Lang Resources & Evaluation (2016) 50:291–319
DOI 10.1007/s10579-016-9356-5

ORIGINAL PAPER

The enrichment of lexical resources through incremental parsebanking

Victoria Rosén[1] · Martha Thunes[1] · Petter Haugereid[1] · Gyri Smørdal Losnegaard[1] · Helge Dyvik[1] · Paul Meurer[2] · Gunn Inger Lyse[1] · Koenraad De Smedt[1]

Published online: 30 May 2016

Abstract Automatic syntactic analysis of a corpus requires detailed lexical and morphological information that cannot always be harvested from traditional dictionaries. Therefore the development of a treebank presents an opportunity to simultaneously enrich the lexicon. In building NorGramBank, we use an incremental parsebanking approach, in which a corpus is parsed and disambiguated, and after improvements to the grammar and the lexicon, reparsed. In this context we have implemented a text preprocessing interface where annotators can enter unknown words or missing lexical information either before parsing or during disambiguation. The information added to the lexicon in this way may be of great interest both to lexicographers and to other language technology efforts.

Keywords Lexical resources · INESS · NorGramBank · Treebanking · LFG · Language research infrastructure · Automatic syntactic analysis

1 Introduction

Parsebanking is the creation of a treebank through automatic parsing of a corpus with a grammar and lexicon. Since this process results in a large number of analyses which can readily be inspected, it provides an excellent testing ground for the development of a lexicon as well as a grammar. As parsing requires fine-grained distinctions which are often overlooked in traditional lexicography, parsebanking

✉ Victoria Rosén
victoria@uib.no

[1] LLE, University of Bergen, PO box 7800, 5020 Bergen, Norway

[2] Uni Research Computing, Uni Research, PO box 7810, 5020 Bergen, Norway

presents a good and until now insufficiently recognized context for enrichment and testing of the lexicon.

The INESS project (Infrastructure for the Exploration of Syntax and Semantics) is developing NorGramBank, a large parsebank for Norwegian.[1] In the process, a grammar and lexicon for Norwegian are being further developed in tandem. Since the parser requires quite detailed morphosyntactic information in order to provide an analysis, the lexicon must be syntactically well informed. In our experience, which will be discussed in some detail in this paper, feedback from the parsebanking process is valuable for testing and improving lexical information.

An example of a lexical property missing in ordinary dictionaries is the *inquit* reading of verbs. Inquit verbs are verbs of saying and related verbs which take a sentential complement as an argument, occur after a quotation, and involve inversion, i.e. with the subject following the verb, as illustrated in example (1). Information about the set of verbs that occur in this construction is necessary for parsing.

(1) Hvordan staver du kjærlighet? spurte Nasse Nøff.
how spell you love asked Piglet
'"How do you spell love?" asked Piglet.'

It is our hypothesis that traditional dictionaries are insufficient sources of lexical information for parsing and that adding unknown words and more precise and complete information about known words will significantly improve parsing. We hope to show that parsebanking is a productive context for discovering and describing words and their morphosyntactic properties.

In the following, we will first explain how the syntax and lexicon mutually inform each other in our parsebanking approach. In Sect. 3, the interface for preprocessing texts will be presented. Section 4 describes how words that are not recognized by the morphological analyzer are treated, while Sect. 5 details the procedure for adding information for known words. In Sect. 6 issues concerned with multiword expressions are presented.

2 Grammar development and incremental parsebanking

Most current manually checked treebanks are produced in part by parsing a corpus. However, not all sentences may automatically get a correct analysis, due to missing coverage in the grammar and lexicon. Many treebanking efforts remedy this problem by means of manual editing of the parses. This may result in analyses which are not compatible with the grammar which was used for parsing. Furthermore, editing the parses directly will not lead to enrichment or correction of the lexicon. In contrast, our approach is based on incremental improvement of the grammar and lexicon during the parsebanking process (Losnegaard et al. 2012;

[1] http://clarino.uib.no/iness

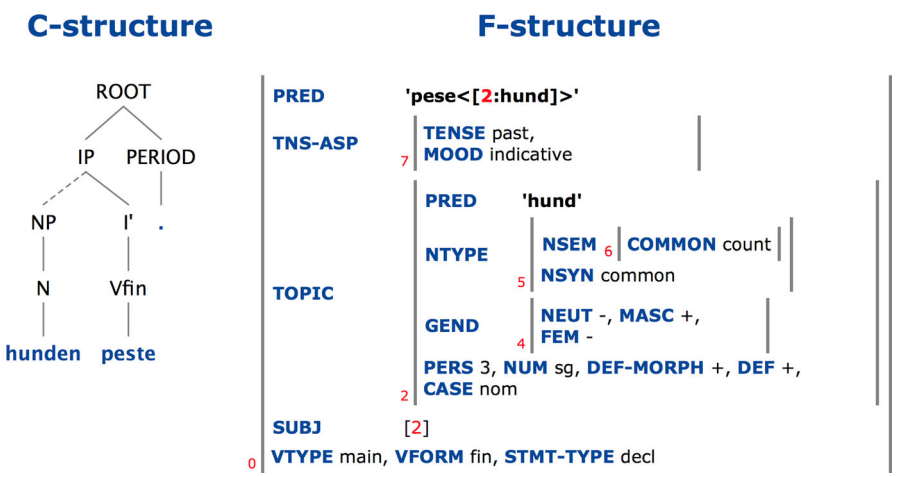


Fig. 1 Analysis of a sentence with an intransitive verb

Rosén and De Smedt 2007). This approach results not only in a manually checked parsebank, but also in a grammar which is fully compatible with the analyses in the parsebank, and moreover, in substantial lexicon improvements, as will be described below.

The grammar used for creating NorGramBank is NorGram, a hand-written broad coverage computational grammar which has been used in several language technology projects (Dyvik 2000; Butt et al. 2002). It is written in the Lexical Functional Grammar (LFG) framework (Bresnan 2001; Dalrymple 2001), which allows for deep analyses of considerable grammatical detail. We use the Xerox Linguistics Environment (XLE) for grammar development and parsing (Maxwell and Kaplan 1993). The analyses produced by XLE with NorGram are disambiguated and stored in the parsebank. Regular reparsing after improvements to the grammar and lexicon provides improvements in coverage. Thus we aim to incrementally produce high quality gold standard treebanks, which in turn are used for training a stochastic disambiguator in order to produce larger fully automatically parsed and disambiguated treebanks. This methodology is similar to and inspired by the LinGO Redwoods treebanking approach (Oepen et al. 2004).

NorGram provides deep syntactic analysis on two levels: constituent structure (c-structure) and functional structure (f-structure). The c-structure is a phrase structure tree showing the linear and hierarchical organization of the phrasal constituents in the sentence. The f-structure is an attribute-value matrix showing grammatical functions and features. This is illustrated for the examples in (2) and (3), showing an intransitive sentence and a transitive sentence, respectively.[2] The c- and f-structure analyses of these examples are given in Figs. 1 and 2.

[2] Since the morphological structure of the words in the examples is not relevant in this article, we have not indicated morphological features in the glosses, but simply used two English words when necessary to render a Norwegian word.

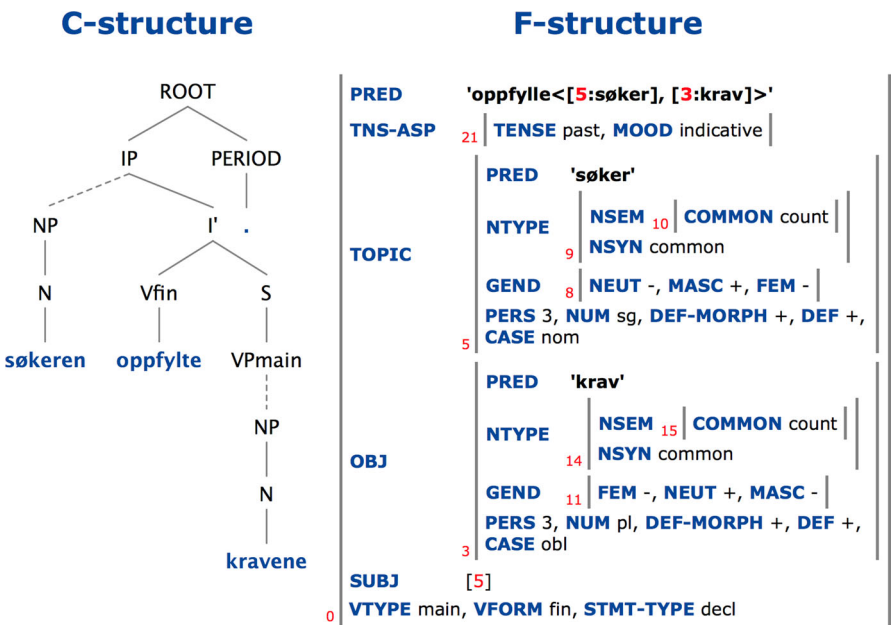


Fig. 2 Analysis of a sentence with a transitive verb

(2) Hunden peste.
the dog panted
'The dog panted.'

