A Corpus-based Study on the Evolution of *There*: Statistical Analysis and Cognitive Interpretation

Gard Buen Jenset

Dissertation presented in partial fulfilment
of the requirements for the degree *philosophiae doctor*

Department of Foreign Languages
University of Bergen
2010

Acknowledgements

First of all, I wish to thank my supervisors, professors Leiv Egil Breivik and Christer Johansson. They have both made unique contributions to this study: Leiv Egil Breivik for leading me into the field of existential *there* in first place, and for his guidance in the field of English historical linguistics; Christer Johansson for his guidance in the fields of statistics, programming, and psycholinguistics. This study would have looked very different without their contribution of skills, knowledge, and advice.

A number of other scholars have lent their time and advice in a number of areas. I would like to thank the following people for their valuable help: Daniel Apollon, Jòhanna Barðdal, Aldo Frigerio, Sandra Halverson, Kari Haugland, Øystein Heggelund, Lars Johnsen, Barbara McGillivray, Tore Nesset, Erik Norvelle, Marco Passarotti, Savina Raynaud, Hans Julius Skaug, Kolbjørn Slethei, and Koenraad de Smedt. Any errors or misrepresentations that might be found in this study are of course my responsibility.

I also wish to thank my parents for their relentless support of my studies.

The following academic institutions have also provided support, for which I am grateful: the Faculty of Humanities of the University of Bergen for providing financial support during my PhD work; Università Cattolica del Sacro Cuore in Milan for a short but productive research stay; and not least my current employer Bergen University College for granting me a generous leave of absence to complete this work.

I would also like to take the opportunity to thank those who have contributed to the production of the corpora that made this project possible. I am also grateful to the Oxford Text Archive for making YCOE available to me. This study could never have been completed without the use of numerous software packages, most of which are freely available through the efforts of their creators, and I owe a great debt to those who have developed these programs.
And, of course, Barbara again.

Bergen, February 2010

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Grammar is the art of speaking.
> Speaking is explaining one's thoughts by signs which men have invented for this purpose.
>
> A. Arnauld and C. Lancelot

The present work addresses the diachronic development of the so-called existential *there* in English. *There*-existentials, or existential constructions with *there* are exemplified below in (1):

(1) a) There is a tree in the backyard.
b) There were many applicants for the job.
c) There appears to be a discrepancy between spending and income.

Using large corpora, I will trace the development of this construction from Old English to Early Modern English, a period stretching from the ninth to the eighteenth century. The status of *there*, even in Present-day English, is disputed: scholars argue over whether the *there* found in (1) has a meaning or not, whether it is motivated by pragmatic, semantic or syntactic factors, and whether it is the subject of its clause or not. Through a detailed study of the diachronic development of existential *there*, this work will shed light on some of the conditions for the use of existential *there* in earlier stages of English and critically evaluate some of the claims made by others about the diachronic development of this construction. I will attempt to explain the diachronic

development with reference to Cognitive Linguistic theory and supporting evidence from other cognitive sciences.

1.1 Some definitions

1.1.1 Existential *there*

Existential *there* is also sometimes referred to as "weak", "introductory", "anticipatory", "expletive" or "dummy", cf. Butler (1980, 4); Nagashima (1992, 1). The term "existential *there*" was coined by Jespersen (1969, 130). A distinction is sometimes drawn between "existential" and "presentational" uses of *there* (see also chapter 2). This distinction is exemplified below, where the first sentence illustrates the existential use, whereas the second illustrates the presentational use:

(2) There is a book on the table (existential)

(3) There came a man into the room (presentational)

In the present dissertation, the term "existential *there*" will be used for the subject-like function of *there* illustrated in (1), as well as for the presentational use in (3). See section 2.2 for a further discussion and justification of this.

Related to this, it is necessary to consider the relationship between existential *there* (broadly defined) and locative *there*, i.e. *there* used as a deictic adverb. The existential use of *there* is normally singled out in English as distinct from the adverb of place *there* (henceforth *locative there*). It is widely recognized that the existential use contrasts with the locative use, as exemplified below in (4) and (5):

(4) *There*$_1$ is a bug in the software. (existential)

(5) a) The nasty little bug ran over *there*$_2$. (locative)
b) *There*$_2$ is the house where I grew up. (locative)

Breivik (1981); (1990); (1997) employs subscript numbers to distinguish the locative and existential uses of *there*: *there*$_1$ corresponds to the existential use in (4) above; *there*$_2$ corresponds to the locative use in (5). In many instances, especially in Old English, it is difficult to determine in all cases whether *there* is used in a locative or existential sense, cf. Butler (1980, 280); Breivik (1990, 181–188). For this reason, the terminology with subscript numbers has not been consistently adopted in the present study. This is also the reason why a study on existential *there* in Old English cannot focus exclusively on this use. Instead, the lexical form *there* is quantitatively described in

context without trying to determine the use in each instance. Thus, existential *there* will be investigated without paying too much attention to whether each occurrence of *there* can best be classified as existential or locative. Although this might seem paradoxical, it is a marriage of convenience between practical and theoretical considerations.

First, theoretical considerations: the approach adopted here is explicitly psychological, that is, it attempts to account for speakers' grammatical knowledge, in the sense described in Croft (2000, 26) (further discussed below). This knowledge I take to be *graded* in the cognitive linguistic sense, where phenomena related to categorization can be represented by good (prototype) examples in radial categories, cf. Rosch (1975); Lakoff (1987). Furthermore, I take this knowledge to be *probabilistic*, in the sense that there is an uncertainty attached to how a given lexical form will be interpreted by the hearer, cf. Jurafsky (2003). Whether probabilistic meanings are also fuzzy, vague, or simply a product polysemy will not be touched upon further here. In any case, such meanings must be considered probabilistic in the broad sense with respect to cognitive processing. Geeraerts (2006b, 99–148) argues that the distinction between polysemy and vagueness is unstable, since meaning arises interpretatively through cognitive processing (Geeraerts, 2006b, 138–141). Thus, the aim of the present work is not so much to identify instances of existential or locative *there*, or to determine exactly when the former emerged; instead, the emphasis is on identifying *contexts and factors* that could potentially serve as clues to listeners and influence the interpretation of the lexical form *there* in one direction or the other.

Second, the practical considerations: this dissertation is based on an analysis of large corpora with automated methods. For each time period (Old, Middle, Early Modern English) there is a corpus of more than a million words, with thousands of examples to be analyzed. It goes without saying that it would be impossible to analyze all these constructions individually to determine the status of every occurrence of *there*. Of course, other approaches would have been possible, such as using the corpora as sources of examples that can be carefully analyzed manually in context. However, such studies have already been carried out, see Butler (1980); Breivik (1990); Nagashima (1992). The ground-breaking aspect of the current dissertation is precisely its use of large amounts of data compared with earlier analyses, combined with the use of sophisticated statistical methods.

1.1.2 Existential Constructions

The terms used to refer to the class of utterances in (1) varies. Jespersen (1969) refers to "existential sentences", whereas Breivik (1990) uses the term "existential clause". In cognitive / construction grammar frameworks they are sometimes referred to as "*there*-constructions" (Lakoff, 1987); (Croft, 2001, 200), or "existential construction with

there, (Langacker, 2008, 496). The theoretical framework of the current study is *Radical Construction Grammar* (RCG), cf. Croft (2001). Following Croft (2001), I will use the term *the English Existential Construction* (EEC) to refer to the construction which the existential *there* appears in, based on a semantic-pragmatic approach to the construction's function, viz. to introduce new information by denoting the existence or occurrence of something. A *construction*, in the RCG sense, is defined in Croft (2001, 16):

> Constructions are objects of syntactic representation that also contain semantic and even phonological information (such as the individual substantive lexical items in the partially schematic idioms, or special prosodic patterns or special rules of phonological reduction as in *I wanna go too*).

I follow the convention laid down in Croft (2001, 51) of using capitalized names for language specific constructions (thus, "Existential Construction" and "existential construction" are different); also, constructions in RCG are properly denoted by the name of the language, hence the abbreviation EEC. For convenience, the general abbreviation EC will be employed, with context taking on the role of identifying it as an instance of the Present-day, Early Modern, Middle or Old English EC.

My definition of an English Existential Construction is merely the contextualized reinterpretation (in a RCG framework) of the definition of *there* found in Biber et al. (1997, 943):

> Existential *there* is a formal device used, together with an intransitive verb, to predicate the existence or occurrence of something (including the non-existence or non-occurrence of something). Most typically, a clause with existential *there* has the following structure:
>
> *there* + *be* + indefinite NP (+ place or time position adverbial)

Since constructions themselves are symbolic units, the categorization of utterances into constructions should rely on both form and meaning properties of the utterances, cf. Croft (2001, 52). As Croft (2001, 53) points out, categorizing constructions is a difficult problem that also faces the speakers and hearers of a specific language, which justifies the reliance on (psychological) research in categorization and taxonomy formation. Thus, the English Existential Construction is the syntactic, semantic, and phonological representation of the linguistic sign(s) used to denote existence or occurrence. I will return to the question of the operational definition of the EC in the next chapter.

1.1.3 Language, grammar, utterance

Notions such as "language", "grammar", and "utterance" are not theory neutral. Consequently, they need to be grounded in some theoretical framework. The definitions below, taken from Croft (2000, 26), are fundamental to the current research project:

- An *utterance* is a particular, actual occurrence of the product of human behavior in communicative interaction (i.e. a string of sounds), as it is pronounced, grammatically structured, and semantically and pragmatically interpreted in its context. ... An utterance as defined here is a spatiotemporally bounded individual. Thus, unlike sentences, only actually occurring tokens count as utterances in our sense

- A *language* is the population of utterances in a speech community. ... [Language] is only the set of actual utterances produced and comprehended in a particular speech community ... it is a spatiotemporally bounded set of actual individuals, not a set of 'possible' individuals

- A *grammar* is the cognitive structure in a speaker's mind that contains her [fn. omitted] knowledge about her language, and is the structure that is used in producing and comprehending utterances ... The grammar of each speaker is acquired on the basis of the subpopulation of the language that she is exposed to [fn. omitted]. Thus, each speaker will have a slightly different grammar.

Some consequences of the above definitions: first, as Croft (2000, 44–51) argues, child-based parameter setting theories of language change, as presented in e.g. Lightfoot (2006), are found inadequate. Second, it becomes necessary to abandon the notion of "grammatically motivated subjects" found in e.g. Butler (1980) and Breivik (1990). The first consequence is discussed in depth in Croft (2000), but see also section 3.4 below. The second consequence follows from the definitions above and will be explained below. Butler (1980, 5) refers to empty, or "dummy", subjects "as *grammatically motivated* because they presumably came into English to meet some new need in the grammar that resulted from language's change in type" (emphasis in the original). Similarly, Traugott (1992, 219) also makes this explicit link when she writes about the use of "*þær* [*there*] constructions" in Old English that "[t]his would be consistent with a construction that was to become obligatory later when subject position had to be filled". Breivik (1990) places more emphasis on the pragmatic discourse motivation for the use of *there* in existential constructions. However, he does suggest that at some point in the diachronic development of English "$there_1$ is inserted as an empty topic in pre-verbal position to satisfy the verb-second constraint" (Breivik, 1990, 298).

All these statements suggest a situation in which an external grammar somehow changes the language and forces (or at least constrains) the speakers' use of the language. However, if the language is the population of utterances produced by the speakers based on their "cognitive structures", the language and grammar cannot change without involving the speakers and their language processing. That is, language change is in the present study seen as taking place "in real time" through particular utterances and their interpretation. As such, the language cannot change "type" (such as from verb-second to verb-third) without directly involving the interplay of utterances and cognitive structures. If the external view of language change is to be taken literally, it requires the change to take place (somewhere and somehow) before the full effects are felt by the speakers. If we instead take a probabilistic view of language processing this, combined with the definitions of language and grammar above, entails a view where grammar is internal and fluent, and where interpretations of utterances are updated and modified based on the speaker's knowledge (i.e. grammar in use as defined above). The question of explanations of language change is dealt with further in chapter 3 below.

1.2 Aim and scope

The scope of the study is the English existential construction, with particular emphasis on the use and non-use of the so-called "existential" *there*, in prose texts from around AD 850 to AD 1700.

The aim of the dissertation is threefold, and relates to empirical, methodological, and theoretical questions:

(i) *Empirical*: to describe in quantitative terms the development of the *there*-existential from Old English to Early Modern English using available corpora

(ii) *Methodological*: to illustrate the usefulness of advanced statistical methods in diachronic corpus linguistics

(iii) *Theoretical*: to offer an integrated diachronic interpretation of the empirical results within the framework of *Radical Construction Grammar* and an evolutionary theory of linguistic change, cf. Croft (2000) and Croft (2001)

All three aims will not be given equal space: since the empirical data are extensive, they will necessarily take up more than a third of the dissertation as a whole.

The methodological assessment is invariable intertwined with the empirical results. While all the statistical methods employed are more or less considered standard in other empirical sciences (such as biology) they are not yet firmly established in linguistics. The methodological aim is properly seen as attempting to find a quantitative stepping

stone, or an intermediary link between the empirical and theoretical aims. The crucial test will be whether the use of advanced statistical methods can offer new insights, in terms of linking empirical description with theoretical interpretation, that would otherwise not have been possible without these methods.

The theoretical inclination of the current work is *Cognitive Linguistics* in broad terms, what Taylor (2002, 4) sums up as theoretical frameworks based on "the belief that language forms an integral part of human cognition, and that any insightful analysis of linguistic phenomena will need to be embedded in what is known about human cognitive abilities". More specifically, I will refer to interpretations using the terminology of *Radical Construction Grammar* (RCG) in an attempt to answer certain basic questions regarding the status of *there* in a principled way. These questions relate to what Coseriu (1987, 150) called the three problems of linguistic change:

> a) the universal problem of linguistic change (why do languages change at all?); b) the general problem of linguistic change (how and under what intra- and extra-linguistic conditions do languages normally change?); c) the historical problem of every individual change, that is, the problem of justifying the creation of a particular tradition and possibly the replacement of an earlier tradition

By combining the empirical results with an RCG analysis, the dissertation will attempt to give an answer to Coseriu's question c) regarding *there*, and to contribute circumstantial evidence for the approach presented in Croft (2000) to Coseriu's question b). The most pregnant question, why languages change at all, will not be touched upon explicitly. However, the evidence presented here together with some of the theoretical background from e.g. Keller (1994) will suggest that this question is perhaps best answered through the two previous ones.

1.3 Material and Method

The approach to language change taken in the present work is explicitly empirical and usage-based, and agrees with the description of the principal aim of historical linguistics as stated in Lehmann (1975, 42): to interpret data. The challenge, as pointed out by Lehmann (1975, 43), is to develop methods sufficiently advanced to allow us to posit explanations for specific phenomena. In the present work, I will attempt to provide explanations based on the integration of RCG theory, my own corpus studies, as well as reference to circumstantial evidence from other studies in experimental cognitive sciences.

1.3.1 Material

The material for the present study is taken from three (manually) syntactically annotated corpora, or *treebanks*, viz. the *York-Toronto-Helsinki Corpus of Old English* (YCOE), the *Penn-Helsinki Parsed corpus of Middle English* (PPME2), and the *Penn-Helsinki Parsed corpus of Early Modern English* (PPEME). Further details on the treebanks can be found in chapters 5, 8, 9 and 10. This approach offers three immediate benefits: first, using existing corpora saves time; second, using existing corpora entails objectivity in that the corpora are not annotated with my study specifically in mind; and third, using existing corpora ensures that the results can be replicated in subsequent studies.

Diachronic linguistics must by necessity be constrained by the amount of material available. The textual material available for historical English is, unfortunately, somewhat uneven. There are some well known gaps in the material, particularly the periods from approximately 1070–1150, and approximately 1250–1350, cf. Rissanen (1990, 357). For Old English, all extant prose is included in the YCOE. For later corpora, more text is available, but there are gaps (notably in early Middle English) which make it difficult to draw generalizations. However, the manuscripts available today do not necessarily have a specific selectional bias towards current linguistic research (although some genres for various reasons may be better represented than others). Consider the history of some of the Old English manuscripts which constitute our basis for saying anything about the language used by people living in England from AD 850 to AD 1150.

Obviously, we must rely on written sources that for some reason or another have been passed on to us. Ker (1957, xlviii–liv) notes that prior to AD 1200, the extant manuscripts containing Old English appear to have been appreciated and read.

The thirteenth and fourteenth centuries saw a devaluation of the old manuscripts, with attitudes towards them ranging, according to Ker, from vague curiosity to considering them useless and without value. This changed after Henry VIII declared himself head of the Church in England, and artifacts from the time when an English vernacular was used in a supposedly independent English Church came to prominence for political reasons. Ker remarks that from 1565 onwards, the old manuscripts were "studied and sought out eagerly by Archbishop Parker [of Canterbury 1559–1575] and his household as a means of promoting the 'Ecclesia Anglicana' " (Ker, 1957, lii). But even with the renewed interest in the Old English manuscripts, they were not immune to accidents: in 1731, a fire destroyed or damaged a number of the manuscripts in the Cotton collection (Ker, 1957, liv). Other texts survived only in the form of fragments after the manuscripts that contained them were cut up by binders (a practice which continued

from Medieval times and up to the nineteenth century) and used to wrap and line books (Ker, 1957, lxi-lxii). Finally, some of the manuscripts in the collection of Archbishop Parker were deliberately purged of texts, presumably because the texts were considered to be imperfect in some respect (Ker, 1957, lxiii).

As this brief overview suggests, our extant sources of Old English can hardly lay claim to exhaustiveness. Essentially, this must be considered an opportunity sample, rather than a proper random sample, cf. Hinton (2004, 50). Similar problems arise for the two later corpora, although for both of them more material is actually available than what has been included in the corpora. One major methodological problem is then how to overcome these shortcomings in the data? The traditional approach is to look at linguistic examples in context. The alternative is to use corpus linguistics and statistics to fill out the gaps and make estimations given what we can infer from the data.

1.3.2 Method: Corpus linguistics

The images brought up in Fillmore (1992) of the "armchair linguist" who suddenly thinks of a neat linguistic example and then writes a paper about it, and the die-hard "corpus linguist" who mindlessly counts frequencies and writes a paper about them, are ironic and powerful. Ironic, because deep inside we know this is not how research is carried out; powerful, because from superficial observation it can easily look like this is the way things are done after all. In the present dissertation, I will take the position that a number of syntactic, semantic and pragmatic phenomena can be studied through large scale corpus investigations using statistical methods. I will also argue that for many questions, such an approach is superior to other approaches for a number of reasons (although it may not always be practically feasible). Although it might seem otherwise, this is in fact an attempt at bridging the somewhat artificial gap between quantitative and qualitative linguistics. The frequencies are not particularly enlightening in themselves. Instead, they must be processed with statistical methods and evaluated against hypotheses and expected frequencies. Only then can the results be properly evaluated, and a major methodological point in the present work is the necessity of (linguistically) informed analysis of the statistical results. The emphasis, then, is on shifting the interpretation away from *both* single examples and raw frequencies, and over to the interpretation of test results and evaluation of hypotheses.

As such, the view defended here diverges from that defended in e.g. Fischer (2007a). Fischer argues that the best practice in historical corpus linguistics is to "check every example in context, which is hardly feasible. In general, what one does is to check a good part of them contextually" (Fischer, 2007a, 45). Fischer is of course correct in stating that corpora, based on specific editions, and possibly annotated, represent several intermediate layers of interpretation between the linguist and the original source

manuscripts (Fischer, 2007a, 45). However, I would argue that this is in many ways an *advantage* over working with the source manuscripts directly. Obviously, it is necessary that the corpus creators make good and consistent choices regarding editions and annotations. If we can take this more or less for granted (give or take some non-systematic errors), the linguist can easily test quantitative hypotheses against a large, objective material. This means that the biases of the individual linguist and his or her aims will only have a minimal impact on the material used for testing the hypotheses. I will attempt to show that by using appropriate statistical methods and comparatively large amounts of data, it is possible to gain a much better idea of the main trends in the material while still capturing the variation in the data.

Corpus linguistics, in the sense of employing processed, electronic texts, is of course not necessary for all branches of historical linguistics. The comparative method for reconstructing historical relationships between languages flourished long before the arrival of the computer. However, historical linguistics must, almost by definition, rely on textual data, and is thus naturally inclined towards quantitative corpus studies.

Similarly, studies in Cognitive Linguistics can be more or less empirical in nature. If there is a continuum between strictly qualitative and strictly quantitative work in Cognitive Linguistics, the present study decidedly leans towards the quantitative end. The strong commitment to corpus frequencies and statistical methods as a basis for drawing conclusions is nevertheless not a rejection of the importance of theory, as reflected in the reliance on RCG. Rather, as pointed out by e.g. Geeraerts (2006a), there is a need to operationalize cognitive theory and test it empirically through corpus studies and experiments. Although there is a growing trend towards more empirical work in Cognitive Linguistics, the field is still dominated by qualitative, introspective studies. As mentioned above, one of the fundamental aims of the current project is precisely to illustrate the value of advanced statistical methods to diachronic and cognitive research.

