarded as general world information, then we cannot classify (13b) as a linguistically predictable translation. In our view, however, the various pieces of information needed in order to identify the possessive relation in (13a) may be included in the linguistic information contributed by the different words in the sentence, and by its syntactic and semantic structure. E.g., lexical information provided for the verb *drikke* may include the information that there is physical contact between the agent and the liquid being drunk, and, further, we assume it is part of the lexical

information associated with the noun *tekopp* that it denotes objects that may contain liquid, preferably tea, and that are typically used by humans in order to drink. Here we will not go into further detail, but we regard it as most likely that such pieces of information can be included in formal language descriptions.[50] Deriving the implicit possessive relation in sentence (13a) requires a thorough linguistic analysis, including inferences performed on various pieces of information, and the NP correspondence *tekoppen – her tea cup* is translationally more complex than prototypical cases, such as the correspondence *haken – her chin* in (11). Correlated with a larger need for inferencing, there is reason to believe that generating the appropriate English target expressions requires a larger amount of processing effort in the non-canonical instances than in the core cases. But as long as all pieces of information needed for deriving the possessive relations are available within the corresponding translational units, such correspondences are in principle linguistically predictable.

Non-computable instances. The recorded data include several occurrences of the inalienability pattern that fall outside the computable domain due to the way in which the investigated texts have been segmented into translational units. This situation is most clearly illustrated by Norwegian-to-English translation. In cases where a translational unit in a Norwegian source text contains a noun phrase marked as an inalienable possession, the English translation of this NP is not linguistically predictable if the extracted source string does not include sufficient information to identify the implicit relation between the possessor and the possessee. Typically in such cases, what is missing within the source string is information about who the anchor is. This is illustrated by the pair of italicised NPs in example (14):

(14a) Sigaretten hang mellom *leppene*, (BV)
'Cigarette-DEF hung between lips.DEF.'
(14b) A cigarette hung from *her lips*,

[50] The framework for lexical representations defined by Pustejovsky (1995) is an example of a formalism that would allow this. In Pustejovsky's approach the meaning of lexical units are represented by structures encoding various semantic properties. Among these, the *telic* aspect of word meaning captures "the purpose or function of a concept" (Pustejovsky 1995: 99), and we assume that information about the typical use of a tea cup could be incorporated in the telic aspect of the meaning of *tekopp*. See Pustejovsky (1995: 99–100), where the telic aspect is illustrated by means of the English nouns *beer* and *knife*.

Given that the definite noun phrase *leppene* ('the lips') in (14a) refers to a body part, the sentence involves an implicit possessive relation, and it can be inferred that the anchor is human and an individual. Otherwise, the source sentence does not contain any information about the possessor of the body part. Since 'lips' is an inalienable concept, possessor marking is obligatory in the English translation. Thus, in order to produce the target sentence it is necessary to search in the linguistic context of (14a) for information about the possessor, so that the appropriate English possessive determiner can be chosen. In the target sentence (14b) the determiner *her* signals that the anchor is a female. This piece of information is available in the immediate context of (14a), given in (15) together with its English translation.[51]

(15a) Hun stemplet mønsteret på bomullsstoffet, de store hendene arbeidet raskt, alt så lett ut. Sigaretten hang mellom leppene, hun knep det ene øyet igjen. (BV)

(15b) She printed the pattern onto the cotton cloth, her large hands moved quickly, everything looked easy. A cigarette hung from her lips, and she screwed up one eye.

Since the translation (14b) indicates that the anchor of the body part is a female, the NP correspondence *leppene – her lips* is an instance of semantic specification in the translational relation. Because the information that is added in the target string cannot be derived from the source unit, *her lips* is not a linguistically predictable translation of *leppene*.[52]

Example (14) may also illustrate how the use of inalienability marking can be influenced by *parole*-related factors in cases where nouns denoting alienable possessions occur in the inalienability pattern. In the Norwegian sentence (14a) the definite NP *sigaretten* ('the cigarette') refers to an object associated with a single user and in physical contact with its possessor. Since 'cigarette' is an alienable concept, the Norwegian NP *sigaretten* is a non-prototypical, context-dependent case of inalienability marking. Another deviation from the canonical pattern in (14) is that the English

[51] The context given in (15) reveals that the anchor is a female individual. In order to identify the person uniquely, an even wider context is necessary.

[52] This is one among several reasons why string pair (14) is a type 4 correspondence.

translation of *sigaretten* is the indefinite noun phrase *a cigarette*, which is not a possessive construction at all. Probably, it is due to stylistic reasons that inalienability marking is avoided on the noun *cigarette*. Since obligatory possessor marking occurs with the body part term *lips*, (14b) is a better translation than the sentence *Her cigarette hung from her lips*.

Example (14) illustrates the significance of the direction of translation when the inalienability pattern occurs in correspondences where no information about the possessor is available in one of the translational units. Among the recorded data, the tendency is that these are correspondences where the English translational unit contains a larger amount of information than the Norwegian does, because the use of obligatory possessor marking in English provides information about the number, person, and gender of the possessor. In Norwegian-to-English translation, this is a non-predictable translational difference as long as no information about the possessor is implicit in the Norwegian source string. As discussed in 6.3.1.1, the translational relation between strings which differ with respect to the degree of semantic specificity may still be classified as computable if the specification is linguistically predictable from a grammatical category or feature specific to one of the languages. But since this is not the case in string pairs like (14) where anchor information is missing in the Norwegian unit, such occurrences are in principle type 4 correspondences, regardless of the direction of translation. Hence, among the recorded data, instances of this kind have been counted among the cases of non-predictable specification and despecification to be presented in 6.3.1.3.[53]

However, it may be argued that in correspondences where the Norwegian string contains no information about the possessor, instances of the inalienability pattern may still be classified as computable in the case of English-to-Norwegian translation; cf. the italicised noun phrases in string pair (16):

[53] Tentatively, we have identified 136 occurrences of the inalienability pattern where anchor information is missing in the Norwegian string, and which are hence classified as non-computable correspondences. 124 of these cases (91,2%) are found among the fiction data, whereas only 24 instances (8,8%) have been identified within the law texts. 85 of the 136 non-computable occurrences of the pattern (62,5%) are prototypical cases involving inalienable possessions, and these have all been identified within the fiction data.

(16a) *His voice* was curt, stern and pure, insisting on standards, (DL)
(16b) *Stemmen* var knapp, streng og klar, innstilt på å følge de vedtatte retningslinjene,
'Voice.DEF was curt, ...'

In (16a) the NP *his voice* refers to an inalienable possession, and through obligatory possessor marking the source expression reveals that the anchor of the described voice is a male individual. The Norwegian translational correspondent, *stemmen* ('the voice'), is semantically less specific, and no information about the possessor is available in the target sentence (16b). In cases of this kind, where an inalienable concept is referred to by a noun phrase carrying possessor marking in the English source text, the deletion of the possessive in the Norwegian translation is predictable from the information that obligatory possessor marking in English corresponds with possessor suppression in Norwegian.[54] Still, for the reasons given above, we have chosen to regard instances of the inalienability pattern of the kind given in (14) and (16) as type 4 correspondences, even if this may appear too strict in relation to the computability issue.

Although predictable differences in the use of possessives do not constitute a frequent phenomenon among the recorded data, they represent, like the case of progressive aspect, an important structural difference between English and Norwegian which must be handled in translation within this language pair. Also, they are both phenomena that highlight the division between predictable and non-predictable translational correspondences.

OCCURRENCE. As already shown by table 6.1 in 6.2.4.2, cases of predictable differences in the use of possessives are not very frequent compared with other subcategories within correspondence types 3 and 4. Within the entire set of recorded data, we have tentatively identified 54 string pairs containing computable instances of the inalienability pattern. Tables 6.8 and 6.9 present further details on how these occur-

[54] Although *stemmen* can be seen as a predictable translation of *his voice*, string pair (16) is a type 4 correspondence due to other semantic differences between source and target string.

rences are distributed along the dimensions of text type and direction of translation, and across the different text pairs.[55]

Table 6.8. Occurrences of predictable differences in the use of possessives, counted within all recorded string pairs, within each direction of translation, and within each text type.

	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T
Across **all data**:	54	1,2
Across all data **E → N**:	15	0,7
Across all data **N → E**:	39	1,7
Across all **law** data:	2	0,1
Across all **fiction** data:	52	1,9

Table 6.9. Occurrences in individual text pairs of predictable differences in the use of possessives.

Legal texts			Fiction texts		
Text pairs	Frequency of string pairs where the subtype is found:		Text pairs	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T		in absolute numbers	in per cent of n_T
AEEA	2	0,1	**AB**	9	1,7
			DL	4	0,5
Petro	0	0,0	**EFH**	16	2,3
			BV	23	3,2

Table 6.8 shows that within the analysed texts this phenomenon is more than twice as frequent among the Norwegian-to-English data than among the English-to-Norwegian ones. However, as in the case of the class of predictable grammatical

[55] The different values of n_T are presented in tables 6.2 and 6.3 in 6.2.4.2.

(de)specification discussed in 6.3.1.1, we cannot draw any conclusions regarding the significance of the direction of translation, since the investigation is so limited.

With respect to the dimension of text type, the results shown in tables 6.8 and 6.9 again show a clear difference between the law data and the fiction data. Only two instances have been identified within the law texts, and these are non-prototypical cases of the inalienability pattern. Probably, this contrast between the two text types primarily reflects differences in content. In the investigated law texts, neither persons nor inalienable possessions are described, whereas all of the selected fiction texts contain stories evolving around a set of characters.

However, the present results do not indicate that the described types of possessive constructions in general do not occur in law texts. Firstly, among the recorded law data we have identified 24 instances of the inalienability pattern which count as non-predictable correspondences because information about the possessor is not available in the Norwegian translational units. All these cases are non-prototypical, as they do not involve inalienable possessions. Secondly, the selected law texts represent a limited set of legal domains, and if data had been collected from, e.g., a text on family law, the analysis would probably have given different results with respect to this translational phenomenon.[56]

The figures presented in table 6.9 also indicate differences between individual narrative texts: a noticeably larger number of occurrences are found in Bjørg Vik's text than in the other fiction texts. It is to be expected that there will be variation between different fiction texts with respect to the extent to which they contain possessive constructions, as well as the extent to which inalienable possessions are referred to, because such factors will be determined by the content of each story.

6.3.1.3 Non-predictable specification and despecification

DESCRIPTION. If there is a non-predictable difference with respect to the amount of linguistically encoded information between a sequence of words and its translational correspondent, this means that the difference cannot be accounted for solely by

[56] Cf. the discussion in 5.5.1.2 on legal sub-domains in relation to the investigated law texts.

means of the linguistic information contained in the source expression, together with information about the interrelations between the grammars and lexicons of the two languages.

Specification. String pair (17) shows an example of non-predictable specification. This is a pair of simple matrix sentences which has been recorded as a type 4 correspondence. (17a) is a passive sentence, whereas (17b) is an active sentence.

(17a) This could be seen through the broken window just above them on the first floor. (DL)

(17b) Hun kunne se dette gjennom den istykkerslåtte glassruten i vinduet rett over dem, i annen etasje.
'She could see this through the broken window-pane.DEF in window.DEF straight above them, in second floor.'

In the translation (17b) we may identify three linguistic signs which express information not present in the original. Firstly, the active target sentence (17b) contains the pronoun *hun* ('she'), whose referent fills the agent role in the situation described. In the passive source sentence (17a) the agent role is not expressed. Secondly, the noun *glassruten* ('the pane of glass'), and, thirdly, the preposition *i* ('in') in (17b) have no correspondents in (17a). Given the purposes of our investigation, the important issue now is to identify what types of information that may account for the semantic specification in the translation.

The addition of the pronoun *hun* follows from the change from passive to active voice in example (17). We cannot know why the translator chose an active sentence in the target text, but the reason is possibly that in this context passive voice is in Norwegian perceived as somewhat too formal for the given kind of literary text, and has been discarded in order to avoid a stylistic effect that does not suit this text type. Passive voice in the English source sentence is, according to the judgment of a native speaker, not regarded as stylistically formal. Hence, the voice change may be seen as a translational choice influenced by the translator's knowledge about how readers of English and Norwegian narrative fiction texts may perceive the use of, respectively, active and passive constructions. Thus, the choice follows from information about stylistic features, which in the given example are text-type specific as well as

language-specific. This is an example of information about textual norms, which, according to the typology of chapter 2, belongs to the given, general sources of extra-linguistic information for translation.[57]

Two observations may support this account of voice change in example (17). Firstly, since voice change in this case concerns the way in which the texts are received by their audiences, it may have been chosen in order to maintain what Koller (1992) describes as pragmatic translational equivalence (cf. 1.4.1.1). Secondly, some results of an investigation documented by Johansson (2007: 197–215) indicate that in English-to-Norwegian translation, passive-to-active conversion is noticeably more common than changes from the active voice to the passive. The study deals with subject changes in Norwegian translations of an English short story and an English scientific article, and among the observations are cases of voice conversion between translationally corresponding sentences. In the overall majority of the analysed sentence pairs, there is no voice alternation, but the changes that have been found, largely involve passive-to-active conversion, and not active-to-passive; cf. Johansson (2007: 200–201).[58] He concludes that a preference for passive constructions appears to be stronger in English than in Norwegian (2007: 214), and our view of the passive construction as stylistically appropriate in the English sentence (17a), but stylistically marked if used in a Norwegian translation of (17a), is compatible with this.

The addition of the pronoun *hun* in (17b) is caused by the change from passive to active, and the use of active voice requires filling the agent role in the described situation, but the passive source sentence (17a) does not contain the information needed to identify a unique discourse referent which the agent role can be anchored to. The use of the pronoun *them* in (17a) indicates that in a possible interpretation of the source sentence, a group of persons is present in the described situation, and this is a candidate referent for the agent role of the relation expressed by the verb *see*.[59]

---

[57] On textual norms, cf. 2.4.2.1–3.

[58] In the case of the short story, there are 12 instances of English passives translated into Norwegian active sentences, whereas only 2 English active constructions are changed into passives. With respect to the scientific article, 86 English passives are changed into the active voice in Norwegian, while 12 English active sentences are converted into Norwegian passives. Cf. Johansson (2007: 201).

[59] Other interpretations are possible: the referent of the pronoun *them* is not necessarily human, nor necessarily animate, but due to the plural form of the pronoun, it must be a group.

However, the choice of the singular pronoun *hun* in the translation (17b) reveals that the agent role is filled by a female individual, and this information is contained in the linguistic context immediately surrounding the source sentence (17a); cf. (18):

(18) "I should think, 1910," said Alice, "look how thick the walls are." This could be seen through the broken window just above them on the first floor. She got no response, ... (DL)

Thus, the information that contributes to the addition of the linguistic sign *hun* in (17b) is not available in the source string (17a), but in a wider linguistic context. According to our typology, this piece of information can be described as linguistic, task-specific, contextual information (cf. 2.4.2.1–2).

With respect to the addition of the linguistic signs *glassruten* and *i* in (17b), this relies on general, extra-linguistic world information: normally, a window includes a pane of glass (*glassrute*), and when the source text says that a window is broken, the translator can infer from world knowledge that it is a pane of glass which is the broken part of the window, and this is the information encoded in the Norwegian expression *den istykkerslåtte glassruten i vinduet rett over dem* ('the broken pane of glass in the window just above them'). Thus, the Norwegian noun *glassruten* specifies which object is broken, and the preposition *i* specifies where this object is located. That the translator has chosen to make the target text more semantically precise than the source text by supplying this description, is probably a consequence of the general tendency of explicitation in translated texts (cf. 5.3.2).

The conclusion for example (17) is that the target sentence is not a linguistically predictable translation, since producing it requires access to world information about windows, as well as to contextual information identifying the agent of the described situation. Neither of these pieces of information are available in the SL expression, nor in the given, general information about source and target language systems and their interrelations. Moreover, since the added information in (17b) is expressed through lexical signs, (17) is an instance of non-predictable lexical specification.

Despecification. Example (19) contains a subcorrespondence exhibiting an instance of non-predictable despecification. (19) is a pair of complex matrix sentences,

and we will focus on an embedded string pair of noun phrases, given in italics, and shown in (20).

(19a) Den gir ikke enerett til undersøkelser i *de områder som er nevnt i tillatelsen* og heller ikke fortrinnsrett ved tildeling av utvinningstillatelse.(*Petro*)
(19b) It does not give any exclusive right to explore in *the areas mentioned in the licence* nor any preferential right when production licences are granted.

(20a) de områder som er nevnt i tillatelsen (*Petro*)
'the areas which are mentioned in license.DEF'
(20b) the areas mentioned in the licence

String pair (20) is recorded among our data since (20a) constitutes a translational unit according to criterion (1c) in 4.3.2: (20a) is a noun phrase containing a relative clause as syntactic complement, and it is extracted together with its translational correspondent (20b), which is a noun phrase containing a nonfinite verb phrase as syntactic complement. There is one semantic difference between (20a) and (20b): in the Norwegian phrase temporal information is expressed by the present tense of the auxiliary verb *er*, whereas no temporal information is linguistically encoded in the English translation. Thus, there is one grammatical sign, present tense, which is found in the original, but not in the translation, and, hence, string pair (20) is an example of despecification. It can also be said to be a minimal example of a type 4 correspondence, since there is only one semantic difference between source and target string.

It should be noted that the temporal difference between (20a) and (20b) is not an instance of non-equivalence in grammatically expressed tense, as presented in 6.2.4.1. In cases of that kind, translationally corresponding verb phrases have conflicting tense markers, whereas in (20) the difference is that one of the verb phrases is without any temporal feature.

Like (17), (20) illustrates the point made in 2.4.2.1 that textual norms can be language-specific. A literal translation of (20a) into English could be *the areas which are mentioned in the licence*, but in relation to the given text type, the nonfinite expression *the areas mentioned in the licence* appears to be the preferred stylistic

choice, for several reasons. Firstly, the expression has a non-personal style since it does not contain any active, finite verb, and this is suitable in a law text, which is a formal document. Secondly, the use of a past participle construction, rather than a relative clause, has the effect of condensing the text, and brevity and precision is in line with the norms governing the domain of law writing (cf. 5.4.2.1). Thirdly, the omission of temporal information in the English past participle construction does not reduce the amount of information conveyed to a recipient with access to the context of (20b). The immediate linguistic context of (20b), as shown in (19b), contains a finite verb expressing present tense. Hence, that the temporal scope of the described state of affairs covers that of the utterance situation is a piece of information derivable from the matrix sentence. This information is also implicit in the extra-linguistic context of the source string (20a): granted that the original is a law text, it may be generally assumed that what is expressed in this act holds simultaneously with the period of its application, which is, in a sense, its time of utterance. This is an assumption concerning the pragmatic function of the law text, and it is in line with the point made in 2.4.1.2 that background information available to the recipient contributes to determining the informational content received from a specific signal.[60]

String pair (20) is an instance of what we have described in 5.2.2 as the *nonfinite-finite pattern* in the language pair English-Norwegian. As pointed out there, translational links between English nonfinite constructions and Norwegian finite clauses exhibit a certain regularity which follows from information about the two language systems, primarily because finite and nonfinite constructions may be associated with corresponding types of syntactic functions in the two languages. This is seen in (20): the relative clause *som er nevnt* in (20a), and the *-ed*-clause *mentioned in the licence* in (20b) are both postmodifiers to the nouns that precede them. Still, (20b) is not an obligatorily chosen translation, because the finite alternative (*the areas which are mentioned in the licence*) also follows from information about the two language systems, and is, moreover, a literal translation, since an NP with a

[60] The assumption is compatible with Bower's observation of "the declarative illocutionary force of a statute which is, pragmatically, always speaking and heard in the reader's present time" (1989: 241). Also, the assumption is congruent with speech acts typical of law texts; cf. 5.4.2.4.

relative clause can in this case share a maximum of the meaning properties of the source expression (cf. 2.3.2). In the given text type, (20b) has the preferred stylistic characteristics of English law texts. The piece of information that governs the choice is information about textual norms, i.e. information about the stylistic norms of law writing. At a more general level, it may also be viewed as information about stylistic norms applying to several formal, non-fictional text types, as correspondences between NP-internal relative clauses in Norwegian and NP-internal past participle constructions in English are not specific to law texts. Since the information that is used to select the nonfinite translation (20b) is not contained in the source string, nor in general, given information about SL and TL and their interrelations, (20b) is not a linguistically predictable translation of (20a). We regard stylistic phenomena as belonging to the level of language use, and, hence, string pair (20) is an example where *parole*-related factors have influenced the choice of translation.[61] The conclusion is that (20) shows a case of non-predictable grammatical despecification, because it is the presence and absence of a grammatical sign that has caused a difference between the two strings in the amount of linguistically expressed information.