(3) Søkeren oppfylte kravene.
the applicant fulfilled the requirements
'The applicant fulfilled the requirements.'

In LFG, the syntax and the lexicon have an important interaction with each other, especially in the treatment of predicate-argument structure. The lexical entry for each verb must specify which arguments a verb requires. For example, in a transitive sentence, the lexical entry for the verb must specify that the verb can take an object. This specification interacts with the syntax in such a way that no grammatical analysis will be assigned to sentences lacking syntactic arguments which the verb specifies, or containing syntactic arguments which the verb does not specify.

The f-structure in Fig. 1 has only a subject but no object. This is in accordance with what the NorGram lexical entry for the verb *pese* 'pant' in (4) requires. In this lexical entry, written in the XLE format, `V-SUBJ` is a mnemonic reference to a template for intransitive verbs. In contrast, the f-structure in Fig. 2 has a subject and

an object, in accordance with the lexical entry in (5). Here `V-SUBJ-OBJ` is a reference to a template for simple transitive verbs.

(4) pese V XLE @(V-SUBJ pese pese)

(5) oppfylle V XLE @(V-SUBJ-OBJ oppfylle oppfylle)

As a result of the ubiquitous ambiguity of natural languages, parsing with a high-coverage formal grammar and lexicon will often return a very high number of alternative analyses for a sentence, whereas normally only one of those analyses will be appropriate in the given context. Some degree of manual disambiguation is unavoidable for the purpose of building a gold standard parsebank, which subsequently may be used for training a stochastic disambiguator. Whereas annotators in our approach never manually edit an analysis, they must verify if the parser has produced a correct analysis, and choose the correct analysis if several possible analyses are produced.

The disambiguation process has been optimized through the use of *discriminants* (Carter 1997; Oepen et al. 2004). The parsebanking system automatically analyzes the forest of alternatives, reducing it to a set of binary discriminants which allow the annotators to efficiently distinguish and select among a high number of alternatives (Rosén et al. 2007, 2009, 2012).[3] While disambiguating, the annotators may discover that the correct analysis is not among the alternatives produced by the parser. In that case they first attempt to diagnose the problem, and often they may solve it by updating the lexicon and reparsing. If the problem persists, a change in the grammar may be necessary, which is reported through the issue tracking system that is integrated into the disambiguation interface.

This potentially continuous approach is scalable: new text can be automatically parsed and disambiguated stochastically by training on the manually disambiguated material. The information which is stored as a result of manual disambiguation is not just the selected analysis, but also the discriminant values chosen by the annotators, along with the rest of the analyses. Hence, when the entire treebank has been reparsed with the updated grammar (which happens with certain intervals), the stored discriminant values can be reapplied to the new set of alternative analyses, which is frequently sufficient to pick out a unique solution again. As mentioned above, this methodology is inspired by LinGO Redwoods (Oepen et al. 2004). What is novel in our approach is that we have designed and implemented discriminants for LFG grammars, and that the entire process is supported through a web-based annotation interface.

The advantage of this parsebanking approach is that the resulting parsebank will always be fully compatible with the grammar. Parsebanks constructed in this way therefore achieve a very high level of consistency. It is also the case, however, that only sentences that are grammatical according to the current grammar will be fully analyzed, while others may receive a fragment parse or may fail to parse.

Earlier we carried out a detailed study of a small subcorpus in order to find out what the main causes of failed analyses are (Losnegaard et al. 2012). This study

[3] For a discussion of interannotator agreement in the disambiguation process, see Dyvik et al. (2013).

found that 21 % of the sentences had a full analysis that was not the correct one. Moreover, the study identified the interventions that were necessary in order to achieve the intended analysis for these sentences. Since the sentences studied initially had some analysis, they all involved words that were recognized, but sometimes not with the correct morphological analysis. We found that 29 % of the failed analyses were caused by syntactic problems, while 71 % were caused by lexical problems. Of the lexical problems, 41 % were caused by missing multiword expressions (MWEs), whereas 41 % were caused by incorrect lexical categories.[4] These numbers indicate that correct lexical information is essential for successful syntactic analysis.

A parsebanking approach of this kind requires a large lexicon with detailed morphosyntactic information. The main basis for the NorGram lexicon has been the NorKompLeks electronic lexicon (Nordgård 2000). This lexicon is an adapted version of two traditional dictionaries of Norwegian: *Bokmålsordboka* (Landrø and Wangensteen 1993) and *Nynorskordboka* (Hovdenak et al. 1986). These dictionaries were developed by the University of Oslo and the Norwegian Language Council (Språkrådet), and in practice they define the official norms for spelling and inflection. *Bokmålsordboka* has approximately 70,000 lemmas, while *Nynorskordboka* has approximately 90,000. The dictionaries contain both etymologies and examples. The web versions are standard works of reference for most Norwegian users, with more than 70 million searches per year between them. The NorKompLeks lexicons added subcategorization frames for the verbs. The NorKompLeks format was converted by means of a program into the format required by XLE.[5] Morphological analysis is handled by finite-state transducers derived from the resource Norsk Ordbank (Norwegian Word Bank), a database which contains inflectional and other information about all entries in *Bokmålsordboka* and *Nynorskordboka*, in addition to further material. However, as we will see below, the lexical information in NorKompLeks and Norsk Ordbank is not always complete and accurate, and needs to be supplemented.

3 Text preprocessing

An important source of texts for NorGramBank is a large repository of OCR-read fiction texts supplied by the National Library of Norway. Because OCR software makes certain errors, such as misinterpreting characters, omitting text, or inserting unwanted material, the documents must be preprocessed before syntactic parsing. Moreover, when a corpus is parsed, there will always be words that are unknown to the morphological analyzer and/or the lexicon. INESS has therefore developed an intelligent browser-based preprocessing interface which facilitates efficient text cleanup and the treatment of unknown word forms (Rosén et al. 2012).

[4] The original paper (Losnegaard et al. 2012) erroneously suggests that 31 % were caused by incorrect lexical categories.

[5] See http://www2.parc.com/isl/groups/nltt/xle/doc/walkthrough.html for an explanation of the XLE lexicon format.

The first step is text cleanup, which involves for example removing superfluous material that does not belong to the text, joining parts of sentences that have erroneously been split, and adding punctuation where it is missing. The interface offers practical editing functions for these cleanup operations.

After text cleanup, the annotators process word forms that have not been automatically recognized. The preprocessing interface presents a list of unknown words. Some of these are errors which must be corrected in the text itself before parsing, such as OCR errors, incidental misspellings, and typos. Other unknown words should be covered in the lexicon. Examples are names, foreign words, neologisms, productive compounds not recognized by the compound analyzer, and words only occurring in MWEs.

Nonstandard words of various types are also added to the lexicon. We distinguish between three main classes: archaic words, systematic misspellings, and forms belonging to nonstandard language varieties. An example of the first class, archaic words, is the plural noun form *fjelle*, in contrast to the current standard spelling *fjell* 'mountains'. The second class, systematic misspellings, includes forms which are produced regularly by one or more authors. An example is the form *tennveske*, which is a common misspelling of *tennvæske* 'charcoal lighter fluid'. Finally, the third class of nonstandard words covers forms that can be ascribed to a particular dialect, technolect, sociolect, or other language variety. An example is *barnehagan*, instead of the standard form *barnehagen* 'the preschool'. The suffix *-an* in the nonstandard variant is used to imitate a dialect pronunciation. Instances of these three nonstandard classes are left unchanged in the text because normalizing them would be to interfere with actual language use.

The important common denominator of all types of unknown words which are not to be corrected is that while these forms fall outside standard dictionaries, it is a prerequisite for successful parsing that they nevertheless be included in our lexicon. Nonstandard words are explicitly marked as such in the lexicon, so that any reuse of the lexicon, for example for generation, would not result in these words being output inadvertently.

4 Adding unknown words during preprocessing

Table 1 presents an overview of the types of unknown words that were added through preprocessing of a subcorpus of NorGramBank of about 42 million words. Among these words, members of the open lexical classes (nouns, verbs, adjectives, and adverbs) account for 39 % of all entries. These are given as the category *Open word class* in Table 1.

The preprocessing interface allows the annotators to add information about unknown words to the lexicon. Noninflecting words such as names and interjections are entered by assigning the appropriate lexical category to each entry. For words belonging to the open lexical classes the annotator specifies an inflectional pattern. Verbs must also be assigned subcategorization frames necessary for parsing. When a word is not recognized because of nonstandard spelling, the annotator must consider whether the spelling deviation concerns the stem or an inflection. Variant

Table 1 Overview of the various types of unknown words added through preprocessing

Category	Instances
Open word class (N, V, A, ADV)	13,095
Last name	6557
Organization or brand name	6502
Place name	4646
Title	2754
Miscellaneous name	2683
Foreign expression	2380
Unclassified	2219
Variant inflectional form	2086
Person name	1776
Interjection	1548
First name, masculine	1180
First name, feminine	861
Taxon name	92

stems are registered with existing standard inflectional paradigms, and variant inflectional forms are registered as deviations from individual, standard inflectional forms. In order to add unknown words to the lexicon in an efficient way, the annotator makes use of a set of predefined options in the preprocessing interface. Each option corresponds to a certain type of entry. Most of these types can be entered by a single mouse click, while the recording of inflecting words and variant inflectional forms requires a few more steps.