Through operational definitions, various hypotheses (presented in chapter 2) will be tested against the corpus data. That is, the methodological core of the present approach is to *not* use the corpora as a source of examples, but to test hypotheses against frequencies and other observable phenomena in the extant textual material. The hypotheses themselves are mostly generated based on previous research involving studies of examples, which illustrates the false dichotomy between quantitative and qualitative methods. Studying examples can lead to hypotheses which in turn can be operationally defined and tested quantitatively. The outcome of those hypothesis-tests are then interpreted in light of RCG and independent evidence from experimental cognitive sciences. The theoretical framework is integrated with the empirical part of the study through the operational definitions of the research questions. Essentially, by defining a language as a population of utterances, it is possible to argue that the texts in the corpora can represent language and grammar, in the sense defined here. This is, of course,

a theory-dependent operational definition, since it would hardly be possible to describe "all" utterances of any language, not even a contemporary one. As such, the sample of utterances used to describe the population can easily be seen as lacking representativeness. To complicate things further, we should not expect a one-to-one relationship between spoken and written language. However, for historical linguistics this is nevertheless our best foundation for estimating diachronic trends in the language. As the present study will show, a combination of large amounts of data and statistical models can yield estimates that are both informative and robust with respect to the main tendencies in language change.

1.3.3 Software

A number of different software tools were employed for the current project. Gries (2009) argues that *R* (R Development Core Team, 2008), being both a statistics and a text-search tool, is all that a corpus linguist needs. To some extent this is certainly true, since R is a remarkably versatile and useful program (and programming language), and R is the main tool employed for the statistical analysis in the present project. However, as Gries (2009) also remarks, certain corpora require specialized software for searching. This holds both for the diachronic and the synchronic treebanks used for the present dissertation.

The YCOE, PPME2 and PPEME treebanks come with their own search software, *CorpusSearch 2.0*.[1] The philosophy behind this search tool is to find syntactic trees that match given criteria, that is, it gives examples rather than frequencies.[2] For this reason it was necessary to further process the output files created by CorpusSearch. The tool chosen for this task was the scripting language *Perl*. This makes it possible to get detailed information on each syntactic tree in the treebanks, from which more fine-grained frequencies can be computed. A more thorough discussion on the issues relating to CorpusSearch and Perl can be found in chapter 5.

For the most part, specific R functions are referred to where they have been used, but one useful general function is mentioned here: most, although not all, tables in the subsequent chapters were created with the `xtable()` function, cf. Dahl (2009), which converts R matrices and data frames into LaTeX format.

[1] Freely available from http://sourceforge.net/projects/corpussearch/.

[2] Some summary statistics for a given search are presented by CorpusSearch at the end of the output file. As the subsequent chapters will make clear, this rather crude information is not sufficiently detailed for the current project.

1.4 Overview of the study

The chapters of the study are organized into three main groups, or parts. The first part is concerned with general background material and foundations for the subsequent investigations. Chapter 2 gives an overview of some main topics dealt with in previous studies of *there*. As an extension of this, a RCG description of the EC is presented, together with the hypotheses that will be tested in subsequent chapters. Next, chapter 3 gives some consideration to the question of explanation and causation in linguistics, and the role of statistics in linguistics. Following this, chapter 4 gives an overview of statistical tests and procedures which will be used in the analysis.

The next part deals with the extraction and description of corpus data. Chapter 5 gives an overview of the datasets extracted from the corpora. The structure of the treebanks is discussed and some comments on data extraction are provided. However, the bulk of chapter 5 is devoted to a brief description of the datasets and their measurement variables, or factors. Two of these factors are treated in more depth in separate chapters. In chapter 6 a measure of syntactic complexity is introduced, since it clearly warrants an explicit justification. Chapter 7 provides an overview of another measurement variable, namely a set of semantic verb classes used in the subsequent analyses. As with the measure of syntactic complexity, a separate chapter is required to give an adequate description of this factor.

In the third part, elements from the background discussions and the datasets are brought together. Chapters 8, 9 and 10 present data from the three treebanks. These chapters present surveys of the data, in addition to testing the hypotheses from chapter 2. Then, chapter 11 attempts to give an overview of the status of existential *there* in early English, as well as discussing possible mechanisms that might have driven the evolution of *there*.

Finally, chapter 12 summarizes the main findings and discusses the goals set out in the present chapter. The appendices provide in-depth information to some of the specific tests carried out and the data collection process.

Chapter 2

Previous studies of *there*

> "Cheshire Puss," she began, "would you please tell me which way I ought to go from here?"
> "That depends on where you want to get," said the cat.
>
> Lewis Carroll

2.1 Introduction

In this chapter I will critically examine some of the major themes and assumptions regarding existential *there* that can be gleaned from the literature on the topic. The discussion will be centered around these themes, and no comprehensive overview of all the relevant literature will be offered. This is motivated by the desire to place the key issues in focus coupled with the existence of such overview treatments elsewhere. Relevant works published up until about 1985 are comprehensively treated in Breivik (1990, 18–113). A summary list of works published up until 1992 dealing with *there* can be found in Levin (1993, 88). Ebeling (1999, 4–6), which is more cross-linguistic in scope, mentions some research published before 1999. Some of the issues dealt with in previous studies are arguably more peripheral to the current one, since they are explicitly situated within other theoretical paradigms such as generative grammar. One such issue is the question of whether existential *there* is base-generated or transfor-

mationally inserted, cf. Breivik (1990, 22–82); Pérez-Guerra (1999, 68–72).[1] Instead, some deeper consideration will be directed at what (based on existing literature) seems to be generally accepted, and what is disputed. Furthermore, some attention will be given to the extent to which some of the issues (whether controversial or not) can be considered a sound basis for an empirical corpus-based investigation of linguistic phenomena. Subsequent chapters will elaborate on some of these matters. The chapter culminates with a presentation of some hypotheses which can be gathered from the existing literature and tested empirically.

2.2 One or two *there*s?

Perhaps the most fundamental question is whether one or two types of $there_1$ should be recognized. In addition to the locative/existential distinction introduced in chapter 1, there is also the possible existential/presentational distinction, as exemplified below:

(1) There is a book on the table (existential)

(2) There came a man into the room (presentational)

Breivik (1990) discusses both types, encompassing "clauses containing existential/locative *be* or an intransitive verb which has included in it the meaning 'be in existence' or 'come into existence' " (Breivik, 1990, 4).

Coopmans (1989, 745) distinguishes the two and argues that *there* in presentational sentences functions as a "true adverbial introducing a particular context for presentational focus", caused by a form of semi-pro drop.

This suggests that there might be issues pertaining to the subject status of *there* depending on which categorization one chooses. In the present work, I follow Breivik's wider definition of what is "existential", through the notion of the EC. This has the advantage of shifting attention to the construction in which the morpheme is used. Whether (1) and (2) in fact are one or two constructions is difficult to determine on empirical grounds alone.

In RCG , the status of *there* is by definition linked to the construction. Croft (2001, 53–55) illustrates the difference between RCG and reductionist approaches to syntax by pointing out that in a reductionist view the categories "Subject" and "Verb" are

[1] Radford (1997, 333–334), which situates itself within the *Minimalist Program*, i.e. the latest incarnation of generative theory, seems to favor an insertion hypothesis where *there* is inserted into SPEC-IP. Whether this is an indication that the question has been resolved or not is considered outside the scope of the present (non-generative) study.

seen as parts of more than one pattern, e.g. both transitive and intransitive constructions. That is, the categories "Sbj" and "Verb" stand for the same categories in the two patterns [Sbj Verb] and [Sbj Verb Obj]. In RCG, these categories are defined by the constructions they appear in, giving [IntrSbj IntrVerb] and [TrSbj TrVerb TrObj]. Thus, instead of a general subject category, RCG distinguishes between e.g. a transitive subject and an intransitive subject. The rationale for this is that it "captures the fact that the distributional categories defined by the roles in the Transitive construction are not identical to those defined by the roles in the Intransitive construction" (Croft, 2001, 54). There are differences with respect to which verbs can occur in the two constructions, which motivates tying the verb (and the syntactic roles) to specific constructions.

The question is, should the two examples above be classified as two separate *constructions* in the RCG sense? Croft (2001, 55–58) sees constructions as inductive generalizations over usage, that is, the existence of a given construction is seen as an empirical question pertaining to speakers' grammars in the sense defined in chapter 1. It also follows from this that the syntactic and semantic differences between the verbs in (1) and (2) are not necessarily an argument against them belonging to the same construction, where "existential" refers to existence/appearance in the broad sense defined by Breivik (1990). The parts of the EC can then be described as follows, where the construction is taken to consist of an existential subject, an existential verb, and an existential NP-argument:

(3) [ExSbj ExVerb ExNpArgument]

Thus, the question is whether there is empirical evidence suggesting that the two sentences in (1) and (2) are sufficiently similar to be classified as belonging to the same construction. In the present work it is assumed that constructions are indeed created inductively through generalizations over categorizations of utterances as suggested by RCG. Since the semantic differences between the two examples in (1) and (2) are rather small (both have to do with existence and/or appearance), they are treated as part of the same EC, cf. (3).

2.3 The meaning of *there*

Most accounts favor a description of existential *there* where it is considered more or less empty, dummy or similarly without meaning. Breivik (1990) refers to existential *there* as a "dummy element", (Pérez-Guerra, 1999, 64) calls it an "expletive or dummy particle whose contribution to the propositional meaning of the sentence [is] null", to take but two examples.

The present study situates itself within RCG which tries to avoid having "empty"

or "dummy" elements, a characteristic it shares with other cognitive-functional frameworks. As a consequence of this, the meaningfullness of *there* follows more or less automatically. However, merely stating that *there* is meaningful in principle is somewhat unsatisfactory. An alternative view is defended by Bolinger who states that "*there* is neither empty nor redundant, but is a fully functional word that contrasts with its absence" (Bolinger, 1977, 121). This view is further developed in Lakoff's extensive case study on locative and existential *there*, cf. Lakoff (1987). However, both these studies rely heavily on the researcher's own native speaker intuitions regarding acceptability judgments, cf. also the comments on Bolinger's study in Breivik (1990, 83–87). Apart from the obvious fact that no native speakers are available for earlier stages of English, introspective judgments carry with them a range of difficult methodological problems, further discussed in chapter 3. Thus, a different approach is needed to resolve this question.

2.3.1 Locative or existential?

There seems to be widespread agreement that the existential use of *there* is derived from the locative use, cf. Bolinger (1977); Breivik (1990) and Breivik (1997); Lakoff (1987); Pérez-Guerra (1999). Most scholars seem to attribute this hypothesis to Jespersen (1924); Jespersen (1969).

A much debated question in the literature is the putative remnants of locative meaning in the existential use of *there* in Present-day English. Lakoff (1987) argues that the Present-day English existential *there* is motivated through its relationship with locative *there*. Specifically, existential *there* is argued to refer to a mental rather than a concrete space. Bolinger (1977) also argues that existential *there* has a trace of its original locative meaning, and that it is "locative in the broadest sense of whatever in space and time can be seen as something 'out there' " (Bolinger, 1977, 120). Breivik (1997) also refers to the notion of a mental space, whereby existential *there* retains some of its original locative meaning, although the nature of the location (and hence the meaning) is very different.

Attractive as these proposals might seem, they are difficult to test on diachronic material with its lack of native speakers. For the present study, I will take the terms "locative *there*" and "existential *there*" to instead refer to the locative and existential *uses* of *there* in various constructions. This shifts the focus from the morpheme to the construction and – in its extreme case – implies that there is no real difference between a locative and an existential *there*, only between the constructions the morpheme *there* occurs in. This solves the problem of ambiguous instances of *there* in Early English: instead of attempting to identify which cases of *there* are existential and which are locative, I will try to identify which contextual (linguistic and other) properties increase the

probability of the existential use of *there*. Of course, such information will be based on the corpus annotation where specific instances of *there* have been classified as locative or existential. But this classification is inescapably contextual and determined by the construction in which *there* is found. This ensures that with large amounts of data it is possible to inductively construct contextual indicators of the existential use of *there*, indicators which themselves are probabilistic, thus neutralizing the annotator-effect for specific ambiguous cases. This means that it is possible to study *there* through generalizations over its use in various constructions, rather than through intuitions about specific example sentences.

2.3.2 Pragmatic function

Related to the question of the meaning of *there* is its proposed pragmatic function. This has especially been discussed in Bolinger (1977) and Breivik (1990); Breivik (1997). Both Bolinger and Breivik hypothesize that the speaker uses $there_1$ to bring something into the awareness of the addressee. Breivik (1990, 150–156) refers to as the "signal function" of *there*. He proposes the plausible hypothesis that this signal function works through a "visual impact constraint" which requires $there_1$ to be inserted in clauses that do not provide detailed information about the physical setting of the situation, that is, clauses that do not bring something before the addressee, either literally or figuratively (Breivik, 1990, 141).

2.4 Locative-derived existentials in a typological perspective

It is necessary to devote some space to the analysis presented in Freeze (1992) and Freeze (2001), so as to clear away some serious misunderstandings. In his typological treatments of existential constructions, Freeze classifies the Germanic languages at least partially based on an erroneous description of existentials in Scandinavian languages. He states that the non-referential pronoun (or existential subject) in Germanic "is typically not locative", and cites examples from Swedish and German with respectively *det* and *es* (both meaning "it") for sentences where English would use *there*. The following Swedish example is provided by Freeze (1992, 573):

(4) Det fanns inget postkontor i den byn.
it find.PASS no post.office in that town
"There was no postoffice in that town."

He goes on to state that

> English is the only language in which I have found a lexically locative existential pronoun in subject position … It differs from that of (most?) other Germanic languages in that its pleonastic pronoun is lexically locative.
>
> Freeze (1992, 574)

I will not go into Freeze's generative analysis of existentials, but simply show that the differences he posits between English and the Scandinavian languages regarding the lexical origins of existential subjects are much smaller than what he claims.

The confusion found in Freeze (1992) and Freeze (2001) seems to stem from the fact that in Swedish and Norwegian it is typically possible to use *either* the equivalent of *it* or the equivalent of *there/here* as an existential subject-like pronoun. At least to some extent, this is subject to dialectal variation.

In his discussion of "dummy" subjects in Scandinavian languages, Breivik (1990, 251–261) gives examples of contexts where the use of a locative-derived existential subject "there" (*der/där*) is interchangeable with "it" (*det*). The example below is taken from Breivik (1990, 253), and det_1 denotes the existential use of "it", interchangeable with der_1 and corresponding to the English *there*.

(5) Der_1 / det_1 er mange folk til stades
"There / it are many people present"

As Breivik points out, the interchangeable use of *det* and *der* has a long history. In present-day Norwegian, *det* is more widely used than *der*, but *der* is frequently found in dialects along the southern and western coast of Norway, as well as in some conservative variants of written Norwegian (Breivik, 1990, 255). The following is an authentic example taken from an op-ed piece in the online version of a major newspaper in Bergen:[2]

(6) Men som alltid når kontrakten virker lovende – **der** finnes en liten skrift.
"But as always when the contract seems promising – **there** is a small print"

In this example the writer clearly uses *der* as an existential subject. This example is not merely an ungrammatical slip of the pen. A search of the *Norwegian Newspaper Corpus*[3], revealed at least 24 instances of the *der finnes* ("there exists") construction in

[2] http://www.bt.no/meninger/kommentar/holsen/Embetsmannsstatens-gjenkomst-966654.html, accessed on November 17, 2009.

[3] http://helmer.aksis.uib.no/aviskorpus/english.page

the Bergen newspaper *Bergens Tidende* in the period from January 2007 to November 2009. This quick search does not say anything about exactly how widespread such usage is, but together with anecdotal evidence from spoken Bergen dialect, it is beyond debate that the use of *der* "there" is a conventional alternative to *det* "it" in some Norwegian dialects.

However, the picture is even more nuanced when we turn to Modern Swedish. Falk (1993, 273) states that the "locative adverbs *här*/*där* 'here/there' also occur in the inverted subject position of main clauses and in the subject position of embedded clauses". The following examples are taken from Falk (1993, 273):

(7) a) Måste här städas till jul?
"Must here be-cleaned for Christmas?"
b) Kan där finnas ormar?
"May there be snakes?"
c) Jag undrar om här måste städas til jul
"I wonder if here must be-cleaned for Christmas"
d) Jag undrar om där kan finnas ormar
"I wonder if there may be snakes"

Such an existential use of *here* might also be a possibility in some Norwegian dialects.[4] In other words, when more dialectal evidence is taken into account, we find not only an "*it*/*there*" variation, but an "*it*/*there*/*here*" variation regarding existential subjects.

Danish, on the other hand, uses only *der* as a formal subject in ECs, whereas *det* is used in impersonal constructions (e.g. with weather verbs such as *rain*, *snow*, etc.), cf. Breivik (1990, 259–260). The Danish example below is taken from Breivik (1990, 259):

(8) Der_1 var mange heste på Tjele.
"There_1 were many horses at Tjele"

The overall picture of Scandinavian existential subjects is one of national, dialectal, and to some extent stylistic variation. What is clear is that there is ample evidence that Scandinavian languages do in fact have one or more existential elements or empty subjects that are obviously derived from locative adverbs. Contrary to the claims made by Freeze that English is the only (or part of a small minority) Germanic language with a existential subject derived from a locative element, Scandinavian languages do indeed use a locative-derived element as an existential subject, alongside *det* "it".[5] See

[4] Leiv Egil Breivk, p.c.

[5] Unless we are prepared to consider the three Scandinavian languages and English as a minority of the Germanic languages.

$$\overbrace{\overbrace{\text{iconic relation 1} + \text{iconic relation 2}}\ \text{indexical relation 1}, \text{indexical relation 2}\ \ldots}^{\text{symbolic relation}}$$

FIGURE 2.1: *A hierarchy of signs, illustrating how symbolic signs can be seen as composed of indexical signs, which are again composed of iconic signs. Reproduced from Deacon (1997, 75).*

e.g. Pfenninger (2009) for a further discussion on the situation in German with respect to *es*. For the equivalent Dutch construction with *er*, see e.g. Grondelaers, Speelman, and Geeraerts (2002); Grondelaers, Geeraerts, and Speelman (2007).

Perhaps the most important question which this Scandinavian variation raises is why English has settled exclusively for *there* in ECs. The specific evolution leading to this situation must surely be one of historical contingency. But it does suggest that it would be pertinent to look for other candidates competing with *there* as EC subjects in the Early English period.

2.5 On symbolic and other signs

Since RCG is a sign-based theory of language, it makes sense to turn the question around: instead of asking "what is the meaning of *there*", we can ask "what kind of sign is *there*"?

Following C. S. Peirce, Deacon (1997, 70) distinguishes three categories of signs, or "referential associations":

- *icon*: the sign is *similar to* its object of reference
- *index*: the sign is *close to, or correlated with* its object of reference
- *symbol*: the sign refers to its object by *convention*

Deacon changes the order of Peirce's hierarchy, and argues persuasively that they constitute a compositional hierarchy, where symbols are composed of indexes, and indexes are composed of icons. Figure 2.1 on page 20 shows the semiotic hierarchy proposed by Deacon (1997, 75).

It is worth noting that these are not essential properties of the signs, but a result of cognitive processing of usage events. As Deacon (1997, 71) puts it: "No particular objects are intrinsically icons, indices, or symbols. They are interpreted to be so,

depending on what is produced in response." Deacon posits that the conventional description of symbolic units in linguistics ("arbitrary pairing of meaning and sound") is in fact a property also held by indexical signs. Symbolic signs proper are singled out by another characteristic, namely their contrast or opposition with other symbolic signs. Deacon (1997, 86) states that

> the relationship that a [word] has to an object *is a function of* the relationship it has to other [words], not just a function of the correlated appearance of both [word] and object. This is the essence of a symbolic relationship.

That is, arbitrariness in language is *not* taken to be exclusively linked to symbolic signs. A sign can be taken to be indexical, yet still conform with the fundamental insight that language is the arbitrary mapping of meaning with form, cf. Saussure (1983); Croft (2001, 9). A corollary of this is that a breakdown in reference will cause an ordered descent, so that a sign which is no longer processed as symbolic will be interpreted as indexical. Deacon in fact uses *there* as an example to illustrate a point, but the mentioning is only in passing and not elaborated on:

> function words like "there" ... derive reference by being uniquely linked to individual contexts, objects, occasions, people, places, and so on
> Deacon (1997, 80)

Based on this, it is possible to suggest a unified model for understanding the difference between two types of *there* in Present-day English:

- a *symbolic*, or deictic use, denoted *there*$_2$ in Breivik (1990)
- an *indexical*, or existential use, denoted *there*$_1$ in Breivik (1990)

This interpretation is also coherent with the analysis presented by Panther and Thornburg (1998, 764), in their study on inferencing in discourse.