Example (20) is another illustration of the point made in 2.4.2.1 that the distinction between the linguistic and the extra-linguistic domains of information is relative to the way in which language systems are conceptually individuated. If we regard the given law texts as created within specific sublanguages of respectively Norwegian and English, then the correspondence between Norwegian NPs with present tense relative clauses as syntactic complement and English NPs with nonfinite past participle constructions as syntactic complement could be included in the translational relation between those two sublanguage systems. This would be a language description tailored to the domain of law texts. As explained above, the information that the temporal scope of the described state of affairs covers that of the utterance situation is, in the case of (20b), derivable from the pragmatic function of the law text. In an assumed sublanguage analysis, this information would be associated in the English sublanguage grammar with the rules specifying the type of complex NP

[61] Leech and Short (2007: 9–11) define 'style' as a phenomenon of language use; cf. 4.2.1.3. In the present study, the importance of *parole*-related factors has previously been mentioned in 2.3.1–2, 3.3.5.2, and 5.2.2.

instantiated by *the areas mentioned in the licence*. Hence, this would then be a piece of linguistic information, available at the level of *langue*, and it would match the temporal information expressed by the present tense verb in (20a), so that (20b) would be a linguistically predictable translation of (20b). However, in the analysis presented above, it is seen as extra-linguistic information related to the level of *parole*. Our point of departure is the domain of general language; it is not limited to restricted domains (cf. 2.3.2), and for this reason we have chosen to analyse the omission of temporal information in (20b) as a non-predictable semantic difference in relation to the source expression (20a), a difference that can be accounted for by extra-linguistic information about textual norms.

Distinctions. The discussions of examples (17) and (20) have illustrated the distinction introduced in 6.3.1 between correspondences where the difference in the amount of expressed information pertains to the presence or absence of lexical signs (*lexical (de)specification*, as in (17)), and cases where the difference pertains to the presence or absence of grammatical signs (*grammatical (de)specification*, as in (20)). Moreover, we have identified special classes of predictable grammatical (de)specification, presented in 6.3.1.1–2. Since all these subtypes are categories that must not necessarily be associated with entire pairs of translational units (cf. 6.3.1), various kinds of (de)specification may occur within one and the same string pair.

With respect to lexical (de)specification, a further distinction has been observed among the compiled data. While the quantitative difference in information is in some cases a deviation between source and target expression in the *number* of lexical signs (cf. (17)), there are other cases showing a difference in semantic *granularity* between translationally corresponding lexical signs. This will be illustrated below.

Clearly, there is great diversity concerning the ways in which translationally corresponding text units may differ with respect to the amount of linguistically expressed information. For one thing, the discussions of (17) and (20) have shown that in some cases the semantic deviation between source and target string appears to be greater than in others. In our view, the semantic differences between (17a) and (17b) are to a larger extent determined by linguistic choices made by an individual translator than what can be seen in (20), where the difference is related to a fairly

systematic pattern in the translational relation between Norwegian and English. Moreover, in some cases where source and target string differ with respect to the presence or absence of lexical signs it may appear as inadequate to describe the non-correspondence simply as lexical (de)specification. The addition of the lexical sign *hun* ('she') in (17b) is a case in point: introducing a linguistically expressed agent in the translation is not merely an addition of one lexical sign; it is a linguistic change with consequences for the level of predicate-argument structure, as well as for the levels of syntactic structure and functions.

This indicates that the topic of (de)specification in translational correspondences concerns a wide range of linguistic phenomena, and the present discussion does not aim at an extensive description of it.[62] The intention is merely to present a few interesting observations, and we do not regard the compiled empirical data as a sufficient basis for a comprehensive investigation of ways in which translationally corresponding text units may differ with respect to the amount of linguistically encoded information. On the basis of the available data we maintain the distinction between grammatical and lexical (de)specification, since these categories appear useful for the purpose of describing certain recurring patterns in translational correspondences, and since the distinction is fairly easy to identify.

Cases of lexical and grammatical (de)specification can alternatively be seen as examples of cross-linguistic denotational differences, which are presented in 6.3.2 as a separate subcategory. The point has previously been made that our categorisation of semantic subtypes within the main types 3 and 4 should be seen as tentative (cf. 6.2.3). However, we want to regard as (de)specification cases where it is unproblematic to identify either a difference in the number of signs expressed in corresponding translational units, or a difference in semantic granularity between translationally corresponding lexical signs.

Granularity. Example (21) illustrates a case where translationally corresponding lexical signs differ in terms of semantic granularity.

[62] Cf. the wide range of phenomena mentioned in connection with explicitation in translation studies; see 5.3.2.

(21a) Fru Bendixen hadde sydd klær til skuespillere i flere filmer. (BV)
'Mrs Bendixen had sewn clothes to actors in several films.'
(21b) Mrs Bendixen had made clothes for actresses in several films.

(21) is a pair of simple matrix sentences, and it contains two instances of cross-linguistic differences in semantic granularity: firstly, the correspondence between *hadde sydd klær* ('had sewn clothes') and *had made clothes* is a case of despecification, and, secondly, the correspondence between *skuespillere* ('actors') and *actresses* is an example of specification.[63]

Concerning the example of despecification in (21), there is a higher degree of semantic granularity in the Norwegian expression than in the English one. While the notion of 'making' clothes is present in both of the expressions, the Norwegian collocation specifies one technique by which clothes are made, whereas the English one is unspecified with respect to production method. Thus, the Norwegian phrase *sy klær* ('sew clothes') is hyponymic to the English *make clothes*.[64] Because the production method is specified, there are fewer possible interpretations of utterances of *sy klær* than of utterances of *make clothes*, and in this manner the Norwegian original provides a larger amount of information than the English translation does. The expression *sew clothes* is a literal, and hence linguistically predictable, translation of the Norwegian expression *sy klær*. To bilingually competent speakers of English and Norwegian the phrase *make clothes* can also appear as an acceptable translation, since it is readily inferred from general world information that when clothes are made, they are most likely produced by sewing, although other techniques are also possible. But *make clothes* cannot be described as a linguistically predictable translation since the target language system in this case offers an alternative expression which equals the source expression with respect to semantic granularity.[65]

In the example of specification in (21), the Norwegian expression *skuespillere* is hyperonymic to the English expression *actresses*. There is a high degree of parallel-

[63] (21) illustrates the point made in 6.3.1 that since the phenomena of specification and despecification must not be associated with entire translational units, both kinds may be instantiated within one and the same string pair.
[64] If *sy klær* is a hyponym to *make clothes*, then *make clothes* is a hyperonym to *sy klær*.
[65] A native speaker has confirmed our judgments of the given English expressions.

ism between Norwegian and English with respect to the word pairs *skuespiller – skuespillerinne* and *actor – actress*. Both the Norwegian morpheme *-inne* and the English morpheme *-ess*, express the meaning component 'female'. In Norwegian, *skuespillerinne* is derived from *skuespiller*, and, in English, *actress* is derived from *actor*. In order to refer to a male actor, the correct choice is *skuespiller* in Norwegian, and *actor* in English, but in both languages the unmarked word here is the word without the derivational morpheme, so that each of the translationally parallel expressions *noen skuespillere* and *some actors* can refer to groups including both male and female members. Thus, the degree of semantic granularity is higher in the words *skuespillerinne* and *actress*, which can only be used about females, than in the words *skuespiller* and *actor*, which may be used without specifying the sex of the referent, at least in the plural form.

However, there is also some deviation between Norwegian and English in relation to these word pairs. In order to refer to a female actress, the most likely choice in Norwegian is *skuespiller*, as *skuespillerinne* is now regarded as a rather archaic word. In English, such connotations are not to the same degree associated with the word *actress*, which is in general use.[66] Thus, in Norwegian singular as well as plural forms of *skuespiller* may refer to both males and females. In English, on the other hand, *actor* in the singular would most likely refer to a male individual, but could also refer to a female individual, and *actor* in the plural refers to groups that may include members of both sexes.

From these observations it follows that in (21a) it is not linguistically expressed whether the referents of the Norwegian plural NP *skuespillere* is a group of males, of females, or of both, whereas the English plural NP *actresses* can only refer to a group of females. In this sense the target expression *actresses* is semantically more specific,

---

[66] With respect to English, some language users prefer gender-neutral terms like *actor* also when referring to female individuals, because the use of non-neutral terms like *actress* is seen as sexist, but this view is not shared by all speakers. In Norway during the 1970ies, it gradually became a widespread opinion that gender-marked expressions were politically incorrect, and after not many years, avoiding non-neutral terms had become conventionalised among most members of the Norwegian language community. For decades now gender-marked terms ending in *-inne* have largely been discarded, so that female individuals are normally referred to by means of the corresponding gender-neutral terms. This has been general practice for so long that if *skuespillerinne* is today not included in the active vocabulary of a Norwegian speaker, the main reason for that is that such words are now regarded as archaic, and the connotations of political incorrectness are not as strong as they used to be.

and has a smaller set of possible interpretations, than the source expression *skuespillere*. Further, the translational correspondence between *skuespillere* and *actresses* falls outside the domain of the linguistically predictable since it is not possible to predict *actresses* as the translation of *skuespillere* on the basis of linguistic information sources alone.[67]

Summing up. The discussion has shown that the shared characteristic of cases of linguistically non-predictable specification and despecification is a quantitative difference between translationally corresponding expressions with respect to the amount of linguistically encoded information. The expressions in question need not constitute entire translational units, but may be subparts of such units. Although we cannot measure this quantitative difference in mathematical terms, we want to correlate it with the way in which information can be measured as a reduction in uncertainty: in cases of (de)specification, a measure of the difference in the amount of expressed information is that the semantically most specific expression, which contains the larger amount of information, has a smaller set of possible interpretations than the semantically least specific expression. Among the compiled data, we have observed that (de)specification may be instantiated as differences between source and target expression with respect to, firstly, the number of lexical signs, secondly, the number of grammatical signs, and, thirdly, the degree of semantic granularity between translationally corresponding lexical signs. We expect that a more extensive empirical investigation of (de)specification would reveal other phenomena as well.

OCCURRENCE. Tables 6.10–13 present tentative results on the frequencies of non-predictable specification and despecification within the recorded correspondences. As previously shown by table 6.1 in 6.2.4.2, non-predictable specification is by far the most common category among the semantic subtypes that we have identified (found in 918 string pairs), and non-predictable despecification (found in 604 string pairs) is the second-most common subtype. That is, specification occurs in 20,7% of all recorded string pairs, and in 41,1% of all type 4 correspondences, while despecifi-

[67] *Actresses* is of course an appropriate translation if the information that *skuespillere* in (21a) refers to a group of females is available in the context of the source text. It is however not, and the translator may have assumed that since the character in question is a woman, the clothing she makes is most likely for females, which resulted in the choice of *actresses*.

cation is identified in 13,6% of all string pairs, and in 27,2% of all type 4 correspondences. The observations to be presented will be fairly general, as a deeper level of detail in the discussion would have required a more fine-grained analysis of the various factors involved in differences between translationally corresponding expressions in the amount of linguistically encoded information.

Table 6.10. Occurrences of non-predictable specification, counted within all recorded string pairs, within each direction of translation, and within each text type.[68]

	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in per cent of n_T	in per cent of n_4
Across **all data**:	918	20,7	41,4
Across all data **E** → **N**:	546	26,0	47,9
Across all data **N** → **E**:	372	15,9	34,5
Across all **law** data:	326	19,0	44,1
Across all **fiction** data:	592	21,7	40,0

Table 6.11. Occurrences of non-predictable specification in individual text pairs.

Legal texts				Fiction texts			
Text pairs	Frequency of string pairs where the subtype is found:			Text pairs	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in % of n_T	in % of n_4		in abs. numbers	in % of n_T	in % of n_4
AEEA	199	25,2	49,1	**AB**	82	15,7	39,4
				DL	265	33,5	50,3
Petro	127	13,8	37,9	**EFH**	104	14,8	28,5
				BV	141	19,9	37,2

[68] The different values of n_T and n_4 are presented in tables 6.2 and 6.3 in 6.2.4.2.

Table 6.12. Occurrences of non-predictable despecification, counted within all recorded string pairs, within each direction of translation, and within each text type.

	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in per cent of n_T	in per cent of n_4
Across **all data**:	604	13,6	27,2
Across all data **E → N**:	240	11,4	21,1
Across all data **N → E**:	364	15,6	33,7
Across all **law** data:	255	14,9	34,5
Across all **fiction** data:	349	12,8	23,6

Table 6.13. Occurrences of non-predictable despecification in individual text pairs.

Legal texts				Fiction texts			
Text pairs	Frequency of string pairs where the subtype is found:			Text pairs	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in % of n_T	in % of n_4		in abs. numbers	in % of n_T	in % of n_4
AEEA	114	14,4	28,1	**AB**	44	8,4	21,2
				DL	82	10,4	15,6
Petro	141	15,3	42,1	**EFH**	127	18,1	34,8
				BV	96	13,5	25,3

Among the figures presented for specification as well as for despecification, the perhaps most interesting result is how the frequency of specification varies according to the dimension of direction of translation: whereas non-predictable specification occurs in 26,0% of the string pairs extracted from English-to-Norwegian translation, it has been identified in merely 15,9% of the correspondences compiled from Norwegian-to-English translation. Due to the limited scope of the present study, this result cannot in general be representative of the dimension of direction, and it is primarily an effect of the high frequencies found in two of the text pairs analysed for

English-to-Norwegian, i.e. the *AEEA* (25,2%) and DL (33,5%).[69] Still, the result is interesting because these two text pairs are not of the same text type, and we will argue below that the large occurrence of specification is caused by different factors in the two text pairs.

With respect to the dimension of text type, the frequency of non-predictable specification is somewhat higher among the fiction data (21,7%) than among the law data (19,0%). Given the strict constraints on legal translation, in particular the norm of avoiding explicitation, it is not surprising to find a larger occurrence of specification in fiction than in law text.[70] The identified difference between the text types is, however, smaller than anticipated, and it is our view that this result may be influenced by occurrences of the nonfinite-finite pattern in English-Norwegian translation, a point we will return to below.

Concerning the results found for non-predictable despecification, the picture is in a sense the opposite of the results for specification. If we consider the percentages given in relation to the total number of string pairs within each subset of the data, we may observe the following: with respect to the direction of translation, the frequency of despecification is higher within the Norwegian-to-English data (15,6%) than within those for English-to-Norwegian (11,4%), and in relation to text type, the frequency is higher among the law data (14,9%) than among the fiction data (12,8%).[71] The lower frequency of despecification in English-to-Norwegian is largely caused by the relatively small numbers of occurrences in the fiction texts representing that direction (cf. table 6.13), which may be a result of individual variation in translators' preferences. As regards the dimension of text type, it is contrary to our expectations to find a higher percentage of despecification within the law data than within those representing fiction, as the norms of legal translation aim at the preservation of meaning. Possibly, this is a result that has been influenced by occur-

[69] Cf. the point made in 5.2.2 that the empirical material is too limited to allow for generalisations.

[70] Cf. the discussions of norms in law texts (5.4.2.1), and of explicitation in legal translation (5.5.1.2).

[71] The latter result is, however, strongly influenced by the greater average string length found across the law data than across the fiction data (cf. table 5.9 in 5.4.2.6). As table 6.12 shows, the frequency of despecification is greater in fiction than in law text, if counted in absolute numbers of string pairs.

rences of the nonfinite-finite pattern.[72] Still, this is tentative, since the difference is small.

The typical situation in occurrences of the nonfinite-finite pattern is that the temporal information expressed by the finite verb in the Norwegian translational unit is absent from the English unit, as illustrated by example (20) above. In English-Norwegian parallel texts, the pattern thus creates correspondences where English nonfinite constructions are semantically less specific than their Norwegian correspondents in the cases where these are finite subclauses. Depending on the direction of translation, such correspondences are instances either of specification or of despecification. As argued in 5.2.2 and illustrated by (20), occurrences of the nonfinite-finite pattern are non-predictable correspondences since they are determined not only by the interrelations between the two languages, but also by *parole*-related factors. The "opposite" pattern may also occur, i.e. cases where a finite structure in English corresponds with a nonfinite construction in Norwegian.[73] This is less common since the use of nonfinite constructions is far more widespread in English than in Norwegian (cf. 5.2.2).

Thus, in English-Norwegian parallel texts, it is to be expected that the frequency of the opposite pattern is considerably lower than that of the characteristic nonfinite-finite pattern. This means that to the extent that translational links between nonfinite and finite constructions may contribute, in this language pair, to the frequency of, respectively, specification and despecification, we expect to find correlations between the direction of translation and the occurrence of each of these phenomena. In particular, we expect to find a higher frequency of specification among the English-to-Norwegian string pairs than among the Norwegian-to-English ones, and a lower frequency of despecification among the English-to-Norwegian correspondences than among the Norwegian-to-English ones. The results presented in tables 6.10 and 6.12 support these predictions.

---

[72] We shall see below that 69,5% of the instances of non-predictable despecification found in the law text pair *Petro* are occurrences of the nonfinite-finite pattern.

[73] An example is given below.

Since the presence or absence of temporal information is only one among several ways in which non-predictable specification and despecification may be instantiated in translational correspondences, it is necessary, in order to gauge the effect of the nonfinite-finite pattern on the frequency of these phenomena among the recorded data, to identify the string pairs where this kind of semantic difference is the only factor that has caused either specification or despecification in the translational relation.

Firstly, this has shown, as expected, that it is very rare to find the opposite pattern where English finite structures correspond with Norwegian nonfinite constructions. Only a few cases have been identified in the two pairs of law texts, and none in the fiction texts. In these string pairs the tendency is that the nonfinite unit in the Norwegian expression is, or includes, a technical expression which, at the level of lexical correspondences, has no direct match in English legal language, so that the finite string in the English expression functions as a paraphrase.[74] A small handful of such cases have been identified for both directions of translation in the law texts.

Secondly, identifying those cases where an occurrence of the nonfinite-finite pattern is the only cause of non-predictable specification, or despecification, shows that within the analysed texts, the pattern leaves a clearer imprint on the law data than on those recorded from fiction. With respect to specification in English-to-Norwegian translation, it is striking to observe in the *AEEA* law text pair that in as much as 63,8% of the identified cases (127 of 199) the change from nonfinite to finite is the only factor that has caused specification. In comparison, the corresponding figures for the two pairs of English-to-Norwegian fiction texts are 37,8% in AB (31 of 82), and 1,9% in DL (5 of 265). With respect to despecification in Norwegian-to-English translation, we have observed in the *Petro* law text pair that in 69,5% of the identified cases (98 of 141) the change from finite to nonfinite is the only factor that has created despecification. In comparison, the corresponding figures for the two pairs of Norwegian-to-English fiction texts are 15,0% in EFH (19 of 127), and 7,3% in BV (7

[74] E.g., in the Norwegian *Lov om petroleumsvirksomhet*, § 38, the nominal expression *skadelidte* ('who have suffered from damage') is translated into an English noun phrase with an embedded relative clause: *the parties that have sustained damage*. What is expressed by the English finite subclause corresponds translationally with the Norwegian segment *-lidte*, which is an inflected form of the past participle verb form *lidd* ('suffered').

of 96). That the pattern has a more visible effect on the law data than on the fiction data indicates that there is a markedly larger variety of linguistic factors involved in (de)specification in the pairs of fiction text than in those of law text. That seems plausible on the background of the larger degree of restrictedness in law text than in fiction text (cf. 5.4.2.1), and it is compatible with the observation presented in 6.2.4.2 that in type 4 correspondences extracted from the law texts, there are normally one or a few semantic deviations between the translational units, whereas in type 4 correspondences extracted from the fiction texts, there tends to be several semantic differences between source and target string.

Having analysed the impact of the nonfinite-finite pattern on the occurrences of (de)specification across the recorded data, we have some basis for supporting the claim made in chapter 5 that in the analysed texts, the pattern is probably the most important factor inducing minimal type 4 correspondences, which have been found to influence the measurement of translational complexity (cf. 5.2.2). In 6.2.4.2 we showed that well over two thirds of the identified minimal type 4 cases are found among the law data.[75] Then we have seen, through the identification of semantic subtypes, that in the two pairs of law texts, non-predictable specification and despecification are far more frequent than other subcategories within the main correspondence class 4. Hence, we will claim that because the nonfinite-finite pattern is responsible for 63,8% of the cases of specification in the *AEEA*, and 69,5% of the cases of despecification in *Petro*, and since 45,7% of all type 4 correspondences extracted from the law texts are minimal cases (cf. 6.2.4.2), we may regard the nonfinite-finite pattern as the factor that has most frequently caused minimal instances of type 4 across the recorded string pairs.

The semantic subtypes of specification and despecification are interesting in relation to the translational phenomenon of explicitation. We have argued in 6.3.1 that our notion of specification does not fully overlap with explicitation, insofar as the latter is understood as expressing explicitly in the translation information which is only implicit in the original (cf. 5.3.2). Correspondences where the presence of the

[75] Tentatively, we have recorded altogether 493 minimal type 4 correspondences, among which 338 are found among the law data, and 155 among the fiction data; cf. 6.2.4.2.

nonfinite-finite pattern is the only factor that has caused specification in English-to-Norwegian translation, do not fall within the notion of explicitation if the temporal information is seen as inaccessible to the English translational unit, which it is according to the principles defining the assignment of correspondence type (cf. 4.3.6.3). However, in real translation, such units are not treated in isolation, and since the piece of temporal information that is absent from the English nonfinite construction is easily accessible to the translator from the linguistic context, it can be regarded as implicit in the English source string. Under this view, such specification may, according to a certain definition, be included in the phenomenon of explicitation.

Thus, the question of whether specification caused by the nonfinite-finite pattern falls within explicitation or not can be reduced to a matter of definition. If, with respect to explicitation, we focus on the tendency that translators make target texts more semantically precise than the originals in order to ensure that the recipient will interpret the target text correctly relative to what the translator judges to be the intended interpretation of the source text, then the cases where specification is created by the regularity of the nonfinite-finite pattern become less interesting in relation to explicitation. What matters more are cases where semantic specification is the result of a translator's choice that has not been influenced so much by interrelations between the language systems. If we disregard the occurrences where it is only the nonfinite-finite pattern that has caused, respectively, specification in English-to-Norwegian correspondences, and despecification in Norwegian-to-English ones, then the result, across all data, is still that non-predictable specification is far more frequent (755 occurrences) than non-predictable despecification (480 cases). Given that explicitation is normal in translation, this is an expected result.