4.1 Open word classes

In Norwegian, words belonging to the open word classes usually have inflection. When a new inflecting word is added to the lexicon, the annotator must specify its set of inflectional forms on the basis of an existing lexical entry with matching inflection. As the new lemma is stored, it thus inherits the lexical category of the existing lemma.

As an example, consider the word form *narrativen* 'the narrative'. This word form can only be a singular definite inflection of a noun, and the context where it occurs, shown in Fig. 3, makes it clear that this is how it is used. The lemma *narrativ*, however, was found only as an adjective; the annotator therefore adds a new noun entry *narrativ* to the lexicon. The procedure is carried out through a dialogue box in the preprocessing interface, as shown in Fig. 3. First, the dictionary entry form of the new lemma, *narrativ*, is entered in the "Base form" field. Next, an inflectional paradigm for the new lemma must be specified, either by selecting one from a drop-down menu of potentially matching lemmas proposed by the system, or by entering the base form of an existing lemma with matching inflection in the "Inflects like" field. In this case *komparativ* is entered, and the interface then presents a pop-up menu with the new word inflected in all patterns that the entered

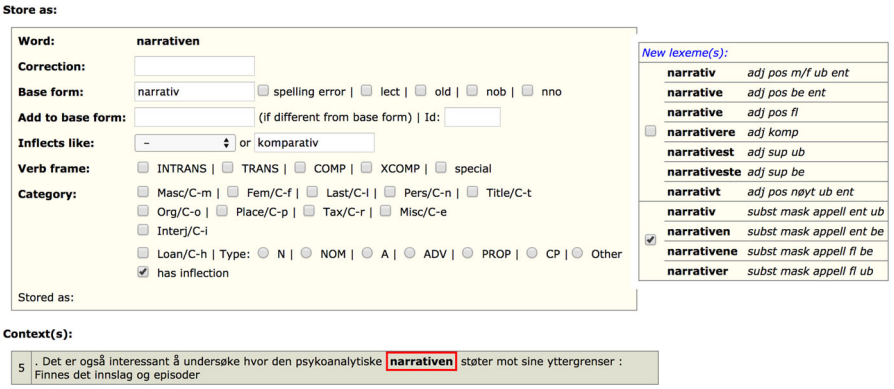


Fig. 3 Adding a noun

base form allows. Since *komparativ* can be both an adjective and a noun, two inflectional patterns are proposed. As shown in Fig. 3, the annotator ticks off the noun, having checked that the suggested set of inflectional forms is correct, and stores the new noun entry with a keyboard shortcut.

Another example of an unknown word form is *synonymiserer* 'synonymizes', illustrated in Fig. 4. This is a productive verbal derivation from the noun *synonym*, inflected in the present tense. In order to add a new verb *synonymisere*, the annotator follows the same procedure as the one described for adding the noun *narrativ*. In this case, a verb with matching inflection, *polemisere*, has been selected from the drop-down menu in the "Inflects like" field, and this creates the proposed set of inflectional forms shown to the right in Fig. 4. Since *synonymiserer* is used intransitively in the given context, the annotator ticks off "INTRANS" in the "Verb frame" field before storing the new verb.

It can often be justified to add misspellings to the lexicon, as mentioned in Sect. 3. An author can for instance use a creative spelling to imitate a dialect pronunciation. An example is *mordern* instead of the standard form *morderen* 'the murderer'. The elided vowel is imitative of an eastern Norwegian accent, and *mordern* was not recognized because it is a nonstandard word. As shown in Fig. 5, *mordern* may be included in the lexicon as a variant of the standard form *morderen*. To achieve this, the annotator first enters the standard lemma form, *morder*, in the "Base form" field. The option "lect" is ticked off to mark the word as a dialect form. By pressing a specific key combination the annotator then opens a new window ("Is a variant of") which lists all standard inflectional forms of the noun *morder*. From this list the annotator picks the word form *morderen*, which has morphological features matching the grammatical properties of the deviating word form. Subsequently, *mordern* is stored as a dialect variant of the inflectional form *morderen*.

Spelling deviations may also occur in the stem, as in the example *tennveske*, mentioned in Sect. 3. Because *tennveske* is a common misspelling of *tennvæske* 'charcoal lighter fluid', it is added to the lexicon as a variant lemma. The procedure

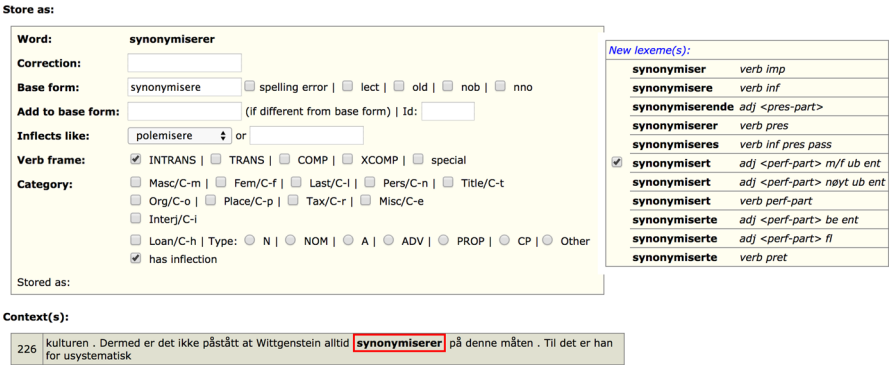


Fig. 4 Adding a verb

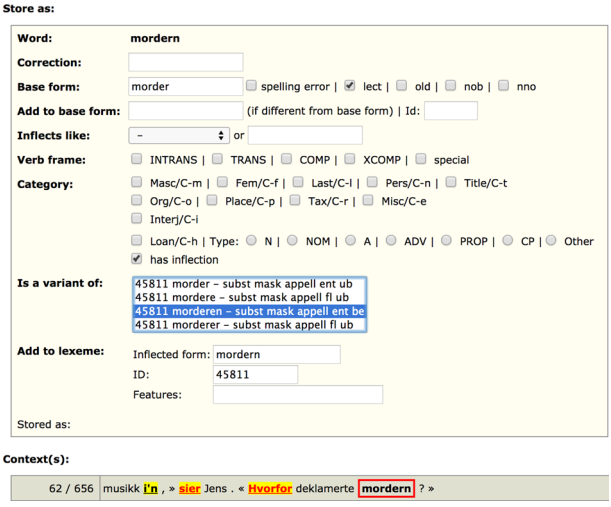


Fig. 5 Adding a variant inflectional form

for doing this is illustrated in Figs. 6, 7 and 8. First, the annotator enters the base form of the new lemma, *tennveske*, in the "Base form" field, and ticks off the option "spelling error", as shown in Fig. 6. Next, the dictionary entry form of the standard lemma, *tennvæske*, is entered in the "Add to base form" field. This will open a new window on the right-hand side, displaying the inflectional forms associated with the standard lemma (also shown in Fig. 6). If the standard base form is categorially ambiguous, the window will list the set of inflectional forms for each category. In this case there is only one set of forms. The annotator must tick off the appropriate standard lemma, and its ID number will then appear in the "Id" field, as shown in Fig. 7. The next step is to specify the inflectional pattern of the new lemma, and this is done according to the normal procedure for inflecting words, as already described for *narrativ* and *synonymisere*. Thus, as shown in Fig. 8, the annotator enters the base form of the standard lemma, *tennvæske*, in the "Inflects like" field, and the

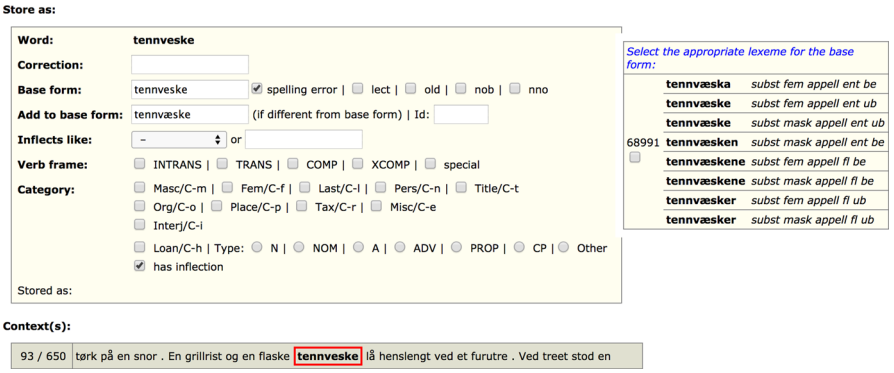


Fig. 6 Adding a variant stem, step 1

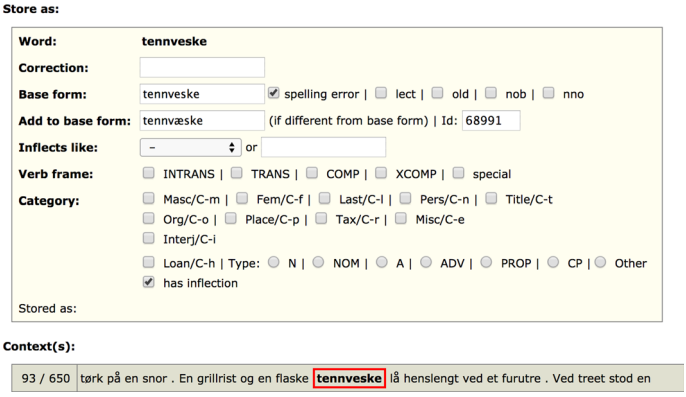


Fig. 7 Adding a variant stem, step 2

interface proposes a set of inflectional forms for the new lemma. Finally, *tennveske* is stored as a new noun entry, and in this way the nonstandard word *tennveske* is included in the lexicon as a variant lemma, associated with the standard entry for *tennvæske*.