2.6 A description of existential *there* and the EC

As the quote from Biber, Johansson, Leech, Conrad, and Finegan (1997) in the previous chapter indicated, the EEC is composed of *there*, an intransitive verb (typically *be*), an NP, and some optional locative element. As discussed in Lakoff (1987, 462–585), it seems likely that the Present-day EEC is structured as a radial category. Furthermore, the central insight in Breivik (1990) and Breivik (1997) that *there* functions as

a presentative signal to the hearer seems to be essentially correct. Moreover, the suggestions from Deacon (1997) and Panther and Thornburg (1998) nicely complement Breivik's pragmatic analysis. Combining these points in a coherent, psychologically based, analysis is straightforward in RCG. Croft (2001, 54) contends that categories within constructions are construction-specific, leading to the distinction between categories such as [INTRVERB] and [TRVERB], to distinguish between verbs in the English Intransitive and Transitive Constructions, respectively. Thus, taking the simple constructed sentence in (9) as example, an RCG analysis can be attempted.

(9) There is a book on the table.

A first decision involves the status of the link between *there* and *be*: is it simply an instance of a constructional collocation? Croft (2001, 180) describes collocations as "combinations of words that are preferred over other combinations which otherwise appear to be semantically equivalent". Based on this, Croft (2001, 182–183) argues that there is a collocational dependency between the two symbolic units *spill* and *the beans* in the *idiomatically combining expression* meaning "divulge information". An idiomatically combining expression is an "idiom chunk" where "the syntactic parts of the idiom (e.g. *spill* and *beans*) can be identified with parts of the idiom's semantic interpretation ('divulge' and 'information')" (Croft, 2001, 181). However, this leaves us with the thorny issue of trying to fix a compositionally identifiable meaning to *there* in (9): it would appear that any meaning attached to *there* in (9) is derived from the construction. Another option would be to analyze the whole construction as partially idiomatic and partially open. Such an approach would rob us of the insight that the construction is flexible in that it allows non-prototypical patterns and verbs, as described in e.g. Lakoff (1987). In a worst case scenario, the idiomatic approach would need to posit separate partially idiomatic constructions for all the combinations of *there* + [EECVERB]. Alternatively, the pattern would need to be so broad that its usefulness would seem questionable: an instance of *there* somewhere in the utterance combined with an NP and a verb. Furthermore, it is not clear how such an approach would incorporate the intuition discussed in Breivik (1990) and Breivik (1997) that *there* does seem to have a pragmatic function. It is possible to recast this pragmatic signal information with reference to Croft's definition of *profile equivalence*, cf. Croft (2001, 257):

(10) *Profile equivalence*: In a combination X + Y, X is the *profile equivalent* if X profiles / describes a kind of the thing profiled / described by X + Y

There arguably profiles the entire construction, that is, it profiles the existence of something somewhere (or signals the existence of something, in Breivik's terms).

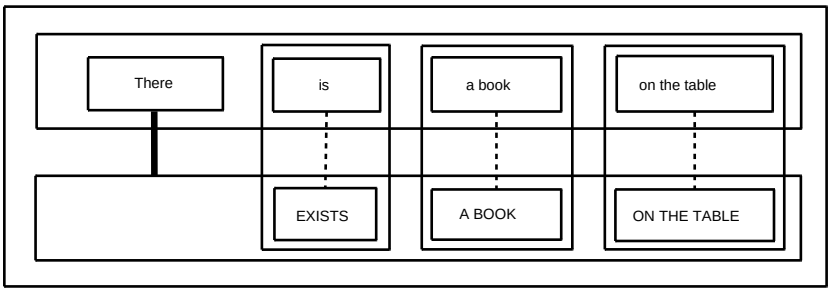


FIGURE 2.2: *Partial analysis of the EEC in a RCG framework. The proposed analysis connects* there *directly with the semantics of the EEC in an indexical relation (bold line). Symbolic relations between form (lower case) and meaning (upper case) are indicated by dotted lines. For ease of exposition the horizontal semantic links between the symbolic elements have been left out.*

In order to incorporate this in a unified sign-based analysis, the suggestions in Deacon (1997) and Panther and Thornburg (1998) seem to be the best option: *there* is analyzed as a *sign*, but as an *indexical* sign rather than a *symbolic* one. This approach allows us to maintain several valuable insights regarding the status and function of *there* in the EEC. At the same time, we are not forced to accept "empty" or "dummy" elements without "meaning" in a sign-based theory of language, cf. also Croft (2001, 6). Thus, I would suggest that *there* in this case can be analyzed as being in an indexical relation[6] to *the whole construction*. This suggestion simply amounts to expanding the conceptual toolbox regarding which units can be posited in a RCG analysis. This indexical sign, when interpreted as such in the context of the EEC, is then analyzed as the profile equivalent of the EEC, which accounts for its signal function in Breivik's analysis. Figure 2.2 on page 23 summarizes this analysis in a diagram. Under the current analysis, it does not make sense to label *there*$_1$ "empty" or "without semantic content". According to the definitions above, *there* has a meaning, viz. the reference inherited through its indexical relation to the EEC: something exists somewhere. This is not the prototypical form of lexical meaning or reference, but nevertheless I maintain that *there*$_1$ is not a "dummy" or "empty" element in Present-day English. However, it has a highly specialized meaning, one which is only activated within the EEC (by necessity, since this is the only context in which existential *there* can appear). It would seem that the motivation for *there*$_1$ being labeled as "empty" by previous studies has been its lack of a semantically compositional meaning and its lack of reference to other lexical

[6]Not to be confused with indexically coded dependencies, cf. Croft (2001, 199).

items for whatever meaning it has. But a RCG-based analysis takes care of the first problem: since words (more or less simple constructions) inherit their meaning from constructions (complex constructions) anyway, the meaning of *there*$_1$ is relatively easy to pinpoint, even if it is not compositionally identifiable. That *there*$_1$ does not have the same referential relationship to other words, as e.g. locative *there* to *here*, is accounted for by classifying *there*$_1$ as an indexical sign in Deacon's terms: such signs are pairings or meaning with form, but since they do not enter into a referential relationship with other words, they are indexical rather than symbolic signs.

2.7 Methods for studying *there*

To highlight some of the differences between the current work and previous studies of there, it is interesting to compare the methodological approaches taken elsewhere with mine. As already mentioned, the basis for the current study is the prose part of three historical treebanks, totalling almost 4.5 million words.

The most extensive previous historical studies in terms of data are Breivik (1990) and Nagashima (1992). In Breivik's study a total of 1 663 main and subordinate clauses is analyzed. Nagashima's study (or rather, collection of studies) covers at least 1 770 clauses. Both figures are based on the summaries presented in the respective studies. Since no relevant electronic corpora were available at the time of the original studies, Breivik (1990) and Nagashima (1992) used manually collected data.

Chapter 3 of Pérez-Guerra (1999) deals with *there* and presents a corpus based study. The actual frequencies are lower than for the two previous studies mentioned; 489 existential *there* sentences from Middle and Early Modern English, as well as 202 Present-day English existential *there* sentences, i.e. a total of 691, cf. Pérez-Guerra (1999, 92). The historical material is taken from the diachronic part of the *Helsinki corpus of English Texts*. This corpus is a precursor to the larger corpora used in the present study, and the total size of the diachronic Helsinki corpus is 1 572 800 words, i.e. about one third of the material available today through the three treebanks.

At the other extreme we find Pfenninger (2009), which is a comparative study of the grammaticalization paths of English and High German existential construcions. In the parts that refer to *there*, Pfenninger makes little or no use of frequency-based argumentation, and instead relies on discussions of a few examples from edited works of source texts or the research literature.

2.8 Linguistic change and what causes it

A classical candidate for change in (more or less) functional linguistics is *grammaticalization*, cf. Hopper and Traugott (1993) and Lehmann (1985).[7] Breivik (1997) mentions grammaticalization as a causal factor in the evolution of *there* and points out that it shares many characteristics with the classical case of *going to* → *gonna*: phonological reduction and occurs in a "highly constrained pragmatic and morphosyntactic [context]" (Breivik, 1997, 41).

The basic idea in grammaticalization is that lexical items, through frequency of use, associations with specific contexts etc. become less "lexical" in nature in that they lose semantic content and take on characteristics of function words. Hopper and Traugott (1993, 128) argue that "[t]he evidence is overwhelming that a vast number of known instances of the development of grammatical structures involved the development of a lexical item or phrase through discourse use into a grammatical item, and then into an even more grammatical item, and that these changes were accompanied by decategorialization from a major to a minor category". But what exactly does "grammatical structure" or "grammatical item" mean in the case of *there*? Surely, we would hesitate to label the grammatical function of subject as more grammatical than the grammatical function of an adverbial of place?

Grammaticalization has been severely criticized by many, perhaps most pertinently in Joseph (2001) and Janda (2001). Their criticism is primarily that grammaticalization is an adequate description of a number of phenomena, but does not constitute an explanation. I wholly concur with this criticism. It is crucial to avoid getting caught in the use of metaphor. "Lexical *content*", "bleaching", and "reduction" must in this context properly be viewed as metaphors expressing a qualitative evaluation of some aspect of linguistic *change* which in itself is neither good nor bad, neither content-giving nor content-stealing. While arguments concerning "bleaching", "loss", "reduction" etc., cannot *define* grammaticalization, the classification or description of $there_1$ as a case of grammaticalization is not necessarily rendered invalid.

Comrie and Kuteva (2005, 201) describe grammaticalization as "a natural process that leads ... to the development of more abstract semantic units and to the development of more complex morphology". It is not clear whether Comrie and Kuteva use *natural process* to mean "a process in/of/by nature" (they refer to it as a "cognitive process") or simply something along the lines of "ordinary, expected, common, spontaneous" etc.

The first interpretation must surely be wrong, unless we define a language as belonging to the natural world, together with such entities as e.g. mountains, rivers, oxy-

[7] The argumentation in this section is to a large part based on Jenset (2008).

gen, and birds. A language must – as it is defined in the present thesis – exist in and through its users. The naturalistic interpretation of Comrie and Kuteva's statement would probably seem too strong under most definitions of what language is, even one with a very strong biological commitment since this would require a biological mechanism (or at least a psychological one, probably with clear implications for underlying biological structures) to either act on language so as to cause grammaticalization, or indirectly provide a propensity for it.

That leaves a weaker interpretation, namely that grammaticalization is a "common, expected, spontaneous" phenomenon. But this robs the term of all explanatory value. It could then be rephrased as "grammaticalization is what happens (for whatever reason) when semantic units become more abstract and morphology more complex". This is merely renaming the phenomenon – it gives no indications as to what could possibly cause e.g. semantic units to become more abstract. It is, of course, possible to claim that what causes this is "grammaticalization". However, this makes the definition circular:

- Grammaticalization is the process by which linguistic units receive a more abstract meaning.
- Grammaticalization is caused by linguistic units attaining a more abstract meaning.

There is of course nothing wrong with the tautological statement that once linguistic units are perceived as more abstract they will tend to be interpreted abstractly; however, it hardly advances our understanding of language. Clearly, some *mechanism* must be introduced, which can cause the observed development. Two such frequently cited mechanisms in are *reanalysis* and *analogy*.

Is existential *there* a case of reanalysis? This is suggested in Breivik (1990, 290–291), citing a previously offered explanation for the equivalent Norwegian construction. The following examples are all taken from Breivik (1990, 290–291):

(11) a) Der_2 bor en gammel mann.
"There_2 lives an old man".
b) Der_2 ligger en bok.
"There_2 lies a book".

(12) a) Der_2 bor en gammel mann, i det huset.
"There_2 lives an old man, in that house".
b) Der_2 ligger en bok, på bordet.
"There_2 lies a book, on the table".

(13) a) Der_1 bor en gammel mann i det huset.
"There_1 lives an old man in that house".

b) Der_1 ligger en bok på bordet.
"There_1 lies a book on the table".

However, Campbell (2004, 284) asserts that "reanalysis depends on the possibility of more than one analysis of a given construction". This means that *before* $there_2$ can be reanalyzed as $there_1$, it must be possible to reanalyze $there_2$ as $there_1$. In other words, the reanalysis must already have taken place before it occurs. Or, more plausibly, it must at the very least be possible to analyze $there_2$ as something *else* than $there_2$. How could such a situation arise? As with grammaticalization, a tautological reference to reanalysis as what happens when a linguistic unit is reanalyzed is possible, but does not provide new insight. Instead, it is necessary to specify under which circumstances something could be reanalyzed as something else.

Appeals to abductive reasoning, as opposed to deductive and inductive reasoning, cf. Hopper and Traugott (1993, 38–40) and Andersen (1973), are possible, but not entirely satisfactory. As Andersen (1973, 775) states, "[a]bduction proceeds from an observed result, invokes a law, and infers that something may be the case". Intuitively this seems reasonable, especially if the law[8] in question receives a probabilistic interpretation, which would be very close to a form of statistical reasoning known as Bayesian statistics, cf. Gelman, Carlin, Stern, and Rubin (2004) and the discussion in chapter 4. The probabilistic (or vague, cf. the comments in section 1.1.1 above) nature of meaning as a result of processing could easily yield alternative interpretations through frequent occurrence in specific contexts, stress reduction etc. But note that these sources of variation are not necessarily caused by grammaticalization.[9] It may well be the case that abductive reasoning is one of the ways human cognition functions. However, although explaining language change or grammaticalization in terms of general cognitive capacities is certainly a goal in cognitive linguistics, such a blanket explanation would reduce to "because humans reason that way", and would anyway be contingent on the empirical issue of whether and to what extent people actually do reason in such a way. Explaining the observed development would require further elaboration on how and under what circumstances such reasoning would have an impact on language, i.e. making predictions.

The question of explanations and causality in linguistics will be discussed in depth in chapter 3. For now, suffice it to say that as it stands, I consider grammaticalization to be a useful and insightful *description* of the evolution of existential *there*. However, I do not agree with its proponents in that it has the potential to *explain* these processes.

[8] Andersen (1973, 775) states it may either be "an established truth" or a new "tentative generalization", i.e. not necessarily a law in the strict sense.

[9] For a good and balanced overview of the problems with cause and effect in grammaticalization see Fischer (2007a, 115–124).

2.9 Hypotheses

Based on the discussion in the present and previous chapter, a number of questions and hypotheses regarding the status of *there* and the EC in earlier stages of English have been outlined. Two overarching questions crystallize themselves:

(i) Is *there* special compared with other (locative) adverbs?

(ii) Does the status of *there* change in some way at identifiable points in history?

To investigate (i), it is necessary to examine the frequencies of *there* and its contexts. In the case of (ii), the frequencies and contexts need to be cross-checked against the chronological development to see whether diachronic factors directly influence *there* and its contexts.

2.9.1 *There* and other adverbs

An important question is why *there* was reanalyzed or evolved into an existential subject. What set *there* apart from other adverbs? Butler (1980, 279) says

> But if Adv – V – S was a common word order in late Middle English, why was it only in existentials that adverbs became reanalyzed as subject pronouns? Since Adv – V – S order deviated from the canonical SVO pattern, one might expect many fronted adverbials to be reanalyzed as subjects. I do not have a satisfactory answer for to this question. I can only suppose that existentials were a recognizable syntactic type with a clear function, and that it was a natural development for a special morpheme to arise to mark it.

This is indeed a good question, and several scholars have discussed the possibility of an existential *þa* ("then") in Old English, cf. Breivik (1990, 289–290) and Enkvist (1972). A working hypothesis for the present investigation is that the reanalysis of any morpheme is context-dependent, i.e. dependent on the *construction* as it is defined in cognitive linguistic terms. This can be further operationally defined as the syntactic context. Thus, it is to be expected that *þa* and *there* occur in different contexts given the fact that Present-day English has an existential *there* but no existential *then*.

A key issue, then, is what kind of corpus characteristics of *there* can plausibly be said to account for the use of this morpheme rather than another one. Again, quoting Butler (1980, 279–280):

> the high frequency of locative and time adverbials in existentials may result from the basic meaning of the existential construction: if existentials

> function to introduce new topics onto some scene or setting, we would expect adverbials often to be present referring to that setting. It is very hard, however, to prove that *þær/there* was more frequent than other locative and time adverbs.

With the currently available corpora, the frequency distributions of locative and temporal adverbs can easily be scrutinized. In chapter 8 dealing with Old English, it will be shown that not only overall frequency, but also frequency by context needs to be considered. This is also suggested by Butler (1980, 281), who writes "[if] locative *there* was the most commonly occurring adverb in early English existentials, then it would seem to be the most likely adverb to become the marker of the existential construction".

Furthermore, Butler (1980, 284) suggests that

> In Middle English, *þær* became increasingly associated with existentials and at the same time became semantically weaker, while SVO developed as the functional word order in simple sentences. Because sentence-initial *there* deviated from the emerging functional word order, it was reanalyzed as a new category type, a subject pronoun.

Breivik (1990, 296) argues that "$there_1$ becomes more and more common from Old English onwards". Combining large corpora and statistical techniques, it is possible to investigate whether there is in fact a diachronic trend, or if changing frequencies are caused by e.g. changing sample sizes.

The prediction that this process mainly took place in Middle English can be tested empirically by using sentences with an adverb, first and foremost *thær*, and *be* as a proxy for the existential construction, since constructions as such are not annotated in the corpora. An additional refinement involves looking at syntactic complexity, based on the hypothesis that existential constructions will have a tendency to have a low complexity. While not perfect, such an approach should cover many, hopefully most, of what can reasonably be described as existential constructions with *there* in Early English. Although there is bound to be some noise in the data (complex existential constructions and "light" locative constructions), the results should be reasonably accurate approximation due to relatively large sample sizes.

2.9.2 Initial adverbials

According to Butler (1980, 278–282):

> As the Old and Middle English periods progressed, adverbial elements came to be sentence-initial more and more consistently, and almost always caused subject-verb inversion. At the same time, the word order of

> simple sentences was becoming more regularly SVO. The fronted adverbs in existentials deviated from this regular SVO order, and at some stage the most frequent of these adverbs, *there*, was reanalyzed as a subject pronoun. (Butler, 1980, 282)

This can be restated into two separate, but related, hypotheses:

- Throughout the Old English period there is an increasing proportion of adverbial elements in initial position;
- Sentences with *there* in initial position should have a lower syntactic complexity.

Butler uses the term "simple sentences" several times, but it is not clear whether he takes it to mean "syntactically simple" or "basic, common". A competing view is posited by Pérez-Guerra (1999, 107–109) who, having studied the post-verbal elements of existential *there* constructions in Early English, argues that the use of existential *there* can at least in part be explained by an extraposition strategy (similar to *it* in *It is difficult to write a dissertation*) governed by the principle of end weight.

Irrespective of whether this is a fair interpretation of Butler, such a hypothesis is nevertheless interesting: does syntactic complexity matter in a putative process of grammaticalization or reanalysis? Butler speculates that the reanalysis of *there* took place in simple sentences first, under pressure from a typological change. The problem of defining syntactic complexity will be addressed in chapter 6.

2.9.3 The status of *there*

If Deacon (1997) is right in suggesting that *there* is indexical when used in the EC, and the RCG analysis summarized in figure 2.2 is accepted, it is possible to posit the following hypothesis:

- if *there* in Old and Middle English is fully symbolic ($\approx$ locative, deictic), it should *not* be tied to one specific context
- conversely: if *there* in Old and Middle English is indexical ($\approx$ empty, existential, non-referential), it is expected to be associated with a particular context as in Present English

The operational definition of the prototypical EC is thus as follows: a non-random probability of the co-occurrence of *there* followed by *be*, and an NP. Note that the indexical status of *there* does not follow from this operational definition; rather, the indexical status of *there* is a theoretical definition from which the operational definition

is deduced. What is being tested is thus a null-hypothesis that the EC in previous stages of the language does *not* behave like the Present-day EC. Or to put it differently, under the null-hypothesis it is not expected that behavior of the *there* and *be* will be like that in Present-day English. The operational definition above allows for non-prototypical instances of the EC in that other verbs than *be* could occur, as well as other adverbs than *there*. Again, under the null-hypothesis that the EC has not changed, the probabilities are expected to be stable. That is, we would not expect more than random variation with respect to the probability existential *there*.

2.10 Summary

Some key points have been emphasized, namely that a unified, sign-based description of existential *there* can be achieved through a RCG analysis. This solves a number of problems such as motivating the distinction between the so-called existential and presentational *there*, the putative meaninglessness of *there*, and also provides a reasonable operational definition of existential *there*.

Regarding methods, it was shown that while introspection can be useful as a way of generating hypotheses, it is difficult to directly evaluate this for extant languages and impossible for historical material. Furthermore, relying on reference grammars, as in the case of Freeze (1992), is hazardous since this approach is only as good as the reference grammar itself. If the grammar books lack vital information regarding usage and variation, the resulting analysis will have less validity than one which is based on empirical data.

Finally, a number of hypotheses based on the research literature were proposed and operationally defined, specifically:

(i) *there* is more frequent than other locative adverbs in Old and Middle English ECs

(ii) *there* is more frequent than temporal adverbs in Old and Middle English ECs

(iii) *there* became fully associated with the EC only in Middle English

(iv) the proportion of adverbs in initial position increased during the Old English period

(v) sentences with *there* in initial position had a lower syntactic complexity than other sentences

(vi) if *there* is used as a fully symbolic sign in Old English, it is not expected to be tied to one specific context

The next chapters will discuss the methodological foundations for testing these hypotheses.