We have pointed out above that the identified difference between the two text types with respect to the frequency of specification is smaller than expected, since we have argued that due to the norms of the legal domain, the level of explicitation in translation is relatively low in parallel law texts (cf. 5.5.1.2). If we exclude the cases where the nonfinite-finite pattern has caused specification within the English-to-Norwegian data, then the average frequency of specification across the two pairs of

law text will decrease from 19,0% to 11,6%, whereas the average frequency of specification across the four pairs of fiction text will merely change from 21,7% to 20,4%. Compared with the results shown in table 6.10, this brings forth a more noticeable text-typological difference concerning the occurrences of non-predictable specification, and this is compatible with our view that the pattern reflects a systematic language difference.

In relation to the direction of translation, we have stated above that the relatively high frequencies of specification in the *AEEA* and DL text pairs are the main reason why this subtype is found to be noticeably more frequent among the English-to-Norwegian data than those of Norwegian-to-English. With respect to the *AEEA* pair of law texts, as many as 127 of 199 cases of specification (63,8%) can be attributed to the nonfinite-finite pattern. Thus, within the *AEEA* data, addition of information is caused by other factors than the pattern in only 72 identified occurrences of specification, i.e. in merely 9,1% of the total number of string pairs compiled from the *AEEA*.[76] If we hold the view that specification created by the nonfinite-finite pattern is marginal to explicitation, this observation supports the assumption presented in 5.5.1.2 that due to the special constraints applying to the translation of supranational law texts, the element of explicitation is more modest in the *AEEA* than in other types of translation. Moreover, this indicates that the nonfinite-finite pattern contributes substantially to specification in the *AEEA*, and since specification has been identified in as much as 49,1% of all type 4 correspondences in this text pair, occurrences of the pattern have most likely contributed to a larger proportion of semantic non-equivalence in the *AEEA* than expected, given the strict norms of legal language.[77]

Concerning the English-to-Norwegian fiction text pair DL, it is interesting that the nonfinite-finite pattern has caused as few as only 5 of 265 occurrences of specification (1,9%). Among all text pairs investigated, DL has the highest frequency of

---

[76] In comparison, the pair of law texts *Petro* exhibits 127 occurrences of specification, i.e. 13,8% of the total number of string pairs. Only 6 of these involve the addition of temporal information through correspondences between Norwegian nonfinite constructions and English finite clauses.

[77] The surprisingly large proportion of semantically non-equivalent string pairs within the *AEEA* data is discussed in 5.5.1.2.

non-predictable specification: it is found in 33,5% of all string pairs. This is not surprising given that DL is the text pair where type 4 correspondences cover the largest amount of the analysed texts (76,4%; cf. table 5.17 in 5.5.2.1). DL is also the text pair where specification has been identified in the largest proportion of type 4 correspondences (50,3%; cf. table 6.11). These observations indicate, firstly, that the nonfinite-finite pattern has had almost no effect on the frequency of specification in DL, and, secondly, that in this text pair there is a considerable element of explicitation which contributes substantially to the large proportion of semantic non-equivalence found among the DL data. In our view, these results reflect the individual preferences of the translator, and they support the discussion of the DL text pair given in 5.5.2.2.

These suggestions appear plausible also in the light of the other English-to-Norwegian fiction text pair, AB, which differs from all other pairs of fiction texts in that type 4 correspondences cover merely 43,9% of the analysed texts (cf. table 5.15 in 5.5.2.2). Within the four pairs of fiction texts, AB also exhibits the lowest number of occurrences of specification (cf. table 6.11). Moreover, in AB the nonfinite-finite pattern has caused 31 of 82 cases of specification (37,8% of the occurrences). This leaves 51 instances of specification involving other factors than the addition of temporal information, i.e. only 9,8% of the total number of string pairs compiled from AB. These observations suggest that in this text pair the translator has been relatively faithful to the original (cf. 5.5.2.2). The latter could also be compatible with the fact that the frequency of despecification is somewhat lower in AB than in any of the other three pairs of fiction texts (cf. table 6.13).

With respect to the data compiled from Norwegian-to-English translation, we have seen above that in the two pairs of fiction texts there is no strong effect of the nonfinite-finite pattern. In this direction we have found its clearest imprint on the data compiled from the law text pair *Petro*, where the pattern has caused 69,5% of the identified cases of despecification. If these cases are disregarded, then despecification occurs in only 4,7% of the total number of string pairs compiled from the *Petro* text pair. It is also noteworthy that since despecification has been identified in as much as 42,1% of the type 4 correspondences compiled from *Petro*, then it is clear that

instances of the nonfinite-finite pattern have contributed noticeably to the proportion of semantically non-equivalent string pairs within the *Petro* data. These observations, together with a frequency of 13,8% for specification, supports the view presented in 5.5.1.2 that the major concern in the translation of *Lov om petroleumsvirksomhet* has been to convey the original content as accurately as possible without omitting any information.

Some conclusions may be drawn on the basis of our observations of non-predictable cases of specification and despecification. Firstly, the results reflect in an interesting way the tendency that the use of nonfinite constructions is more widespread in English than in Norwegian. In chapter 5 we have seen that there is a somewhat higher degree of translational complexity across the English-to-Norwegian data than in the opposite direction (cf. 5.3.1), and in this section we have seen that that result is influenced by the nonfinite-finite pattern, since it has caused a considerable number of minimal type 4 correspondences. Secondly, the recorded data show that specification is far more frequent than despecification, also if we disregard the cases that can be attributed to the nonfinite-finite pattern. Thirdly, the extent to which the nonfinite-finite pattern contributes to, respectively, specification and despecification is considerably greater within the pairs of law texts than within those of fiction. Fourthly, having taken into account to what extent the nonfinite-finite pattern influences the level of explicitation in the translated texts, the results show considerable differences between the individual text pairs. Within the law texts, the *AEEA*, which is the most strongly norm-governed case, exhibits the most modest degree of explicitation. A stronger tendency is found in the *Petro* text pair, not so far below the level observed in a couple of the pairs of fiction texts. Within the latter text type, the results primarily reveal considerable variation among the individual text pairs, probably reflecting different preferences on the part of the translators, as discussed in 5.5.2.2.

6.3.2 Denotational non-equivalence

Sections 6.3.2.1–3 will deal with translational correspondences where we find some kind of discrepancy between source and target string with respect to denotation. The

characterisation of this category is inspired by the notion of denotational equivalence defined by Koller (1992: 216). In his approach denotational equivalence pertains to the extra-linguistic state of affairs described by the source text (cf. 1.4.1.1).

'Denotation' is defined by Löbner (2002: 25) in the following way: "The **denotation** of a content word is the category, or set, of all its potential referents."[78] Lyons (1977: 235–238) discusses denotation in a translational perspective, which we will return to below. He provides a definition of the concept that differs slightly from Löbner's: according to Lyons (1977: 207), 'denotation' applies to lexemes, and it means "the relationship that holds between that lexeme and persons, things, places, properties, processes and activities external to the language-system." The term *denotation*, as used by Löbner (2002), overlaps with the term *denotatum* in the work of Lyons, who applies the latter to "the class of objects, properties, etc., to which the expression correctly applies" (1977: 207). The difference between these two approaches is one of expression rather than of kind.

It is important to keep apart the two concepts of 'denotation' and 'reference'. According to Löbner (2002: 5), 'reference' is the very general notion that an expression is used for something. Lyons draws the distinction in the following manner: *denotation* refers to a relation that "holds independently of particular occasions of utterance" (1977: 208), whereas *reference* covers "the relationship which holds between an expression and what that expression stands for on particular occasions of its utterance" (1977: 174).

Denotation is not applied to any kind of linguistic entity. Löbner (2002: 25) ties denotation primarily to content words. It is normally not associated with sentences, but Löbner (2002: 25–26) points out that the denotation of a sentence would be understood as "the set, or category, of all situations to which the sentence can potentially refer." Lyons (1977) applies denotation primarily to lexemes, and he considers whether it can be applied also to predicative and referring expressions. In the case of predicative expressions, such as *(be) a crook*, Lyons (1977: 214) concludes that "the denotation of '(be) a crook' is the intension of the class whose

[78] In this definition the term *category* refers to a kind of entities; cf. Löbner (2002: 20).

extension is the denotatum of 'crook'." In connection with referring expressions, he argues that it is problematic to speak of denotation (1977: 214–215). In the case of personal and demonstrative pronouns, it is difficult because the conditions governing the application of such expressions are so strongly linked to language use. With respect to descriptive noun phrases (e.g. *the red car*) Lyons (1977: 215) argues that it is not the phrases but the associated predicative expressions (*(be) a red car*) which have denotation.

In a discussion of the denotation of lexemes in the context of translation, Lyons (1977: 236) points out how difficult it is to answer the question "what constitutes semantic equivalence between lexemes from different languages." The question may be hard to decide due to cultural differences, and due to discrepancies in the judgments of bilingual speakers. Lyons (1977: 236–237) argues that semantic equivalence between lexemes of two different languages is tied to the applicability of the lexemes, i.e. equivalence with respect to what entities the lexemes may apply to and in what situations they may apply. Since the denotation of a lexeme is part of the conditions governing its applicability, denotational equivalence is part of semantic equivalence. In Lyons' view "denotational equivalence is relatively independent of the cultural context" (1977: 237), and hence he regards the investigation of denotational equivalence as the most fruitful approach to studying semantic equivalence between lexemes of different languages. However, as he points out, it is often difficult for the translator to find denotationally equivalent lexemes in the target language for source language lexemes. If there is only a partial overlap, rather than a full match, between the denotata of translationally corresponding lexemes, then it is difficult to see them as denotationally equivalent. Lyons (1977: 238) concludes that "[t]he meanings of words ... are internal to the language to which they belong," and that each language has its own structure, not only with respect to grammar and phonology, but also as far as vocabulary is concerned.

In the present study it is a frequent phenomenon among the recorded data that denotational equivalence does not hold between translationally corresponding expres-

sions.[79] In contrast to Lyons' view of denotational equivalence, we want to apply the notion not only to lexemes but also to other linguistic entities. This is more in line with Koller's notion of denotational equivalence, which deals with the described state of affairs. In our framework, *denotational non-equivalence* covers deviations with respect to the denotational properties of translationally corresponding expressions. This involves the denotata of words of lexical categories, and it includes properties such as the modalities expressed by modal verbs, the potential for temporal linking expressed grammatically by temporal categories, and more. Thus, we want to apply 'denotational non-equivalence' to a fairly heterogeneous group of phenomena, and it is a deliberate choice to leave the list of relevant properties open, as it is our view that it may be wrong to assume this to be a finite set.

When translationally corresponding expressions are non-equivalent, they differ with respect to properties which contribute to the *propositional potential* of each of them. We speak of propositional potential in order to distinguish these properties from the propositional content which is expressed by a specific utterance of a sentence. This is convenient because sentences which differ with respect to denotation may in a given context express the same proposition.[80] The division drawn between propositional potential and propositional content is a parallel to the type-token distinction applied in our empirical analysis (cf. 4.3.6 with subsections). Thus, our notion of propositional potential applies to linguistic types, while propositional content applies to linguistic tokens. For the sake of convenience, we will in the following use the term *proposition* when referring to propositional content as opposed to propositional potential.

Allwood et al. (1977: 21) illustrate the fact that different sentences may express one and the same proposition, depending on the situations of utterance: "The sentence *It's Monday today* uttered on a Monday expresses the same proposition as *It was Monday yesterday* uttered on a Tuesday." Thus, the two sentences in this example may express the same proposition provided that they are uttered in the appropriate,

[79] The frequency has previously been shown by table 6.1 in 6.2.4.2.

[80] As noted in 2.4.2.1, we understand 'proposition' as designating "what a sentence says about the world" (Allwood et al. 1977: 20).

but different, contexts. Since the sentences are denotationally non-equivalent, they do not express the same proposition if they are uttered in the same context, and hence they are different with respect to propositional potential. Although the example shows that a pair of denotationally non-equivalent sentences may, depending on the context, express the *same* proposition, two sentences will differ with respect to propositional potential as long as there are any contexts in which they would express *different* propositions. Hence we define 'propositional potential' in the following way: the propositional potential of a sentence is a function from utterance situations to propositions, a function that gives the set of propositions that may be expressed by uttering a sentence in appropriate contexts.

We have here used the notion of propositional potential in connection with sentences, but in our study it is also applied to a substantial number of other syntactic categories which we accept as translational units (cf. 4.4.3). 'Proposition' is a concept borrowed from logic, and it applies basically to sentences. It would be wrong to say that all the different translational units we have identified can express propositions. But the various units have properties which, if they appear in a sentence, will contribute to the propositional content of that sentence when uttered, and these properties are what we want to include in the notion of propositional potential. There is an analogy to this point in Löbner (2002: 24) where he states that "it is not only content words that shape the descriptive meaning of the sentence. Functional elements such as pronouns and articles or tense, a grammatical form, contribute to the proposition as well ..."

Considering the points brought forward by Lyons' observations regarding cross-linguistic deviations in semantic structure, it is obvious that even between translationally linked expressions that constitute linguistically predictable correspondences, there is not necessarily full equivalence with respect to their denotational properties. In natural language, full denotational equivalence could normally be expected only in the case of translational correspondences between domain-specific technical terms. Thus, in the recorded string pairs where instances of denotational non-equivalence have been identified, the denotational deviation between translationally related expressions is large enough to exclude the correspondence from being predictable on

the basis of the interrelations between the two language systems. In the analysed parallel texts the large majority of the identified cases of denotational non-equivalence between translationally corresponding expressions have been classified as non-predictable. There is a subclass of correspondences exhibiting a certain kind of difference in denotational properties that we regard as linguistically predictable; this is presented in 6.3.2.1.

6.3.2.1 Predictable denotational differences

DESCRIPTION. Within the recorded data we have identified a category involving translationally interrelated noun phrases that deviate with respect to whether they refer to sets or to individuals. That is, if the two expressions in such correspondences are seen as linguistic types, then one of them denotes a set, whereas the other denotes an individual. In this sense the two expressions are denotationally non-equivalent and do not contribute in the same way to the propositional potential of the translational units in which they are included. As the examples will illustrate, this class comprises cases where the noun phrases share a certain type of semantic property, that of expressing generic reference.

The notion of 'generic reference' can be understood as the relation between an expression and an entire type, or a whole class of entities, represented by that expression. Generic reference occurs when referring expressions are used in utterances that predicate something about a class, or a type.[81] Quirk et al. (1985: 265) point out that since generically referring expressions apply to entire classes, the semantic distinctions expressed by the nominal categories of number and definiteness are not so important in relation to generic reference: "Singular or plural, definite or indefinite, can often be used without appreciable difference in meaning in generic contexts." Thus, all possible forms of English noun phrases can have generic reference. The situation is similar in Norwegian: Faarlund et al. (1997: 292) state that both singular and plural, and definite as well indefinite noun phrases may occur with generic reference in Norwegian; even bare nouns may do so.

[81] According to Lyons (1977: 193–194), expressions with generic reference are used in sentences that assert *generic propositions*, which are timeless and describe a class.

Although English and Norwegian have in common that generically referring expressions may be either in the singular, plural, definite, or indefinite form, it is likely that there are different factors in the two languages determining the choice of form in specific instances of generic reference. In the analysed parallel texts, we have observed deviations with respect to grammatically expressed number between translationally related noun phrases used in generic contexts. Cf. example (22), which is a pair of complex noun phrases:

(22a) dwellings that had simply grown together like incrustations and agglomerations of shells on *rocks* (AB)

(22b) boliger som rett og slett hadde vokst sammen lik lag og klumper av skjell på *en klippe*
'dwellings which simply had grown together like layers and lumps of shells on a rock'

(22) is recorded as a string pair because each noun phrase contains a finite, relative clause as syntactic complement (cf. 4.3.2.3). In spite of the number difference between the plural indefinite NP *rocks* and the singular indefinite NP *en klippe*, we have classified the correspondence between them as linguistically predictable because we regard the semantic deviation between the them as falling within the domain of the linguistically predictable. In example (23) there is a similar difference between the singular expression *innretningen* ('the installation') in (23a) and the plural *the installations* in (23b).[82] In this case the phenomenon involves definite NPs:

(23a) *Innretningen* må ikke volde urimelig ulempe for rettighetshaver. (*Petro*)
'Installation.DEF must not cause unreasonable inconvenience for licensee.'

(23b) *The installations* must not cause unreasonable inconvenience to the licensee.

Since the NPs *rocks*, *en klippe*, *innretningen*, and *the installations* in (22) and (23) are generically referring, they are pairwise coreferential. We regard them as contributing

[82] The difference in definiteness between the indefinite *rettighetshaver* and *the licensee* in (23) illustrates a phenomenon to be discussed in 6.3.3.1.

to differences in propositional potential between the translationally corresponding sentences, because the NPs are denotationally non-equivalent if seen in isolation.

It falls outside the scope of the present project to investigate in detail what factors determine the form of generically referring expressions in respectively English and Norwegian. For instance, we have not studied whether such NP correspondences will always exhibit equivalence with respect to definiteness. At this point we will merely observe that within the recorded data, the tendency is that when translationally related NPs in generic contexts differ with respect to grammatically encoded number, then the English NP is a plural expression, and the Norwegian one is in the singular form. There are also several cases where a singular English NP corresponds with a plural Norwegian NP, but the former pattern is dominating among the identified instances. Hence, we regard this as a predictable regularity included in the interrelations between the two language systems. In our view, there are two reasons why such correspondences can be viewed as linguistically predictable. Firstly, in such cases, the information that generic reference is expressed is available within the source string; this is included in task-specific linguistic information about reference relations holding between expressions in the source sentence and entities in the world (cf. 2.4.2.2). Such information is given through the relevant interpretation of the source string. Secondly, the set of linguistically predictable TL expressions that may express generic reference is small and finite in these cases; it is limited to the set of possible word forms of predictable noun correspondents. In this sense the category of predictable denotational non-equivalence parallels predictable grammatical (de)specification, described in 6.3.1.1.

OCCURRENCE. As shown by table 6.14, we have tentatively identified 115 string pairs exhibiting predictable denotational differences of the kind described above. This is merely 2,6% of the total number of recorded correspondences. Although this is not a very frequent subtype, the distribution of its occurrences reveals a clear contrast between law and fiction texts, and it also indicates an interesting difference between the two pairs of law texts.

Table 6.14. Occurrences of predictable denotational differences, counted within all recorded string pairs, within each direction of translation, and within each text type.[83]

	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T
Across **all data**:	115	2,6
Across all data **E → N**:	23	1,1
Across all data **N → E**:	92	3,9
Across all **law** data:	102	6,0
Across all **fiction** data:	13	0,5

Table 6.15. Occurrences of predictable denotational differences in individual text pairs.

Legal texts			Fiction texts		
Text pairs	Frequency of string pairs where the subtype is found:		Text pairs	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T		in absolute numbers	in per cent of n_T
AEEA	15	1,9	**AB**	8	1,5
			DL	0	0,0
Petro	87	9,4	**EFH**	1	0,1
			BV	4	0,6

As table 6.14 shows, predictable denotational differences are considerably more frequent within the law data than within those recorded from fiction texts. This is not surprising given that law texts to a large extent describe generalised situations, where generic contexts will be the normal case. However, we find that this result does not merely reflect a contrast between law and fiction, but rather a difference between, on the one hand, technical texts in general and narrative fiction texts on the other hand.

[83] The different values of n_T are presented in tables 6.2 and 6.3 in 6.2.4.2.

Since story-telling involves the rendering of specific situations, it is likely that the description of generalised situations will occur more frequently in technical texts than in narrative fiction texts.[84]

Table 6.14 shows that, within the recorded data, there are fewer occurrences of predictable denotational differences in English-to-Norwegian translation than in the opposite direction. Since only very few cases have been identified within the fiction text pairs, the difference along the dimension of direction can merely be seen as an effect of the difference between the two pairs of law texts, as shown in table 6.15. Instances of this semantic subtype are considerably more frequent within the *Petro* data than within those compiled from the *AEEA*. This may reflect a larger element of domain-uniqueness in *Lov om petroleumsvirksomhet* than in the *EEA Agreement*, which we argued for in 5.5.1.2. It is likely that a strong degree of domain-specificity in a text can be correlated with certain linguistic features. A number of occurrences of this subtype were also found in a technical, English-Norwegian parallel text that was included in the empirical basis for the study presented in Thunes (1998).[85] This text pair, which describes a firewater system on an oil rig, resembles the *Petro* text pair in at least two respects: firstly, each text pair is strongly tied to one restricted domain, respectively, petroleum technology and the law regulating petroleum activities, and, secondly, both text pairs include a substantial number of headings, normally realised as complex noun phrases. In the *Petro* text pair there is a clear tendency that many of the identified cases of predictable denotational differences occur in translationally corresponding headings, as illustrated by (24), where the indefinite singular *undersøkelsestillatelse* corresponds with the indefinite plural *exploration licences*.

(24a) Tildeling av undersøkelsestillatelse m.v. (*Petro*)
'Granting of exploration-licence etc.'
(24b) Granting of exploration licences etc.