4.2 New compounds

Norwegian is a language with extensive productive compounding. Since compounds are written as single graphical words and compounding may be done on the fly, many legitimate compounds cannot be listed in the lexicon. Therefore an automatic compound analyzer is run on the text prior to preprocessing in order to identify compounds that are not already in the lexicon. Although the analyzer recognizes many compounds, the analysis of potential compounds is nevertheless restricted in order to prevent overgeneration.

Allowing compound constituents of less than three letters is generally considered a risk in automatic compound analysis; if such short constituents are allowed in

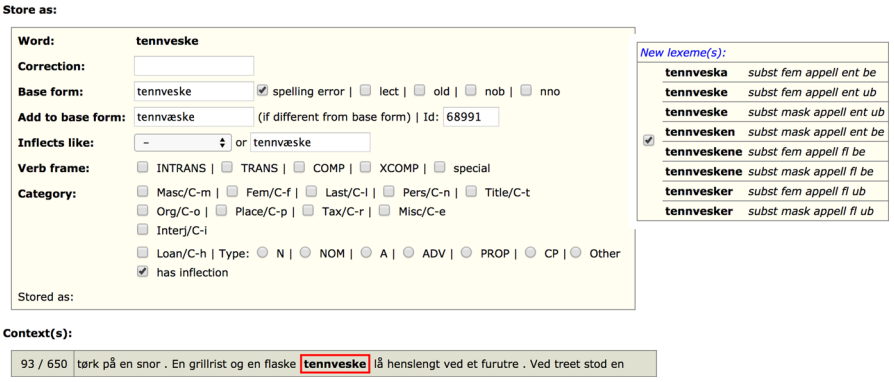


Fig. 8 Adding a variant stem, step 3

general, many typos and misspelled words may be erroneously analyzed as compounds. We implement this restriction and allow short elements only if they are listed specially due to their observed occurrence in compounds.

Furthermore, some of the combinations that the compound analyzer allows have certain constraints imposed on them. For noun + adjective compounds, only a few nouns that occur frequently as the first element in compounds are allowed; examples are *kjempe* 'giant', *drit* 'shit', and *rekord* 'record'. This explains why the compound *avisgrå* 'newspaper gray' was not recognized. This example, and numerous others, such as *guttegærn* 'boy crazy', *silkehvit* 'silk white' and *helseriktig* 'health correct' ('healthy'), show that this constraint is too strong. For adjective + verb compounds, the verb is restricted to only being a past participle, which is the reason why *blekpudre* 'pale powder' ('powder something to make it pale') was not recognized. Again, however, this restriction seems too strong, since there are many examples of other forms of verbs than the past participle in this type of compound: *ansvarliggjøre* 'responsible make' ('make responsible'), *finpiske* 'fine whip' ('whip until fine'), *hardkode* 'hard code', etc.

Another reason why compounds are not recognized is that special forms which are only used in compounds are missing from the lexicon. An example is *engleflokk* 'angel flock' ('flock of angels'), where *engle* is a variant of *engel* 'angel'. Other examples are *billedramme* 'picture frame', where *billed* is an archaic form of *bilde* 'picture', and *faktafeil* 'facts error' ('factual error'), where *fakta* is the plural of *faktum* 'fact'. Although compounds with such special forms occur in the dictionaries that were used as sources, the specific first elements themselves were missing.

Finally, some compounds are not recognized because one or both of their constituents are misspelled. Examples are *hårshampo*, a misspelling of *hårsjampo* 'hair shampoo', and *cafébord*, a misspelling of *kafébord* 'café table'.

Compounds which are not recognized by the compound analyzer are presented to the annotator in the same way as other unknown words. The annotator can then add them to the lexicon as simplex words, while at the same time marking their internal structure by inserting the character + between the elements. This internal structure is not added to the lexicon used for parsing, but is recorded in a separate list in order

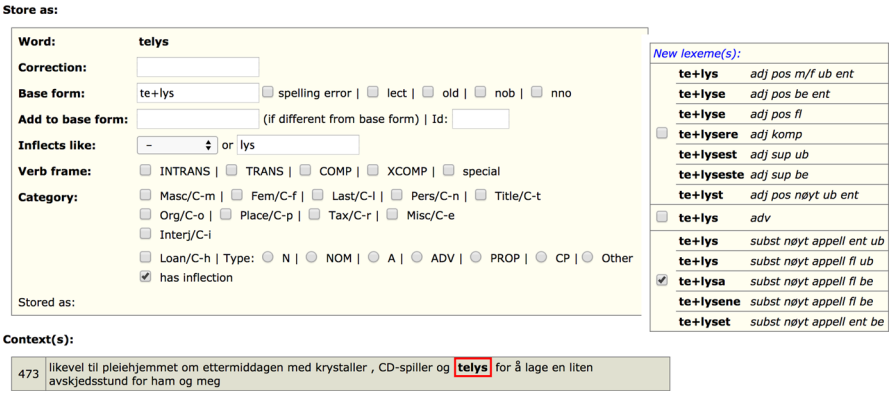


Fig. 9 Adding a compound

to enable the discovery of frequent compound elements and compound types that are not already accounted for by the compound analyzer. The screenshot in Fig. 9 illustrates how the unknown compound *telys* 'tea light' is added to the lexicon by the annotator. For the second element, which determines the inflection of the compound, the relevant paradigm is indicated in the same way as described for other words in Sect. 4.1. For the first element, an abbreviation for the lexical category is written in parentheses before the +, except for nouns, in which case no category is indicated (as in this example). Since the string *lys* is categorially ambiguous, the annotator must tick off which pattern or patterns are to be registered. In Fig. 9, the annotator has selected the noun, and not the adjective or adverb.

Table 2 gives an overview of the most common compound types that have been registered by annotators in this way. The column headed CA (for compound analyzer) shows which types the compound analyzer currently allows: noun + noun, noun + adjective, adjective + noun, adjective + verb and verb + noun. The reason why only these five were allowed initially was that they were assumed to be the most frequent types; allowing too many possible combinations could lead to many incorrect analyses of unknown words. The overview of types that were actually found shows that there are several additional frequent types that should be considered for incorporation into the compound analyzer. A detailed study of the individual examples in the different categories will help to determine which new types should be added to the compound analyzer, as well as which frequent short elements should be allowed.

4.3 Other types of unknown words

A particularly frequent type of unknown words is names. These are typically missing from dictionaries. From Table 1 it appears that unknown last names, organization or brand names, and place names are very common. Since names are normally invariable, they can simply be assigned a part of speech.

Table 2 Overview of the most common compound types recognized during preprocessing

Type	CA	Example	Instances
Noun + noun	✔	*te* + *lys* 'tea light'	2434
Noun + adjective	✔	*avis* + *grå* 'newspaper gray'	1096
Adjective + adjective		*blå* + *brun* 'blue brown'	730
Adjective + noun	✔	*fin* + *kåpe* 'nice coat'	263
Preposition + noun		*av* + *knapp* 'off button'	218
Preposition + adjective		*gjennom* + *korrupt* 'through corrupt'	190
Preposition + verb		*av* + *beite* 'off graze'	182
Noun + verb		*dybde* + *bore* 'depth drill'	153
Verb + noun	✔	*ete* + *fest* 'eat party'	151
Adjective + verb	✔	*blek* + *pudre* 'pale powder'	118
Verb + adjective		*drikke* + *klar* 'drink ready'	59

Foreign words are often used in Norwegian sentences. Sometimes they are spontaneous uses of a word from another language, most often English. Other times they are well-established loan words in Norwegian, but have not yet made their way into standard dictionaries. An example of the spontaneously used English word *alien* is shown in (6).