Chapter 3

Method and methodology: From corpus frequencies to language

> We must not shun speculation *per se* but only untestable and barren speculation.
>
> Mario Bunge

3.1 Introduction

In this chapter I will discuss some problems related to method, methodology, and explanations. As such, this chapter functions as an epistemological background to the dissertation, presenting the justification for the operational definitions and methodological decisions made elsewhere, including definitions of sample and population. I present a sketch to an epistemology of language which justifies claims about "mechanisms" of change in linguistics. "Epistemology" is here understood broadly, as "issues having to do with the creation and dissemination of knowledge in particular areas of inquiry", cf. the first paragraph of Steup (2005). An attempt is made to relate this to the three fundamental methodological assumptions of the study:

(i) Central aspects of language can be quantified and measured;

(ii) General cognitive (i.e. individual) factors play a role;

(iii) Studies of language change requires usage-based data.

The following topics will be dealt with in the sections below. First, the role of statistics as a tool in linguistics is discussed and some possible counterarguments against its use are dismissed (section 3.2). Second, the crucial question of population and sample is considered in section 3.3. Section 3.4 deals with the complex issue of explanations in linguistics, including whether and how it is possible to make claims about causation in studies of language change. Finally, these themes are attempted brought together in a consistent epistemological approach to language change (section 3.5).

3.2 Statistics in linguistics

The place of statistics and frequencies in linguistics is not obvious, but then, it is not obvious that frequencies do *not* play a role either. The view that statistics does belong in linguistics is by no means new. Consider for instance the statement in Guiraud (1959, 15) below:[1]

> La linguistique est la science statistique type; les statisticiens le savent bien; la plupart des linguistes l'ignorent encore
> (Linguistics is a typical statistical science; the statisticians know this well; most linguists still ignore it)

Below I will argue that statistics and frequency based arguments are eminently suited for linguistics, but that frequencies and statistics need to be handled with some care.

3.2.1 Views on statistics

There are probably almost as many views on statistics in linguistics as there are linguists. Below I discuss three views that seem particularly relevant to the present study. These views can be categorized as follows:

(i) statistics is potentially useful, but not very convenient or practical;

(ii) statistics cannot as a matter of principle contribute to the goals of linguistics;

[1] I am grateful to Kolbjørn Slethei for bringing this quote to my attention.

(iii) statistics is useful, but should be limited to certain types of problems.

As the following discussion will reveal, all three views can be rejected.

Argumentation from convenience

Bloomfield (1933, 37) noted that a detailed statistical study of language use would be very informative, especially for studies of language change. However, having stated this, he immediately dismisses it as being unnecessary, since language is a convention-bound activity, and all the linguist really needs is to describe the norms that govern this activity. Bloomfield's motivation appears to be mainly pragmatic, since he refers to the simplicity of the latter method compared with the former (Bloomfield, 1933, 37). Given the lack of what we today would take for granted – easily accessible corpora, cheap and fast desktop computers, advanced and user-friendly statistics software – it was not unreasonable to take this view (although some studies showed the usefulness of statistics in linguistics already before the Second World War). Mair (2004) presents a modified version of this argument, when he argues that in cases where "superficial" statistical analyses come up short (Mair only refers to raw frequencies and proportions), the linguist should turn to "philological methods" (i.e. looking at examples in context) as a backup (Mair, 2004, 134). The alternative of turning from "superficial" to more advanced statistical analyses is not discussed by Mair, but would serve as an adequate (and perhaps superior) alternative. Note, however, that given the technical research infrastructure available today, this switch from simple to sophisticated statistics is mainly conceptual and its most demanding aspect is perhaps the proper operational definition of quantifiable hypotheses. Thus, Mair's insistence on using simple quantitative arguments as a first choice and then turning to qualitative methods as a secondary option (rather than refining the hypotheses and the quantitative methods), comes across as a variant of the convenience argument.

Argumentation from principle

A more hostile view of statistics has been maintained by Chomsky, who has claimed that there is no such thing as "corpus linguistics" (Aarts, 2000, 5). In his early writing, Chomsky discussed a statistical approach to language as an alternative to what he calls the "grammatical" approach, albeit a less desirable one. Consider the following quote from *The logical structure of linguistic theory* (*LSLT*) which gives one of the most well-known linguistic examples ever:

> The grammatical approach thus contrasts with a statistical approach that leads to an ordering of sequences from more to less probable, rather than a

> sharp division into classes within which no such graduations are marked. This literally correct statement of two different approaches can be misleading. It would be easy to picture the grammatical approach as an attempt, motivated by the complexity of the statistical data, to impose a rough approximation to the full statistical variation, with all sequences higher than a certain probability being assigned to *G* and all other to $\bar{G}$. But this would be a gross misconception. We have already noted that if our theory is to begin to satisfy the demands that led to its construction, then *G* will have to include such sentences as **11**, while such sentences as **12** are assigned to $\bar{G}$
>
> **11** colorless green ideas sleep furiously
> **12** furiously sleep ideas green colorless
>
> But clearly these strings are not distinguished by their assigned probabilities. If probability is to be based on an estimate of frequency in some English corpus, then this probability will be zero in both cases. Nor can they be distinguished, in some more sophisticated way, in terms of the probabilities of their parts.
>
> Chomsky (1975, 145)

Despite its later publication date, *LSLT* was written in the 1950s, and the views and limitations of that era are clearly seen in Chomsky's writing. Chomsky's argument is that both a syntactically well-formed string with nonsense semantics and a word-salad string will have the same probability in a corpus. Just like the lack of large, easily accessible electronic corpora made it difficult to analyze large samples of language data at the time, the lack of computational power also hindered statistical generalizations to *unseen* data, although such procedures had already been used during the Second World War, cf. Gale and Sampson (1995). Pereira (2000, 1242) notes that there is an important unstated underlying assumption in Chomsky's argument, namely that a probabilistic model *would assign a probability of zero to unseen events*. To illustrate this, consider the difference between a population and a sample (such as a corpus).

A sample is an integer subset of a population ranging from 0 to N, with properties that can be known. Population frequencies can only be *estimated* as the probability of randomly selecting a given type from the population, so that population estimates are probabilities ranging from 0 to 1, and an easy way of estimating the population probabilities is to divide sample frequencies by sample size N. This approach is known as the *maximum likelihood estimation* for a population frequency (Gale and Sampson, 1995, 217). To put it simply, we find a percentage of the sample and use that to infer the properties of the population. However, as Gale and Sampson (1995, 218) note, a problem with the maximum likelihood estimate is that any type which for some reason

is not included in the sample is assigned a probability of 0. This has two consequences: rare types (or types which for some other reason were left out of the sample) are estimated as non-existent, i.e. the probability is too small, whereas the actually observed types in the sample get an estimated maximum likelihood which is too large, precisely because some types were left out.

However, as mentioned above, there are statistical methods which can be used to estimate the unseen types in sample; some of which had already been published before Chomsky published his early work in Generative Grammar, as noted by Pereira (2000, 1242). The technicalities of such methods fall outside the scope of the current discussion (but see section 8.3 for one variant); Gale and Sampson (1995) contains a very readable outline of one way of solving this problem, based on earlier work by Turing and Good. Pereira (2000) applied such estimation techniques to "colorless green ideas sleep furiously" vs. "furiously sleep ideas green colorless" and found that the first (i.e. grammatical) sentence has an estimated probability which is 200 000 times higher than the second, ungrammatical sentence.[2] In brief, using a fairly simple statistical model (based on the probability of neighboring words), it is possible to distinguish between Chomsky's two classical sentences using sample frequencies, rather than by native speaker intuition.

It seems warranted to conclude that Chomsky was mistaken with respect to the usefulness of statistics, both in the late 1950s and today. A wholly different issue is whether the dichotomy between "grammatical" and "ungrammatical" sentences is valid. For recent proposals for dispensing with grammaticality judgments in linguistics, see e.g. Sampson (2007); Stefanowitsch (2007). The next section will discuss the methodological problems inherent to grammaticality judgments in more depth.

Argumentation from limitation of scope

A perhaps more widely held view of statistics in linguistics is shared by among others Sampson (2001) and Talmy (2007). This view entails that, while valid as a means of investigating syntactic phenomena like word order and collocations, statistics has no or only a limited place in the study of semantics. Sampson puts it like this:

> even if some areas of linguistics can be made more scientific than has been usual recently, one might expect to find other areas which cannot be treated scientifically at all. The outstanding example, to my mind, is word meaning.
>
> Sampson (2001, 181)

[2] Pereira (2000) uses newspaper texts as the basis for the estimates, thus the question of linguists discussing Chomsky's examples should not bias the estimates.

Talmy (2007) admits, at least implicitly, that other methods can be used to study meaning, but argues that a non-quantitative method such as introspection is the superior tool.

> Not only is meaning the aspect of language that linguistic introspection is best at, but, in addition, introspection has the advantage over other methodologies in seemingly being the only one able to access it directly.
>
> Talmy (2007, xiii)

The contexts of the two arguments are somewhat different. Chapter 11 of Sampson (2001) aims at rejecting formalist (or more specifically, generative) treatments of semantics in terms of compositional truth-conditional features, whereas Talmy (2007) defends the place of introspection among other methods (such as corpus linguistics or experimental approaches) in cognitive linguistics. The net effect of both lines of argumentation is the same, namely that there are some fields of study in linguistics which are unsuited for quantitative, statistical methods. There are two problems with this view, as pointed out in Geeraerts (2006a), related to theory and method.

As a method, introspection has certain weaknesses, as attested by the fact that grammaticality judgments are not consistent. Hill (1961) conducted experiments where native speakers were asked to judge the grammaticality of isolated sentences, and found that judgments varied considerably. Now, grammaticality judgments and semantic introspection are not necessarily the same thing, but Talmy (2007, xiii–xiv) explicitly mentions easy access to grammaticality judgments as a virtue of introspection as a method, alongside access to meaning. In fairness, it should be mentioned that Talmy also points out that grammaticality judgments do vary considerably across individuals (Talmy, 2007, xiv). However, the main point is this: Talmy defends introspection as a method which can be applied profitably to both semantics and grammaticality, but since we know that introspection about grammaticality is problematic as a measure of norms or consensus, we also have reason to doubt the reliability of introspection applied to meaning.

The argument just outlined concerns the *accuracy* of introspection, that is, how good are intuitions at classifying grammaticality and meaning: native speakers have diverging opinions about what is grammatical or not and what words mean. Another problem concerns the *reliability* of qualitative methods, that is, how good is the method at finding relevant evidence? It is interesting to note the comments by Kroeber and Chrétien regarding their dispute with Meillet over the classification of Germanic and Italo-Celtic. When the qualitative and quantitative analyses diverge, which evidence weighs heavier? Kroeber and Chrétien (1937, 97) suggest that the linguist working only with his intuition easily becomes biased when he

> observes a certain affiliation which is real enough, but perhaps secondary; thereafter he notes mentally every corroborative item, but unconsciously overlooks or weighs more lightly items which point in other directions.

Thus, both when it comes to finding and judging relevant evidence, there seem to be valid reasons to doubt approaches that rely only on introspection and intuition.

It is perhaps not obvious why this constitutes a problem; taken in isolation it may not be a problem at all. However, scientific and scholarly research is, on the whole, *not* conducted in isolation. On the contrary, it is part of one or more fields of research, theoretical schools and disciplines, etc. This implies a communal and cumulative view of research, where the aim is to gradually increase knowledge by testing (i.e. attempting to reject) hypotheses. In this situation, relying exclusively (or mainly) on introspection becomes a problem, since it is no longer possible to choose between competing alternatives in a principled way. Simply put, theoretical linguistics today lacks a standardized method for choosing between alternative hypotheses and theories (Geeraerts, 2006a, 25–26). As Geeraerts (2006a, 43–44) points out, this represents a problem for linguistics in itself, but it also represents a problem for linguistics as a cross-disciplinary endeavor. Relying on a subjectivist, non-objective, method makes it difficult for linguistics to interact with neighboring disciplines such as psychology; see Gibbs (2007) for an elaboration of this argument.

Using corpora as a source of usage-based sentences to analyze introspectively – what Geeraerts (2006a, 36) calls "corpus-illustrated research" – only partially solves the problem. As Geeraerts (2006a, 37) points out, the world of language is rarely black or white. Consequently, we need a way of precisely distinguishing between gradients of language use. As the example in the previous case regarding Chomsky's argument illustrated, we also need to take care of the relationship between the sample and the population and make suitable allowances for what might be left out of the sample. Thus, I take the proper view of statistics in linguistics to be that defended in Geeraerts (2006a): that a crucial step must be taken *after* the introspection has taken place, namely the operational definition of hypotheses followed by empirical testing.

> *advanced corpus research needs to pay specific attention to the operationalization of hypotheses*. If we interpretatively assume that a linguistic entity has this meaning or that function, what can we expect about its observable behavior in actual language use? (Original emphasis)
>
> Geeraerts (2006a, 37)

This is of course a potentially challenging task. However, it is possible, also in semantics. The so-called *distributional hypothesis* operationalizes semantic similarity

between entities A and B as "a function of the similarity of the linguistic contexts in which A and B can appear" (Lenci, 2008, 3). This hypothesis is considered one of the foundations of the corpus analyses in subsequent chapters. Furthermore, together with the definition of grammar from Croft (2000, 26) as a usage-based cognitive structure, the distributional hypothesis makes the grammatical/ungrammatical distinction less relevant. Instead, the emphasis is placed on the occurrence of language in context and in use (Lenci, 2008).

In conclusion, whether statistics is the appropriate way of testing a hypothesis does not depend on which field of research it has been formulated in, but in which way the hypothesis itself was formulated.

Some final remarks

As the preceding discussion has highlighted, there are a number of ways in which to argue that statistics either do not belong in linguistics or only has a limited role to play. The easiest argument against statistics in linguistics both to understand and to refute is the convenience argument by Bloomfield. Today, the data and the statistical tools for processing them are literally at the linguist's fingertips. Although statistics – like all other techniques used in research – must be *learnt*, this is hardly a valid reason for not using it.

The second argument, prominently promoted by Chomsky, viz. that statistical corpus studies are incapable of distinguishing between "grammatical" and "ungrammatical" utterances (i.e. those belonging to G and $\bar{G}$, following Chomsky's terminology) was shown to be incorrect. Once the nature of the problem is understood, it is possible to make reasonable estimates about likely and unlikely utterances in a corpus, only based on frequencies and statistics. Although Chomsky explicitly rejects a probabilistic interpretation of grammaticality based on absence of corpus evidence, it was shown that statistical methods can also be applied to estimate the likelihood of utterances that are not present in a corpus. Furthermore, statistical estimates are explicitly probabilistic, which has the added advantage of allowing for fuzzyness and gradient phenomena.

The third argument, that statistics can be used for some problems, but not for others intuitively appears more reasonable. If word meaning is not fixed or embedded in words, it is not immediately obvious how this can be measured in a corpus. At first glance, this leaves introspection as the only possible method.

However, introspection is an unreliable method, and the problems it raises are immediately apparent when two opposing introspective judgments stand against each other, or when attempts are made to use results based on introspection in linguistics within other fields. Instead, an alternative was outlined based on Geeraerts (2006a).

Introspection does of course belong in linguistics, but not as a means of testing hypotheses. Rather, the subjective interpretations must be operationalized and tested empirically, using appropriate empirical data from e.g. corpora or experiments which are then evaluated using statistical methods.[3] Note that this is not merely a "last resort" to escape the methodological problems concerning introspection. In fact, a large body of research suggests that distributional properties are a good source of information with respect to semantics, cf. the overview in Lenci (2008). Consequently, the proper treatment of empirical data is crucial, as argued in the following section.

3.2.2 Statistics vs. frequency

Let us assume that the case has been sufficiently argued for the use of empirical data in linguistics. After all, the number of linguists who work with corpora or other empirical material rather than relying on introspection alone is substantial, although perhaps not as large as it ought to be. As shown in Geeraerts (2006a, 34–38) and Sampson (2005), it seems likely that introspection is still the dominant method. However, simply reporting raw frequencies or percentages does not constitute an optimal treatment of the empirical material.

The impression from reading corpus linguistic research literature, is that to the extent that numbers and arguments based on frequencies are used, the study is often limited to raw frequencies or percentages based on them. Such a use of numbers is problematic for a number of reasons. Consider Chomsky's classical argument that *I live in New York* is bound to appear more often in a corpus of American English than *I live in Dayton Ohio*, simply because more people are likely to say or write the former than the latter, as quoted in McEnery and Wilson (2001, 10). If we look at raw frequencies, this might very well be the case, but then, there are no good reasons why raw frequencies should be the main scientific fact for corpus linguists. This is what Stefanowitsch (2005, 296) calls *the observed frequency fallacy*:

> Observed frequencies of occurrence represent relevant facts for scientific analysis.

This fallacy of raw frequencies is matched by what Stefanowitsch in the immediately following sentence calls *the expected frequency epiphany*

[3] I admit that this statement may need some exceptions as it stands. Related language-based disciplines such as philology may need to operate in different ways, e.g. due to extremely sparse data, or attempts to understand texts in a cultural context. However, Apollon (1990) argues persuasively for the use of exploratory statistical methods also in philology as an augmentation of qualitative methods.

> Observed frequencies of occurrence must be evaluated against their expected frequencies of occurrence before they become relevant facts for scientific analysis.

As Stefanowitsch points out in the same paragraph, this is of course not exceptional to corpus linguistics. All empirical quantitative academic disciplines base their argumentation not on raw frequencies of observations, but on statistical inferences from those observations. The reasons for this are obvious. How can we tell whether an utterance, a collocation or a word occurs many or few times? This is analogous to asking "how big is a big number?"; the answer must obviously refer to some relevant comparison.

However, as Crawley (2005, 2) points out, "everything varies". Places and times differ, and measurements at different times and places (even on the same individuals, where this is possible) will differ too, even if it is only on a very small scale. For this reason, finding that, say, a word occurs with different raw frequencies in two different corpora is not in itself very interesting. If we accept Crawley's statement, it would be more surprising to get identical frequencies than different ones. The question, then, is to separate the everyday background (or random) variation from the interesting variation which is related to a given research question.

The only tool which can reliably distinguish between random variation and effects worth noting is statistics. The main tendency in a dataset can be distinguished from random variation by comparing the observed frequencies and the expected frequencies (the technical aspects are discussed in chapter 4). This provides a whole range of advantages over working with raw frequencies. Four of these advantages are listed here, although the list is not meant to be exhaustive. First, tendencies in datasets can reliably be distinguished from background variation; second, statistical methods allow us to estimate the magnitude of such a tendency through effect size measures; third, it is possible to estimate the accuracy of the results through confidence intervals; fourth, it is possible to estimate the combined effects of several variables simultaneously in a consistent manner. In contrast, a raw frequency (or a percentage from a small sample) provides no reliable way of estimating the size of an effect or the accuracy of the estimate. Even detecting the trend can be difficult, especially with large and complex datasets.

Thus, raw frequencies and percentages based on them are of limited usefulness in empirical research. Instead, it is necessary to compare raw frequencies with what would be expected under a null hypothesis (i.e. do statistical testing), look at the *size* of any statistically significant effect, and estimate the accuracy of the results (e.g. by using confidence intervals). Only then can we begin to make inferences from the sample to the population. The next section discusses these two crucial concepts.

3.3 Population and sample

"Sample" and "population" in corpus linguistics are of course never given entities. Rather, they are the constructs of the corpus linguist, as pointed out in e.g. Halliday (1992) and Rissanen (1992). Consequently, it becomes necessary to carefully define the sample and the population which results will be generalized to.

Gorard (2003, 56–57) asks the interesting question why we would want to use a sample at all. He argues that in general, it is often preferable to work with data from the whole population (i.e. a complete dataset), rather than to introduce additional error and bias through a sampling procedure (Gorard, 2003, 57). Sometimes such an approach is possible, e.g. when working with a very small population. There are also good reasons to use a sample, though: in many cases the population will be so large as to make a complete dataset impossible (a good example here would be linguistics: how to capture *all* utterances in a language?). Also, many statistical tests are explicitly based on sampling theory, as noted by Gorard (2003, 57). Doing e.g. a Pearson chi-square test on a complete dataset raises some interesting questions regarding the validity of the test, since the underlying logic of the test is based on the appropriateness of the χ^2 distribution as a model for the unknown population being generalized to. If the complete population is known, and if this population does *not* follow the χ^2 distribution, then it is not clear what exactly is demonstrated by using the test.[4]

Two key properties of samples that loom large in introductions to the topic are *randomness* and *size*, cf. e.g. pp. 48–51 and 103–109 of Hinton (2004). To this can also be added *independence of observations*. Below, these three properties are discussed in detail, in relation to corpus linguistic data.