[84] However, technical texts may also include elements of narration.
[85] The parallel text in question is *Firewater Supply and Distribution System. Section 2 System description. Gullfaks A 71* and its Norwegian translation; cf. Thunes (1998: 32). Unfortunately, that report does not provide quantitative data for the various semantic subtypes.

We have not studied systematically to what extent the cases of predictable denotational differences fall within such headings, but the same tendency was observed in the firewater text pair, and this may indicate a text-typological similarity in relation to *Petro*. Similar cases are identified in the *AEEA* text pair, but there they are less frequent. Most likely this reflects the fact that there are fewer headings in the *AEEA* texts than in the other two, but it could possibly also be an indication of a stronger degree of domain-specificity in the *Petro* text pair than in the *AEEA*.

Although the large majority of the identified cases of predictable denotational non-equivalence are included among the law data, the phenomenon is also found in general language use as represented by the fiction texts. Hence, we will regard this as a type of correspondence that can be attributed to the interrelations between the two language systems, although its occurrence seems to be correlated with descriptions of generic situations.

6.3.2.2 Non-predictable denotational differences

DESCRIPTION. As explained in 6.2.4.1, denotational non-equivalence is a common denominator in a large and varied set of semantic subtypes. Some of these cases exhibit characteristic properties on the basis of which they may be identified as separate subclasses of denotational non-equivalence. As shown by table 6.1 in 6.2.4.2, only modest numbers of occurrences have been identified for most of the subclasses of denotational non-equivalence, and the majority of these subtypes are briefly presented in 6.2.4.1. The second most frequent subclass, involving coreferential noun phrases, will be discussed in 6.3.2.3. The largest set of cases where denotational non-equivalence has been identified is negatively defined in the sense that there is no particular denotational property with respect to which the translational units do not correspond, and the aim of this section is to illustrate the heterogeneity of this class.

Example (25) is a pair of corresponding matrix sentences:

(25a) From a side room came the sound of soft drumming. (DL)
(25b) Fra et av rommene hørte de dempete trommer.
'From one of rooms.DEF heard they subdued drums.'

By means of world knowledge it is possible to see that the sentences (25a) and (25b) may, depending on context, describe corresponding situations, but they do not express the same proposition, and they are denotationally non-equivalent in several ways. A central difference is that whereas (25a) depicts an event of sound emission, (25b) includes both an event of sound emission and one of sound perception. The perceiving individuals referred to in (25b) by the third person plural pronoun *de* ('they') are not referred to in (25a). Since the perception event is not described in the English sentence, it can be inferred from the source string only if the information that perceiving individuals are present is available from the preceding context.[86] This means that the translator has made use of task-specific information from the preceding linguistic context in order to produce a target sentence where the syntactic subject refers to a group of individuals not mentioned by the source sentence.

Moreover, the location of the source of the described sound is referred to by denotationally non-equivalent expressions, respectively *a side room* in (25a) and *et av rommene* ('one of the rooms') in (25b). The English phrase specifies the location as a side room relative to the described scene, while the Norwegian correspondent presents the location merely as a room. The expressions referring to the source of the sound are also denotationally non-equivalent: in (25a) *the sound of soft drumming* focuses on the drumming activity that creates the sound, and is semantically more precise than the translationally corresponding expression *dempete trommer* ('subdued drums') in (25b), which describes the instrument, but leaves implicit the information about the activity. The translator may, however, assume that this piece of information is available to the target text reader from general world information.

Example (26) is a pair of corresponding subclauses extracted from law text:

(26a) at forholdene blir lagt til rette, slik at fagforeningsvirksomhet blant egne ansatte og entreprenørens og underentreprenørens personell kan foregå i samsvar med norsk praksis (*Petro*)

[86] Since the Norwegian sentence describes explicitly both events of sound emission and perception, string pair (25) also contains an occurrence of non-predictable specification, and we regard it as an example of explicitation.

(26b) that the circumstances permit trade union activities to take place among his own employees and the personnel of contractors and sub-contractors in accordance with Norwegian practice

String pair (26) contains an embedded correspondence where the source text is an adverb phrase with a finite subclause as syntactic complement, and the target text is a nonfinite clause with infinitival verb phrase (cf. table 4.3 in 4.4.3). We will here focus on an instance of denotational non-equivalence found in the matrix correspondence, shown in (27):

(27a) at forholdene blir lagt til rette, (AdvP:4) (*Petro*)
'that circumstances.DEF becomes put to the-right ADVERBIAL'
(27b) that the circumstances permit (CPinf:4)

In string pair (27) there is denotational non-equivalence between the translationally corresponding verb phrases *blir lagt til rette* and *permit*. A linguistically predictable English translation of the Norwegian verbal expression *legge <object> til rette* could be *arrange <object> in order to permit*. Instead, the translator has chosen the verb *permit*, which has an argument structure that differs from that of the source text correspondent.

The semantic relation expressed by the Norwegian construction *legge X til rette* assigns the agent role to its first argument and the patient role to its second argument. In the semantic structure of the matrix sentence (27a) the patient role is filled by the referent of the subject *forholdene* ('the conditions/circumstances'), and the agent role is empty because (27a) is a passive sentence, and hence the first argument to the verbal relation is not expressed. The English verb *permit* in (27b) expresses a relation where the agent role is assigned to its first argument, and the patient role to its second argument. In the semantic structure of the target sentence (27b) the agent role is filled by the referent of the subject *the circumstances*, and the patient role is filled by the type of situations referred to by the nonfinite construction (represented by the label *CPinf*) which is the syntactic object in (27b). We may say that whereas (27a) de-

scribes a process where the patient is arranged for a certain purpose, (27b) describes a situation where the agent permits something to happen.[87]

Clearly, given the denotational deviations between source and target string in this example, (27b) is not a linguistically predictable translation of (27a). However, with access to task-specific information from the surrounding linguistic context shown in (26a), it is possible to infer that the result of the process expressed by (27a) will be the situation described by (27b). Thus, in string pair (27) the choice of target text is influenced by task-specific inferencing about extra-linguistic facts described by the surrounding context. That a linguistically predictable translation has not been chosen can most likely be ascribed to the translator's judgment of idiomatic language use: the predictable translation suggested above would not have given an elegant target sentence.[88] Such information about the stylistic effects of specific TL expressions is part of the general, given information sources that are available to the translator prior to the translation task; it is included under information about textual norms (cf. 2.4.2.1). Since it is a type of information derived through practice with language use, we regard it as separate from the information about the target language system itself.[89]

OCCURRENCE. As previously shown by table 6.1 in 6.2.4.2, instances of non-predictable denotational differences constitute an important subclass among the recorded data because it is relatively frequent, compared with other subtypes. It has been identified in 9,8% of the total number of string pairs, and, in absolute numbers, it is tentatively the third-most frequent subtype (433 occurrences; cf. table 6.16). In a given case, two corresponding units of translation may exhibit denotational non-equivalence with respect to a range of different properties (cf. the examples discussed above), or, in so-called minimal cases, with respect to a single property only. Since the quantitative data are based on a count of how many string pairs that exhibit at

---

[87] The example is not described as a case of non-equivalence in argument structure, because there is no direct correspondence between the argument structures of, respectively, (27a) and (27b); cf. 6.2.4.1.

[88] Suggested literal translation: ... *that the circumstances are arranged in order to permit trade union activities to* ...

[89] If, in an alternative approach, such information about stylistic effects of specific expressions would be regarded as information about the language system, then the translation (27b) would still be linguistically non-predictable, but the unsuitability of the suggested literal translation would be predictable.

least one kind of semantic discrepancy included in the given category, the data cannot reflect any differences between cases where the denotational deviation is small and cases where it is large. During the recording of string pairs, it was, however, quite clear that instances of large deviations are far more common among the fiction data than among the law data.[90]

Table 6.16. Occurrences of non-predictable denotational differences, counted within all recorded string pairs, within each direction of translation, and within each text type.[91]

	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in per cent of n_T	in per cent of n_4
Across **all data**:	433	9,8	19,5
Across all data **E** → **N**:	269	12,8	23,6
Across all data **N** → **E**:	164	7,0	15,2
Across all **law** data:	98	5,7	13,2
Across all **fiction** data:	335	12,3	22,7

Table 6.17. Occurrences of non-predictable denotational differences in individual text pairs.

Legal texts				Fiction texts			
Text pairs	Frequency of string pairs where the subtype is found:			Text pairs	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in % of n_T	in % of n_4		in abs. numbers	in % of n_T	in % of n_4
AEEA	65	8,2	16,0	**AB**	35	6,7	16,8
				DL	169	21,3	32,1
Petro	33	3,6	9,9	**EFH**	58	8,3	15,9
				BV	73	10,3	19,3

[90] This observation is commented on towards the end of 6.2.4.2.

[91] The different values of n_T and n_4 are presented in tables 6.2 and 6.3 in 6.2.4.2.

The results shown by tables 6.16 and 6.17 will only be briefly commented on as they are in general compatible with various tendencies that have already been discussed in connection with other subtypes.

Firstly, in relation to the total number of string pairs, non-predictable denotational differences are more than twice as frequent among the fiction data than among those recorded from the law texts. This confirms the previously observed tendency that the extent to which semantic equivalence holds between translationally corresponding strings is greater in law text than in fiction, because the strong norms controlling legal language use contribute to the preservation of meaning in translational relations.

Secondly, cases of non-predictable denotational differences occur more frequently in the data compiled from English-to-Norwegian translation than among those extracted from Norwegian-to-English parallel texts. Again, this reflects the fact that there is a larger proportion of string pairs exhibiting semantic equivalence within the data compiled from the *Petro* text pair than within the *AEEA* data. In 5.5.1.2 we have described factors that may have contributed to this difference.

Thirdly, it is again our view that the results shown for each of the fiction text pairs reveal differences between individual translators (cf. table 6.17). The relatively high frequency of non-predictable denotational differences in the text pair DL in comparison to the other three fiction pairs, probably reflects that in the former case the translator has been less true to the original than in the latter cases (cf. 5.5.2.2).

6.3.2.3 Denotational non-equivalence between coreferential noun phrases

DESCRIPTION. Within the class of correspondences involving non-predictable denotational differences, we have identified a group of special cases where the denotational deviaton concerns translationally corresponding noun phrases which are coreferential. The examples in this section will illustrate that the phenomenon may involve individually as well as generically referring NPs, and that the denotational deviation between the corresponding phrases may be full as well as partial.

The phenomenon is illustrated by the italicised expressions in string pair (28).

(28a) Bert reached into *a cupboard* and took out a thermos flask the size of a bucket, (DL)

(28b) Bert tok ned en termosflaske på størrelse med et spann fra *en hylle*,
'Bert took down a thermos-bottle on size with a bucket from a shelf,'

There are several semantic differences between sentences (28a) and (28b); we will here concentrate on the denotational non-equivalence between the translationally corresponding expressions *a cupboard* and *en hylle* ('a shelf'), which causes the following difference in propositional potential between source and target sentence: whereas (28a) describes the situation where a person named Bert takes a thermos out of a cupboard, the translation (28b) describes the situation where the same person takes a thermos down from a shelf. The two events are partially identical as both include the moving of a thermos from specific locations by the agent Bert. The italicised expressions *a cupboard* and *en hylle* refer to these locations. Both phrases are indefinite, single NPs, referring to specific entities which are introduced as new discourse referents. We regard *a cupboard* and *en hylle* as translational correspondents since the objects they refer to fill the same role within the content that is shared between source and target sentence.

The nouns *cupboard* and *hylle* ('shelf') are denotationally non-equivalent; they even have disjoint denotata, and they are not translationally related in any linguistically predictable way. While *cupboard* denotes a class of objects which have doors, the denotata of *hylle* are objects without doors. But inside cupboards there are usually one or more shelves, and this is possibly the reason why *a cupboard* has been translated into *en hylle*. The translator may have inferred, on the basis of the description in (28a), that the exact location from which the thermos is taken is a shelf inside the cupboard. This is a probable interpretation of (28a), although not the only possible one, as the thermos may also have been taken from the bottom of the cupboard. If we assume that the former interpretation lies behind the choice of target expression, then the translation has been influenced by general background information about the world, and by task-specific extra-linguistic information produced through reasoning about the situation described by the source sentence. Since the nouns *cupboard* and *hylle* have disjoint denotata, it may appear difficult to regard the

NPs *a cupboard* and *en hylle* as coreferential, but since the referents of these phrases fill a shared role in original and translation, we nevertheless want to regard this pair of NPs as an instance of denotational non-equivalence between coreferential noun phrases.

Example (29) is taken from one of the pairs of law text:

(29a) The Contracting Parties shall endeavour to promote the dialogue between *management* and labour at European level. (*AEEA*)

(29b) Avtalepartene skal bestrebe seg på å fremme dialogen mellom *arbeidsgivere* og arbeidstagere på europeisk nivå.
Contracting-parties.DEF shall endeavour themselves on to promote dialogue.DEF between employers and employees at European level.'

As pointed out in 6.3.2.1, law texts are to a large extent descriptions of types of situations, where generic contexts are the normal case. The italicised expressions in (29), *management* and *arbeidsgivere* ('employers'), are indefinite noun phrases with generic reference. They are coreferential since they refer to the same role in the situation type described by the sentences (29a) and (29b), but they are denotationally non-equivalent as the nouns *management* and *arbeidsgiver* apply to different classes of objects. The denotata may overlap: in a workplace the management can be identical with the employer, but this is not necessarily true. Thus, from background world information it follows that the translationally corresponding NPs *management* and *arbeidsgivere* in (29) may be coreferential, but from linguistic information it follows that they are not denotationally equivalent. Hence, the Norwegian expression *arbeidsgivere* is not a linguistically predictable translation of *management* in example (29), and there is a difference in propositional potential between source and target sentence.

The two examples in this section illustrate that there is heterogeneity within the class of correspondences involving denotational non-equivalence between coreferential noun phrases. The instance presented in (28) possibly reflects the preferences of a translator who tends to deviate from the original with respect to the linguistically encoded meaning, which results in a translation with violations of denotational

equivalence as defined by Koller (1992: 216).[92] The occurrence in (29) may indicate a cultural difference within the domain of professional life, since the source sentence focuses on the management aspect, and the translation on the employer function, of the shared referent of the NPs *management* and *arbeidsgivere*.[93]

In our view it is not so important whether the denotational deviation amounts to partial overlap, or disjunction, between sets of denotata in the cases falling within this semantic subtype. The important criterion is that due to the denotational deviation in question, the correspondences are not included in the domain of linguistically predictable translations. Moreover, with respect to this category we apply a somewhat special understanding of 'coreferential', i.e. reference to entities that fill a shared role in translationally corresponding expressions.

OCCURRENCE. As previously shown by table 6.1 in 6.2.4.2, instances of denotational non-equivalence between coreferential noun phrases are fairly frequent, compared with other subtypes. They are identified in 6,8% of the total number of string pairs, and, in absolute numbers, it is tentatively the fourth-most frequent subtype (304 occurrences; cf. table 6.18).

The results presented in tables 6.18 and 6.19 will not be discussed in detail, primarily because the occurrences of this subtype seem to be fairly evenly distributed across the text types as well as across the directions of translation, especially if we consider the percentages calculated in relation to the numbers of string pairs. On the basis of the results given in table 6.19 for individual text pairs, we may trace some tendencies similar to those observed in connection with a closely related subtype, i.e. non-predictable cases of denotational deviation (cf. 6.3.2.2). Firstly, denotational non-equivalence between coreferential noun phrases is more frequent in the *AEEA* than in the *Petro* text pair, which we regard as another consequence of the factors that may account for the difference between these two text pairs concerning the proportion of semantic equivalence. Secondly, the variation found among the pairs of

[92] It may not be a coincidence that example (28) is taken from the text pair DL, which is the text pair exhibiting the largest proportion of semantically non-equivalent correspondences; cf. 5.5.2.2.

[93] In Norwegian this domain could be described as *arbeidslivet* ('the work life'), but there is no direct lexical correspondent in English to the Norwegian noun *arbeidsliv*. In the *AEEA* corpus the Norwegian collocation *arbeidslivets parter* occurs as the translation of the English expression *economic and social partners*.

fiction texts probably reflects differences between the individual translations concerning the degree of faithfulness to the original (cf. 5.5.2.2).

Table 6.18. Occurrences of non-predictable denotational differences between coreferential NPs, counted within all recorded string pairs, within each direction of translation, and within each text type.[94]

	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in per cent of n_T	in per cent of n_4
Across **all data**:	304	6,8	13,7
Across all data **E** → **N**:	158	7,5	13,9
Across all data **N** → **E**:	146	6,3	13,5
Across all **law** data:	131	7,6	17,7
Across all **fiction** data:	173	6,3	11,7

Table 6.19. Occurrences in individual text pairs of non-predictable denotational differences between coreferential NPs.

Legal texts				Fiction texts			
Text pairs	Frequency of string pairs where the subtype is found:			Text pairs	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in % of n_T	in % of n_4		in abs. numbers	in % of n_T	in % of n_4
AEEA	85	10,7	21,0	**AB**	43	8,3	20,7
				DL	30	3,8	5,7
Petro	46	5,0	13,7	**EFH**	29	4,1	7,9
				BV	71	10,0	18,7

With respect to the issue of faithfulness to the original, we have in 6.3.2.2 commented on the relatively high frequency of the more general category of non-predict-

[94] The different values of n_T and n_4 are presented in tables 6.2 and 6.3 in 6.2.4.2.

able denotational differences in the text pair DL, and on that background it is surprising that the same text pair shows a very low frequency of denotational non-equivalence between coreferential NPs. Without a more detailed study of the texts, this cannot be accounted for, but a possible explanation is that due to a relatively large element of denotational non-equivalence, the DL text pair contains fewer pairs of coreferential noun phrases. However, if we consider the other fiction pairs, it is not straightforward to find a correlation between the general type of denotational non-equivalence and the particular kind pertaining to coreferential NPs. In both AB and BV the frequency of the latter type is considerable, whereas the frequency of the former kind is lower in AB, but very similar in BV. In EFH, both subtypes show rather modest frequencies. We prefer to regard the relatively large occurrence of non-predictable denotational differences between coreferential NPs in DL as reflecting a lower degree of faithfulness to the original than in the other pairs of fiction texts, a factor that may have reduced the general frequency of coreferential NPs in that text pair.

6.3.3 Referential differences

We have previously explained that in linguistically predictable translational correspondences, source and target string are semantically equivalent, which means that they must be equivalent in terms of compositional semantic properties (cf. 3.3.4.1).[95] This requirement includes equivalence with respect to referential properties. 6.3.3 with subsections will describe cases where translationally interrelated noun phrases do not correspond with regard to referential properties, mainly due to differences in the marking of definiteness.

The grammatical category of definiteness applies to nominal expressions; it signals whether the referent of a noun phrase is already accessible, or not, in the universe of discourse shared by speaker and hearer, or sender and recipient. The category of definiteness is grammaticalised in English as well as in Norwegian: in both

[95] Cases of compositional non-equivalence in type 3 correspondences constitute an exception to this. These are cases where expressions, seen as units, are linguistically predictable translations of each other even if there is not local semantic equivalence between corresponding subparts of the two units (cf. 6.2.4.1).

languages, noun phrases are marked either as definite or as indefinite. If an NP is uniquely referring, its form is definite, and if it is non-uniquely referring, its form is indefinite. Whereas the referent of a definite NP "can be identified uniquely in the contextual or general knowledge shared by speaker and hearer" (Quirk et al. 1985: 265), an indefinite noun phrase typically introduces a new referent in the discourse. Hence, the opposition between the definite and the indefinite functions as a marker of the distinction between referential givenness and referential newness.

Our analysis of parallel texts has not identified a large number of occurrences where translationally corresponding noun phrases are non-equivalent with respect to the marking of definiteness. However, the observed cases are interesting in two respects: firstly, their distribution indicates clear text-typological differences, and, secondly, within the identified occurrences, there seems to be a division between predictable and non-predictable instances.

6.3.3.1 Predictable differences in the use of definiteness

DESCRIPTION. Within the recorded data, we have identified quite a few cases where translationally corresponding noun phrases differ with respect to the marking of definiteness and where this difference can be accounted for by means of purely linguistic information, i.e. information about source and target language systems, and their interrelations. The formal difference between the translationally linked NPs can be seen as a surface criterion for this subtype of correspondences. Since the opposition between the definite and the indefinite normally signals a difference between unique and non-unique reference, these cases are included in the more general category of referential non-equivalence. However, within the predictable subset of this category, the non-equivalence between the corresponding NPs primarily concerns the marking of definiteness, because, as the examples will demonstrate, their referential properties will usually match each other in spite of the formal difference. Also, the examples will illustrate that there are various kinds of linguistic aspects that cause differences in the marking of definiteness.

Instances of predictable differences in the use of definiteness have almost exclusively been identified in the pairs of law texts. Within the fiction texts, only three occurrences have been found; one of them is shown in example (30):

(30a) "Det eneste som mangler nå," sier faren, "er *koleraen*. (EFH)
'The only which lacks now, says father.DEF, is cholera.DEF.'
(30b) "All we need now," says his father, "is *cholera*.