(6) «Jeg skulle ikke være noen alien for deg,» sa Auguste.
I should not be some alien for you said Auguste
'"I'm not really an alien for you," said Auguste.'

Example (7) contains both the English loan *air conditioning* and the name *American Bar*.

(7) Han gikk inn på American Bar, som reklamerte med air conditioning.
he went in on American Bar which advertised with air conditioning
'He went into American Bar, which boasted air conditioning.'

Missing lexical entries like these are easily added to the lexicon when they are identified in the preprocessing step. In this case, *American Bar* was entered as an organization name, and *alien* and *air conditioning* were entered as loans.

A particularly productive part of speech is interjections; especially writers of fiction are very creative in the way in which they write interjections. *Bokmålsordboka* has an entry for the interjection *hysj* 'hush' which also includes the alternative

spelling *hyss*. There are several occurrences of this interjection in the fiction texts of NorGramBank, and many of them do not have either of the two standard spellings. The following eight variants of *hysj/hyss* have been registered so far: *hysjjj*, *hyssj*, *hysssj*, *hysssjjj*, *hysssssj*, *hysssssjjj*, *hysst*, *hyyyyysssjjj*. These examples show that the spelling of this interjection is unpredictable and to a large extent determined by the way in which an author chooses to express it in a given context. For parsebanking purposes, the challenge is that each time a new spelling is encountered, it is displayed in the preprocessing interface as an unknown word. The INESS interface makes it possible for annotators to add new variant spellings to an existing interjection in the lexicon.

In conclusion, as the annotator processes the unknown words, these words and the necessary information about them are added to the lexical resources exploited by the parser.

5 Known words with missing or incorrect information

For the parser, it is not sufficient that words are known. It is also essential that the information about them is complete and accurate. Even though the NorKompLeks lexicon is a rich resource, in parsing we still often find that it lacks necessary lexical information that we need in order to analyze even quite common words. We need information about *inter alia* lexical category, inflection, subcategorization, countability and compound structure. We also need lexical entries for MWEs, which are seriously underrepresented in the resources we have used as a basis for our work.

Even though the NorKompLeks lexicon has added subcategorization frames for the verbs in *Bokmålsordboka* and *Nynorskordboka*, many quite common frames are not included. Table 3 gives an overview of the types of lexicon updates made by annotators during disambiguation. As shown in this table, the most frequent type of lexicon update concerns subcategorization frames for verbs. These instances cover a large number of different types of verb frames, which are sorted into six categories: MWE frames, intransitive readings, intransitives with expletive subject, transitive readings, verb complement readings, and inquit readings. New verb frames involving MWEs account for over half of these cases. In Table 3 the six groups of verb frame types, as well as the other types of lexicon updates, are listed in descending order of frequency.

Consider the case of particle verbs, a frequent type of MWE. The sentence in (8) illustrates a use of the reflexive particle verb *flate seg ut* ‘flatten out’. It is also possible to use this verb without the reflexive, as illustrated in (9).

(8) Stien flater seg ut.
the path flattens itself out
‘The path flattens out.’

Table 3 Overview of lexicon updates made by annotators for known words

Type of lexicon update	Instances
Verbs	
MWE frames	379
Intransitive readings	83
Transitive readings (incl. ditransitive)	68
Inquit readings	53
Verb complement readings	40
Intransitives with expletive subject	26
Nouns	
Mass readings	211
MWE frames	75
Added count nouns	31
Added title readings	12
Adverbs and prepositions	
Added adverb readings	93
Added prepositions	8
Adjectives	
MWE frames	42
Added adjectives	6

(9) Fjellet flater ut.
the mountain flattens out
'The mountain flattens out.'

The lexicon already had a subcategorization frame for the usage in (8), but not for that in (9), resulting in an incorrect analysis for the latter. The Norwegian word form *flater* is categorially ambiguous: it can be either a verb or a noun. Therefore, the only analysis found by the parser for (9) was that of a noun phrase, where the word form *flater* was analyzed as the plural indefinite of the noun *flate* 'surface' functioning as an apposition to the noun *fjellet* 'the mountain', and with *ut* 'out' analyzed as an adverbial adjunct to the noun. Figure 10 shows the c- and f-structures for this unintended analysis of (9).[6] After the missing subcategorization frame had been added to the lexicon, the sentence was reparsed. As the c-structure in Fig. 11 shows, *flater* is now analyzed as a present tense verb (with the lexical category Vfin), and *ut* as a particle (PRT).

Adding this argument frame involves making an addition to a lexical entry. Example (10) shows the lexical entry of *flate* with this addition. The XLE notation $\{\ldots|\ldots\}$ is a disjunction specifying alternatives. The first line of the disjunction

[6] Here and in the following examples of f-structures, we use the simplified XLE format where features other than predicates and functions are not displayed.

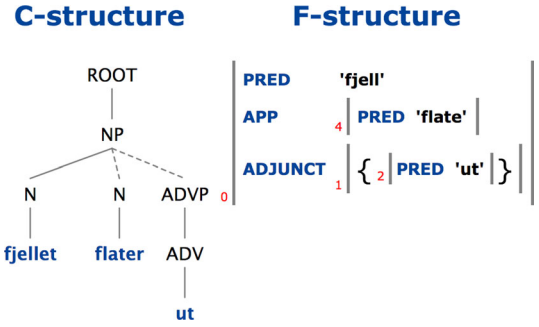


Fig. 10 Analysis of (9) before lexical update

provides the previously available reflexive frame, whereas the second line gives the new frame without the reflexive.

(10) flate V XLE { @(V-SUBJ-OBJrefl-PRT flate flate ut)
| @(V-SUBJ-PRT flate flate ut) }

(11) V-SUBJ-PRT (P S PRT) =
@ (CONCAT P `* PRT %FN)
{ (^ PRED)='%FN<(^ SUBJ)>'
~(^ PASSIVE)=+
| { (^ PRED)='%FN<NULL>(^ SUBJ)'
| (^ PRED)='%FN<(^ OBL-AG)>(^ SUBJ)'
(^ PASSIVE)=c +
(^ PRESENTATIVE-TYPE)=passive
(^ SUBJ PRON-TYPE)=c expl_
| (^ PRED)='%FN<(^ OBJ)>(^ SUBJ)'
(^ PRESENTATIVE)=+
~(^ PASSIVE)=+
(^ SUBJ PRON-TYPE)=c expl_
~(^ OBJ DEF)=+
(^ CHECK _PRT-VERB)=+
(^ PRT-FORM)=c PRT.

The new line in the lexical entry in (10) refers to the XLE template `V-SUBJ-PRT` in (11), which already existed and was not added as part of this modification. The second line in (11) builds the predicate name 'flate*ut' by concatenation. The next line in the template starts a disjunction which specifies three alternatives: regular active, as in example (9), impersonal passive, as in (12), and active presentative, as in (13).

(12) Det flates ut.
It is flattened out
'There is flattening out.'

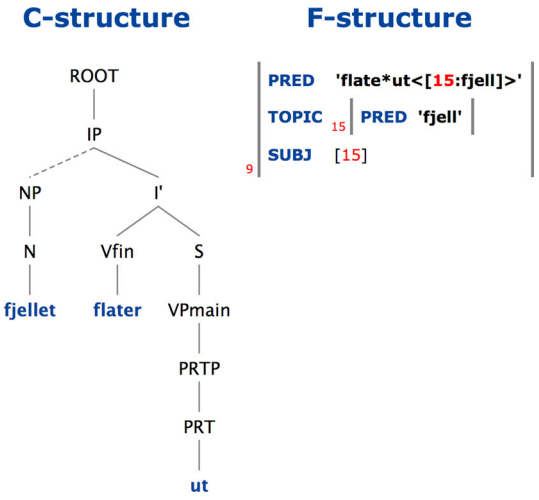


Fig. 11 Analysis of (9) after lexical update

(13) Det flater ut en fugleflokk.
it flattens out a bird flock
'There is a flock of birds flattening out.'

Table 3 also shows that the second and third most frequent types of lexicon updates for verbs are new intransitive and transitive readings, respectively. Thus, parsebanking draws attention to the fact that many verbs exhibit variation along the dimension of transitivity, and it reveals that this variation is not fully captured by standard dictionaries. For the sentence in (14) parsing initially failed when the lexicon contained no intransitive reading for *avslå* 'decline'. After this intransitive reading was added, the sentence was successfully reparsed.

(14) Men bestefar avslo.
but grandfather declined
'But grandfather declined.'

Adding an inquit reading is another frequent type of update in lexical entries for verbs. As mentioned above, inquit verbs are verbs of saying and related verbs that may occur in this function, and in the analyzed texts a large variety of verbs are used in inquit clauses. This is not surprising, since the text material contains many fiction texts with numerous passages of dialogue as well as internal monologue. The addition of an inquit reading in the lexical entry for a verb involves adding a subcategorization frame specifying that the verb takes a sentential complement as one of its arguments as well as a feature allowing it to occur in the syntactic position typical of inquit verbs.