3.3.1 Randomness

Almost regardless of which test statistical test or model is used, there is usually an underlying assumption of random sampling from a population. A random sample is by definition a sample in which every member of the population has equal probability of being included in the sample. This technical definition is different from what is often meant by "random" in everyday language. As Kilgarriff (2005, 264–265) points out, "random" is not "arbitrary"; in the real world truly random events (in the mathematical sense) are exceedingly rare.

Since the present study deals with historical data, which can be considered both

[4] In such a case, there are a number of alternatives available, such as Fisher's exact test for null hypothesis testing and correspondence analysis for a more exploratory approach. This is discussed in more depth in chapter 4.

sparse and skewed,[5] the question of sampling is obviously something which requires careful consideration. It is doubtful whether the York-Toronto and Penn-Helsinki corpora can be considered random selections of language use in previous times. It was mentioned in chapter 1 that a number of historical accidents played a part in determining what kind of material we have available today from e.g. Old English. The arbitrariness which caused some of the *Cotton* manuscripts to survive and others to perish in 1731 makes it more difficult to assess variation, since it reduced the size of the sample. It did not constitute some kind of "random" selection mechanism, in the sense that every manuscript had an equal probability of being preserved (see section 1.3 above). Furthermore, it is also doubtful whether we can consider the sentences and words in the manuscripts (and by extension, the corpora) as a random sample of everyday language in Medieval England. How, then, is it possible to justify the use of statistical methods?

Baayen (2001, 2008a) and Evert (2007) discuss the randomness requirement from the perspective of word (n-gram) studies. Both Baayen and Evert argue that although the words in the corpus are obviously not a random sample, we are nevertheless justified in assuming that this is the case for the purposes of statistical testing and modeling.

According to Evert (2007, 180), we can think of the language as a vast library, which although it is not random in itself, can be approached with statistical models assuming randomness, because selecting

> a particular corpus as the basis for study – among all the other language fragments that could also have been used – is like picking an arbitrary book from one of the shelves in the library. It is this choice which introduces an element of randomness into corpus frequency data.

According to Evert, language is thus not *inherently* random, but singling out a small subset of the language for study based on what happened to be included in a corpus introduces an element of random selection. However, this does not completely solve the problem.

Evert (2007, 189) also remarks that the discrepancy between the unit of investigation (words, sentences) and units of sampling (entire documents or large fragments) may distort the picture, since many linguistic phenomena tend to cluster together. This is what Church (2000) discusses as the case of "the two Noriegas". Church shows that the first mentioning of a content word depends strongly on the number of documents it occurs in, whereas the second mentioning is independent of frequency (Church, 2000, 183), but is dependent on previous occurrence; that is, having seen an infrequent word

[5] In effect, the historical corpus material must be considered an opportunity sample from a statistical point of view.

in a document, it would be surprising *not* to spot the word again in the same document. That is, even if we were to sample truly randomly at the level of documents, we would expect to find systematic dependencies between word occurrences internally in the documents. The strength of the effect must obviously be taken into account given the research questions being investigated, since Church notices that the effect is stronger for content words than for function words (Church, 2000, 186).

Kilgarriff (2005) argues against the use of statistical testing on corpus data, for two reasons. First, statistical null hypothesis tests rest on the assumed plausibility of a null hypothesis of randomness, which is doubtful in the case the of linguistics. Second, since all hypothesis tests have a known bias towards significance given large amounts of data, it is almost always possible to find significant p-values given modern corpora. Gries (2005), while acknowledging Kilgarriff's methodological point, defends statistical testing. Crucially, Gries (2005) conducts a number of experiments which show that once the proper corrective steps are taken (post-hoc correction for multiple testing, effect size measures), then simple null hypothesis tests such as chi-square tests give reliable results. Chapter 4 will return to this question in more detail, suffice it to say that the *technical* aspect of Kilgarriff's argument can, to some extent, be met.

Another question is how meaningful and interesting it is to blindly apply null hypothesis tests e.g. to compare frequencies of words in two different corpora. Such a simple hypothesis is probably vastly underspecified, and a more prudent approach would be to develop and operationalize the hypotheses further before starting the statistical testing. Thus, the *methodological* aspect of Kilgarriff's argument is related to the discussion above regarding operationalization of hypotheses in corpus linguistics as advocated by Geeraerts (2006a).

In summary, it has been shown that although the corpora do not constitute a true random sample in the strict sense, there are still good reasons to consider the assumptions of null hypotheses tests fulfilled. This holds in particular when the proper technical corrections are applied, to avoid spuriously significant results. Furthermore, there is a methodological aspect to statistical testing on corpus data, as well as a technical one. Making the operational definitions of the hypotheses more precise and specific ensures that statistical tests can be applied in a more systematic and targeted way, rather as "exploratory" devices for comparing frequencies. Finally, as a more practical consideration: since the material in the corpora was never selected (or de-selected) and collected with *there* in mind, there is no reason to believe that there exists a *systematic* bias in the material which should affect the behavior of *there* and the Existential Construction. Furthermore, *there*, both as a deictic adverb and existential pronoun, is probably closer to function words than fully lexical words, thus avoiding the full force of the "Noriega effect" discussed in Church (2000).

3.3.2 Size

As stated above, it is generally agreed upon in statistics that in order to claim representability for the sample it must be random. But random sampling is not in itself a guarantee that the sample will accurately represent the population. Size is also crucial, something which is often overlooked. The problem with small samples is what Tversky and Kahneman (1971) call "the law of small numbers". This is a logical fallacy which proceeds as follows:

> *The fallacy of the law of small numbers:* The law of large numbers guarantees that a very large sample will be representative of the population it is drawn from; therefore this is true of small samples as well.

The first part is of course true, a large sample will tend to be fairly representative because it is large enough to capture a reasonable amount of the actual population variation. The fallacy involves the belief that "the law of large numbers applies to small numbers as well" (Tversky and Kahneman, 1971, 106).

As Tversky and Kahneman (1971) point out, this is merely a variant of the classical gambler's fallacy, that is, a vulgarized version of the law of large numbers: since with a fair coin the number of heads and tails will be approximately the same when the number of tosses approach infinity, this law will also hold over, say, the next ten tosses as well. The problem with small samples arises first and foremost when studying a phenomenon whose "magnitude is small relative to uncontrolled variability, that is, the signal-to-noise ratio ... is low" (Tversky and Kahneman, 1971, 106).

This seems to be a reasonably good description of linguistic corpus data. It is well known that many linguistic phenomena follow a distribution with a small number of items occurring frequently, accompanied by a large number of rare events, cf. Baayen (2001). Under such circumstances, it becomes necessary to question the signal-to-noise ratio, that is, the natural sampling variation (which is sometimes brushed aside as "performance error"). Careful consideration of the sample variation in relation to the sample size is absolutely necessary if any kind of statistical testing is to have explanatory adequacy. Unless the sample is large enough to plausibly capture the variation in the population, any hypothesis-testing procedure with accompanying explanations are, as Tversky and Kahneman (1971, 108) point out, likely to be "an exercise in explaining noise". Thus, random sampling is no guarantee that the sample is a good approximation to the population if that sample is too small; despite its randomness, the sample *might* be highly misleading if it is too small.

Thus, while random sampling is desirable, so are large samples. It is not obvious what makes a sample "large", but for the present investigation the limitations are set by the amount of material which is available. The three largest existing corpora for Old,

Middle, and Early Modern English prose are used in their entirety, which amounts to about 4.5 million words. Although this is small compared with some synchronic corpora, it is not possible to extend the sample size further within the scope of a PhD-project (and not possible at all in the case of Old English where all extant prose is included in the corpus). Consequently, the stance adopted here is one of pragmatism where the sample is considered large enough, until empirical results prove otherwise.

3.3.3 Independence

Finally, it is necessary to briefly consider the assumption of *independence*. Most statistical tests and models assume that data points are independent of each other, that is, the same "individual" (as defined for a given study) should not contribute to several data points simultaneously.

As Crawley (2005, 13) states: "Repeated measures through time on the same individual will have non-independent errors because peculiarities of the individual will be reflected in all of the measurements made on it". In other words, the measures are correlated with each other. Such non-independent errors lead to *pseudoreplication*, which might again lead to spuriously significant results (Crawley, 2005, 14).

In historical linguistics, data are typically analyzed as if they were independent, although they are in fact part of a time series. One example is the source of data, viz. documents, where peculiarities from one manuscript might be repeated in later ones through exposure or copying. Linguistic constructions can also be considered temporally non-independent, since they are acquired on the basis of previous examples.

There are various ways of dealing with non-independent errors and pseudoreplication. In chapter 4 below two such ways, mixed effects models and correspondence analysis, will be presented.

3.4 Explanations in linguistics

Having discussed the place of statistics in linguistics, it is necessary to consider under which circumstances it will be possible to admit the statistical results as evidence regarding the population in question. Furthermore, the question of what might constitute an adequate explanation in diachronic linguistics will be discussed. Finally, some thought will be given to the notion of causation, before an outline of the methodological foundations of the present study is presented.

Not everyone have been equally optimistic regarding the possibility of finding explanations in historical syntax. It is interesting to note the comment made by Hirt (1934, vi):

> der Grund für die mangelnde Teilnahme an der Syntax liegt ... m.E. darin, daß wir auf dem Gebiet der Syntax in vielen Fällen keine Erklärungen finden. Man stellt eine Reihe von Tatsachen zusammen, man weiß aber nichts damit anzufangen. ... Es gibt keine Antwort und das ist unbefriedigend.
> (the reason for the lack of treatment of syntax lies ... in my opinion in that we in the area of syntax in many cases cannot find any explanations. A number of facts are assembled, but one does not know what to do with them ... There is no answer and that is unsatisfactory).

What constitutes acceptable evidence and explanations in linguistics is theory dependent. I agree with Fischer (2007b, 251–252) that theory provides both a basis for, and a necessary guideline to, empirical research in historical linguistics. Consequently, it is warranted to briefly consider what a "theory" is. Following McCawley (1982, 2), I take a theory to be:

(1) an ontology combined with a conception of what propositions are meaningful and what their relationship to possible facts is.

In other words, a "theory" is defined as a coherent, metaphysical framework, which relates ontological considerations to models ("meaningful propositions") and hypotheses ("relationships to possible facts"). As mentioned in chapter 1, the theoretical framework of the present study is RCG and the evolutionary theory of language change described in Croft (2000). This comprises the metaphysical framework and some of the models for the study. In chapter 2 a number of hypotheses were presented, but to fully integrate these hypotheses into the theoretical foundations of the study, it is necessary to justify their explanatory potential. That is, testing the hypotheses empirically does not amount to an explanation before the hypotheses can be explicitly linked to the theory.

Thus, the acceptability of a given explanation must rest on two pillars, first its empirical justification and second on its coherence and theoretical validity. Below, I will attempt to justify on theoretical grounds the prerequisites for positing an empirical causal explanation in historical syntax.

3.4.1 Evidence and explanations

We can broadly distinguish two main types of explanations in diachronic linguistics: child-based and functional explanations. This is of course a simplification. Previously, there have been attempts at constructing law-based explanations in diachronic

linguistics, such as the Neogrammarian sound laws, cf. Campbell (2004); Lightfoot (2006, 26–36); McMahon (1994, 17–24). As McMahon (1994, 21) notes, the principal Neogrammarian explanation for sound change was mechanistic and physiological, an approach which did not translate well into syntax (McMahon, 1994, 107). Alternatively, structuralist explanation tends to view language as having a function within a system where distinctions need to be preserved, cf. McMahon (1994, 24–32). However, what the system actually represents is quite abstract, cf. Saussure (1983), and it leaves the speaker with a minimal role to play in language change. Rather, the individual becomes subservient to changes in an abstract system to which he or she must adhere without much active participation. This promotes what Lightfoot (2006, 37) calls an external view of language, where languages "were seen as external objects floating smoothly through time and space". Chomsky, from early publications such as Chomsky (2002) or Chomsky (1959) and later, argued that the individual's psychology also needs to be taken into account. It seems to me that today most formal and functional explanations in linguistics make use of an active individual within a system (with various degrees of psychological commitments), cf. also the comments in Fischer (2007a, 57–58). Thus, the Neogrammarian and Saussurean types of explanations will not be pursued further. Instead, the two types alluded to initially will be considered in detail, before an alternative is discussed:

- Child-based: The child as language learner has an initial state (or "Universal Grammar" (UG)). Deriving descriptive generalizations about language from the principles of this initial state constitutes an explanation of the generalization.

- Functional: Language is used to serve communicative purposes such as to reach objectives and/or to reduce ambiguity in communication. Stating the usefulness/-functional purpose of a descriptive generalization about language constitutes an explanation of the generalization.

Both of these two types of explanation are somewhat problematic, as the following discussion will reveal, although I agree with the comment made in Fischer (2007a, 82–83) that both approaches can be useful as heuristic devices.

Child-based explanations

As argued in Croft (2000, 44–53), child-based parameter setting theories are poor frameworks for linguistic explanations. First, in synchronic linguistics the so-called "poverty of stimulus argument" (see further discussion below) only holds in principle under very specific conditions, and remains to be documented empirically. Of more

concern here, though, is the diachronic case, where Lightfoot is a major proponent of child-based explanations, cf. e.g. Lightfoot (2006, 2007).

Child-based parameter setting diachronic explanations are not plausible for a number of reasons. Croft (2000, 50–53) discusses the empirical problems brought about by having the grammar (or grammatical competence, also known as "Internal language" or I-language) of the child acting as the agent of change. These problems include uniformity of use assumptions and the problems posed by gradual change spanning long periods. Later versions of child-based explanations have been modified somewhat, so that "cues" in the linguistic environment (i.e. in the "External language" or E-language, otherwise known as "performance") need to be activated at a certain "threshold" for children to notice them and base their grammar-learning on them. Similarly, this gradual "activation" and de-activation of cues can in itself be gradual before an abrupt change takes place, which supposedly explains "why children at a certain point converged on a different grammar" (Lightfoot, 2006, 100), see also the discussion in Fischer (2007a, 104–115). To be able to posit both gradual and abrupt change at the same time, cf. (Lightfoot, 2006, 135), the child-based parameter setting approach needs to separate grammars from their use. However, this argument hinges on the plausibility of a cognitive (or even biological) grammar which can be separated from use and possibly also from general cognitive abilities, an assumption which is highly problematic.

According to the theory of an innate proto-grammar or "Universal Grammar" (UG), cf. Chomsky (1986), Hauser, Chomsky, and Fitch (2002) or **?**, UG describes an initial state of the language learner, before being exposed to language-specific input. As such, this initial state described by UG is taken to be a "species characteristic, common to all humans" (Chomsky, 1986, 18). It would take too much space to go into the details of this line of argumentation. However, a few brief points need to be made.

The nativist views espoused by Chomsky have come about as his suggested solution to the so-called "logical problem of language acquisition", i.e. that children seem to learn language amazingly fast and with little (and, according to Chomsky, degenerate) input. As chapter eight of Cowie (1999) explains, this argument is really two arguments: an empirical one stating that the nature of linguistic exposure is such that children do not get the proper input to generalize a grammar from; and a logical one, stating that children logically cannot learn a language/grammar from contextual input. As the literature reviewed in Cowie (1999) shows, the empirical evidence is still not conclusive, but on the whole it does seem plausible that children do find relevant input in their linguistic surroundings. As a consequence, its proponents have placed more emphasis on the *logical* variant of the statement, i.e. that children cannot construct a grammar inductively no matter what the empirical input might be. Initial support from this came from Gold (1967), who proved – under certain restricted and idealized conditions – the limits for learning a formal language. Hauser et al. (2002, 1577) still refer

to Gold's paper and state that the chance of finding domain independent general learning mechanisms which can acquire language "seems rather dim". However, Horning (1969) effectively showed that in the presence of noise (i.e. Chomsky's degenerate input), a probabilistic Bayesian approach (see chapter 4) could adequately handle noise albeit at a higher computational cost. The main point, however, is that the formal proof provided by Gold only holds under very specific circumstances. When slightly different premises are laid down, Horning (1969) provides a formal proof that inductive learning is possible, despite noisy input.

Furthermore, as noted by Stewart and Cohen (1997, 250–253), the nativist arguments of Chomsky run into biological problems as well. More specifically, they reduce to "preformationism", that is, the idea "that every phenomenon must already exist, in rudimentary but potentially complete form, in a precursor" (Stewart and Cohen, 1997, 250), see also e.g. Cowie (1999, 41). There is currently no known evidence of a biological, or psychological "Language Acquisition Device" as proposed by Chomsky, cf. Chomsky (1986). Pullum (2009, 17) dryly observes that the current Minimalist Program proclaims to concern itself with "biolinguistics", but somehow manages to "live with the fact that the real biologists and neurophysiologists are not getting involved".

Thus, the so-called "logical problem of language acquisition" does not provide evidence for the existence of such an identifiable biological grammar separated from use, neither formally nor empirically. Appeals to grammar as identifiable individual "competence" distinct from "performance" also quickly runs into problems. Hill (1961) showed how grammaticality judgments vary considerably, which means that a description of a "grammar" based on the linguist's own introspections is trivial in the sense that it cannot describe more than a single such competence-grammar (or the linguist's perception of it). But it gets worse. Even if there would be some overlap (or even much overlap) in grammaticality judgments, we would in fact need one "grammar" for every speaker of a language, because the grammaticality judgments would almost certainly differ at some point. We would then have to either accept that speakers' judgments vary (hence they have slightly different grammars), or we would arbitrarily have to assign the disputed phenomena to a "peripheral" role. Neither seems particularly attractive, hence the need for individual grammars, which amounts to building a map of the world on a one to one scale. Clearly, this is hardly a basis for making claims about what separates "competence" from "performance", let alone specifying psychological or even biological components of such "competence".

Chomsky has repeatedly asserted that he is interested in the grammar of an *idealized* "speaker-listener" in a homogenous speech community, cf. e.g. Chomsky (1965, 3); Chomsky (1986, 17). However, it is not clear exactly what identifies the idealized speaker. In Chomsky (1961), in a comment on Hill (1961), Chomsky claims that if tests show that native speakers of English cannot distinguish between his examples (such as

"colorless green ideas sleep furiously" vs. "furiously sleep ideas green colorless"), then the wrong test has been chosen (Chomsky, 1961, 227). In other words, only tests that *support* Chomsky's definition of the grammar of an idealized speaker are admissible as evidence. This is made even more explicit in footnote 17, (Chomsky, 1961, 227), where Chomsky explicitly states (with more than a hint of sarcasm) that those who do not agree with his grammaticality judgments are simply wrong and not likely to enjoy much success in studying syntax. Thus, it seems that we cannot but interpret the phrase "the grammar of the idealized speaker / hearer" as meaning anything other than "the linguist's / Chomsky's intuitive interpretation of the grammar of the idealized speaker / hearer". Since this interpretation is still based on the linguist's introspective judgments (shown to be unreliable), the inevitable conclusion must be that appeals to idealizations have no merit as a foundation for one of the central claims of child-based explanations of language change. A further, more dramatic step would be to claim that the idealization has only a "theoretical" status. It is not clear to me, however, how this would serve to make linguistics an empirical discipline with transparent standards for explanations and evidence capable of interacting with neighboring disciplines.

In summary, it would appear that there is no real independent evidence for an internal grammar or linguistic competence which on principled grounds can be separated from language use, either synchronically or diachronically. The fact that individual judgments of grammaticality are highly variable[6] suggests that use and competence are not easily – or fruitfully – separated as linguistic phenomena. Furthermore, claims about first language acquisition – both empirical and logical – which are meant to support a division between an internal psychological language and an external usage language are not adequately supported. Thus, there is no good reason to separate language use from language competence. Consequently, child-based parameter setting explanations of language change are left without their supporting evidence. A move to save the situation with appeals to "idealizations" and "models of grammar" is conceivable. Under such a view the explanation would be a "theoretical" explanation of what happens in an idealized grammar, but this quickly renders the proposed "explanation" so schematic and theoretical that it is no longer clear what it is supposed to explain. As Fischer (2007b, 270) points out, grammars are theoretical *constructs*, i.e. they are inferred or constructed for some purpose, but there is little evidence that they exist as more than idealized models, something which leaves them as poor candidates for explanations in linguistics. Instead, the grammar (or model) itself is in need of an

[6] As mentioned above, there are of course also many points of agreement. This is not the point. If the Chomskyan argumentation holds, it is not clear why we would find any divergences in grammaticality judgments *at all*, since competence is supposed to be distinct from the actual posited source of variation, namely usage. In light of this, it would seem that a prototype-based organization in terms of central and peripheral usages or constructions would capture the observed facts much better.

explanation.

Functional explanations

Functional explanations, although intuitively appealing have their own problems. According to Hovorka et al. (2002, 174), a functional explanation is "provided by the end state or goals of a phenomenon". That is, to describe the functional communicative outcome of a change is to explain the change. This makes it possible to posit a functional linguistic explanation in terms of communicative function or grammatical simplification, cf. Mithun (2003), or social factors such as group identification, cf. chapter seven of Croft (2000). Aitchison (1991, 117–118) mentions functional explanations as an expression of speakers' "needs", which certainly seems plausible with respect to new words for new technical or social phenomena, but it is not clear how this can be successfully applied to syntax. An example highlights this problem. Aitchison (1991, 119–120) discusses the case of multiple negation as a result of a desire for emphasis and vividness, i.e. a functional communicative explanation for syntactic change with such examples as *It ain't no cat can't get in no coop*.