In the source sentence (30a) the definite form of the Norwegian NP *koleraen* signals that this is a uniquely referring phrase, whereas its English translational correspondent in (30b), the indefinite NP *cholera*, is not marked as uniquely referring. The English noun *cholera* is the linguistically predictable translation of the Norwegian noun *kolera*. Moreover, in the italicised NP correspondence in (30), the indefinite form of the English NP is also predictable from linguistic information: *cholera*, like other names of diseases in English, appears normally without the definite article, as in (30b); cf. Quirk et al. (1985: 279)[96]. This means that the indefinite form of *cholera* in the translation (30b) is predictable from information about the lexeme *cholera* combined with general information about English grammar. It also means that with respect to the noun phrase *cholera* in (30b), the distinction between unique and non-unique reference is neutralised.[97] In this sense the translationally corresponding phrases *koleraen* and *cholera* have non-equivalent referential properties, if the strings in which they occur are considered independently of their respective contexts. Since the expression *cholera* is neutral in relation to the distinction between unique and non-unique reference, it is, depending on context, potentially a uniquely referring expression, and this confirms its status as a linguistically predictable translation of the uniquely referring expression *koleraen* in (30a).[98]

---

[96] Quirk et al. (1985: 279) point out exceptions to this rule: "… *the* is often used, in a more traditional style of speech, for some well-known infectious diseases: (*the*) *flu*, (*the*) *measles*, (*the*) *mumps*, (*the*) *chicken pox*; also (*the*) *hiccups*."

[97] Hence, this may also be seen as a case where the English expression is, in a predictable way, grammatically less specific than the Norwegian one; cf. 6.3.1.1.

[98] The topic of the context preceding sentence (30a) is the spreading of infectious diseases in the slums of late 19th century London. Thus, the referent of *koleraen* may be uniquely identified by combining task-specific,

The majority of the cases of predictable differences in the marking of definiteness are identified among the data compiled from the *Petro* text pair (cf. table 6.21). These occurrences mainly follow a pattern where a Norwegian indefinite noun phrase is translated into an English definite noun phrase, which can be illustrated by (31):

(31a) *Rettighetshaver* blir *eier* av den petroleum som produseres. (*Petro*)
'Licensee becomes owner of the petroleum which produce.PASSIVE.'
(31b) *The licensee* becomes *the owner* of the petroleum which is produced.

Since example (31) is taken from a pair of law texts, both sentences (31a) and (31b) describe a type of situation, and the noun phrases contained in them are generically referring expressions, which pick out roles of that situation type.[99]

First, we may consider the use of indefinite form in two Norwegian noun phrases in (31a). This concerns the singular indefinite NP *rettighetshaver* ('licensee') and the complex expression *eier av den petroleum som produseres* ('owner of the petroleum which is produced'), where the singular indefinite noun *eier* is the syntactic head of the phrase. Both NPs pick out unique referent roles in the described situation, although they do not have definite form. That is, in formal terms, the syntactic heads of these two NPs (*rettighetshaver* and *eier*) can be described as *bare nouns*. This particular use of bare nouns is explained by Faarlund et al. (1997: 288): in Norwegian, indefinite common nouns with no premodifying determiner may in certain contexts behave as proper names and refer uniquely, and as typical examples they mention titles referring to leader positions, and certain legal titles. They also point out that indefinite form paired with unique reference is common in "official" or "administrative" uses of Norwegian; this applies to the case shown in (31a). Hence, the use of bare nouns in the mentioned NPs in (31a) is a marked choice: it is marked in relation to the normal way of signalling unique reference, which is the use of the definite suffix (*rettighetshaveren* and *eieren*), and it is also marked in relation to the normal way of signalling non-unique reference, which is the use of the indefinite article (*en*

extra-linguistic information contained in the preceding context with general background information about diseases.

[99] Cf. the comments in 6.3.2.1 and 6.3.2.3 on generic reference i law texts.

rettighetshaver and *en eier*). This particular use of bare nouns represents a linguistic feature specific to the given text type, and noun phrases with these characteristics are a recurrent phenomenon in the Norwegian law text *Lov om petroleumsvirksomhet*. Due to the regularity of this pattern, and because it does not only appear in texts of the legal domain, it is, in our view, part of the Norwegian language system, and not merely an effect of textual norms specific to a technical field.

Next, we may consider the translational relation between *rettighetshaver* in (31a) and the definite NP *the licensee* in (31b). On the basis of the formal properties of the Norwegian NP, it is possible to derive the information that *rettighetshaver* is a uniquely referring expression. Then it follows from general information about English grammar that in this type of contexts bare nouns are not used in the same way as in Norwegian, and, hence, it is linguistically predictable that the target expression requires the definite marker (*the licensee*) in order to express unique reference.

Then, we may consider the translational relation between *eier av den petroleum som produseres* in (31a) and the definite NP *the owner of the petroleum which is produced* in (31b). Again, the form of the Norwegian source expression shows that it is uniquely referring. In this case the source string also provides further linguistic information which identifies the referent of the NP, since the postmodifying preposition phrase *av den petroleum som produseres* restricts the meaning of the entire NP. Also, the lexical meaning of *eier* carries with it the presupposition that there is only one owner (or one group of owners) for a given object. Thus, there are three linguistic aspects of the source expression which, together with general information about English grammar, predict the use of definite form in the target expression in order to signal unique reference.[100]

As already noted, the translation pattern illustrated by the NP correspondences discussed in example (31) is relatively frequent within the data compiled from the *Petro* text pair. There are some occurrences among the *AEEA* data, and otherwise it is absent from the remaining text pairs (cf. table 6.21). Although it seems to be the

[100] The English NP *the owner of the petroleum which is produced* shows an instance of the phenomenon described by Quirk et al. as "cataphoric reference". This is understood by them as "the use of the definite article in a context where what follows the head noun, … enables us to pinpoint the reference uniquely" (1985: 268).

dominating pattern within its class, it does not account for all the cases that have been identified of predictable differences in the use of definiteness. Within the *AEEA* data, the most frequent pattern is illustrated by the italicised NP correspondence in example (32):

(32a) Moreover, they shall facilitate *cooperation within the framework of this Agreement*. (*AEEA*)
(32b) De skal videre lette *samarbeidet innen rammen av denne avtale*.
'They shall further ease cooperation.DEF within frame.DEF of this agreement.'

The indefinite NP *cooperation within the framework of this Agreement* in (32a) is translated into the definite NP *samarbeidet innen rammen av denne avtale* in (32b). Again, the source sentence, taken from a law text, describes a situation type, and the English NP is generically referring. The indefinite form of its syntactic head (*cooperation*) is determined by a rule of English grammar: according to Quirk et al. (1985: 282), the "zero article" is used when plural nouns, and mass nouns like *cooperation*, refer generically. This rule seems here to override the demand for the definite article in cases where the postmodification of a noun serves to identify its referent: in (32a) the semantic contribution of the postmodifying preposition phrase *within the framework of this Agreement* is to identify the type of cooperation that is referred to. Thus, through linguistic information expressed in the source sentence it is clear that the noun phrase *cooperation within the framework of this Agreement* picks out a unique role in the situation described by (32a). The choice of definite form in the translation *samarbeidet innen rammen av denne avtale* is predictable from general information about Norwegian grammar: as noted above, definite form is the unmarked way of signalling unique reference in Norwegian.[101]

OCCURRENCE. As previously observed, the distribution of instances of predictable differences in the use of definiteness shows a clear difference between the data compiled from law texts and the fiction data, and this is further illustrated by the results presented in tables 6.20 and 6.21. Whereas 134 cases have been identified

[101] Given the text type, the indefinite NP *samarbeid* is also a possible translation; cf. example (31). As noted in 6.3.2.1, all forms of Norwegian nouns may be used to express generic reference.

among the pairs of law texts (i.e. in 7,8% of the string pairs), only 3 occurrences are found in one of the pairs of fiction texts. This is in line with the observation made above, that although these cases may be explained by referring to regularities of the two language systems, they reflect aspects that are not found in all text types, and almost not in the investigated fiction texts.

Table 6.20. Occurrences of predictable differences in the use of definiteness, counted within all recorded string pairs, within each direction of translation, and within each text type.[102]

	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T
Across **all data**:	137	3,1
Across all data **E** → **N**:	33	1,6
Across all data **N** → **E**:	104	4,5
Across all **law** data:	134	7,8
Across all **fiction** data:	3	0,1

Table 6.21. Occurrences in individual text pairs of predictable differences in the use of definiteness.

Legal texts			Fiction texts		
Text pairs	Frequency of string pairs where the subtype is found:		Text pairs	Frequency of string pairs where the subtype is found:	
	in absolute numbers	in per cent of n_T		in absolute numbers	in per cent of n_T
AEEA	33	4,2	**AB**	0	0,0
			DL	0	0,0
Petro	101	11,0	**EFH**	3	0,4
			BV	0	0,0

[102] The different values of n_T are presented in tables 6.2 and 6.3 in 6.2.4.2.

Table 6.20 seems to indicate a difference in frequency correlated with the direction of translation. Due to the very low number of investigated text pairs, this merely reflects the difference in frequency between the two pairs of law texts. Cases of predictable differences in the use of definiteness are noticeably more frequent among the *Petro* data than within those compiled from the *AEEA*. As indicated above, correspondences between Norwegian indefinite NPs and English definite NPs constitute the most frequent pattern identified in the *Petro* text pair is (cf. example (31)): tentatively, 96 occurrences of this kind are identified within the *Petro* data, whereas merely 9 cases are found in the *AEEA* text pair. To the extent that this pattern may be correlated with other text-typological aspects, this difference between the two pairs of law texts may, as in the case of predictable denotational differences, indicate a larger element of domain-specificity in *Petro* than in the *AEEA* (cf. 5.5.1.2 and 6.3.2.1). However, verifying this assumption requires a further study of texts associated with a larger variety of domains.

The pattern that is most frequent in the *AEEA* is seen in correspondences between English indefinite NPs and Norwegian definite NPs. Possibly, this is related to a tendency observed in connection with non-predictable referential differences between English and Norwegian (cf. 6.3.3.2). This, too, needs to be studied further.

6.3.3.2 Non-predictable referential differences

DESCRIPTION. As shown previously by table 6.1 in 6.2.4.2., non-predictable cases of differences in the marking of definiteness are not frequent among the recorded data. Apart from differences in definiteness, it is not easy to identify characteristics that may identify such correspondences as a class. In general, since the opposition between definite and indefinite marks the distinction between given and new information, translational correspondences involving referential non-equivalence will create differences between original and translation concerning what elements of information that are already known or not, to the reader, as well as to characters described in the text. The examples will illustrate that due to mismatches with respect to definiteness, information may be lost or added, and there may be stylistic deviations between source and target expression. The cases are varied, and their common denominator is

linguistically non-predictable differences in the marking of definiteness between translationally corresponding noun phrases.

An example is the italicised NP correspondence in example (33):

(33a) Det var den store billedbibelen, med *alle kopperstikkene*. (EFH)
'That was the big picture-Bible, with all engravings.DEF.'
(33b) That was the big illustrated Bible full of *engravings*.

The Norwegian definite noun phrase *alle kopperstikkene* ('all the engravings') in (1a) is translated into the English indefinite noun phrase *engravings* in (1b). The definite form of the source expression signals that it is a uniquely referring NP, and hence a linguistically predictable translation into English would be the definite expression *all the engravings*.[103] In this case there is neither any rule of English grammar, nor any aspect of the relationship between source and target languages, which can predict the target correspondent of *alle kopperstikkene* to be an indefinite noun phrase expressing non-unique reference, and concerning this NP correspondence, this is as far as we get given our approach to the classification of translational correspondences, since it studies relations between strings viewed as system units.

In order to understand the consequences of the difference in definiteness between *alle kopperstikkene* and *engravings*, it is necessary to consider the contexts, in (34), from which (33a) and (33b) are extracted. This is to make an excursion, but it may throw light on the translational mismatch illustrated in (33).

(34a) Det fantes en bok oppe i stuen også, som Jason så meget i – men den var annerledes. Det var den store billedbibelen, med alle kopperstikkene. Moren pleide å lese i den for ham. (EFH)

[103] It could perhaps be argued that another linguistically predictable translation would be *all its engravings*, where the possessive determiner *its* expresses the part-whole-relation holding between the referent of *engravings* and the referent of *the big illustrated Bible*. This is a non-prototypical instance of the inalienability pattern described in 6.3.1.2. The part-whole relation is implicit in the source sentence, and may be inferred from the facts that an illustrated Bible contains pictures and that engravings are a kind of picture, together with the semantic contribution of the Norwegian preposition *med* ('with'), which expresses a relation of inclusion. These pieces of information may be regarded as purely linguistic, and they may also be seen as truly on the borderline between linguistic and extra-linguistic information; cf. the discussion in 2.4.2.1.

(34b) There was another book up in their living room which Jason often looked at – but that book was different. That was the big illustrated Bible full of engravings. His mother used to read to him out of it.

The sequences in (34) are taken from a passage in Fosnes Hansen's novel where the story deals with special illustrated books that are kept in the home of the boy Jason, the protagonist, and which make a strong impression on him during his childhood. In the Norwegian original the illustrated Bible is introduced in the discourse by means of the definite NP *den store billedbibelen*, which signals that the referent of this phrase is a known entity, and thus the reader is invited into Jason's mind, where the bible is known. This stylistic effect is enhanced by the use of definite form in the succeeding NP *alle kopperstikkene*. The latter phrase provides additional information about the bible, and its definite form signals that its referent, too, is a known entity. Thus, the use of definiteness creates the effect of sharing Jason's knowledge about the bible with the reader.

A fully parallel stylistic effect is not achieved in the English translation (33b), since the translator has chosen the indefinite NP *engravings* as the translation of *alle kopperstikkene*. The indefinite form of *engravings* signals that the phrase contributes new information, but as the preceding context reveals that Jason knows the described bible well, its engravings cannot be new to him. The referential newness of *engravings* is newness in relation to the reader, not in relation to the character Jason. Thus, the choice of indefinite form weakens the reader's experience of looking into Jason's mind. Since the non-correspondence between the phrases *alle kopperstikkene* and *engravings* pertains to factors concerning the recipients of the texts, it gives rise to a case where what Koller terms pragmatic equivalence (1992: 216) is not achieved between source and target text. Thus, a consequence of referential non-equivalence between the two NPs is that something is lost in the translation.

Example (35) is a case where conversion from indefinite to definite form adds something in the translation:

(35a) Hun brettet sammen *frottéhåndklær*, (BV)
'She folded together towels,'
(35b) She folded up *the towels*,

In (35) the Norwegian indefinite noun phrase *frottéhåndklær* in (35a) is translated into the English definite NP *the towels* in (35b).[104] The indefinite form of *frotté-håndklær* expresses referential newness, and, as in the case of the NP correspondence discussed in example (33), the definite form of the target expression *the towels* cannot be predicted on the basis of the information that is linguistically encoded in the source sentence, together with general, given linguistic information about the two languages. If (35) is considered in relation to a wider context, given in (36), we may see that the definite form of *the towels* has added something in the translation.

(36a) ... da moren til Hildegun kom fra tørkebåsen med kurven full av laken og dynetrekk. Hildegun og Brita sto i den smale entréen og trakk sengetøyet mellom seg, det luktet vår av tøyet, de ble lattermilde og klessnippene glapp ut av hendene på dem. Tøyet er rent, sa moren. Hun brettet sammen *frottéhåndklær*, hun arbeidet fort og sint. (BV)
(36b) ... and then her mother appeared from the drying-room with a basket full of sheets and bedcovers. Hildegun and Brita stood in the narrow entrance hall and stretched the sheets between them. The washing smelt of spring. They were in a giggly mood and the corners of the sheets slipped out of their hands. "Mind those things, they're clean," said the mother. She folded up *the towels*, working quickly and angrily.

Although the indefinite form of *frottéhåndklær* signals new information, the reader can easily infer from general information about the world that the referent of this NP is included among the items contained in the linen basket. The definite form of the translationally corresponding NP *the towels* expresses that the referent of the phrase is already known, although towels have not been introduced earlier in the discourse. This creates an effect similar to what we observed in connection with example (33), where the definite form of the NP *alle kopperstikkene* invites the reader into the mind of the protagonist. The use of definite form in *the towels* in (35) signals that the

[104] Other semantic deviations between (35a) and (35b) will not be discussed here.

objects referred to are known to the characters Hildegun and Brita. In the case of (35) this stylistic effect appears in the target text, and not in the original, as in (33), and hence the difference in definiteness marking adds something to the translation.

Finally, we may consider an example where a non-predictable difference in the marking of definiteness creates a change in meaning in addition to a mismatch concerning referential properties. In (37) the relevant noun phrases are italicised:

(37a) They already knew that the Council, to prevent squatters, had sent in *the workmen* to make the place uninhabitable. (DL)

(37b) De hadde hørt at kommunen hadde sendt *en arbeidsgjeng* hit for å gjøre huset ubeboelig og hindre okkupasjoner.
'The had heard that Council.DEF had sent a work-gang here for to do house.DEF uninhabitable and prevent squattings.'

In the English original the noun phrase *the workmen* introduces the referent of this expression into the discourse. The use of definite form signals that the referent is given information, and the reader will infer that the characters referred to by the pronoun *they* in (37a) already know who the described workmen are. In contrast, in the Norwegian translation the use of the indefinite form in the corresponding phrase *en arbeidsgjeng* signals that the NP introduces a new referent in the universe of discourse. This implies that nothing is yet known about the group of workers referred to, and in this sense the difference in referential properties between the corresponding noun phrases creates a slight change of meaning, and something is lost in the Norwegian translation.

OCCURRENCE. Within the recorded data, we have only identified 55 string pairs exhibiting non-predictable referential differences between translationally corresponding noun phrases. Tables 6.22 and 6.23 provide more information on the distribution of these cases.

Table 6.22. Occurrences of non-predictable referential differences, counted within all recorded string pairs, within each direction of translation, and within each text type.[105]

	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in per cent of n_T	in per cent of n_4
Across **all data**:	55	1,2	2,5
Across all data **E → N**:	14	0,7	1,2
Across all data **N → E**:	41	1,8	3,8
Across all **law** data:	6	0,4	0,8
Across all **fiction** data:	49	1,8	3,3

Table 6.23. Occurrences of non-predictable referential differences in individual text pairs.

Legal texts				Fiction texts			
Text pairs	Frequency of string pairs where the subtype is found:			Text pairs	Frequency of string pairs where the subtype is found:		
	in abs. numbers	in % of n_T	in % of n_4		in abs. numbers	in % of n_T	in % of n_4
AEEA	2	0,3	0,5	**AB**	6	1,2	2,9
				DL	6	0,8	1,1
Petro	4	0,4	1,2	**EFH**	16	2,3	4,4
				BV	21	3,0	5,5

The most interesting result to be read out of tables 6.22 and 6.23 is that occurrences of this subtype are mainly identified within the data extracted from fiction texts, which contain 49 of the 55 identified instances. This clearly reflects a text-typological difference. As the discussion of examples has shown, non-equivalence concerning the referential properties of translationally corresponding expressions will in-

[105] The different values of n_T and n_4 are presented in tables 6.2 and 6.3 in 6.2.4.2.

fluence the interpretation of a text, and it is to be expected that such mismatches are not compatible with the strict norms governing the translation of law texts.

Within the fiction data, non-predictable referential differences are somewhat less frequent among the English-to-Norwegian string pairs than among those representing the opposite direction of translation. Since the numbers are anyway very small, it is impossible to generalise. Nevertheless, within the identified cases of this subtype, we have observed the tendency that definite NPs matched translationally by indefinite ones are more frequent in the Norwegian texts, originals as well as translations, than in the English ones. Johansson presents a similar observation in his study of subject changes in English-to-Norwegian translation (2007: 197–215).[106] On the basis of correspondences between subject NPs, he claims (2007: 214) that English seems to prefer indefinite noun phrases to a larger extent than Norwegian does. Although his investigation is restricted to NPs in subject position, our observations of differences in definiteness marking between translationally corresponding NPs are compatible with his conclusion. Possibly, the correspondences we have identified indicate that when translationally related noun phrases of English and Norwegian differ with respect to the marking of definiteness, the choice of target expression may be influenced not only by *parole*-related factors involved in the preferences of the translator, but also by aspects of the target language system. However, a further study of this falls outside the scope of the present project.

6.4 Summary

At the end of chapter 3 we pointed out that in order to clarify the division between computable and non-computable translation it is useful to discuss both literal and non-literal translation. Literal, or linguistically predictable, translation is covered by the definitions of correspondence types 1–3 in chapter 3. In this chapter we have discussed certain phenomena which are recurrent among the recorded data, and which involve some kind of semantic deviation between translationally corresponding units. We have sorted these phenomena into classes, described as subtypes of the

[106] This study has previously been commented on in 6.3.1.3.

main correspondence types. A few of these classes involve systematic, and hence linguistically predictable differences between the two languages; these are seen as subcategories of type 3. The majority of the classes involve non-predictable semantic deviations, and are thus subcategories of type 4. The discussions of subtypes of 3 have aimed to illustrate the limit of the linguistically predictable domain, and the presentations of subtypes of 4 have exemplified phenomena that cannot be included in literal translation.

6.2. with subsections discusses the identification of semantic subtypes. In terms of string length, correspondences of types 3 and 4 together cover about 90% of the analysed parallel texts. Hence, it has been of interest to describe recurrent types of semantic deviations within this part of the empirical material. The classification of semantic subtypes resembles previous approaches to shifts in translation, but also differs from them in important ways. Our subtype description is a data-driven classification of relations between source expressions and their existing translations, and it has emerged from observations made for one language pair only, and within a small selection of texts and text types (cf. 6.2.1). On the basis of the framework presented in chapters 2 and 3, we have a set of five different dimensions which apply to the sorting of translational correspondences: degree of translational complexity, linguistic predictability, informational content, amount of information, and semantic phenomena (cf. 6.2.3).