(15) Hva mener du med det? stotret hun.
what mean you with that stammered she
'"What do you mean by that?" she stammered.'

The sentence in example (15) was initially given a partial analysis by the parser. That is, the word sequences *Hva mener du med det* and *stotret hun* were respectively identified as sentence units, but no complete analysis was found, because the lexicon entry for the verb *stotre* 'stammer' contained only an intransitive reading. An inquit reading was added to the entry, and after reparse the sentence *Hva mener du med det?* was successfully analyzed as a sentential complement to the inquit verb.

Table 3 also presents numbers for lexicon updates concerning nouns, adverbs, prepositions, and adjectives. With respect to nouns and adjectives, the data indicate that also in these categories there is a considerable need for adding subcategorization frames involving MWEs. Moreover, Table 3 shows that adding mass readings for nouns is another frequent type of lexicon update. Traditionally, dictionaries for Norwegian do not provide information on the distinction between mass terms and countables, but this information is required for producing correct syntactic analyses with NorGram. By default, the noun entries in the NorGram lexicon are therefore count nouns, and mass readings are added as they are encountered in the corpus.

One could imagine a number of automated procedures that create new lexical entries with modified subcategorization frames or features on the fly. A procedure that has been implemented in NorGram is similar to the Universal Grinder (Pelletier 1975), which produces mass noun readings from count nouns. In order to prevent overgeneration, the grinder is only applied in cases where the parser does not produce any analysis for a sentence; in these cases, all count nouns get mass readings as alternatives, and the sentence is automatically reparsed.

One could also imagine similar procedures for verbs. However, for verbs, there are many possible subcategorization frames, and allowing automatic postulation of unattested frames would easily lead to overgeneration. Therefore we only produce new subcategorization frames manually for cases that are present in the corpus.

Table 3 shows that lexicon updates involving new readings of adverbs constitute another frequent type. This illustrates how the lexical category of a given word must often be more fine-grained than what is provided by the lexicon. In the case of adverbs, there is only one large class with the part of speech ADV in the original lexicon. However, different types of adverbs vary considerably in their syntactic distribution, and it is therefore necessary to classify them into subcategories in order to account for this distribution. NorGram distinguishes between 24 categories of adverbs based on syntactic position, usually named according to their typical semantic contribution.

For instance, between the finite verb and the object there are adverb positions with ordering constraints for ADVatt (attitude adverbs like *dessverre* 'unfortunately), ADVprt (particle adverbs like *vel* 'I suppose'), ADVcmt (commitment adverbs like *egentlig* 'actually'), ADVneg (negation adverbs like *ikke* 'not'), and others. Example (16) illustrates that particle adverbs (*vel*) occur before commitment adverbs (*egentlig*), which occur before negation adverbs (*ikke*).

(16) Jeg har vel egentlig ikke noe å legge til.
I have I suppose actually not something to lay to
'I actually have nothing to add, I suppose.'

We also distinguish between ADVdeg (degree adverbs like *ganske* 'quite', which modify adjectives and other degree adverbs) and ADVdegloc (locational degree adverbs like *langt* 'far', modifying locative adjuncts); see example (17), where *langt* modifies the preposition *fra* 'from'.

(17) ganske langt fra vannet
quite far from the lake
'quite far from the lake'

These examples illustrate that a descriptively adequate treatment of Norwegian needs to distinguish between different classes of adverbs motivated by their syntactic distribution. Such distinctions are not only relevant for parsing, but also for other purposes, such as language learning.

6 Multiword expressions

As already noted in the previous section, MWEs are involved in many of the necessary lexical updates. The term MWE is frequently used in computational linguistics[7] and refers to the idiomatic, often non-literal part of the language. The notion of *idiomaticity* has been applied in numerous and various ways, but is generally associated with properties such as *lexical and grammatical fixedness* (or *frozenness*), *convention*, and *non-compositionality* (Nunberg et al. 1994; Moon 1998; Cowie 1998; Sag et al. 2002; Baldwin and Kim 2010). Non-compositionality refers to a situation where the linguistic properties of an expression cannot be fully derived from the properties of its component words and the way in which these normally combine, and it is central to many of the problems encountered in parsing of MWEs. Traditional dictionaries often list idioms as examples, but fail to provide information about their idiomaticity.

MWEs, and in particular MWEs that are idiosyncratic at the linguistic level (*lexicalized MWEs* in the terminology of Sag et al. 2002), present a great challenge for parsing because they exceed word boundaries, have unpredictable or irregular morphosyntactic properties, and are sometimes discontiguous.[8] The most immediate problem with MWEs, however, simply concerns recognizing them as such (Losnegaard et al. 2012). Although there are a considerable number of MWE

[7] See for instance http://mwe.stanford.edu/, http://multiword.sourceforge.net/, http://typo.uni-konstanz.de/parseme.

[8] For a thorough account of MWEs and automatic analysis we refer to Sag et al. (2002).

entries in NorKompLeks (more than 2500 prepositional verbs, 1800 particle verbs and almost 400 fixed expressions), these are far from sufficient for accounting for all of the MWEs occurring in our corpus.

Besides particle verbs, which were already discussed in Sect. 5, phrasal verbs include prepositional verbs. Moreover, not only verbs, but also nouns and adjectives may take prepositional arguments. NorKompLeks only provides this kind of subcategorization frame for verbs. Such frames are added to the NorGram lexicon by augmenting the relevant predicate-argument structures. Examples are *rette på* 'adjust', *mening med* 'point of', and *opptatt med* 'concerned with', as illustrated in (18), (19) and (20), respectively.

(18) Han rettet på parykken og snudde seg langsomt
he straightened on the wig and turned himself slowly
mot henne.
towards her
'He adjusted his wig and slowly turned to face her.'

(19) Hva var da meningen med å sette meg i slik
what was then the meaning with to put me in such
forlegenhet?
embarrassment
'What was the point of embarrassing me like that?'

(20) Hun ble veldig opptatt med å børste kakesmuler av
she became very busy with to brush cake crumbs off
kåpa si.
the coat her
'She became very concerned with brushing cake crumbs off of her coat.'

In constructions with selected prepositions, the verb, noun or adjective will as a rule keep its original meaning, while the meaning of the preposition is semantically bleached and does not contribute to the semantics of the overall construction to any large extent. An example of this is *le av* 'laugh of' ('laugh at'), where the verb retains its main sense 'laugh', and the preposition introduces as an argument the participant causing the laughter. Insofar as the preposition conveys some modification of the main predicate, this change in meaning will be idiosyncratic and fairly transparent. The meaning of the expression is thus not fully compositional, and the preposition to be used is not fully predictable. In example (18), the inherent and concrete meaning of the verb *rette* 'straighten' is preserved while the addition of the preposition *på* 'on' invokes the more specialized and figurative meaning 'make right', 'adjust'.

In this respect, constructions with selected prepositions are situated somewhere between institutionalized MWEs (i.e. MWEs that are linguistically regular but whose

component words have a high frequency of cooccurrence) and semantically transparent idioms; the relation between the main predicate and the preposition can, more than anything, be viewed as a special case of lexical preference. Whether these are to be treated as special constructions in the dictionary or grammar, dealt with compositionally, or accounted for as a valence property of the main predicate is a decision for the lexicon and grammar developer. The selected preposition is treated not as a lexical word, but as a grammatical word which is analyzed as an incorporated element in the predicate and whose main function is to signal the semantic role of an argument.

Other types of MWE frames that have been added to the lexicon during parsebanking are fixed expressions and verbal idioms. Fixed expressions are invariable expressions that do not necessarily have a normal syntactic buildup (Sag et al. 2002). It is thus the expression as a whole, and not the individual words, that must be assigned a lexical category. An example of a fixed MWE is *på kryss og tvers* 'crisscross', as in (21).

(21) Hvorfor er månen overstrødd av sprekker og rygger
why is the moon sprinkled of cracks and ridges
på kryss og tvers?
on cross and across
'Why is the moon completely crisscrossed by cracks and ridges?'

The prepositional phrase *på kryss og tvers* has a coordination of a noun and an adverb. Such coordinations are not licensed by the grammar rules, and the expression thus caused a fragment analysis prior to the addition of the MWE to the NorGram lexicon. Being a completely invariable prepositional phrase that allows no lexical variation, internal modification, or inflection, the MWE is added to the lexicon as a word-with-spaces entry, and appears in the c-structure as one node, as if it were a single word. Because of its syntactic properties, this particular MWE is classified as a locative adverb. The lexicon entry is given in (22), where the backquotes escape the spaces and treat them as regular characters in a single graphical word. Figures 12 and 13 show the c- and f-structures respectively for the example sentence.

(22) på` kryss ` og` tvers ADVloc ∗
@(LOCADVERB på-kryss-og-tvers på-kryss-og-tvers)

Conventional dictionaries usually provide limited information about MWEs, and their treatment is sometimes incomplete or incoherent. Often the expressions are not given as separate lexical entries, but occur as examples in the definitions of single-word entries. This information is difficult to extract when constructing an electronic lexicon. The case of *på kryss og tvers* exemplifies this problem.