On the face of it, the account in Aitchison (1991, 119) seems reasonable. Assuming a simple negated proposition as starting point, it can be made more "emphatic" or "vivid" by adding more negatives. Then, in the course of time, this heaping up of negations becomes the standard, and more negatives need to be added, etc. Note that there are several unstated empirical assumptions in this line of argumentation:

- Assumption: language processing is based on long term frequencies of use.
- Assumption: "emphaticness" and "vividness" are psychological entities, experienced by language users.
- Assumption: long term / high frequency use of an emphatic element will cause the emphatic effect to diminish.
- Assumption: the perceived loss of an emphatic function will spur a functional need for a new element to fill the same role.

I do not want to argue that these assumptions necessarily are wrong, but merely that they are left implicit by the functional argumentation. The problem with them is that they are bundled together and taken for granted, when they in fact require specific empirical justification. However, if the chain of causality whereby communicative needs spur syntactic change breaks down, then the functional explanation is left with a problem.

What is lacking are the specific *mechanisms* which explicitly cause or generate these effects. *How* does long term use affect vividness? And how do speakers perceive the "loss" of a distinction? This is discussed in detail by Lass (1997), on which the following account is based.

Lass (1997, 359–360), in his discussion of functional explanations of language change, offers some possible causal mechanisms for how the speakers maintain the functionality of the language system when facing the potential loss of a distinction.

(2) *Precognition*: speakers anticipate the potential problems caused by the loss of a distinction and take corrective action before the distinctions break down.

However, as Lass (1997, 359) points out, this raises the interesting question of how the speakers are able to muster such insight, since linguistic change typically proceeds by cumulative variation over a long time span, possibly outside the life span of any one single speaker. This is the problem alluded to by Johansson (1997, 169), when he states that "it is impossible to explain the arrival of a distinction from its function since that function cannot exist until the distinction is made, unless the future determines the past". If we dismiss the plausibility of language users with prophetic foresight, some other mechanism is needed, as suggested by Lass:

(3) *Repair*: speakers allow system-changes to take place; then, when a distinction has broken down, some corrective measure is taken to re-institute a functional distinction.

But as Lass points out, this explanation immediately begs the question of the necessity of such a functional explanation in the first place. In short, a change takes place and goes to its completion, which causes a breakdown of (presumably) functional differences. It is reasonable to assume that this takes some time. The members of the speech community have lost a distinction, whose motivation was communicative function, which presumably leaves them with a dysfunctional language system. The speakers are now required to realize their mistake, and start instituting a new change to plug the functional gap. However, assuming again language change to be a slow process, the members of the speech community now face a long (and potentially confusing) period wherein they must try to restore full functionality to their language. As Lass (1997, 360) argues, such an explanatory mechanism does not appear consonant with linguistic data. As an illustration, he gives the example of the loss of the second person singular – plural distinction in English, which has subsequently been restored in some dialects (*you* vs. *yous* or *ya'll*). There is scarce evidence that the dialects where his distinction has not been reintroduced suffer from a poorer communicative system than the ones that have reintroduced it.

For several generations of speakers to manage without such a distinction before it is reintroduced in a new form hardly suggests a pressing functional motivation. It is simply not very plausible that a presumably functional (why else would it be reintroduced?) distinction can disappear and leave the speakers with a dysfunctional language until they agree on the new form this distinction should take. Unless, as Lass (1997, 360) points out, we are prepared to accept that the functional loss is felt as gradually more and more problematic by each generation. This, however, casts some doubt about the functional underpinnings of the explanation, since the plausibility of the functional explanation is based on the perceived needs of the speakers which (at least at some fundamental level) must be based on some common psychological or cognitive traits. In other words, what is a felt as a necessary distinction for speakers at time t should be perceived as equally necessary at time $t + x$ for any value of x.

The argumentation outlined here is broadly similar to that of McMahon (1994, 146) when she criticizes the explanations put forth by Vennemann regarding word-order change: if there is a typological (functional) pressure for languages to stay consistent so that initial variation in SOV order causes SVO order via an intermediate topicalization or TVX stage, then why not simply fix the preferred order at SOV? Vennemann's argument would imply that the strong functional pressure is invoked only when needed, i.e. on a purely *ad-hoc* basis.

To sum up, functional explanations, while somewhat more appealing than child-based explanations, do carry with them a set of problems. First, as pointed out by Keller (1994), actions and intentions are easily confused. Second, the proposed mechanisms by which the functional pressures acts on language is not clear (except perhaps through intentional action, which leads us back to the first objection). And third, functional explanations are *ad-hoc* in the sense that a functional explanation for a given phenomenon can usually be invoked in most cases to account for a proposed change. However, it is not not obvious how it helps us distinguish between different proposed explanations (if we assume that one of them is the right one), cf. the desired properties of quantitative corpus linguistics set out by Geeraerts (2006a). In the next section, an alternative will be outlined.

Evolutionary explanation

An alternative is to take an evolutionary approach. A number of works proposing such an approach to language have been published, cf. Croft (2000); Keller (1994); Johansson (1997); Pagel, Atkinson, and Meade (2007); Lieberman, Michel, Jackson, Tang, and Nowak (2007); Fitch (2007). Perhaps one of the more developed frameworks is that presented in Croft (2000), which is an evolutionary explanation based on utterance selection. The presentation below is largely based on Croft (2000). As Croft (2000,

20–25) acknowledges through his citations, much of this is taken from other works on biological evolution or generalized selection. However, Croft (2000) gives a clear overview of the issues involved and explicitly relates them to language change in an evolutionary perspective. Of these issues, there are two which are fundamental in any evolutionary theory: selection and reproduction.

Most of us are probably familiar with evolution in biology, particularly DNA replication, however its application in linguistics may not be obvious. Croft (2000, 10) puts it as follows:

> Evolution is recognized as a process that occurs with certain types of entities. The process is probably best understood as it occurs with populations of biological organisms; that is evolutionary biology. The hypothesis is that language change is an example of the same process, or a similar process, occurring with a different type of entity, namely language.

Clearly, under such a hypothesis it is crucial to determine the properties that allow for this conceptualization. In other words, does language have the same key properties as biological organisms? The answer, according to Croft (and others, cf. the references at beginning of this section), is yes, for a certain conceptualization of what language is. But let us first consider what constitutes evolution.

A general characterization of evolution is that it consists of two steps: replication of individuals and selection of individuals through interaction with their environment (Croft, 2000, 22). As Croft (2000, 23) notes, the replication should leave the replicated structure largely intact, but should also allow for alternations in that structure, i.e. both normal and altered replication. Put differently, once we have reproduction with variation (i.e. imperfect copying or imitation) and environmental attrition, then evolution follows.

Can these characteristics be recognized in language? Croft (2000, 26) argues that a *language* is a population of utterances in a speech community, stressing that what constitutes the population are actually occurring utterances, rather than "possible" ones. An *utterance* is a spatiotemporally bounded occurrence of human communicative behavior, as it occurs in its context, either in speech or writing (Croft, 2000, 26). Furthermore, he argues that a *grammar* is the cognitive structure which contains the speaker's knowledge of the language (i.e. the population of utterances), based on the subpopulation of the language which a particular speaker has been exposed to. That is, the grammar is an acquired psychological entity, shaped by previous exposure to language as well as general cognitive capacities involved in producing, reproducing, and comprehending utterances (Croft, 2000, 27).

In short, language can be considered a kind of cultural evolutionary system. As such, it is, as pointed out by Gould (1978), "Lamarckian" rather than "naturalistic" or

"Darwinian", since inheritance happens through successive acquisition (rather than by DNA). It is possible to subsume all forms of evolution (both cultural and naturalistic) under a general theory of selection. Croft (2000, 22) discusses four key terms that constitute the working parts of such a general theory of selection:

(i) "replicator": an entity that passes on its structure in successive replication, i.e. what is being replicated;

(ii) "interactor": an entity that interacts as a whole with its environment in a way which causes replication with variation;

(iii) "selection": a process which causes differential extinction and perpetuation of the replicator, i.e. some form of environmental attrition;

(iv) "lineage": an entity which persists through time (in the same or altered state) because of replication.

The replicator in the case of language is a structured spatiotemporally bounded individual which Croft (2000, 28) refers to as a "lingueme", to distinguish types from tokens. Thus, the replicator is the lingueme, embodied in an utterance (or conversely, an utterance is composed of one or more linguemes). Replication is caused by means of the interactor, i.e. the speaker, who produces utterances. For various reasons (including social expectations, linguistic and cognitive processing factors etc.), this replication will tend to vary, that is, speakers ensure replication with variation or altered replication. The selection phase occurs when circumstances (be they functional, social, or psychological) arise under which linguemes are either replicated or not, i.e. the continuation or discontinuation of conventions. Finally, a lineage is a "summary of all replications" of a replicator over a period of time (Croft, 2000, 32).

The scenario is not unrealistic. In a study of lexical change Pagel, Atkinson, and Meade (2007) used corpus frequencies and statistical estimation techniques to predict rates of lexical evolution. Pagel et al. (2007, 719) argue that

> humans are capable of producing a culturally transmitted replicator that, perhaps because of the purifying force of spoken word frequency, can have a replication accuracy as high as that of some genes.

The culturally transmitted replicator in this case is a word, i.e. a linguistic sign. But given the fundamental assumptions of RCG that all linguistic units are constructions, there should be no principled distinction between a one-word construction and a multi-word construction – both are pairings of form and meaning.

Replication and selection

For the present study, the lineage (or lineages) being studied is that of *there*. In chapters 8, 9, and 10 I present an investigation of the structure of and variation among the utterances. To attempt an explanation of how *there* could split into two lineages (cf. 11) it is necessary to consider how the replicator could have been modified through selection. The question of selection is discussed in more depth below.

It is crucial to keep in mind the proper workings of an evolutionary selection mechanism. In particular, we need to guard against the fallacy that an evolutionary selection mechanism *actively* selects the best suited individuals for reproduction. This is, of course, not the case. Gould (1993, 146) summarizes Darwin's key idea that *natural selection*:

> builds adaptation negatively—by eliminating all creatures that do not vary fortuitously in a favored direction, and preserving but a tiny fraction to pass their lucky legacy into future generations.

The mechanism is consequently not one of finding the optimal system, but of removing what is dysfunctional or poorly adapted. An important task, then, is to identify the mechanism(s) by which this selection takes place.

To properly explain the linguistic selection process, some kind of causal mechanism is needed. As Croft (2007) points out, the selection process in the case of language is of course mediated, not self-replicating. Croft (2007, 141) stresses the social-functional aspect of replication when he writes that "speakers' interactions with their environment – what is to be communicated and above all who they are speaking to – causes selection of linguistic structures in utterances". Although not without merit, it seems to place too heavy a burden on social interaction and communicative goals to let them serve as selectional mechanisms in all cases of linguistic replication. As discussed above, functional explanations must rely on combinations of arbitrariness and speaker intentions to bring about distinctions and variation in the replicator. However, there is sufficient evidence around suggesting that linguistic and cognitive processing should also be considered as serious selectional mechanisms, cf. Aitchison (2003); Hawkins (2004).

Such potential selectional mechanisms can be found among general cognitive capacities that are regularly invoked in cognitive linguistics. Examples include memory, perception, attention and categorization, cf. Croft and Cruse (2004, 3). An immediate advantage of positing such psychological phenomena as causal selective mechanisms is domain independence and testability. The effects and properties of memory, attention etc. can be tested and evaluated independently of their effects in language, but it is also possible to test language specific effects. Conversely, communicative goals are

much harder to attest and investigate independently of the linguistic effects they are supposed to explain.

A possible problem with such a psychological perspective is that it might entail what Aitchison (2003, 741) calls the "one-off nature of various interesting speculations". In other words, is this a real explanation or simply what Cowie (1999, 39) calls "blatant ad hockery"? Without any further modifications, a psychological explanation of diachronic phenomena could easily take an invalid logical form, known as "affirming the consequent":

1. if P, then Q.
2. Q.
3. Therefore P.

This is logically invalid, since there could be other conditions under which Q is true (a valid form is found when the "if" in 1. is substituted with "if and only if", i.e. P is the *only* condition under which Q is true.). This could e.g. take the form of "if processing cost affects the use of construction x, construction x will decline in a diachronic corpus. x declines in a diachronic corpus, therefore it is affected by the processing cost". However, there could easily be a number of *other* reasons why x would decrease in the corpus, including imperfect sampling, changing literary styles, genre differences, dialectal differences or language contact. (Note that this counter-argument does not affect psychological mechanisms only, it applies equally to any of the other historical or social mechanisms discussed so far.)

Merely finding an association in some corpus and positing a plausible psychological explanation for it is not sufficient for making deterministic causality claims. Thus, although we can take all manner of methodological freedom in re-naming correlation, it is not obvious that statistics can provide easy access to causality on its own, as discussed in Goldthorpe (2001) and Gorard (2003).[7] Rather, what is needed is some sort of non-deterministic notion of causality.

Clearly, to properly motivate an evolutionary explanation in linguistics, we need some kind of conception about *causation*. This topic will be dealt with further below.

3.4.2 Causation

It is perhaps not obvious that "causation" has any place in diachronic linguistics at all. In other disciplines, things are perhaps a little less controversial. Gorard (2003, 158)

[7] At least outside the context of experiments proper, but see discussions in Goldthorpe (2001) and chapter seven of Gorard (2003).

gives the example of smoking and lung cancer. Few today would object to the proposition "smoking causes lung cancer". At the group level, we can observe a consistent and stable correlation between smoking and lung cancer. For this to amount to a convincing explanation, it is necessary to add a plausible *cause* which *generates* the lung cancer. In the case of smoking and lung cancer, this has been done in the form of experiments which have isolated carcinogens in the tobacco smoke, pathological studies of lung tissue etc. However, what could plausibly amount to a "cause" in linguistics? Can something really "cause" language change in a strong sense of the word?

The problem lies in the deterministic nature of the deductive-nomological (D-N) model under which the hypothetical causal mechanism is supposed to operate. Itkonen (1981, 695) rejects law based D-N models in diachronic linguistics and instead insists on "pattern explanation" models, where "the explanatory coherence obtains between the circumstances (including other actions performed by the agent), the ends, and the means". However, there is a problem with this view. As Gould (1978) points out, the fact that an explanation is coherent does not necessarily imply that it is true. It is possible to arrive at a coherent and consistent stories which can plausibly be related to variables such as those mentioned by Itkonen, but without some kind of empirical test we have no way of choosing between proposed explanations since all of them could equally well be true, which is unsatisfactory.

There is a more problematic aspect of Itkonen's position, though. The statements from Itkonen (1981, 695) seem to suggest that speakers change their language not only purposely, but rationally (or intentionally, i.e. for some purpose), but not consciously, i.e. not according to any plan – this is particularly suggested by footnote 4 in Itkonen (1981, 695). Keller (1994, 10) argues that a view such as that argued by Itkonen (1981) creates more problems than it solves. As Keller (1994, 10) remarks:

> what does it mean to say that something is done unconsciously; when 350 million people are involved? ... As long as the logic underlying the relation between a collective and its correspondent individual statement is not clarified, such a collective statement explains nothing.

No doubt, part of the problem stems from a confusion of the terms *intentional*, *planned* and *conscious* (Keller, 1994, 10). Essentially, he argues that *intentional* and *planned* are not synonyms, and that *intentional* and *unconscious* are not antonyms (Keller, 1994, 13). The intention with which we do something does not necessarily amount to a conscious plan about the intention to do something, and the conscious achievement of an intention might well involve a number of unconscious components, as the construction of a relative clause (presumably done unconsciously) to speak consciously (Keller, 1994, 12). That is, Keller (1994, 13) suggests that language change is neither intentional, planned, nor conscious. The problem with Itkonen's view is, as pointed out

above, that there is an insufficient connection between the unconscious intentionality of the individual and the collective patterns (which can only be a-conscious). Merely relating a plausible but hypothetical speaker intention with an observed pattern does not constitute an explanation, unless we can prove that the intentional action was *planned* (which Itkonen seems to reject). This leaves the proposed explanation as little more than what Gould (1978) terms "story telling".

Causation, correlation, and manipulation

What is needed, then, is rather some non-deterministic causal model which is capable of relating the individual (with conscious and unconscious intentions) to a group level *and still maintain a causal link between the two*. A classical candidate is statistical correlation.

This is what Goldthorpe (2001, 2) refers to as "causation as robust dependence". Itkonen (1981) is also generally positive to statistical explanations based on correlation. However, it is a truism in statistics that correlation does not imply causation. Karl Pearson considered causation a "fetish", which ought to be replaced altogether by the more precise term correlation, as quoted in Goldthorpe (2001, 1). The obvious reason for this is the problem of "lurking factors". Assuming that a statistically significant association has been established between variables x and y where x is temporally prior to y, we still do not know whether x caused y or whether *both* x and y are caused by an unknown variable z (Goldthorpe, 2001, 2).

Early statisticians' scepticism toward causation was also in part influenced by what is known as *Simpson's paradox*, cf. Pearl (2000, 3), also known as "aggregation bias" (Gill, 2006, 316). An example from Gill (2006, 315–316) illustrates this. Suppose that a study is conducted on a group of participants (with equal proportions of men and women) from a social welfare program, where half the participants receive some kind of job training, whereas the other half receives no training. Suppose further that the figures show that on average, those who received job training had slightly better chances of finding employment. However, suppose that breaking the numbers down by gender, i.e. men and women who received training and no training, shows that both men and women who received the training did *worse* at finding employment than men and women who had no training.

This sounds counterintuitive, but the effect is well documented. For the paradox to arise, it is necessary for job training and employment to be correlated with each other, *and* for one subgroup (say, men) to be correlated with *both* job training and employment. In other words, if more men than women received training and got jobs, the unequal weighting leads to the paradoxical effect where receiving job training is confounded with gender. The effect is not present for the whole group, since aggregating

the subgroups is not the same as adding the averages from the two subgroups. Put differently, the confounding factor is canceled out (Gill, 2006, 316).

Thus, not only is correlation and causation distinct, but even identifying the correct *correlations* can in some situations be tricky, making any attempts at drawing causal implications even more precarious. In short, standard statistical analyses of observational data formally leave "black boxes" of causality (Goldthorpe, 2001, 9). Nevertheless, causation as robust statistical dependence has some merit, but as Goldthorpe (2001, 3–4) remarks, it is decidedly more successful in disciplines that make predictions, such as economy, meteorology, and to some extent biology. Behavioral disciplines such as sociology (and linguistics), where the emphasis is on *explaining* rather than *predicting*, are at a disadvantage under such a conceptualization of causation.

Some theorists of science have proposed to redefine causation as variables which can – at least conceptually – be experimentally manipulated, cf. e.g. Holland (1986), Rubin (1986, 1990) and comments in Glymour (1986) and Goldthorpe (2001). Without going too much into detail, suffice it to say that this raises a number of problems, not least for diachronic linguistics where it is not clear at all which factors can conceptually be experimentally manipulated. Another problem, as pointed out in Glymour (1986) and Goldthorpe (2001), is that it shifts our attention from looking for the causes of effects (e.g. as established through statistical correlations), to establishing the effects of causes (e.g. establishing the effect of providing a placebo pill vs. an active drug to groups of experimental subjects).

However, as pointed out in Goldthorpe (2001) this presents a difficulty for *all* behavioral or observational sciences, since the approach requires an experimental design, and furthermore leaves no room for purposive intention (as distinct from, but not necessarily opposed to, planned intention, cf. above). A statement such as *she did well on the exam because she studied for it* is thus not an admissible causal explanation, as pointed out by Goldthorpe (2001, 7), since there is no experimentally manipulable variable involved, only the student's goals. But this is a highly counter-intuitive notion of causality, since we would like to be able to state that studying for the exam actually influenced the outcome.

In short, some non-deterministic causal mechanism which is observable (but not necessarily manipulable) and which leaves room for purposive intentions, is needed. The following section outlines such an alternative.

Generative processes

An alternative, non-deterministic causality is discussed in Goldthorpe (2001), where it is referred to as "causation as a generative process".[8] As he points out, this is not so much a type of statistical thinking, as "what must be *added to* any statistical criteria before an argument for causation can convincingly be made" (Goldthorpe, 2001, 8), (original emphasis). As noted above, a statistical analysis of observational data formally leaves "black boxes" of causality, since they measures correlation, not causation. Instead, some theoretically based process which produces or generates the statistical correlations must be referred to.

Goldthorpe (2001) schematically outlines this as follows:

- establish the explanandum, e.g. through statistics;
- hypothesize a generative process;
- test hypotheses.

In other words, first the phenomenon to be investigated needs to be established; both whether it exists and to what extent it exists. Here, statistics clearly play an important role, although subject-matter input is also essential. The measurement variables for the statistical investigation must necessarily be informed by domain specific knowledge, both with regard to data collection and numerical description. In some (or perhaps most) cases, some sort of hypothesis testing might be required already at this stage, merely to establish the explanandum, e.g. "does word-order really change in Old English main clauses, or are the differences in frequencies just random variation?". Only then, when the explanandum is sufficiently established, can the search for generative processes start.