Through these dimensions we have tentatively identified groups of related subtypes, presented in 6.2.4.1. The largest, and most general, class consists of correspondences involving some kind of difference in informational content between translationally related expressions. Correspondences not included here are string pairs of type 3 exhibiting compositional non-equivalence in the translational relation between expressions which, seen as units, are linguistically predictable translations of each other even if there is not local semantic equivalence between corresponding subparts of the two units. The wide category of differences in informational content is divided into three groups of subtypes: (i) classes of correspondences involving differences between source and target string in the amount of linguistically expressed information, (ii) classes of denotational non-equivalence between translationally

corresponding expressions, and (iii) classes of referential differences between correspondents. Within each of these three groups of subtypes we distinguish between linguistically predictable and non-predictable cases.

6.3 with subsections provides more detailed discussions of a selection of the subtypes included in the groups (i)–(iii). Priority has been given to certain subtypes which are relatively frequent, and to types that may reveal differences between the two investigated text types, and, to some extent, between the two directions of translation. As explained in 5.2.2, the limited scope of our investigation makes it impossible to generalise in relation to these dimensions, but interesting tendencies may be observed within the empirical data.

Subtypes involving differences between translational correspondents in the amount of linguistically expressed information are presented in 6.3.1 with subsections. We distinguish between specification, where the translation contains more information than the original, and despecification, where the target text contains less information than the source text. Moreover, we draw a line between lexical and grammatical (de)specification, and within lexical (de)specification we distinguish between cases where source and target expression differ in the number of lexical signs, and cases showing a difference in semantic granularity between translationally corresponding lexical signs.

In the group of subtypes where translational correspondents contain different amounts of information, the predictable classes primarily involve progressive aspect in English, which is not matched in Norwegian, and differences between the two languages in the use of possessive determiners (cf. 6.3.1.1–2). These phenomena give rise to cases where given expressions are semantically more specific than their translational correspondents because of the use of certain grammatical markers. Since the semantic differences in such cases follow from information about SL and TL and their interrelations, we classify such correspondences as predictable grammatical (de)specification. Cases of these kinds are relatively infrequent across the recorded data, but they are interesting because they reflect structural differences between English and Norwegian which must be handled in translation within this language pair. Also, these phenomena highlight the division between predictable and non-

predictable translational correspondences, and their distribution within the recorded data reveals a clear text-typological difference, as the majority of the identified cases occur in the fiction texts.

In the analysed texts, the most frequently identified subtype is the category of non-predictable specification, and the second-most frequent subtype is non-predictable despecification (cf. 6.3.1.3). Since explicitation is normal in translation, it is an expected result that specification occurs more frequently than despecification. The cases recorded of these two important subtypes also reflect the tendency that the use of nonfinite constructions is more widespread in English than in Norwegian. In particular, the phenomenon described in 5.2.2 as the nonfinite-finite pattern has caused a considerable number of type 4 correspondences where a difference in the amount of linguistically expressed information is caused only by the absence or presence of grammatical tense. Such cases have contributed to a higher degree of translational complexity in the English-to-Norwegian data than in those extracted from Norwegian-to-English (cf. 5.3.1–2). Given the difference in restrictedness between the two investigated text types, we would expect both specification and despecification to be more frequent in the fiction texts than in the law texts, but the frequencies of these subtypes almost indicate the opposite. In our view, the explanation for this is that the extent to which the nonfinite-finite pattern contributes to, respectively, specification and despecification is considerably greater within the pairs of law texts than within those of fiction.

The topic of 6.3.2 with subsections is denotational non-equivalence between translational correspondents. This category covers deviations in denotational properties, and the notion is applied in a wide sense, covering a range of semantic properties that contribute to the propositional potential of linguistic expressions (cf. 6.3.2). Included here is a class of correspondences where translationally linked noun phrases are generically referring, but differ with respect to number. We regard this as a linguistically predictable semantic deviation, since it seems to be correlated with generic contexts. It occurs mainly in the law texts. Non-predictable denotational differences constitute a heterogeneous type, and they are less frequent in the law texts than in the

fiction texts, which is to be expected, given that the norms of legal language will contribute to the preservation of meaning in translational relations.

Finally, 6.3.3 with subsections deals with correspondences between translationally related noun phrases which differ with respect to the marking of definiteness. The predictable instances occur mainly in the law texts, in contexts where the NPs are generically referring. As in the case of predictable denotational differences, they reflect regularities of the two language systems, but as they tend to occur in generic contexts, they are very rare in the investigated fiction texts. Non-predictable correspondences between NPs which differ in the marking of definiteness create deviations concerning referential newness and givenness. That such cases are rare in the law texts, in comparison to the fiction texts, is probably another indicator of the difference in restrictedness between the two text types.

PART V
SUMMING UP

7 Conclusions

7.1 The research questions

This thesis opened by stating two research questions, focussed on the topics of computability and text types, respectively. Firstly, we wanted to investigate to what extent it is possible to automatise translation in a selection of English-Norwegian parallel texts. By this we understand the computing of translations with no human intervention, and we assume an approach to machine translation based on linguistic knowledge. In order to answer this question, we have applied a measurement of translational complexity to the parallel texts. Secondly, we wanted to find out if there is a difference in the degree of translational complexity between the two text types, law and fiction, included in the empirical material. This chapter will draw some conclusions on the basis of our study, and these will be centred around three topics: our framework, the method, and the results. At the end, we indicate a possible extension of our analytical approach.

7.2 The framework

The present work is a product-oriented approach to complexity in translation. We have studied intersubjectively available relations between source texts and existing translations, and the scope of our investigation does not include aspects related to translation methods, or to the cognitive processes behind translation.

The notion of 'information' is a key issue in our analysis of parallel texts. Our understanding of 'information' is taken from information theory, where information is a quantitative notion, an objective commodity that exists independently of interpretive processes. Following Dretske (1981), we keep the quantitative information concept distinct from the semantic notion of 'informational content', which is the

message conveyed by a signal. The informational content is determined by the existence and transmission of a specific amount of information, and it is influenced by background information available to the recipient of the signal. These two notions have been applied in our analysis of translationally corresponding text units. Thus, we distinguish between information *about* a linguistic expression, and the informational content contained in it.

For the purpose of analysing translational complexity, we have developed a typology of information sources for translation. Distinctions in this typology are drawn in a way meant to reflect the kinds of information sources which are relevant in order to account for the observable relations between source and target expressions. The information sources are sorted along three different dimensions, each containing a binary division. Firstly, we distinguish between linguistic and extra-linguistic information; secondly, between general and task-specific information, and, thirdly, within the linguistic domain, between mono- and bilingual information.

The most important distinction in the typology is that between linguistic and extra-linguistic information, as it is associated with the limit of computability in the translational relation. In accord with Dyvik (1998, 1999, 2005), we assume that the translational relation between the inventories of simple and complex linguistic signs in two languages is predictable, and hence computable, from information about source and target language systems, and about how the systems correspond. That is, computable translations are predictable from the linguistic information coded in the source text, together with given, general information about SL and TL and their interrelations. As defined in chapter 2, a target language expression b_{L2} is a linguistically predictable translation (LPT) of a given source expression a_{L1} provided that b_{L2} shares a maximum of the meaning properties of a_{L1}, taking into account differences between the two language systems. Thus, a predictable translation is normally semantically equivalent with the source expression.[1] Following Dyvik (1999), we have defined 'literal translation' to be the same as linguistically predictable, or computable, translation.

[1] In chapter 6 we have discussed certain types of minor semantic deviations which fall within the domain of the linguistically predictable, because they follow from information about SL and TL and their interrelations.

Further, in the present approach the division between the linguistic and extra-linguistic domains of information defines the limit of computability in the translational relation. We have argued that, in principle, information about a language system is a finite domain, whereas the extra-linguistic domain is not. Hence, we may distinguish between linguistic and extra-linguistic information by delimiting the given language system, and, in line with Dyvik (2003), we thus relate the distinction to the way in which language systems are conceptually individuated. This, in turn, will be influenced by the purpose for which the language description is meant to be applied, and by empirical facts about language use.

Given these assumptions, we regard non-computable translations to be correspondences where it is not possible to predict the target expression from the information encoded in the source expression, together with given, general information about SL and TL and their interrelations. Non-computable translations require access to additional information sources, such as various kinds of general or task-specific extra-linguistic information, or task-specific linguistic information from the context surrounding the source expression.

In our approach, 'translational complexity' is associated with the notion of a 'translation task', i.e. the task of producing a particular target expression on the basis of the information encoded in the given source expression together with other information sources. Then, the degree of translational complexity in a given translation task is determined by the types and amounts of information needed to solve it, the accessibility of these information sources, and the effort required when they are processed.

For the purpose of measuring the complexity of the relation between a source text unit and its target correspondent, we apply a set of four correspondence types, organised in a hierarchy reflecting divisions between different linguistic levels, along with a gradual increase in the degree of translational complexity. In type 1, the least complex type, the corresponding strings are pragmatically, semantically, and syntactically equivalent, down to the level of the sequence of word forms. In type 2, source and target string are pragmatically and semantically equivalent, and equivalent with respect to syntactic functions, but there is at least one mismatch in the sequence of

constituents or in the use of grammatical form words. Within type 3, source and target string are pragmatically and semantically equivalent, but there is at least one structural difference violating syntactic functional equivalence between the strings. In type 4, there is at least one linguistically non-predictable, semantic discrepancy between source and target string. I.e., type 4 covers correspondences where the translation cannot be predicted from the source expression together with information about source and target language and their interrelations. The type hierarchy, ranging from 1 to 4, is characterised by an increase with respect to linguistic divergence between source and target string, an increase in the need for information and in the amount of effort required to translate, and a decrease in the extent to which there exist implications between relations of source-target equivalence at different linguistic levels.

Correspondences of types 1–3 constitute the domain of linguistically predictable, or computable, translations, where there is semantic, and possibly also syntactic, equivalence between source and target expression. Type 4 correspondences belong to the non-predictable, or non-computable, domain, where semantic equivalence is not fulfilled. To translate is to make a choice among alternative expressions in the target language. In predictable correspondences the chosen translation falls within the LPT set of the source expression, which is constrained by the interrelations between the two language systems, whereas in non-predictable correspondences the selected target expression falls within a larger and less clearly delimited set of possible translations, where various *parole*-related factors decide which among the target alternatives is the most appropriate choice.

7.3 The method

In the present approach, translationally corresponding strings are extracted from parallel texts, and each string pair is assigned one of the types defined by the correspondence hierarchy. The analysis is applied to running text, omitting no parts of it. As explained in chapter 4, the finite clause is chosen as the primary unit of analysis, because we have tried to find out how far it would be possible to automatise the translation of the selected texts. Another concern has been to be able to delimit trans-

lational units on the basis of surface syntactic structure. The main syntactic types among the recorded string pairs are matrix sentences, finite subclauses, and lexical phrases with finite clause as syntactic complement. Since syntactically dependent constructions like finite subclauses occur as translational units, the data include nested correspondences where a superordinate string pair contains one or more embedded string pairs.

The identification of translational units, as well as the classification of each extracted correspondence, have been done manually. The assignment of correspondence type to string pairs is an elimination procedure where we start by testing each correspondence for the lowest type and then move upwards in the hierarchy if the test fails. The analysis is an evaluation of the degree to which linguistic matching relations hold in each string pair. We assume that disambiguation of the source expression is not part of the translation task; hence, type assignment applies to the correspondence between an identified source string, given its relevant interpretation, and the parallel unit in the target text. Further, type assignment is done solely on the basis of the information encoded linguistically in the two strings. In cases of nested string pairs, embedded units are treated as opaque items, identified only by their syntactic category and function within the superordinate string, and the classification of a superordinate correspondence is done independently of the degree of complexity in embedded string pairs. Otherwise, it is a general principle that a string pair is assigned the correspondence type of its most complex non-opaque subpart.

As we have seen in chapters 3 and 5, there is a clear tendency for the language pair English-Norwegian that the lower correspondence types (1 and 2) are found in string pairs involving relatively short and syntactically simple units, whereas longer and more complex correspondences are normally of the higher types. Thus, there is a correlation between the size of the translational units and the complexity measurement for an entire text pair, and we may say that the measurement is relative to the chosen units of analysis.

Defining smaller units of translation than we have done could have resulted in a lower degree of complexity for individual text pairs if it had uncovered a larger number of the least complex correspondence types between units below the level of

finite clauses. But the usefulness of reducing the size of the units of analysis is limited: the smaller the units, the greater the frequency of zero correspondences — unless a text pair consists entirely of word-by-word correspondences.

It is anyway necessary to define standardised units of translation in our analysis. One could envisage an approach where the choice of unit would be determined by the parallel texts themselves, i.e. by identifying as small units as possible, provided that they have correspondents in the other text. This would be an investigation of linguistic tokens, not of types, as we have done in the present study. Also, it would yield a text pair-specific analysis, and would be more relevant to a study of the particular translation in relation to its original. For the purposes of the present project, it has been necessary to define a certain set of syntactic categories as units of extraction, since we wanted to analyse properties of text types and of a language pair, to the extent it has been possible within the limited set of data.

In our view, the finite clause has proved to be an appropriate translational unit. It is the primary unit that an MT system must handle in order to be useful, and the choice of the finite clause has not resulted in many instances of zero correspondences among the recorded string pairs. In some cases of minimal type 4 correspondences the finite clause may have appeared as a too limited unit of extraction. This has been illustrated by, e.g., string pairs involving non-finite constructions and finite subclauses, which have been assigned the most complex correspondence type only because temporal information is not encoded in the nonfinite expression. In such cases choosing the matrix sentence as the unit of analysis might yield a more accurate complexity measurement for an individual correspondence, but applying it in general would substantially increase the average length of the extracted strings, and would result in a higher degree of translational complexity for the entire text pair.

As we have made clear, our preference for the finite clause as the basic unit of analysis is motivated by the main purpose of this investigation, and we do not claim that this is necessarily the best option when analysing parallel text by means of the correspondence type hierarchy. Rather, the choice of unit must suit the aims of the analysis. As mentioned in chapter 1, the hierarchy has been adapted as a model for describing and analysing translational correspondences in three different studies

dealing with English-Portuguese. Tucunduva (2007) has chosen the noun phrase as the unit of analysis; in Silva (2008) the unit is mainly the finite clause, but non-clausal constructions, like headings, are also extracted when they are syntactically independent. In Azevedo (in progress) the unit of analysis is the verse in sonnets, which may consist of one or more finite clauses, or even of sub-sentential syntactic constructions.[2]

In chapter 6 we made the point that since about 90% of the analysed texts are included in correspondences of types 3 and 4, it is of interest to make more fine-grained analytical distinctions within these two main categories. As we have seen, categories of cross-linguistic deviations, structural as well as semantic, which resemble our set of semantic subtypes are identified also in other approaches to the analysis of translational correspondences, cf. Merkel (1999), Cyrus (2006), and Macken (2010).

As pointed out in chapter 1, the four main correspondence types are in principle language-pair independent, and so far the method has been applied to two language pairs, English-Norwegian, and English-Portuguese. However, some of the subcategories presented in chapter 6 are clearly specific to the relationship between English and Norwegian. This holds mainly for the predictable classes of deviations (e.g. differences in the use of possessives), which is natural since the predictable categories reflect interrelations between these two language systems. The subtypes that involve more general kinds of divergences, such as non-predictable specification and despecification, have a wider application which is not limited to the language pair English-Norwegian.

The fact that only a small corpus of about 68 000 words has been analysed in the present study, raises the question whether the present approach could be applied to large parallel corpora. Since the method is time-consuming, and implemented manually, scaling up would require either automatisation or using a team of annotators. With respect to the latter, it would be a challenge to secure consistency in the analysis, since the classification relies on linguistic judgments. As regards automati-

[2] I am indebted to Marco Antonio Esteves da Rocha, of the Federal University of Santa Catarina, for information on the studies presented in Tucunduva (2007), Silva (2008), and Azevedo (in progress).

sation, Merkel (1999: 209) comments on this in relation to his own model of translational correspondences. The identification of structural and semantic deviations in his analysis of parallel text is, like in our study, carried out by manual tagging of the data. He suggests that correspondence phenomena pertaining to syntactic structure and function might be analysed automatically by the aid of parsing tools, but he regards it as problematic to "decide semantic relationships" without a human annotator (1999: 209). We hold a similar view in relation to the present approach: cases of the two least complex correspondence types might be identified automatically, but to recognise occurrences of the semantic subtypes, as well as to distinguish between instances of the main types 3 and 4, may require deep-level linguistic analyses that are hard to automatise. Thus, we regard it as difficult to extend our method to large-scale processing of parallel texts, a point we will return to in 7.6.

We mentioned in chapter 1 that Hasselgård (1996) applied the correspondence type hierarchy, as defined by Dyvik (1993), to a small-scale study of word-order differences between English and Norwegian. She pointed out that for her purposes the main correspondence types constitute a too coarse-grained approach to contrastive language analysis (1996: 122–123). The methodology developed for the present study is not identical to the one applied by Hasselgård (1996), and our refinement of the classification model pertains to semantic phenomena, not to word order, but we agree with her view that for certain purposes it is necessary to draw finer distinctions than those given by the main types.

Then, in the studies where the correspondence type hierarchy is applied to English-Portuguese (Tucunduva 2007, Silva 2008, Azevedo in progress), the method has been found to provide a satisfactory approach to contrastive language analysis because it offers a consistent way of describing all pairs of translationally matched units in a given body of parallel text. Azevedo (in progress) presents a working hypothesis that in comparison to the notion described by Baker (1993) as *universal features of translation*, the correspondence type hierarchy offers a more adequate descriptive approach to parallel corpora.[3] We do not want to question the notion of translation

[3] Baker (1993: 242) describes universal features of translation as "patterns which are specific to translated texts."

universals, and in our view the contributions of Tucunduva (2007), Silva (2008), and Azevedo (in progress) illustrate that Baker's universals and our correspondence types, respectively, are associated with different purposes. Translation universals apply to the description of properties of translation, as opposed to other kinds of language use (cf. Baker 1993: 235), whereas the correspondence type hierarchy is developed for the characterisation of relations between translationally matched strings, and is suitable for analysing running parallel texts.

7.4 The results

The complexity measurements presented in chapter 5 for the investigated text pairs do not seem very promising in relation to the primary research question, the automatisation issue. At least, it appears likely that it would not be of any benefit to apply machine translation in the cases where more than 50% of the analysed text pairs are included in non-computable correspondences. This is the case for one pair of law texts, and three pairs of fiction texts.

But in relation to this, we might challenge our view that the chosen parallel texts, where the translations are produced by humans, provide an appropriate standard for judging whether automatisation is worthwhile. Since it is generally accepted that the application of MT tools in translation requires post-editing to secure the quality of the final product, the human-created target texts are problematic as a gold standard for automatic translation because they represent an ideal for the end result, and not for the raw output of an MT application. The chosen norm is probably an unrealistic, and perhaps also unfair, goal for MT development, especially since high-quality translation without post-editing, or revision, is uncommon also when the translator is human. Still, we have used this standard because our task has not been to evaluate the products of real systems, and because we wanted the complexity measurements of this study to show to what extent we assume that an ideal, rule-based system could simulate the given translations, with no human intervention, and purely on the basis of information about the two languages and their interrelations.

On this background, we may take a second look at the two cases where relatively large proportions of the analysed texts are included in computable correspondences.

As regards the law text translated from Norwegian into English, our estimate is that 60,9% of the parallel texts involve literal translation, and for the extract of André Brink's novel translated into Norwegian, the corresponding figure is 56,1%. In these text pairs it would be highly interesting to find out to what extent the string pairs of type 4 have been classified as such because of only one, or very few, semantic deviations between source and target units. That is, if the semantic difference between two corresponding strings is small, then the major part of the correspondence would involve literal translation, and it might be unproblematic for a post-editor to correct that subpart of the machine output that does not meet the standard. In such cases what Jurafsky and Martin (2009: 931) describe as the *edit cost* of post-editing would probably be low.

If post-editing amounts to simple corrections of linguistic errors that are few and easy to spot, then the *editing distance* between the machine output and the standard is small, and automatic translation may be useful.[4] On the other hand, if there are many errors in the output, and, if the revision also requires syntactic and/or semantic reorganisation of the automatically generated sentences, and maybe even careful considerations of the appropriateness of various target alternatives, then the editing distance is large, and it is perhaps more cost effective to do a fully manual translation.

The types of recurrent, non-predictable semantic deviations between translational units discussed in chapter 6 may indicate kinds of challenges that the post-editor will be faced with, i.e. types of properties that should be observed in the translation, but which cannot be predicted from the source expression without taking into account contextual information, and/or various kinds of extra-linguistic information. Here we cannot discuss this in detail, but we will assume that time-consuming decision making can be involved for instance in cases where the post-editor must apply background information to produce a target expression that deviates denotationally from the source expression, but may create a communicative effect in the target text audience similar to the effect of the source text. Also, we expect the editing distance

[4] The term *editing distance* is borrowed from information theory. According to Jurafsky and Martin (2009: 108), "[t]he minimum edit distance between two strings is the minimum number of editing operations (insertion, deletion, substitution) needed to transform one string into another."

between a literal, machine-generated translation and a target string with multiple semantic deviations in relation to the original to be considerably greater than the distance between a literal translation and a target expression exhibiting only a minimal semantic difference in relation to the source string.