In *Bokmålsordboka*, the phrase *på kryss og tvers* occurs as an example both under the entry for *kryss* 'cross' and the entry for *tvers* 'across', but does not exist as an

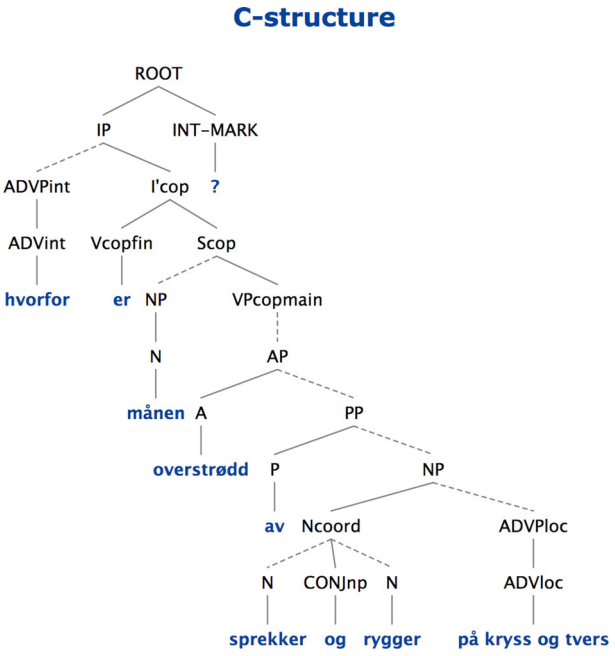


Fig. 12 C-structure analysis of (21)

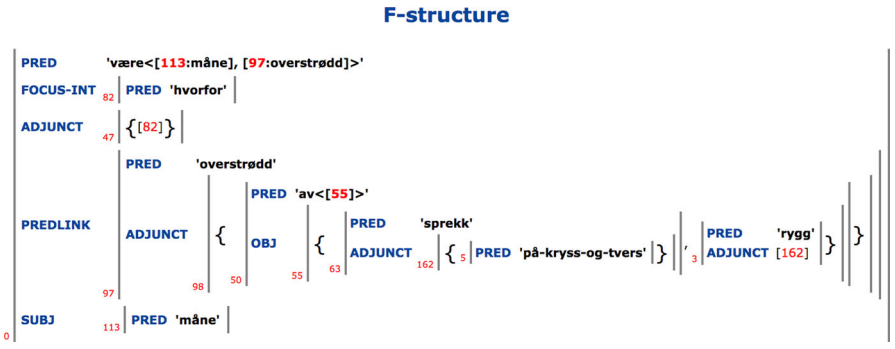


Fig. 13 F-structure analysis of (21)

entry of its own. The entry for *kryss*, whose main category is specified as a noun, has a sense partition named *kryssende bevegelse* 'crossing movement', implying that 'crossing movement' is a specific meaning pertaining to *kryss*. For this sense, the example phrase *gjennomsåke området på k- og tvers* 'search the area in every direction' is given without further information. In the entry for *tvers*, whose main category is an adverb, a sense partition states that the word is also used as a noun, but no information is given about its meaning as such. For this sense, the expression *på kryss og t-* is given along with a definition: 'i alle retninger' ('in all directions').

Although the MWE is referenced twice in *Bokmålsordboka*, neither entry explicitly refers to it as an expression. The information included in the two entries varies and the structures of the subentries are also different. As a consequence, *på kryss og tvers* is not listed in NorKompLeks, and the MWE was added to the NorGram lexicon by an annotator during disambiguation of the treebank. Adding lexical entries for hitherto unanalyzed MWEs is thus an important factor for increasing parsing coverage. Moreover, the addition of words with spaces to the lexicon during parsebanking results in a coherently classified inventory of fixed MWEs.

The NorGram lexicon also includes verbal idioms. These are idioms with a verbal core and a selected object; they have limited variability and have a non-compositional meaning. Some expressions are monovalent and only require a subject. Examples are *finne sted* 'find place' ('take place', 'happen') and *klage sin nød* 'complain one's distress' ('complain', 'pour out one's troubles'), which both consist of a verb and a selected nominal object. Another type is exemplified by *falle på kne* 'fall on knee' ('go down on one's knees', 'grovel'), with a verb and a selected oblique object in the form of a prepositional phrase.

Verbal idioms are added to the lexicon as lexical frames under the verb entry, as shown in (23), (24), and (25). Their organization in the lexicon is mainly based on subcategorizational properties: the core structure of the idiom, i.e. its fixed (or selected) components, and the semantic arguments it requires.

(23) finne V XLE { ...
| @(VPIDIOM-INDEFOBJ finne finne sted)}

(24) klage V XLE { ...
| @(VPIDIOM-DEFOBJ klage klage nød)
(^ OBJ SPEC POSS POSS-TYPE)
(^ OBJ NUM)=c sg
| ... }

(25) falle V XLE { ...
| @(VPIDIOM-PSELOBJ falle falle på kne)}

In each idiom frame, the verb predicate is extended with the fixed components of the idiom. Morphosyntactic restrictions on idiom components, such as definiteness and number for nouns, temporal or aspectual constraints, restrictions on passivization, etc., are regulated by special templates. The two examples of monovalent MWEs with a selected nominal object, *finne sted* 'happen' and *klage sin nød* 'complain', differ with respect to the definiteness of the object. The entries thus call the templates `VPIDIOM-INDEFOBJ` and `VPIDIOM-DEFOBJ`, respectively.

The template `VPIDIOM-INDEFOBJ` in XLE format is shown in 26. The second line builds a new predicate by concatenating the verb predicate (P) and the object predicate (OP). The predicate-argument structure is assigned on the third line,

showing which arguments are required and also which ones are considered semantic arguments of the verb. Only semantic arguments are included in the argument list; these are placed between angle brackets. In this case, only the the subject is a semantic argument, while the selected, non-thematic object (i.e. *sted* in this example) is listed outside the brackets. The last two lines are constraints on the definiteness and number of the object, which must be indefinite and singular.

(26) VPIDIOM-INDEFOBJ (P S OP) =
@ (CONCAT P `* OP %FN)
(^ PRED)='%FN<(^ SUBJ)> (^ OBJ)'
(^ OBJ PRED FN)=c OP
~(^ OBJ DEF)=+
(^ OBJ NUM)=c sg

When certain morphological properties are particular to a MWE, these are specified directly in the MWE entry. The MWE *klage sin nød*, for instance, has special restrictions on the determiner of the noun *nød*, which is mandatory and must be a possessive, and the number of the noun, which must be singular.

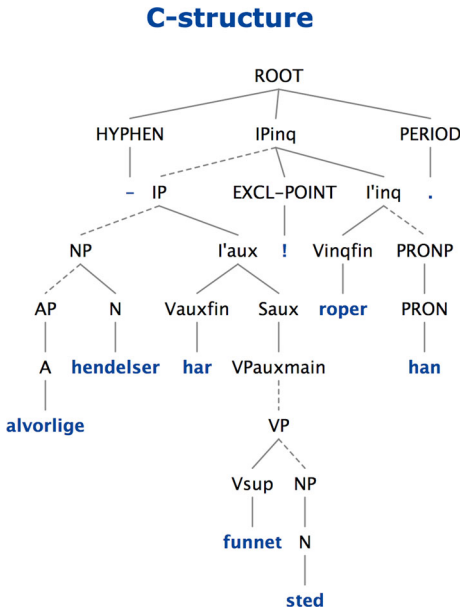
In the syntactic analysis, the verbal idiom is represented as a combined predicate in the predicate-argument structure of the f-structure, while the c-structure reflects the flexibility of the expression by representing each component of the MWE as a separate node. The c- and f-structure analyses of example (27) are shown in Figs. 14 and 15, respectively.

(27) – Alvorlige hendelser har funnet sted! roper han.
serious incidents have found place shouts he
'"Serious incidents have occurred!" he shouts.'

Other idioms may subcategorize for both a subject and a complement. A complement may be nominal (OBJ), clausal (COMP) or infinitival (XCOMP).[9] Examples are *sette pris på OBJ|COMP|XCOMP* 'put price on' ('appreciate'), *få tak i OBJ* 'get grasp in' ('get hold of'), *gjøre et (stort) nummer av OBJ|COMP|XCOMP* 'do a (big) number of' ('make a big deal about'), and *legge merke til OBJ|COMP* 'lay mark to' ('notice'). All of these have the syntactic structure verb, selected noun and selected preposition. A divalent idiom with a different syntactic pattern is *bringe OBJ på bane* 'bring OBJ on field' ('bring (something) up'), where the fixed elements are the verb and a selected prepositional phrase.