At the next step, hypothesizing generative processes, some account must be offered of the variety of agent goals, actions and interactions that arise in situations relevant for the explanandum. That is, some sort of central trait, or common denominator, must be proposed, which captures what is common to the observable variation (which has then been quantified through standard statistical methods) which is found in the sample. Goldthorpe (2001) does not mention it explicitly, but it seems to me that some kind of noise-signal ratio considerations would also be appropriate at this stage. If the variation in the sample is too great, a proposed generative mechanism (i.e. the signal) needs to be strong enough to break through, or be detectable despite, this variation (i.e. noise). However, step two also has a second component, namely that the identified actions must be shown to result in the statistically established explanandum.

[8] Note that "generative" in this context is completely unrelated to the how the term is used in Chomskyan or other forms of generative linguistics.

It is instructive to draw attention to the care with which Goldthorpe (2001, 12) points out that a necessary link must be established between the proposed explanatory factors constituting the central tendencies in situations and contexts, and the statistically identifiable group level results:

> a case must be made to show how these central tendencies in action would, if operative, actually give rise, through their intended and unintended consequences, to the regularities that constitute the *explananda*. (Original emphasis)

As Goldthorpe (2001, 12) notes, this serves to shift the focus from the adequacy *in principle* of an explanation, to its *empirical* validity. Crucially, it is necessary to investigate the implications that follow from an explanation: what other effects should be empirically directly or indirectly observable? As such, this approach is directly comparable with the position defended by Geeraerts (2006a), namely that introspection and theory need to be interpreted into empirical hypotheses, which was discussed in section 3.2.1 above. In such a framework, what is to be explained, then, are "facts about the probabilities or expectation values of outcomes rather than individual outcomes themselves" (Woodward, 2003, § 3.3).

The third step consists of testing the hypotheses regarding the generative processes. For this step, Goldthorpe emphasizes the implications that follow from the proposed explanation. Direct or indirect tests may be required to establish whether, or to what extent, something generates the explanandum. Goldthorpe (2001, 13) also reminds the reader that the *kind* of causation might be only probabilistic or some form of corroboration of the hypotheses, not a once-and-for-all proof. Again, consider the lung cancer example. The mechanisms by which tobacco smoke provides carcinogens that act upon the organism and thus causes lung cancer are clearly understood and well demonstrated. The specific risks of developing lung cancer for any one individual who is exposed to tobacco smoke, may well be probabilistic and of different magnitudes. However, the chain of causality itself is clearly established. Conversely, causes of test results, of the kind *she did well on the exam because she studied for it* are more difficult to establish in the same sense, even if the generative process were to be well corroborated. Thus, in line with e.g. Gorard (2003, 156), I will accept a probabilistic notion of causation, even if it means being "unable to decide whether this worked because the world is actually non-determinist, or because it is too complicated to explain fully". Thus, integrating a probabilistic approach with the notion of generative processes enables the combination of a partial identification of multiple or complex causes, in a non-deterministic manner.

As this section has shown, an account based on generative processes as causal mechanisms provides a reasonably well-defined concept of causality. Moreover, this concept is consonant with statistical approaches for data description and hypotheses

testing to the extent that they can be said to complement each other. It is clear that theory plays an important role in such an approach, but not to the detriment of empirical data. Again, the relationship is one of complementarity, as called for by e.g. Geeraerts (2006a). An important corollary of this view of causation is, as noted by Goldthorpe (2001, 9), that causation cannot simply be established through general statistical procedures. There is also need for specific subject-matter input to the analysis, in order to properly link the proposed causal generative mechanism to the statistical correlations. This crucial link between generative processes and statistical patterns in groups will be further explored in the next section.

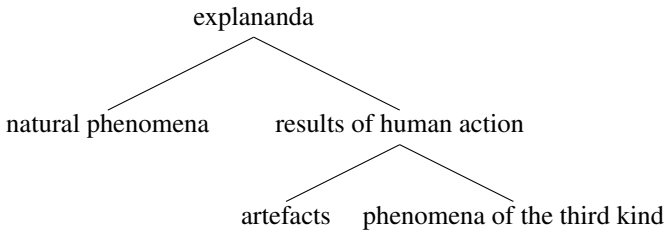
Invisible hands

Keller (1994, 15–18) discusses a series of photos of the structures formed by spectators around two groups of street artists in front of the Pompidou Centre in Paris. The audience form two circles around the performers, and as the photos in Keller (1994, 16–17) show, the spectators gradually organize themselves into two adjacent circles, rather like a number eight where the two circles have been disconnected. As Keller points out, the circles or structures come about not by planning but as an *emergent* phenomenon. That is, they result from the individuals wanting to see as much as possible of the show, not get in each others' way etc., but without intentional regard for the circular structures which emerge as a by-product of the individuals' actions. It is easy to see how the notion of a generative process can be brought into such an explanation. The goals of the spectators constitute the purpose which leads to the actions that generate the circles. Specifically, the spectators move and adjust their position relative to each other and the show, by moving their legs, which gives us a perfectly plausible mechanism which causes the formations.

With language, things are less straightforward, and it is necessary to consider more closely what language is. Keller (1994, 54–57) argues that language is neither a natural phenomenon nor a man-made artifact. This does not mean that language is not a result of human action, only that it is not a result of human *intentions* or plans. Keller (1994, 56) divides the world into three categories. First, there are natural phenomena such as e.g. mountains, the weather, or kidneys, i.e. things that are *not* the goal (or result) of human actions. Second, there are artifacts such as e.g. houses, programming languages, or a cup of espresso, i.e. things that *are* the result of human actions. And third, there are things which are the result of human actions but *not* the goal of their intentions, e.g. natural languages, inflation, or a path cutting across a lawn where people take shortcuts. This is what Keller calls a *phenomenon of the third kind*.

The following figure, reproduced from Keller (1994, 57) provides a schematic taxonomic relationship between phenomena of the third kind and other phenomena:

(4)



In order to explain phenomena of the third kind, Keller invokes *invisible hand* explanations. Interestingly, such a form of explanation fits eminently well with the generative process notion of causality outlined above.

Keller (1994, 65) argues that "cultural phenomena cannot be explained *exclusively* by reference to causality, but an explanation in the domain of cultural sciences can certainly have causal parts" (original emphasis). He goes on to say that

> phenomena of the third kind are always composed of a micro-domain which is intentional and a macro-domain which is causal by nature. The micro-domain is constituted by the individuals or their actions involved in the generation of the phenomenon; ... The macro-domain is the structure generated by the micro-domain

In other words, a phenomenon of the third kind is a causal entity or consequence of a collection of individual intentional actions, where the individual intentions are at least in part directed to the same goal (Keller, 1994, 65).

Keller (1994, 68–69) points out that in *principle* the macro-level is independent from any one specific micro-level explanation. That is, the third-kind effect of language change can come about for a number of reasons, including functional, social or cognitive factors. However, this does not entail that the choice of micro-level causal explanation is trivial, quite the contrary. Merely labeling something as a phenomenon of the third kind still does not constitute an explanation proper. Keller (1994, 69) insists that the two level structure with a micro- and a macro-level can only plausibly be established when we can deduce the macro-level from the micro-level of individual actions. Thus, third kind phenomena are not macro-level only, they are constituted by *both* levels. Keller (1994, 70) lists their following three essential attributes:

(i) they are procedural by nature;

(ii) they are constituted by a micro-level and a macro-level;

(iii) they have something in common with artifacts as well as with natural phenomena.

Based on this, three steps are necessary to constitute an invisible-hand explanation (Keller, 1994, 70):

1. a description of motives, goals and intentions as well as the general conditions of the individuals who generate the phenomenon in question;

2. a description of the process that explains the generation of the structure in question through the individual actions;

3. a description of the structure generated by the actions.

The three requirements laid out above highlight the shortcomings of many functional explanations. Point three (description of the structure) is accounted for by grammatical description, it is the explanandum or the thing to be explained. The functional-pragmatic motivation essentially performs the role of point one above, but point two – the deduction of the macro-level from the micro-level through some generative process – is often left unspecified. Essentially, an explanation of a phenomena of the third kind without this crucial step two is left as a possibly plausible, but weakly supported conjecture since one part of the explanans is missing.

Keller (1994, 72) sets out two criteria for an adequate invisible hand explanation: the *premises* should be plausible (i.e. reasonable given what we know about the world and the phenomena under scrutiny) and the *process* by which the macro-level structure is generated should be cogent (i.e. there should be a relevant link between the micro- and the macro-level). As was shown above, there are severe difficulties with the plausibility of the premises posited in both the functional and the child-based explanations. Furthermore, the generative process posited by child-based explanations is not cogent in the sense that it fails to establish a link between the micro- and the macro-level, cf. Croft (2000, 44–53), whereas functional explanations often specify a less than fully plausible *intentional* generative process.

Causation and grammaticalization

The present section discusses an example to show how and why the notions discussed so far are necessary for a coherent, corpus-based empirical approach to diachronic linguistics. In chapter 2, it was argued that although grammaticalization theory is a useful descriptive device, it cannot be considered as an explanation of the phenomena it describes. In the present section the argument is elaborated through an example presented in **?**.

Bybee (2003) makes the plausible argument that grammaticalization of a construction is affected by *type* (i.e. class) and *token* (i.e. individual unit) frequencies. However,

the actual mechanisms she refers to are somewhat diffuse. She states that the verb *cunnan* ("can") in Old English was "fairly frequent" (an unsubstantiated and unquantified empirical claim) and as a consequence of this lost "semantic force and specificity" (Bybee, 2003, 608). It is not clear exactly what "semantic force" refers to, but the context indicates that this could be something felt or experienced by the speaker. The same goes for passages such as this: "Perhaps *cunnan* is beginning to bleach and grow too weak to stand alone in such contexts" (Bybee, 2003, 609). Again, the context seems to imply that this is experienced by speakers, although it is never spelled out.

As mentioned in the discussion of grammaticalization theory in chapter 2, it is crucial to keep in mind that such metaphors as semantic "bleaching" or "weakening", useful though they might be, do not have an explanatory value in themselves. If a plausible case can be made for a situation where individual speakers experience this as a process, then we have established a linguistic-psychological explanandum, but we still need a causal mechanism (which can be cognitive or functional in nature). Bybee makes the assumption that the entrenchment of a stored unit is based on its text frequency, and that higher frequencies of *cunnan* (i.e. higher degree of entrenchment) entail "bleaching" (Bybee, 2003, 610–611). To illustrate this, she uses seemingly arbitrarily chosen (and unspecified) frequency thresholds which denote "high" frequency (token frequencies from five to 30 are all referred to as "high" in her paper). However, why high token frequencies should lead to a semantic "bleaching" is left unspecified.

Thus, irrespective of the adequacy of a grammaticalization-based description of the diachronic development of *cunnan*, the explanatory claims made by Bybee are left hanging in the air since no plausible mechanism for generating or causing "bleaching" (i.e. semantic change) is proposed. High frequencies are claimed by Bybee to cause "bleaching", but the link between high frequencies (if the reported frequencies in fact are high – as it stands this is impossible to judge) and semantic representation in the speaker (presumably), is never made explicit. This illustrates two deficiencies: first, the empirical claims regarding frequencies are not dealt with properly; and second, causal mechanisms are implied, rather than stated explicitly.

In subsequent chapters it will be shown how statistical methods can be employed to precisely quantify high and low frequency phenomena. Furthermore, specific proposals will be made regarding how cognitive mechanisms generate grammaticalization effects in the case of *there*, and what their links to frequencies are. Below, the basics of the approach is sketched out.

3.5 Explaining language change

Thus far, it has been shown that a probabilistic model of language is appealing for a number of reasons. First, a probabilistic model is able to cope with the distinction between group level patterns which may emerge from possibly very different individual behaviors. Second, the probabilistic view does not depend on whether language is *inherently* probabilistic, or merely very complex. Either way, a probabilistic description of causation is perfectly acceptable.

However, if we want to explain what causes language change, we must add something more to this approach. Gorard (2003, 158) gives the example of smoking and lung cancer. At the group level, we can observe a consistent and stable correlation between smoking and lung cancer. For this to amount to a convincing explanation, it is necessary to add a plausible *cause* which *generates* the lung cancer. In the case of smoking and lung cancer, this has been done in the form of experiments which have isolated carcinogens in the tobacco smoke, pathological studies of lung tissue etc.

Given the cognitive linguistic underpinnings of the current project, a natural place to start looking for such causes is precisely in human cognition. Chapter 11 will return to the question of the specific explanatory potential of cognitive factors in explaining the case of *there*. For the present the *principle* of seeking cognitive explanations for group level, diachronic phenomena is of greater importance than the specific causes and processes.

A reasonable counter-argument might be that such cognitive explanations to diachronically attested corpus phenomena present a confusion of levels of description. It could be argued that the group level correlations attested in the corpus are not directly relevant to the individual. Instead, the correlation patterns in the corpus could come about as an indirect, emergent property of behavior which was not originally directed at this outcome. Take the example of an ant trail or a formation of birds. Neither of these phenomena are natural objects, nor are they consciously constructed. Rather, they are what Keller (1994) calls a "phenomenon of the third kind", or consequences of other (intended) actions. Such phenomena are neither fully naturalistic nor fully created, but emerge at a macro-level based on what happens at a micro level. This constitutes what he calls an "invisible hand explanation" (Keller, 1994, 68). This approach explains its explanandum by reference to the causal consequence of individual intentional actions (Keller, 1994, 71), but the macro-level effect is a non-intended consequence of all the individual actions taken together (Keller, 1994, 64).

However, we still need some kind of causal mechanism to account for how the macro-level emerges from the micro-level, as Keller (1994, 71–73) also points out. In other words, a claim that corpus patterns are phenomena of the third kind caused by some action *without* a plausible mechanism of generation or causality is merely

restating two independent facts: A occurs at the micro-level and B occurs at the macro-level. Formally, this would constitute a *non-sequitur* argument without some additional supporting fact which demonstrates why the one follows from the other. Invoking some "black box" mechanism, of the form that A occurring at the micro-level causes B at the macro-level by means of an unknown, but logically necessary function f quickly causes more problems than it solves. In other words, unless one wishes to claim that the association patterns observed at the macro-level of the corpus are uninteresting epiphenomena without relevance to the object of study, some crucial connection must be set up between the micro- and the macro-level.

3.6 Summary

Two main methodological points have been argued in the present chapter, one relating to statistics, the other relating to language change.

The statistical point deals with the instruments with which language is investigated. It was shown that there are few inherent restrictions on the use of statistics in linguistics, and that quantitative methods have certain advantages over non-quantitative methods such as introspection, although this does not remove the role of introspection. Rather, the two methods are seen as complementary since introspection must precede statistical testing, and the statistical results must anyway be interpreted. Regarding the use of frequency-based arguments in linguistics, it was argued that raw frequencies and percentages can only be of limited value, and that more sophisticated statistical testing must be used to complete a frequency-based line of argumentation, not least to guard against unwarranted positive results.

An important methodological consideration here is that the use of statistical tests should be *motivated* and *focused*. Gelman and Hill (2007, 549) emphasize that causal inference should be estimated in a targeted way, not as the byproduct of a single large analysis. This advice applies to any statistical test. If the researcher throws enough tests at a large enough data set, something will eventually come out as statistically significant. This is a typical example of how to do "cherry picking", that is, selecting data, tests, or methodological approaches that are bound to support the researcher's views. As Rudman (2003, 28) remarks, this is what is sometimes known as "the sharpshooter's fallacy" which is "analogous to the way a gunslinger might empty his six-shooter into the side of a barn and then draw the bull's-eye around the bullet holes".

The second methodological point argued above, is that to talk about language change or grammaticalization, it is necessary to precisely define *what* is changing and to have some conception of which *mechanism(s)* might be responsible for the observed change. This point more or less follows from the previous one: if quantitative methods

are seen as necessary tools for investigating grammaticalization, then it is necessary to have a clear idea of what is being tested. However, statistical methods can only tell whether a change has taken place and how large it is once we have properly defined the measurement variables. The next step, then, is choosing the appropriate means to quantitatively investigate the hypotheses. This topic will be dealt with in the next chapter.

Chapter 4

Tools: Statistical tests

> I fear that the first act of most social scientists upon seeing a contingency table is to compute chi-square for it.
>
> Frederick Mosteller

4.1 Introduction

Besides the more general methodological points dealt with in the previous chapter, it is also important to take into account the specific considerations involving individual statistical tests and procedures.

The present chapter will discuss some statistical tests and measures, as a necessary background for interpreting the statistical analyses presented in subsequent chapters. However, the background information aspect is only part of the motivation for this chapter. An overall goal is to show how the choice of statistical methods is intertwined with the research questions to be answered and the data to be investigated. The choice of one statistical test over another is not merely a matter of convenience, or of finding a specific "prescribed" test for a given data type. In fact, the choice of statistical test is intimately linked to the research questions, the data, and the operational definition of the hypotheses to be tested. Choosing one test over another has profound influence on the conclusions which can be drawn and consequently to the entire research project. As such, careful consideration is needed, not only as to which tests are available or

convenient to use, but of the whole interplay of data, sampling, hypotheses, and the population to which inferences can be drawn.

Statistics, being a numerical undertaking, might sound dull. However, besides being necessary when dealing with large amounts of data, it can also be an excellent tool for revealing new aspects of the data. Far from being reductive in nature, statistics actually improves our understanding of the variation in the data; a variation which Francis Galton – as quoted in Limpert et al. (2001, 342) – called the "charms of statistics". But the choice of test must match both the data and the research questions.

4.2 Statistics – an overview

The present section gives a brief background summary on statistics, with emphasis on views on the epistemological role of statistics. No firm conclusion will be drawn, since these questions are more properly viewed as pre-empirical questions to be settled by the researcher before the actual testing is carried out. All the views presented below have their own justifications, and rather than finding the "correct" version of statistics, the role of the researcher in this question is to make his or her stance clear when presenting research questions and results.

Very broadly, we can distinguish three broad themes in statistics:

(i) Statistics as data exploration;

(ii) statistics as null hypothesis testing;

(iii) statistics as degree of belief in an outcome.

Note that these themes are not inherently tied to specific statistical tests. They are more a question of how the tests and their results are viewed. Most tests and measures can serve in all three roles, although some are more tied to specific themes than others.

4.2.1 Data exploration

A very basic distinction in statistics is that between *descriptive* and *inferential* statistics. The former seeks to describe a *sample*, whereas the latter is used to make inferences about a *population* based on a sample (typically through *null hypothesis testing*, see below). *Descriptive statistics* includes the use of measures of central tendency, such as mean, median, mode, etc. for understanding the properties of the sample. Different descriptive statistics are suitable for different kinds of data or *data levels*, cf. Hinton (2004, 21–22). The three data levels usually recognized in statistics are:

- Nominal data: can only be counted, ordering is unimportant;
- ordinal data: can be counted and ordered;
- continuous data: can be counted, ordered, and measured on a scale where consecutive numbers are of equal intervals.

Continuous data can be further subdivided into *interval* and *ratio* data, the difference being that the former can have an arbitrarily assigned zero value, whereas in the latter case a measurement of 0 really means absence of the measured property. Take the difference between 0 degrees Celsius and 0 kilometers per hour. In the case of 0 kph there is no movement, whereas with 0 °C there is still considerable amounts of heat around (only at −273.15 °C or 0 degrees Kelvin would there be complete absence of heat). With interval and ratio data it is possible to calculate means and standard deviations (see below), since any difference between two observations on the scale is of the same magnitude – the difference between 3 and 4 kmph is the same as that between 120 and 121 kmph, and both can be further subdivided into smaller units. However, it is not a trivial, technical decision to classify *data* as "discrete" or "continuous". Kempthorne (1966, 14) notes:

> But what about the so-called continuous data? It seems to me that while it may be quite reasonable to examine the model of a continuously distributed random variable, one must acknowledge that one never actually observes such random variables, and that all observations are in fact discrete, with a grouping error actually specified by the scientist.

The quote from Kempthorne (1966) is useful to keep in mind also when dealing with other data levels: these classifications are of course not a direct observation of reality, but idealized models, selected and classified by scientists for some purpose, and always in need of explicit interpretation. In linguistics, continuous data are rarely found and we typically deal with nominal data, that is, data which are counted but for which no arithmetic operations such as calculating a mean is possible (or at least requires an explicit justification).

In addition to the measures of central tendency mentioned above, which are assumed to be known, useful descriptive statistics include measures of spread such as a *standard deviation* (SD) and a *standard error of the mean* (SE). A standard deviation is a standardized measure of *distance from the mean* for a given observation (Hinton, 2004, 14–17). The standard deviation is simply the square root of the sum of squared observation deviances from the sample mean, divided by the size of the sample.[1] The

[1] This is a simplification. For technical reasons, the denominator is in fact the size of the sample minus one, cf. Hinton (2004, 16).

standard error of the mean is a measure similar to the standard deviation, but whereas the standard deviation relates individual observations to a population mean, the standard error of the mean relates a *sample mean* to the population mean. The standard error of the mean can be estimated since, as Hinton (2004, 55) points out, a population of means can be proven (through the central limit theorem) to approach a normal distribution as long as the samples are large enough, even if the individual responses themselves are not normally distributed.