In the discussions of recurrent non-predictable semantic deviations, we have seen certain minimal cases where we would regard the editing distance between assumed machine output and the given standard to be very small. One example could be the translation from Norwegian into English of prototypical instances of the inalienability pattern, more specifically in cases where the source unit contains no information about the possessor, which creates the problem of selecting the correct possessive determiner in the target expression. If we treat inalienability as a lexical property, this could be handled, e.g., by tagging the relevant noun as an inalienable in the output, and the post-editor would easily choose the right possessive on the basis of the surrounding context. Moreover, the editing distance would also be small in certain occurrences of the nonfinite-finite pattern, where a Norwegian finite subclause should preferably be translated into an English nonfinite construction, and the literal translation would yield a finite subclause in the English string as well.

These considerations of editing distance in relation to minimal type 4 correspondences lead over to the text type issue, which is our second research question. Chapter 5 showed that, on average, there is a lower degree of translational complexity across the pairs of law texts than across those of fiction, and this was primarily explained by referring to a fundamental difference in restrictedness between these two text types. If we consider that part of the data which falls outside the computable domain in each type of texts, we have observed in chapter 6, firstly, that there is a larger variety, as well as a higher frequency of semantic deviations within the type 4 correspondences extracted from fiction than within those compiled from law texts, and, secondly, that there is a higher frequency of minimal type 4 cases within the law data than within the string pairs extracted from fiction.

This makes the topic of minimal type 4 correspondences relevant. We argued in chapter 5 that such correspondences have the effect of concealing relations of semantic equivalence between source and target strings among the recorded data. At this

point it is more interesting that minimal type 4 cases involve translation tasks that are almost computable, and where automatic translation may generate a result that can be revised to a high-quality translation with very little effort. Naturally, there may be time-consuming cases of considering various target alternatives also when there is only one semantic property in the automatically produced literal translation which is regarded as contextually inappropriate. As explained in chapter 6, in the investigated text pairs well over two thirds of the minimal type 4 correspondences have been identified in the law texts. Also, we found a clear tendency that occurrences of the nonfinite-finite pattern had most frequently caused minimal cases of type 4 correspondences.

With respect to the investigated pairs of law texts, this means that we tentatively regard them as representing a text type where tools for automatic translation may be helpful, provided that the effort involved in post-editing is smaller than that of manual translation. This is perhaps most likely the case for the text pair *Petro*, where we assume that 60,9% of the parallel texts involve computable translation tasks. In the *AEEA* pair of law texts the corresponding figure is merely 38,8%. In that case the potential helpfulness of automatisation would be even more strongly determined by the edit cost. Possibly, translating the *EEA Agreement* is a task for computer-aided translation, rather than for MT. Surely, due to their repetitive character, as shown in chapter 4, both pairs of law texts are cases where e.g. translation memory tools would be useful.

Our careful optimism in relation to the automatisation of law text translation is not only inspired by the findings of the present investigation, but also by the recent emergence of a research field combining insights and methods from artificial intelligence, human language technology, the law, legal informatics, and studies of legal language. E.g., under the heading *Semantic Processing of Legal Texts*, Francesconi et al. (2010) have compiled a set of contributions dealing with topics such as information extraction from legal texts, the construction of legal knowledge resources, semantic indexing, summarisation, and translation evaluation for the legal domain. Furthermore, Johnsen (2010), and Johnsen and Berre (2010) discuss the semantic modelling of law text with reference to Norwegian. Contributions like these

indicate that there is progress in relation to the development of automatic analysis of law text. Moreover, since the language of law is highly specialised and norm-controlled it is, in its own right, of interest to the field of language technology as a testing ground for applications developed for the processing of natural language, translation included.

As regards the investigated fiction texts, it is our view that post-editing of automatically generated translations would be laborious and not cost effective also in the case of the text pair showing a relatively low degree of translational complexity, and this is mainly because the proportion of minimal type 4 correspondences is smaller in the fiction texts than in the law texts. In our opinion, the translation of fiction is not a task for MT, since it demands the linguistic intuitions of a skilled human translator. This is, however, not to say that literal translations are necessarily avoided in manual translation, and they will be chosen in cases where they appear as the preferred alternative considering all information available to the translator. This illustrates a principled difference between human and automatic translation, as conceived in our framework. The machine generates a literal translation because it does not have access to other sources of information than what is needed for producing a linguistically predictable target text. The human translator, on the other hand, creates a literal translation when it appears to be the most appropriate choice also on the basis of information falling outside the given, general linguistic sources.

7.5 Relevance of the study

The relevance of the present study for rule-based MT follows from the definition of the correspondence type hierarchy, since it is designed according to assumptions about how translations can be computed on the basis of formal descriptions of source and target language systems and their interrelations. In chapter 1 we pointed out that because statistical machine translation depends on the availability of relevant and sufficient information about translational correspondences, we assume that the results of our analysis are relatable not only to the linguistic approaches to MT, but also to some extent to the statistical ones. Moreover, it has become the general view that there is a limit to how far the purely statistical methods can reach in terms of

translation quality, and for more than a decade research efforts have been put into hybrid approaches where statistical techniques are combined with some kind of semantic and/or syntactic processing. In our view, it seems unlikely that automatic translation can do without linguistic information, especially in the light of the pervasive ambiguity of natural language expressions. A further idea on the relevance of our method for machine translation will be presented in 7.6.

Although the computability issue has been our primary concern, the present contribution is also pertinent to translation studies, and to contrastive linguistic research. A side-effect of the complexity measurements is that the analysis provides certain indications of to what extent the individual translations are faithful to the corresponding originals. As regards the results discussed in chapter 5, the division between computable and non-computable correspondences has revealed differences among the text pairs concerning the extent to which the recorded source and target units are semantically equivalent. This can be related to the dimension of faithfulness to the original text, but only in a certain degree. As explained in chapter 5, the syntactic extraction criteria, in combination with the classification principles, have in some cases concealed relations of semantic equivalence, and, as shown by the discussions of the pairs of law texts, there may be certain extra-linguistic factors that contribute to semantic non-equivalence in individual string pairs although larger sequences of translationally corresponding texts convey the same informational content.

The discussions of semantic subtypes in chapter 6 shed some more light on the issue of faithfulness to the source text. In general, the subtypes can be seen as descriptions of ways in which corresponding strings differ with respect to linguistically encoded meaning, and thus the frequencies of the non-predictable subtypes may indicate to what extent a translation is faithful to the original. Since semantic deviations are noticeably more frequent, and more varied, within the fiction data than within the law data, and since instances of the nonfinite-finite pattern appear to have caused the majority of the cases of non-predictable specification and despecification in the law texts, it is primarily for the analysed fiction texts that the complexity measurements reveal differences concerning the degree of faithfulness to the original. This is to be expected, given the norms of legal translation. Thus, there is a clearer tendency in the

fiction texts than in the law texts that semantic deviations between translational units are triggered by the translators' individual choices rather than by textual and translational norms, or by systematic differences between English and Norwegian.

The present work may also be of relevance to translation studies as a way of describing interrelations between source and target texts. It is then noteworthy that our study of translational complexity aims not only to identify deviations between corresponding units, but also relations of linguistic equivalence, which are captured by correspondence types 1, 2, and 3. In this respect, the present approach can be seen as a response to Chesterman's (2005: 27) position that in translation studies more effort should be put into developing "typologies of similarity" along with the "typologies of differences (shifts)."

In 7.3 we have already discussed the usability of our analytical approach in contrastive language studies. Since the empirical material includes translations as well as originals for both languages, the present contribution is not only an investigation of target texts in relation to source texts, but can also be seen as a limited cross-linguistic study. In this respect we find it interesting that the analysis of string pairs to some extent reflects interrelations between the English and Norwegian language systems. This pertains, e.g., to the discussions of progressive aspect, of the inalienability pattern, and of the nonfinite-finite pattern. Moreover, our study has shown that in order to understand why a particular expression has been chosen, either by the source text writer, or by the translator, it is often fruitful to observe an interplay between the levels of *langue* and *parole*. The structure of the language system defines the alternatives available for the encoding of a certain semantic content, and aspects related to the specific utterance may influence the choice when there is more than one possible expression. This was seen, e.g. in chapter 6 in discussions of correspondences between nonfinite constructions and finite subclauses, and between passive and active sentences. In both cases textual norms appeared to be influential for the choices of expression, and we saw that for given text types, the manifestations of such norms may vary between different languages.

7.6 Further application

We will end this study by suggesting a possible extension of the analytical framework. As pointed out in 7.3, it is not unproblematic to apply the present method to a large parallel corpus. In our view, the approach could be useful as a diagnostic tool for the feasibility of machine translation in relation to specific text types. That is, by applying the method to limited selections of parallel texts of the same type, it would be possible to estimate to what extent the target text could be generated automatically. If the proportion of assumed computable correspondences would exceed a chosen threshold, it might be worthwhile to tune an MT system for the given language pair to the text type in question, for instance by developing lexicon modules covering the relevant subject domain.

But, as discussed in 7.4, it is not only the proportion of computable correspondences which may indicate whether automatic translation could be helpful for a given text pair; this is also determined by the editing distance between potential machine output and a given target text norm, and in this respect we have discussed the importance of minimal type 4 correspondences. In our view, it would be fruitful to extend the classification model by integrating a fifth correspondence type to be assigned to the minimally non-computable string pairs. As we evaluated the outcome of the complexity measurements for the analysed texts, we saw a need for calculating the proportion of such correspondences in terms of string length within each text pair. In principle, this could have been counted manually, but it would have been very time-consuming. To calculate it automatically would have required the implementation of a fifth category of string pairs in the software used for recording translational correspondences.

Moreover, in order to decide whether automatic translation could be feasible for a given text type it would also be relevant to consider what kinds of challenges the post-editor would face, firstly, when improving the output for the minimally non-computable translation tasks, and, secondly, when editing the automatic translations in cases which are non-computable due to several factors. The analyst would have to draw a conclusion for the given text type by considering all aspects that may add to the burden of post-editing. In our view, the present study is a reminder that the task of

translation presents certain challenges that appear as probably too complex for machines, and which are certainly also non-trivial for humans.

References

Primary sources

Law texts:

Agreement on the European Economic Area. Articles 1–99. 1992. The Norwegian Royal Ministry of Foreign Affairs.

Avtale om det Europeiske Økonomiske Samarbeidsområde. Artikler 1–99. 1992. The Norwegian Royal Ministry of Foreign Affairs.

Lov om petroleumsvirksomhet. §§1–65. 1994. The Norwegian Petroleum Directorate.

Act relating to petroleum activities. Sections 1–65. 1994. The Norwegian Petroleum Directorate.

Fiction texts:

Brink, André. 1984. *The Wall of the Plague*. Pp. 13–23. London: Faber and Faber.

Brink, André. 1984. *Pestens mur*. Pp. 11–20. Translated by Per Malde. Oslo: H. Aschehoug & Co (W. Nygaard) AS.

Hansen, Erik Fosnes. 1990. *Salme ved reisens slutt*. Pp. 15–28. Oslo: J. W. Cappelens Forlag AS.

Hansen, Erik Fosnes. 1996. *Psalm at Journey's End*. Pp. 7–18. Translated by Joan Tate. New York: Farrar, Straus and Giroux.

Lessing, Doris. 1985. *The Good Terrorist*. Pp. 5–15. London: Jonathan Cape.

Lessing, Doris. 1985. *Den gode terroristen*. Pp. 5–15. Translated by Kia Halling. Oslo: Gyldendal Norsk Forlag AS.

Vik, Bjørg. 1979. *En håndfull lengsel*. Pp. 9–23. Oslo: J. W. Cappelens Forlag AS.

Vik, Bjørg. 1983. *Out of Season and Other Stories*. Pp. 1–13. Translated by David McDuff and Patrick Browne. London: Sinclair Browne.

Other texts used to provide translation examples discussed in the dissertation:

Grafton, Sue. 1990. *"D" is for Deadbeat*. London: Pan Books Ltd.

Grafton, Sue. 1993. *"D" for druknet*. Translated by Isak Rogde. Oslo: Tiden Norsk Forlag.

Secondary sources

Abbott, H. Porter. 2002. *The Cambridge Introduction to Narrative.* Cambridge: Cambridge University Press.

Agirre, Eneko and Philip Edmonds (eds). 2006. *Word Sense Disambiguation. Algorithms and Applications. Text, Speech and Language Technology* 33. Dordrecht: Springer.

Aijmer, Karin, Bengt Altenberg, and Mats Johansson (eds). 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4–5 March 1994. Lund Studies in English* 88. Lund: Lund University Press.

Aijmer, Karin and Bengt Altenberg. 2002. Zero translations and cross-linguistic equivalence: Evidence from the English-Swedish Parallel Corpus. In: Breivik and Hasselgren (eds), 2002, 19–41.

Allwood, Jens, Lars-Gunnar Andersson, and Östen Dahl. 1977. *Logic in Linguistics*. *Cambridge Textbooks in Linguistics.* Cambridge, New York, and Melbourne: Cambridge University Press.

ALPAC. 1966. *Language and Machines: Computers in Translation and Linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Research Council. Publication 1416, National Academy of Sciences, Washington D.C.

Alsina, Alex. 1996. *The Role of Argument Structure in Grammar. Evidence from Romance*. Stanford: CSLI Publications.

Alt, Franz L. (ed). 1960. *Advances in Computers* 1. New York: Academic Press.

Alves, Fabio (ed). 2003. *Triangulating Translation: Perspectives in process oriented research. Benjamins Translation Library* 45. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Azevedo, Flávia. In progress. Investigating the problem of codifying linguistic knowledge in two translations of Shakespeare's sonnets: a corpus-based study. Doctoral dissertation. Federal University of Santa Catarina, Florianópolis.

Baker, Mona. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. In: Baker et al. (eds), 1993, 233–250.

Baker, Mona (ed). 1998. *Routledge Encyclopedia of Translation Studies.* London and New York: Routledge.

Baker, Mona (ed). 2010. *Critical Readings in Translation Studies.* London and New York: Routledge.

Baker, Mona, Gill Francis, and Elena Tognini-Bonelli (eds). 1993. *Text and Technology. In Honour of John Sinclair.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Bar-Hillel, Yehoshua. 1960. The Present Status of Automatic Translation of Languages. In: Alt (ed), 1960, 91–163.

Baron, Irène, Michael Herslund, and Finn Sørensen (eds). 2001. *Dimensions of Possession. Typological Studies in Language* 47. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Barton, G. Edward, Robert C. Berwick, and Eric Sven Ristad. 1987. *Computational Complexity and Natural Language.* Cambridge, Massachusetts, and London, England: The MIT Press.

Barwise, Kenneth Jon and John Perry. 1983. *Situations and Attitudes*. Cambridge, Massachusetts: The MIT Press.

Bateman, John and Michael Zock. 2003. Natural Language Generation. In: Mitkov (ed), 2003, 284–304.

Beeby, Allison, Doris Ensinger, and Marisa Presas (eds). 2000. *Investigating Translation. Selected Papers from the 4th International Congress on Translation, Barcelona, 1998. Benjamins Translation Library* 32. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Behrens, Bergljot. 1998. *Contrastive discourse: An interlingual approach to the interpretation and translation of free ING-participial adjuncts.* Doctoral dissertation. University of Oslo.

Behrens, Bergljot. 1999. A dynamic semantic approach to translation assessment. ING-participial adjuncts and their translation into Norwegian. In: Doherty (ed), 1999, 90–111.
Bhatia, Vijay K. 1997. Translating Legal Genres. In: Trosborg (ed), 1997, 203–214.
Bhatia, Vijay K. 2010. Specification in legislative writing: accessibility, transparency, power and control. In: Coulthard and Johnson (eds), 2010, 37–50.
Biber, Douglas. 1988. *Variation across speech and writing.* Cambridge, New York, and Melbourne: Cambridge University Press.
Biber, Douglas. 1989. A Typology of English Texts. *Linguistics* 27:1, 3–43.
Blum-Kulka, Shoshana. 1986. Shifts of Cohesion and Coherence in Translation. In: House and Blum-Kulka (eds), 1986, 17–35.
Borin, Lars (ed). 2002. *Parallel Corpora, Parallel Worlds. Language and Computers: Studies in Practical Linguistics* 43. Amsterdam and New York: Rodopi.
Bowers, Frederick. 1989. *Linguistic Aspects of Legislative Expression.* Vancouver: University of British Columbia Press.
Breivik, Leiv Egil and Angela Hasselgren (eds). 2002. *From the COLT's mouth ... and others'. Language and Computers: Studies in Practical Linguistics* 40. Amsterdam and New York: Rodopi.
Bresnan, Joan. 2001. *Lexical-Functional Syntax.* Oxford: Blackwell.
Brower, Reuben A. (ed). 1966. *On Translation*. Cambridge, Mass.: Harvard University Press.
Cao, Deborah. 2007. *Translating Law. Topics in Translation* 33. Clevedon, Buffalo, and Toronto: Multilingual Matters Ltd.
Cao, Deborah. 2010. Translating legal language. In: Coulthard and Johnson (eds), 2010, 78–91.
Casanova, Pascale. 2010. Consecration and accumulation of literary capital: Translation as unequal exchange. Translated by Siobhan Brownlie. In: Baker (ed), 2010, 287–303.
Catford, John C. 1965. *A Linguistic Theory of Translation. An Essay in Applied Linguistics. Language and Language Learning Series.* Oxford: Oxford University Press.
Chappell, Hilary and William McGregor (eds). 1996a. *The Grammar of Inalienability. A Typological Perspective on Body Part Terms and the Part-Whole Relation. Empirical Approaches to Language Typology* 14. Berlin and New York: Mouton de Gruyter.
Chappell, Hilary and William McGregor. 1996b. Prolegomena to a theory of inalienability. In: Chappell and McGregor (eds), 1996a, 3–30.
Chesterman, Andrew. 1997. *Memes of Translation. The Spread of Ideas in Translation Theory. Benjamins Translation Library* 22. Amsterdam and Philadelphia: John Benjamins Publishing Company.
Chesterman, Andrew. 2005. Problems with Strategies. In: Károly and Fóris (eds), 2005, 17–28.
Chesterman, Andrew, Natividad Gallardo San Salvador, and Yves Gambier (eds). 2000. *Translation in Context. Selected Contributions from the EST Congress, Granada 1998. Benjamins Translation Library* 39. Amsterdam and Philadelphia: John Benjamins B.V.
Comrie, Bernard. 1985. *Tense. Cambridge Textbooks in Linguistics.* Cambridge, New York, and Melbourne: Cambridge University Press.
Correia, Renato. 2003. Translation of EU Legal Texts. In: Tosi (ed), 2003, 38–44.
Coulthard, Malcolm and Alison Johnson (eds). 2010. *The Routledge Handbook of Forensic Linguistics.* London and New York: Routledge.
Cruz, Peter de. 1995. *Comparative Law in a Changing World.* London: Cavendish Publishing Limited.

Cyrus, Lea. 2006. Building a Resource for Studying Translation Shifts. In: *Proceedings of the Fifth International Conference on Linguistic Resources and Evaluation (LREC-2006)*, 1240-1245. Genoa, Italy.

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity. Studies in Language Companion Series* 71. Amsterdam and Philadelphia: John Benjamins B.V.

Dahl, Östen, and Maria Koptjevskaja-Tamm. 1998. Alienability Splits and the Grammaticalization of Possessive Constructions. In: Haukioja (ed), 1998, 38–49.

Dahl, Östen, and Maria Koptjevskaja-Tamm. 2001. Kinship in grammar. In: Baron et al. (eds), 2001, 201–225.

Dahl, Willy. 1995. *Stil og struktur. Linjer i norsk fiksjonsprosa gjennom to hundre år.* Bergen: Eide forlag.

Dalrymple, Mary. 2001. *Lexical Functional Grammar. Syntax and Semantics* 34. San Diego, California: Academic Press.

Dijk, Teun A. van (ed). 1997. *Discourse as Structure and Process. Discourse Studies: A Multidisciplinary Introduction 1*. London, Thousand Oaks, and New Dehli: Sage Publications.

Doczekalska, Agnieszka. 2007. Production and application of multilingual law. The principle of equality of authentic texts and the value of subsequent translation. In: Kredens and Goźdź-Roszkowski (eds), 2007, 57–66.

Doherty, Monika (ed). 1996. *Information Structure: A Key Concept for Translation Theory. Linguistics* 34:3 (Special Issue). Berlin and New York: Mouton de Gruyter.

Doherty, Monika (ed). 1999. *Sprachspezifische Aspekte der Informationsverteilung. Studia Grammatica* 47. Berlin: Akademie Verlag.

Dorr, Bonnie J., Pamela W. Jordan, and John W. Benoit. 1998. *A Survey of Current Paradigms in Machine Translation.* Technical report. University of Maryland, College Park.

Dretske, Fred I. 1981. *Knowledge and the Flow of Information.* Oxford: Basil Blackwell.

Dyvik, Helge. 1990. *The PONS Project: Features of a Translation System. Skriftserie fra Institutt for fonetikk og lingvistikk* 39, B. University of Bergen.

Dyvik, Helge. 1993. Text Pair Mapper. Unpublished manuscript. University of Bergen.

Dyvik, Helge. 1995. Exploiting Structural Similarities in Machine Translation. *Computers and the Humanities* 28, 225–234.

Dyvik, Helge. 1998. A translational basis for semantics. In: Johansson and Oksefjell (eds), 1998, 51–86.

Dyvik, Helge. 1999. On the complexity of translation. In: Hasselgård and Oksefjell (eds), 1999, 215–230.