The treatment of idioms in the source dictionaries is not more consistent than that of other MWEs discussed above. The expression *finne sted*, for instance, is listed in *Bokmålsordboka* under the entry for *sted* 'place' as an example under a sense

[9] Since the semantic arguments are not lexically fixed parts of the idiom, these are represented here in terms of their syntactic functions. Alternative realizations are given as disjuncts.



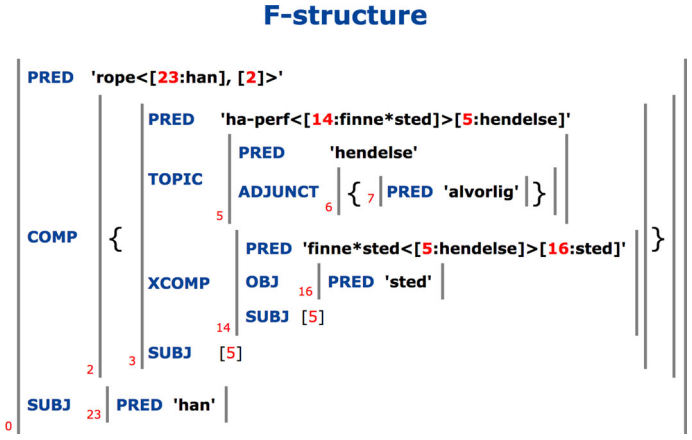


Fig. 14 C-structure analysis of (27)

Fig. 15 F-structure analysis of (27)

partition labeled *by*, *bygd*, *strøk* (‘town, village, district’). Although it is given with a definition, ‘foregå’ (‘happen’), it occurs together with other examples illustrating the concrete sense, and no information is given on idiomaticity or variability. Under the verb *finne*, however, we find more explicit information about the expression. This entry has a separate subentry with the heading *i uttrykket* ‘in the expression’, which is the most common way of marking expressions as such. The idiom is listed under this heading along with its definition ‘hende, skje’ (‘happen, occur’).

7 Evaluation and conclusion

Correct lexical information is essential for successful syntactic analysis, but we have found that lexical resources derived from dictionaries lack much necessary information because they are typically not tested in parsing.

In order to measure the impact of preprocessing on coverage, we analyzed all parsed sentences with a tokenizer and a morphological analyzer that were built solely on the basis of the lexicon before any text preprocessing. As detailed earlier, a sentence can only be successfully parsed if all its word tokens are recognized by the morphological analyzer. Taking MWEs into account, there might be several ways of tokenizing a sentence, and at least for one specific tokenization, all tokens should be recognized for a possible successful parse. Therefore, when for a given sentence there was no tokenization such that all tokens were recognized by the morphological analyzer, we concluded that the sentence would not have been analyzed without the additional extracted morphology. The measured difference in coverage was quite significant: among the 3,312,452 parsed sentences, 219,933 (6.6 %) would not have gotten any analysis without preprocessing. Moreover, many more sentences would have gotten a full analysis, but not the correct one, because of insufficiencies in the lexical resources, as discussed above.

In conclusion, we found that authentic text contains a wide variety of word forms which are not included in traditional dictionaries. Furthermore, traditional dictionaries do not cover all ways in which words are used, for example with respect to their subcategorization or in MWEs. In our parsebanking efforts, which are mainly aimed at high quality treebanks and compatible grammars, we find that the secondary result of tested and updated lexical resources that help overcome the limitations of traditional dictionaries is substantial and deserves more attention.

Although some nonstandard words may not be desirable in a lexicon for language generation, they are useful for parsing where missing items can cause failure. Information about nonstandard words and new compounds can also be useful for other applications such as automatic proofreading. Some information which we add, such as valency, mass terms and MWEs, may in modified form be included in dictionaries and language teaching materials.

One possible approach to the issue of missing lexical information would consist of using more information sources (gazetteers) and making informed guesses. Although our lexicon already includes large lists of named entities, a named entity recognizer might spot a few more potential names. However, since we are developing a gold standard parsebank, any guesses would have to be manually checked anyway. In this context, the benefit of checking a guess over adding an unknown word is small.

Good practice in lexicon development presupposes the involvement of trained annotators, but also the use of a sophisticated preprocessing interface which promotes efficiency and consistency. In the present study we have described how the INESS preprocessing interface (Rosén et al. 2012), in its further developed form, has been useful in enriching the Norwegian lexicon. The software of our interface accommodates in principle any language, but the system would have to be

adapted to the specific lexical categories, morphology, subcategorization, etc. of other languages.

The INESS project is building up a richer lexical resource for Norwegian and will continue to do so during the remainder of the project. The resulting reusable lexical resource will be made available upon completion of the INESS project in 2017.

Acknowledgments The work reported on in this article was carried out in the INESS project, which is funded by the Research Council of Norway and the University of Bergen. We are grateful to three anonymous reviewers for their constructive comments and suggestions.

References

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing, Chapter 12* (2nd ed.). Boca Raton, FL: CRC Press.

Bresnan, J. (2001). *Lexical-functional syntax*. Malden, MA: Blackwell.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., & Rohrer, C. (2002). The parallel grammar project. In J. Carroll, N. Oostdijk, & R. Sutcliffe (Eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics (COLING)* (pp. 1–7). Taipei: Association for Computational Linguistics.

Carter, D. (1997). The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the fourteenth national conference on artificial intelligence, Providence, RI* (pp. 598–603).

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford: Clarendon Press.

Dalrymple, M. (2001). *Lexical functional grammar, volume 34 of syntax and semantics*. San Diego, CA: Academic Press.

Dyvik, H. (2000). Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian: Basic properties of a lexical-functional description of Norwegian syntax]. In Ø. Andersen, K. Fløttum, & T. Kinn (Eds.), *Menneske, språk og felleskap* (pp. 25–45). Oslo: Novus forlag.

Dyvik, H., Thunes, M., Haugereid, P., Rosén, V., Meurer, P., De Smedt, K., et al. (2013). Studying interannotator agreement in discriminant-based parsebanking. In S. Kübler, P. Osenova, & M. Volk (Eds.), *Proceedings of the twelfth workshop on treebanks and linguistic theories (TLT12)* (pp. 37–48). Sofia: Bulgarian Academy of Sciences.

Hovdenak, M., Killingbergtrø, L., Arne, L., Sigurd, N., Magne, R., & Dagfinn, W. (Eds.). (1986). *Nynorskordboka: definisjons- og rettskrivingsordbok*. Oslo: Det norske samlaget.

Landrø, M. I., & Wangensteen, B. (Eds.). (1993). *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Oslo: Universitetsforlaget.

Losnegaard, G. S., Lyse, G. I., Thunes, M., Rosén, V., De Smedt, K., Dyvik, H., et al. (2012). What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. In J. Hajič, K. De Smedt, M. Tadić, & A. Branco (Eds.), *META-RESEARCH Workshop on Advanced Treebanking at LREC2012, Istanbul, Turkey* (pp. 69–76).

Maxwell, J., & Kaplan, R. M. (1993). The interface between phrasal and functional constraints. *Computational Linguistics*, *19*(4), 571–589.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press.

Nordgård, T. (2000). Nordkompleks: A Norwegian computational lexicon. In *COMLEX 2000 workshop on computational lexicography and multimedia dictionaries* (pp. 89–92). Patras: University of Patras.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491–538.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, *2*(4), 575–596.

Pelletier, F. J. (1975). Non-singular reference: some preliminaries. *Philosophia*, *5*(4), 451–465.

Rosén, V., & De Smedt, K. (2007). Theoretically motivated treebank coverage. In *Proceedings of the 16th Nordic conference of computational linguistics (NoDaLiDa-2007)* (pp. 152–159). Tartu: Tartu University Library.

Rosén, V., De Smedt, K., Meurer, P., & Dyvik, H. (2012). An open infrastructure for advanced treebanking. In J. Hajič, K. De Smedt, M. Tadić, & A. Branco (Eds.), *Meta-research workshop on advanced treebanking at LREC2012, Istanbul, Turkey* (pp. 22–29).

Rosén, V., Meurer, P., & De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In T. H. King & M. Butt (Eds.), *The proceedings of the LFG '07 conference* (pp. 397–417). Stanford: CSLI Publications.

Rosén, V., Meurer, P., & De Smedt, K. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In F. Van Eynde, A. Frank, G. van Noord, & K. De Smedt (Eds.), *Proceedings of the seventh international workshop on treebanks and linguistic theories (TLT7)* (pp. 127–133). Utrecht: LOT.

Rosén, V., Meurer, P., Losnegaard, G. S., Lyse, G. I., De Smedt, K., Thunes, M., et al. (2012). An integrated web-based treebank annotation system. In I. Hendrickx, S. Kübler, & K. Simov (Eds.), *Proceedings of the eleventh international workshop on treebanks and linguistic theories (TLT11)* (pp. 157–168). Lisbon: Edições Colibri.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science. Proceedings of the third international conference on computational linguistics and intelligent text processing* (Vol. 2276, pp. 189–206). Berlin: Springer.