Dyvik, Helge. 2003. Translations as a Semantic Knowledge Source. Unpublished manuscript. University of Bergen. URL, last accessed on 19th of May 2011: http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transknow.pdf.

Dyvik, Helge. 2005. Translations as a semantic knowledge source. In: Langemets and Penjam (eds), 2005, 27–38.

Edmonds, Philip and Graeme Hirst. 2002. Near-Synonymy and Lexical Choice. *Computational Linguistics* 28:2, 105–144.

Eiter, Thomas, Adil El Ghali, Sergio Fernàndez, Stijn Heymans, Thomas Krennwallner, and François Levy (eds). 2010. *Proceedings of BuRO 2010: 1st International Workshop on Business Models, Business Rules and Ontologies.* Brixen, Italy.

Ericsson, K. Anders and Herbert A. Simon. 1984. *Protocol Analysis. Verbal Reports as Data.* Cambridge, Massachusetts: The MIT Press.

Ericsson, K. Anders and Herbert A. Simon. 1993. *Protocol Analysis. Verbal Reports as Data.* Second revised edition. Cambridge, Massachusetts, and London, England: The MIT Press.

Eschbach, Achim and Wendelin Rader (eds). 1980. *Literatursemiotik I. Methoden — Analysen — Tendenzen*. Tübingen: Gunter Narr Verlag.

Faarlund, Jan Terje, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Fabricius-Hansen, Cathrine. 1996. Informational density: a problem for translation and translation theory. In: Doherty (ed) 1996, 521–565.

Fabricius-Hansen, Cathrine. 1999. Information packaging and translation: Aspects of translational sentence splitting (German - English/Norwegian). In: Doherty (ed), 1999, 175–214.

Fabricius-Hansen, Cathrine and Dag T. T. Haug (eds). Forthcoming. *Big Events, Small Clauses: The Grammar of Elaboration*.

Fenstad, Jens Erik, Per-Kristian Halvorsen, Tore Langholm and Johan van Benthem. 1987. *Situations, Language and Logic. Studies in Language and Philosophy* 34. Dordrecht: Reidel.

Francesconi, Enrico, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia (eds). 2010. *Semantic Processing of Legal Texts. Where the Language of Law Meets the Law of Language. Lecture Notes in Artificial Intelligence* 6036. Berlin and Heidelberg: Springer.

Gale, William A. and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 9:1, 75-102.

Gentzler, Edwin. 2001. *Contemporary Translation Theories.* Second edition. *Topics in Translation* 21. Clevedon, Buffalo, Toronto, and Sydney: Multilingual Matters Ltd.

Gomard, Kirsten and Sven-Olaf Poulsen (eds). 1978. *Stand und Möglichkeiten der Übersetzungswissenschaft*. *Acta Jutlandica* 52. Aarhus: Det Lærde Selskab.

Grishman, Ralph and Richard I. Kittredge (eds). 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Gunderson, Keith (ed). 1975. *Language, Mind, and Knowledge.* Don Mills, Ontario: Burns and MacEachern Limited.

Halverson, Sandra L. 2000. Prototype effects in the 'translation' category. In: Chesterman et al. (eds), 2000, 3–16.

Harris, Zellig. 1968. *Mathematical Structures of Language.* New York: Wiley-Interscience.

Hasselgård, Hilde. 1996. Some methodological issues in a contrastive study of word order in English and Norwegian. In: Aijmer et al. (eds), 1996, 113–126.

Hasselgård, Hilde. Forthcoming. Possessive absolutes in English and their Norwegian correspondences. In: Fabricius-Hansen and Haug (eds), forthcoming.

Hasselgård, Hilde and Signe Oksefjell (eds). 1999. *Out of Corpora. Studies in Honour of Stig Johansson. Language and Computers: Studies in Practical Linguistics* 26. Amsterdam and Atlanta, GA: Rodopi.

Hatim, Basil and Ian Mason. 1990. *Discourse and the Translator. Language in Social Life Series.* London and New York: Longman.

Haukioja, Timo (ed). 1998. *Papers from the 16th Scandinavian Conference of Linguistics. Publications of the Department of Finnish and General Linguistics of the University of Turku* 60. University of Turku.

Heine, Bernd. 1997. *Possession. Cognitive source, forces, and grammaticalization.* Cambridge: Cambridge University Press.

Herslund, Michael and Irène Baron. 2001. Introduction: Dimensions of possession. In: Baron et al. (eds) 2001, 1–25.

House, Juliane. 1997. *Translation Quality Assessment: A Model Revisited. Tübinger Beiträge zur Linguistik* 410. Tübingen: Gunter Narr Verlag.

House, Juliane and Shoshana Blum-Kulka (eds). 1986. *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies.* Tübingen: Gunter Narr Verlag.

Huang, Yan. 2007. *Pragmatics. Oxford Textbooks in Linguistics.* Oxford and New York: Oxford University Press.

Hurtado Albir, Amparo and Fabio Alves. 2009. Translation as a cognitive activity. In: Munday (ed), 2009, 54–73.

Hutchins, William John. 1986. *Machine Translation: Past, Present, Future.* Chichester: Ellis Horwood Ltd.

Hutchins, William John and Harold L. Somers. 1992. *An Introduction to Machine Translation.* London: Academic Press Limited.

Hutton, Christopher. 2009. *Language, Meaning and the Law.* Edinburgh: Edinburgh University Press.

Ide, Nancy and Yorick Wilks. 2006. Making Sense about Sense. In: Agirre and Edmonds (eds), 2006, 47–73.

Izquierdo, Isabel García and Josep Marco Borillo. 2000. The Degree of Grammatical Complexity in Literary Texts as a Translation Problem. In: Beeby et al. (eds), 2000, 65–74.

Jääskeläinen, Riitta. 1999. *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating. University of Joensuu Publications in the Humanities* 22. University of Joensuu.

Jääskeläinen, Riitta. 2000. Focus on Methodology in Think-aloud Studies on Translating. In: Tirkkonen-Condit and Jääskeläinen (eds), 2000, 71–82.

Jääskeläinen, Riitta and Sonja Tirkkonen-Condit. 1991. Automatised Processes in Professional vs. Non-Professional Translation: A Think-Aloud Protocol Study. In: Tirkkonen-Condit (ed), 1991, 89–109.

Jakobsen, Arnt Lykke. 2003. Effects of Think Aloud on Translation Speed, Revision and Segmentation. In: Alves (ed), 2003, 69–95.

Jakobson, Roman. 1959. On Linguistic Aspects of Translation. In: Brower (ed), 1966, 232–239.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In: Johansson and Oksefjell (eds), 1998, 3–24.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies. Studies in Corpus Linguistics* 26. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Johansson, Stig and Signe Oksefjell (eds). 1998. *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. *Language and Computers: Studies in Practical Linguistics* 24. Amsterdam and Atlanta, GA: Rodopi.

Johansson, Stig, Jarle Ebeling, and Signe Oksefjell. 1999/2002. *English-Norwegian Parallel Corpus: Manual.* Department of British and American Studies, University of Oslo.

Johnsen, Åshild. 2010. *Forstå det den som kan. Semantisk modellering av juridisk regelverk med bruk av SBVR — en brobygger mellom jus og IT.* Master's thesis. University of Oslo.

Johnsen, Åshild and Arne-Jørgen Berre. 2010. A Bridge between Legislator and Technologist — Formalization in SBVR for Improved Quality and Understanding of Legal Rules. In: Eiter et al. (eds), 2010, 29–39.

Jurafsky, Daniel S. and James H. Martin. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, New Jersey: Prentice-Hall Inc.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Second edition. Upper Saddle River, New Jersey: Pearson Education.

Károly, Krisztina and Ágota Fóris (eds). 2005. *New Trends in Translation Studies. In Honour of Kinga Klaudy.* Budapest: Akadémiai Kiadó.

Kay, Martin, Jean Mark Gawron, and Peter Norvig. 1994. *Verbmobil. A Translation System for Face-to-Face Dialog. CSLI Lecture Notes* 33. Stanford: Center for the Study of Language and Information.

King, Margaret. 1986. *A Tutorial on Machine Translation.* Pre-COLING'86 Tutorial Program. August 25 to 29, 1986. Bonn.

Kittel, Harald, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, and Fritz Paul (eds). 2004. *Übersetzung — Translation — Traduction. Ein internationales Handbuch zur Übersetzungsforschung / An International Encyclopedia of Translation Studies / Encyclopédie internationale de la recherche sur la traduction.* Volume 1. Berlin and New York: Walter de Gruyter.

Kittel, Harald, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, and Fritz Paul (eds). 2007. *Übersetzung — Translation — Traduction. Ein internationales Handbuch zur Übersetzungsforschung / An International Encyclopedia of Translation Studies / Encyclopédie internationale de la recherche sur la traduction.* Volume 2. Berlin and New York: Walter de Gruyter.

Kittredge, Richard I. 1987. The Significance of Sublanguage for Automatic Translation. In: Nirenburg (ed), 1987, 59–67.

Kittredge, Richard I. and John Lehrberger (eds). 1982. *Sublanguage. Studies of Language in Restricted Semantic Domains.* Berlin and New York: Walter de Gruyter.

Kjær, Anne Lise. 2007. Legal translation in the European Union: A research field in need of a new approach. In: Kredens and Goźdź-Roszkowski (eds), 2007, 69–95.

Klaudy, Kinga and Krisztina Károly. 2003. Implicitation in Translation: An Empirical Justification of Operational Asymmetry in Translation. Paper presented to the 10th International Conference on Translation and Interpreting, *Translation Targets*, 11–13 September 2003.

Koller, Werner. 1979. *Einführung in die Übersetzungswissenschaft.* Heidelberg: Quelle & Meyer.

Koller, Werner. 1992. *Einführung in die Übersetzungswissenschaft.* Second revised edition. Heidelberg: Quelle & Meyer.

Koller, Werner. 1995. The Concept of Equivalence and the Object of Translation Studies. *Target* 7:2, 191–222. Amsterdam and Philadelphia: John Benjamins B.V.

Koot, Hans van de. 1995. The Computational Complexity of Natural Language Recognition. A Tutorial Overview. *Lingua* 97:1, 37–80. Amsterdam: Elsevier Science B.V.

Kredens, Krzysztof and Stanisław Goźdź-Roszkowski (eds). 2007. *Language and the Law: International Outlooks. Łódź Studies in Language* 16. Frankfurt am Main: Peter Lang.

Krings, Hans P. 1986. *Was in den Köpfen von Übersetzern vorgeht.* Tübingen: Gunter Narr Verlag.

Kussmaul, Paul. 1997. Text-Type Conventions and Translating: Some Methodological Issues. In: Trosborg (ed), 1997, 67–83.

Landers, Clifford. 2001. *Literary Translation. A Practical Guide. Topics in Translation* 22. Clevedon, Buffalo, Toronto, and Sydney: Multilingual Matters Ltd.

Langemets, Margit and Priit Penjam (eds). 2005. *Proceedings of the Second Baltic Conference on Human Language Technologies.* Institute of Cybernetics, Tallinn University of Technology, and Institute of the Estonian Language, Tallinn.

Laurén, Christer, Johan Myking, and Heribert Picht. 1997. *Terminologi som vetenskapsgren.* Lund: Studentlitteratur.

Leech, Geoffrey. 2008. *Language in Literature. Style and Foregrounding.* Harlow: Pearson Education Limited.

Leech, Geoffrey and Mick Short. 2007. *Style in Fiction. A Linguistic Introduction to English Fictional Prose.* Second edition. *English Language Series.* Harlow: Pearson Education Limited.

Lehrberger, John and Laurent Bourbeau. 1988. *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation. Studies in French and General Linguistics* 15. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Leuven-Zwart, Kitty M. van. 1989. Translation and Original. Similarities and Dissimilarities, I. *Target* 1:2, 151–181. Amsterdam and Philadelphia: John Benjamins B.V.

Leuven-Zwart, Kitty M. van. 1990. Translation and Original. Similarities and Dissimilarities, II. *Target* 2:1, 69–95. Amsterdam and Philadelphia: John Benjamins B.V.

Lévy-Bruhl, Lucien. 1914. L'expression de la possession dans les langues mélanésiennes. *Mémoire de la Société de Linguistique de Paris* 19: 96–104.

Löbner, Sebastian. 2002. *Understanding Semantics. Understanding Language Series.* London: Hodder Education.

Locke, William N. and Andrew D. Booth (eds). 1955. *Machine Translation of Languages: Fourteen Essays.* New York: Technology Press of MIT, and Wiley and Sons.

Lörscher, Wolfgang. 1991. Thinking-Aloud as a Method for Collecting Data on Translation Processes. In: Tirkkonen-Condit (ed), 1991, 67–78.

Lyons, John. 1977. *Semantics: 1.* Cambridge, New York, and Melbourne: Cambridge University Press.

Macken, Lieve. 2010. *Sub-sentential alignment of translational correspondences.* PhD thesis. Antwerp: University Press Antwerp.

Mattila, Heikki E. S. 2006. *Comparative Legal Linguistics.* London: Ashgate.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics. An Introduction.* Second edition. Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-Based Language Studies. An advanced resource book. Routledge Applied Linguistics*. London and New York: Routledge.

Merkel, Magnus. 1999. *Understanding and enhancing translation by parallell text processing. Linköping Studies in Science and Technology. Dissertation No.* 607. Linköping University.

Merkel, Magnus, Mikael Andersson, and Lars Ahrenberg. 2002. The PLUG Link Annotator — interactive construction of data from parallel corpora. In: Borin (ed), 2002, 151–168.

Miestamo, Matti. 2006. On the Complexity of Standard Negation. In: Suominen et al. (eds), 2006, 345–356.

Mitkov, Ruslan (ed). 2003. *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press.

Munday, Jeremy. 2008. *Introducing Translation Studies. Theories and applications.* London and New York: Routledge.

Munday, Jeremy (ed). 2009. *The Routledge Companion to Translation Studies.* London and New York: Routledge.

Newmark, Peter. 1981. *Approaches to Translation.* Oxford: Pergamon.
Nida, Eugene A. 1975. *Language Structure and Translation.* Stanford, California: Stanford University Press.
Nirenburg, Sergei (ed). 1987. *Machine Translation. Theoretical and Methodological Issues. Studies in Natural Language Processing.* Cambridge: Cambridge University Press.
Nirenburg, Sergei (ed). 1993. *Progress in Machine Translation.* Amsterdam: IOS Press, Inc.
Nirenburg, Sergei, Harold L. Somers, and Yorick Wilks (eds). 2003. *Readings in Machine Translation.* Cambridge, Massachusetts, and London, England: The MIT Press.
Nordrum, Lene. 2007. *English Lexical Nominalizations in a Norwegian-Swedish Contrastive Perspective.* Doctoral dissertation. University of Gothenburg.
Ochs, Elinor. 1997. Narrative. In: van Dijk (ed), 1997, 185–207.
Palumbo, Giuseppe. 2009. *Key Terms in Translation Studies.* London and New York: Continuum.
Popper, Karl R. 1979. *Objective Knowledge. An Evolutionary Approach.* Oxford: Clarendon Press. First edition 1972.
Pustejovsky, James. 1995. *The Generative Lexicon.* Cambridge, Massachusetts, and London, England: The MIT Press.
Pym, Anthony. 2005. Explaining Explicitation. In: Károly and Fóris (eds), 2005, 29–43.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman.
Renouf, Antoinette and Andrew Kehoe (eds). 2006. *The Changing Face of Corpus Linguistics.* Amsterdam and New York: Rodopi.
Renouf, Antoinette and Andrew Kehoe (eds). 2009. *Corpus Linguistics. Refinements and Reassessments.* Amsterdam and New York: Rodopi.
Robinson, Douglas. 1991. *The Translator's Turn.* Baltimore: The Johns Hopkins University Press.
Robinson, Douglas. 1998. Free translation. In Baker (ed), 1998, 87–90.
Sager, Juan C. 1994. *Language Engineering and Translation. Consequences of automation. Benjamins Translation Library* 1. Amsterdam and Philadelphia: John Benjamins Publishing Company.
Sager, Juan C., David Dungworth, and Peter F. McDonald. 1980. *English Special Languages. Principles and practice in science and technology.* Wiesbaden: Oscar Brandstetter Verlag KG.
Sampson, Geoffrey and Diana McCarthy (eds). 2004. *Corpus Linguistics: Readings in a Widening Discipline.* London: Continuum.
Šarčević, Susan. 1997. *New Approach to Legal Translation.* The Hague, London, and Boston: Kluwer Law International.
Šarčević, Susan. 2007. Making multilingualism work in the enlarged European Union. In: Kredens and Goźdź-Roszkowski (eds), 2007, 34–56.
Schane, Sanford. 2006. *Language and the Law.* London and New York: Continuum.
Searle, John R. 1975. A Taxonomy of Illocutionary Acts. In: Gunderson (ed), 1975, 344–369.
Seiler, Hansjakob. 2001. The operational basis of possession. A dimensional approach revisited. In: Baron et al. (eds), 2001, 27–40.
Shannon, Claude E. 1948. A Mathematical Theory of Communication. Reprinted with corrections from *The Bell System Technical Journal* 27, 379–423, 623–656.
Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication.* Urbana, Ill.: University of Illinois Press.
Shieber, Stuart M. 1993. The Problem of Logical-Form Equivalence. *Computational Linguistics* 19:1, 179–190.

Silva, Norma Andrade da. 2008. *Análise da tradução do item lexical evidence para o português com base em um corpus jurídico.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Smith, May-Britt Marthinsen. 2004. *Initial -ing clauses in English and their translation into Norwegian.* Master's thesis. University of Oslo.

Suominen, Mickael, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen, and Kaius Sinnemäki (eds). 2006. *A Man of Measure. Festschrift in Honour of Fred Karlsson on his 60th Birthday.* Special supplement to *SKY Journal of Linguistics* 19. The Linguistic Association of Finland.

Swales, John M. 1990. *Genre Analysis. English in academic and research settings.* Cambridge, New York, and Melbourne: Cambridge University Press.

Tiersma, Peter M. 1999. *Legal Language.* Chicago and London: The University of Chicago Press.

Tirkkonen-Condit, Sonja (ed). 1991. *Empirical Research in Translation and Intercultural Studies. Language in Performance* 5. Tübingen: Gunter Narr Verlag.

Tirkkonen-Condit, Sonja and Riitta Jääskeläinen (eds). 2000. *Tapping and Mapping the Processes of Translation and Interpreting. Benjamins Translation Library* 37. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Thunes, Martha. 1994. *Transfer and Interlingua in Machine Translation. A Comparison of Two Implementations. Skriftserie fra Institutt for fonetikk og lingvistikk* 44, B. University of Bergen.

Thunes, Martha. 1998. Classifying translational correspondences. In: Johansson and Oksefjell (eds), 1998, 25–50.

Toolan, Michael. 2001. *Narrative. A critical linguistic introduction. The INTERFACE Series.* Second edition. London and New York: Routledge.

Tosi, Arturo (ed). 2003. *Crossing Barriers and Bridging Cultures. The Challenges of Multilingual Translation for the European Union.* Clevedon, Buffalo, Toronto, and Sydney: Multilingual Matters Ltd.

Toury, Gideon. 1995. *Descriptive Translation Studies and beyond. Benjamins Translation Library* 4. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Trosborg, Anna (ed). 1997a. *Text Typology and Translation. Benjamins Translation Library* 26. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Trosborg, Anna. 1997b. Text Typology: Register, Genre and Text Type. In: Trosborg (ed), 1997, 3–23.

Tucunduva, Camila de Andrade. 2007. *Translating completeness: a corpus-based approach.* Master's thesis. Federal University of Santa Catarina, Florianópolis.

Vander Linden, Keith. 2000. Natural Language Generation. In: Jurafsky and Martin 2000, 763–798.

Venuti, Lawrence (ed). 2000. *The Translation Studies Reader.* London and New York: Routledge.

Vinay, Jean-Paul and Jean Darbelnet. 1995. *Comparative Stylistics of French and English: A Methodology for Translation.* Translated and edited by Juan C. Sager and M.-J. Hamel. *Benjamins Translation Library* 11. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Vinje, Finn-Erik (ed). 1990a. *Språket i lover og annet regelverk. CompLex* 2/90. Oslo: Tano A.S.

Vinje, Finn-Erik. 1990b. Moderne norsk lovspråk og annen juristprosa. In: Vinje (ed), 1990a, 9–76.

Vinje, Finn-Erik. 1995. *Lovlig språk. Om språk og stil i lover og annet regelverk.* Second edition. Oslo: Justisdepartementet.

Wahlster, Wolfgang (ed). 2000. *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin, Heidelberg, and New York: Springer Verlag.
Weaver, Warren. 1949. Translation. In: Locke and Booth (eds), 1955, 15–23.
Wendel, W. Bradley. 2005. Professionalism as Interpretation. *Northwestern University Law Review* 99:3, 1167–1233.
Wienold, Götz. 1980. Das Konzept der Textverarbeitung und die Semiotik der Literatur. In: Eschbach and Rader (eds), 1980, 201–209.
Wilss, Wolfram. 1977. *Übersetzungswissenschaft. Probleme und Methoden.* Stuttgart: Klett.
Wilss, Wolfram. 1978. Methodische Aspekte des Übersetzungsprozesses. In: Gomard and Poulsen (eds), 1978, 15–26.
Witczak-Plisiecka, Iwona. 2007. Linguistic aspects of deontic *shall* in the legal context. In: Kredens and Goźdź-Roszkowski (eds), 2007, 181–199.