This overview is purposely kept brief, introducing only some basic terms that will be referred to later. For a more thorough introduction to basic concepts in descriptive statistics, the reader is referred to the general statistics literature, such as Hinton (2004) or Oakes (1998). On a final note, it should also be pointed out that procedures such as *correspondence analysis* can also be used for descriptive and exploratory purposes. This is dealt with in more depth in section 4.6.5 below.

4.2.2 Null hypothesis testing

This is probably the most familiar conception of statistics, forming the basis of most introductory statistics courses. This entails setting up a *null hypothesis* denoting the most conservative situation, typically of the form "no association, no effect", and comparing this with an *alternative hypothesis*, i.e. the research question, of the type "there is an association between x and y". The testing is done by comparing *observed values* (counts, proportions, means) with *expected values*, where the expected values are based on a mathematical distribution with known properties (such as the t, normal, and χ^2 distributions). Based on the comparison between the observed and expected values, a *test statistic* is computed for the data set. When this test statistic is compared with the number of *degrees of freedom* (more or less related to the size of the data set), a p-value can be obtained as a measure of how close the observed values are to the expected ones. Before testing, an *alpha* (α) level is chosen, a cut-off point or threshold for when we will reject the null hypothesis. The alpha is typically set somewhat arbitrarily to 0.05, or 5%, which gives a theoretical error rate (i.e. making the wrong choice between the null hypothesis and the alternative hypothesis) of 5% (or one in twenty) of the cases.

It is important to note a few things regarding this way of treating data. The list is not exhaustive, but these are some central concerns regarding all hypothesis testing. First, the results are dependent on the choice of the distribution. If the chosen distribution used by the test is not appropriate for the data, the results are likely to be distorted, see e.g. comments by Kilgarriff (2005) and the reply by Gries (2005). Second, all such tests are heavily influenced by the number of observations. That is, having more observations makes it easier to reject the null hypothesis purely due to size. Third, the tests assume that the individual data points are independent, that is, having observa-

tion A_i as the outcome somewhere in the dataset should not influence the outcome of observation A_j. Fourth, the *size* of the p-value is of limited importance, beyond the fact that it is either above or below the specified alpha level, i.e. the p-value is a continuous entity, but it represents a *binary* decision of rejecting or not rejecting the null hypothesis. And finally, as pointed out in Cohen (1994), the p-value needs to be properly evaluated against the alternative hypothesis. Researchers are typically interested in questions like "given these data, what is the probability that the null hypothesis of randomness / no association is true?", whereas the hypothesis tests answer the question "given that the null hypothesis is true, what is the probability of these or more extreme data?" (Cohen, 1994, 997). Formally, the null hypothesis test states the probability of the observed data, given the null hypothesis or $P(data \mid H_0)$; whereas the researcher wants the probability of the null hypothesis given the data or $P(H_0 \mid data)$. Why these two questions are not equal is discussed further in Cohen (1994). Most importantly, the null hypothesis test establishes the falseness of the null hypothesis, *not* the truthfulness or accuracy of the data or any predictions about the data.[2]

These properties of null hypothesis testing have led a number of scholars to advocate the abandonment of this form of testing, cf. Kilgarriff (2005); Berger and Sellke (1987); Johnson (1999). This does seem a little too strong a reaction and the present study takes a more pragmatic approach. There are a number of ways of mitigating the problems concerning null hypothesis testing. This involves (but is not restricted to) carefully considering aspects of the study such as sample size, type of data (and the type of study), sample-population relations against the aims of the study; and based on this determine the appropriate levels for significance and effect size and consequently interpret the results rather than simply reject or not reject a null hypothesis, cf. Gries (2005); Kempthorne and Doerfler (1969, 247); Rosenthal et al. (2000, 25–30); Upton (1992).

4.2.3 Degree of belief

Looking at probabilities as degrees of belief is closely related to what is broadly known as *Bayesian statistics*, after Thomas Bayes (1701–1761). Bayesian statistics are sometimes referred to as *subjectivist*, as opposed to the *objectivist* or *frequentist* statistical thinking which is traditionally invoked for the view of statistics as null hypothesis testing. The two approaches agree that probabilities are nonnegative numbers bounded by

[2]There is of course an exception to this, as Cohen (1994, 999) discusses. When a central aspect of a theory makes a specific prediction, this prediction can be challenged through null hypothesis testing. Such hypothesis testing of the "strong" form, with specific predictions, are rare in the humanities and social sciences though.

0 and 1, but differ in their interpretation of what probabilities mean. Bod (2003, 12) summarizes this as follows:

> According to the objectivist interpretation, probabilities are real aspects of the world that can be measured by *relative frequencies* of outcomes of experiments. The subjectivist view ... interprets probabilities as *degrees of belief* or *uncertainty* of an observer rather than as having any external significance. (Original emphasis)

However, this does simplify things somewhat, as Gelman, Carlin, Stern, and Rubin (2004) point out. In fact, the frequentist conception of probabilities as relative frequencies over a long (possibly infinite) sequence of identical repetitions of the event being measured is problematic if the event cannot be at least conceptually embedded into such a sequence (Gelman et al., 2004, 12). To take an example from Gelman et al. (2004, 12): what is the probability of Colombia winning, if Colombia plays Brazil in soccer tomorrow? If we conceptually embed this event in a series of infinite Colombia–Brazil soccer matches, we can think of the probability of Colombia winning as a relative frequency. However, this require us to assume (at least under a strict interpretation) that the soccer matches are identical in all relevant aspects. This is surely not the case, since players and managers come and go; weather conditions change; skills, talent, motivation, teamwork, amount of training and injuries play a role; etc.

The guiding principle in Bayesian statistics is that through probabilities it is possible to, in a systematic way, say something about unobserved events based on partial knowledge, i.e. data (Gelman et al., 2004, 12). The point offered here is not that Bayesian statistics is superior to frequentist statistics, or that Bayesian statistics will be made use of in the following chapters (in a technical sense, it will not). Rather, the point is that from a conceptual and epistemological standpoint, the distinction between frequentist and Bayesian statistics is blurry, since both rely to some extent on subjective belief and both make use of frequencies. As such, it is sometimes more profitable to think of probabilities in broadly Bayesian terms as measures of uncertainty (or degrees of belief), rather than as constituting hard facts regarding long term relative frequencies. This point becomes particularly poignant when dealing with entities that cannot readily be construed as taking part of infinite, identical series of events, as is often the case in linguistics.

4.3 Some common tests

In the following some common tests of statistical significance will be discussed. These are basic parametric and non-parametric significance tests taught in introductory statis-

tics courses, and are fairly familiar sights in publications dealing with quantitative data. Their motivation and computation will only be discussed in passing. Some examples are critically discussed related to these tests. These examples have been chosen because they highlight some interesting use (or potential use) of the given test, and because they occur in presumably widely read handbooks and introductory books, and thus should be familiar to a wide audience.

4.3.1 The t-test

The t-test, or *Student*'s t-test (after William S. Gosset's pseudonym), is an established way of investigating differences in means, that is, it is assumed that the data are interval level.[3] The t-test can be used for paired tests (i.e. two measurements of each individual in one sample) or for two-sample tests. There are several ways of carrying out a t-test, but this will normally be taken care of by a statistics program. The following formula, taken from Hinton (2004, 84) outlines the procedure in broad terms:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}. \tag{4.1}$$

The formula amounts to taking the mean from group one $\bar{X}_1$ and subtracting the mean from group two $\bar{X}_2$ (i.e. the difference in sample means between the two groups), and dividing it by the standard error of the difference between the sample means $S_{\bar{X}_1 - \bar{X}_2}$. The resulting test-statistic, the t-value, can then be considered in light of the degrees of freedom (roughly speaking the number of observations), to obtain a p-value which will then indicate rejection or non-rejection of the null hypothesis given the alpha level. As with the calculation of the t-value itself, these steps are normally taken care of by the statistics program. If the prerequisites for using a t-test are filled (interval or ratio data, equal variances and normal distribution of data), this will indicate whether a real difference exists between the groups being compared, or whether any observed differences are merely random variation. For a further introduction and technical details, see standard introductory books such as Hinton (2004) or Oakes (1998).

4.3.2 The chi-square

The *chi-square*, or the *Pearson chi-square test of independence*, is perhaps the most commonly used statistical null hypothesis test in linguistics, due to its relative lack of assumptions about the data. Formally, the test requires count data with independent

[3] Additional requirements include approximate normal distribution of the data and equal variances. A variant of the classic t-test, the Welch test, compensates for unequal variances.

observations and expected frequencies of at least five or more, cf. Hinton (2004, 258). The test is formally a test of independence between rows and columns, and its logic is simple enough that a more in-depth discussion can be useful as background to later interpretations. The formula is as follows:

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right), \tag{4.2}$$

where "O" denotes observed frequencies and "E" denotes expected frequencies. The notion of expected frequencies might need some further explication.

If we denote the row sum, or row total for a given row j as T_j and the column sum, or column total for a given column i as T_i and let n denote the total number of observations in the table, then, for any given cell in the table we can calculate the expected frequency as follows, cf. Hinton (2004, 255):

$$Expected_{ij} = \frac{T_j \times T_i}{N} \tag{4.3}$$

The logic of this is that two events occurring together are independent if their observed probability equals the product of the probability for each of the events separately, cf. Bilisoly (2008, 117–118); Gill (2006, 317–319). Since the expected frequencies are generated by multiplying probabilities, they provide a baseline for evaluating the observed frequencies against.

That is, if the null hypothesis of equal proportions is true, the observed values should be very close to the expected values. Whether the deviation from expected values is significant or not is assessed by comparing the χ^2 value (i.e. the test statistic) with a table of critical values for a given number of degrees of freedom.

As pointed out by e.g. Mosteller (1968) in his comments on the results presented by **?** from their investigations of the *Brown Corpus*, the chi-square test is sensitive to large amounts of data. Although McEnery and Wilson (2001, 84) rightly warn that the Pearson chi-square test is unreliable with small samples, they are wrong in stating that this is its "main disadvantage". Mosteller (1968, 2) comments on the large number of significant p-values from Kučera and Francis's work, and says that we might not expect the words in the Brown Corpus to be distributed randomly in the first place, but that

> this is not the point, the real point is that the authors [Kučera and Francis] have a lot of data. They can therefore magnify modest deviations.

As discussed in section 4.2.2 above, this is a problem which pertains to most null hypothesis tests, to the extent that some authors dismiss null hypothesis tests as merely testing whether the researcher has gathered enough data to achieve significance, cf.

Cohen (1994) or **?**. Consequently, it is crucial to maintain a healthy scepticism to claims based on low p-values from tables with a large number of observations and / or a large number of categories (note that the formula for the chi-square is partly influenced by the number of cells in the table, thus having many categories will make it easier to reject the null hypothesis than in a simple 2×2 case). This is particularly pertinent when the hypotheses being compared are somewhat unspecific and the sampling is non-random.

Gorard (2003, 142) likens such omnibus uses of the chi-square test to "finding the outlines of animals in the stars in the night sky, but with the added appeal of a statistically significant omnibus chi-square test". In other words, the researcher feels confident that she can "easily see the pattern in the data anyway" (even if the actual interpretation of a chi-square test on a large table can be hopelessly complicated), but with the seducing rhetorical effect of a low p-value. Needless to say, this is *not* good empirical research.

Further caveats and an example

As mentioned above, there are certain criteria which need to be fulfilled for the Pearson chi-square test to be applicable. The following discussion is based on Guy (2003), published in **?**, which (as a handbook example) should be of general interest. Table 4.1 on page 82 gives observed frequencies of speakers in three age groups who delete coronal stops in semi-weak past tense forms (such as *told*, *kept*, *lost*, etc.) to varying degrees. The total number of observations in this case is 34, i.e. fairly low. This is not necessarily a problem if the data (i.e. participants of the experiment) are sampled randomly, and there is no risk of an inflation effect. However, the data in 4.1 seriously violate the conditions of use for the Pearson chi-square in two important respects. First, the expected frequencies, which I have included next to the observed frequencies from Guy (2003, 380), reveal that only two out of nine cells in the table have expected frequencies above five. Expected frequencies in the five to ten range is recommended as an absolute minimum in most statistics textbooks, cf. e.g. Hinton (2004, 258). The reason is simply that with very small expected frequencies, small increases in observed units will lead to large proportional increases. This might be resolved using another test, such as Fisher's exact test, but with the Pearson chi-square it leads to questionable results.

Another and perhaps more serious problem is the discrete categorization of two continuous variables, viz. deletion rate and age. As noted by Fleiss (1994, 250), when a continuous random variable is dichotomized, chi-square based measures are heavily influenced by where the cut points are set. Simply put, the researchers in this case have probably influenced the p-value substantially by setting the boundaries for "high",

"low" and "medium" deletion rates, and dividing up the age groups in "0–18", "19–44", and "45+". The point here is not that the conclusions drawn by Guy are unwarranted; they might very well be correct, but as it stands we have no evidence for this. The results of the Pearson chi-square test he presents are inadmissible as evidence, since the test is not applicable.

This example reveals that by not adhering to assumptions of the test, and furthermore by discretizing a continuous variable, the researchers in this case have probably removed themselves from the original data to such an extent that it is no longer clear what they are testing: the p-value is *unreliable* because of low expected frequencies; it is *artificial* in the sense that the discrete nature of the categories established by the researchers are probably influencing the results more than the original observations warrant.

TABLE 4.1: *Frequencies of coronal stop deletion in semi-weak past tense forms for different age groups and deletion rate, from Guy (2003, 380). Next to the observed frequencies from the source are the added expected frequencies, computed with R.*

	0–18	Exp	19–44	Exp	45+	Exp
High (> 0.75)	7	(1.65)	0	(2.47)	0	(2.88)
Med (0.75 - 0.60)	1	(3.29)	9	(4.94)	4	(5.76)
Low (< 0.60)	0	(3.06)	3	(4.59)	10	(5.35)

4.3.3 Fisher's exact test

Fisher's exact test (also known as the *Fisher-Irwin* or *Fisher-Yates* exact test), is a significance test for 2×2 tables[4] containing nominal data, i.e. it is used much like a chi-square test. The use of Fisher's exact test in corpus linguistics has in recent years been vigorously advocated by Anatol Stefanowitsch and Stefan Th. Gries, cf. Stefanowitsch and Gries (2003); Gries and Stefanowitsch (2004) *et seq*, based on an earlier suggestion by Pedersen (1996). See Fleiss et al. (2003, 55–57) for a very good technical introduction to Fisher's exact test.

The first thing to point out is probably that the test is not "better" than the Pearson chi-square test in the sense that it is more "exact". "Exact" in the sense of the test refers to the calculations involved (discussed below), not the precision of the result.

[4]For larger tables a chi-square test is usually used for computational reasons.

TABLE 4.2: *Numbering of contingency table cells.*

	Col_1	Col_2
Row_1	obs_{11}	obs_{12}
Row_2	obs_{21}	obs_{22}

The p-values from Fisher's exact test and a Pearson chi-square test tend to be similar and point to the same result (i.e. both are either significant or not significant), and as discussed above, the *precise* size of a p-value is uninteresting from the point of view of a null hypothesis test. The important point is whether the value is above or below the specified alpha level.

Fisher's exact test starts by assuming that the marginal sums (i.e. the row and column sums in the contingency table) are kept constant. This way, there is a finite number of ways to arrange the observations within the table and still get the same row and column sums. With this restriction, it is possible to calculate the exact probability of the given table by taking the product of the factorial of row and column sums and dividing them by the product of the factorial of the table total and the factorial of each cell. This is shown in formula (4.4), from Fleiss et al. (2003, 56):

$$P_{obs} = \frac{Rowsum_1!Rowsum_2!Colsum_1!Colsum_2!}{Total!Obs_{11}!Obs_{12}!Obs_{21}!Obs_{22}!}. \tag{4.4}$$

See table 4.2 for the labeling of the cells in a contingency table with four cells.

The next step is to carry out a similar calculation for the other possible (hypothetical) combinations of observations that will give the same row and column sums. Then, the probability of the observed table is *added to* the probabilities from the other hypothetical tables which (if any) are equal to or lower than the probability of the observed table. This summed probability is the *exact probability* of the table.

Fisher's exact test is usually recommended for cases where the assumption of expected frequencies do not hold for the Pearson chi-square test, that is, the expected frequencies in the table are below the conventional threshold of 5. As noted by Faraway (2006, 75), Fisher's exact test is useful in such situations because the test statistic of the chi-square test only follows the χ^2-distribution approximately, and the approximation becomes particularly difficult for tables with small expected frequencies where the proportional differences become unduly large, see also Hinton (2004, 258).

For corpus data, the frequencies are usually large enough that a chi-square test can be used. However, as will be shown in chapter 8, Fisher's exact test and its measure

of association – the *odds ratio* (see section 4.4.2 below) – is a useful supplement when comparing one case to a pool of all other cases, as with e.g. *there* compared with all other adverbs.

Fisher's exact test is not unproblematic in the context of corpus linguistics. For one thing, it assumes fixed marginal total frequencies, an assumption which may work well under experimental conditions, but which is questionable in observational disciplines such as linguistics. Pedersen and Bruce (1996) suggest a purely descriptive approach to the use of the test, rather than trying to measure population parameters in corpus linguistics. Kempthorne (1979) – and later Upton (1992)[5] – recommends its use for small to medium-sized experimental studies.

Stefanowitsch and Gries use Fisher's exact test with an explicitly psycholinguistic interpretation, cf. Stefanowitsch (2006), where the identification of a collostruction in the corpus material is merely a part of a larger analysis frame. Stefanowitsch and Gries use the p-value from Fisher's exact test as a rank-based measure of association for their analysis, as discussed in Stefanowitsch and Gries (2003, Note 6). This is a somewhat different use of the p-value from the usual null hypothesis test situation, and is not discussed further here.

4.4 Effect size

Why do we need effect size? Recall that the p-value as used in a null hypothesis test is a binary variable, which is either above or below the previously specified alpha-threshold of rejecting or not rejecting the null hypothesis. If a test comes out as significant for a given alpha level, this shows that the likelihood is high that a real difference exists somewhere in the data. However, very often we need to know whether this difference is large or small, and for this the p-value cannot help us.

However, as **?**, 123–127 points out, there is an inherent "soft correlation noise" or "crud factor" inherent in behavioral sciences. Essentially, everything is correlated with everything else at some level. This certainly seems to hold true for language use, since no-one would seriously suggest that language units are used randomly (or even arbitrarily), cf. Kilgarriff (2005). This presents a serious problem since there is, as far as I know, no published research on the size of such a crud factor in linguistics. Hence, having a general scale (even one that gives a possibly false sense of accuracy) is useful, since it gives an identifiable label to the association, rather than simply stating whether or not it is negligible. Perhaps more importantly, such a scale allows for easier comparisons of effect sizes between studies. In short, it is of interest to know whether

[5] Upton first argued against its use altogether in 1982, believing it to be just as conservative as the Yates' corrected chi-square, but later changed his mind regarding the Fisher exact test in 1992.

an effect or association could be confused with the crud factor, or whether it is large enough to plausibly be interpreted. The question will be touched upon below, after a brief presentation of some effect size measures for nominal data.

The question of what constitutes a respectively small, medium or large effect size cannot be settled without considerations of data and Cohen (1988, 224) recommends that such conventions (useful though they are) should be treated with some caution.

The measures presented below are the most relevant for the data types investigated in the current study, but other effect size measures will be introduced where relevant in subsequent chapters; in any event the basic conceptual principles remain the same. Finally, it should also be pointed out that regression coefficients (see section 4.6 below) also constitute effect size measures, since they measure the *magnitude* of an effect, not merely its statistical significance. For further discussions on effect sizes, see Cohen (1988); see Cohen (1988, 215–227) for measures that pertain specifically to contingency tables.

4.4.1 *Phi* and *V*

Fleiss et al. (2003, 97) state that a χ^2 statistic is an excellent measure of the *significance* of an association, but that it has no value as a measure of the *degree* of association. They go on to say that "a test statistic is constructed under a null hypothesis and is thus not designed to estimate a nonnull association" (Fleiss et al., 2003, 98). But the size of a nonnull (i.e. a significant) association is very often precisely what we are interested in. The p-value only tells us whether two categories are related, not how strong or weak this association is. For this we need some kind of measure of effect size.

For 2×2 tables, ϕ (Phi) is a recommended effect size measure. The formula for ϕ is given below:

$$\phi = \sqrt{\frac{\chi^2_u}{n}}. \tag{4.5}$$

Thus, ϕ is calculated as the square root of the (uncorrected, i.e. without Yates' continuity correction[6]) chi-square value divided by the total number of observations in the table. This gives a coefficient which varies between 0 and 1, measuring the association between row and column variables. For larger tables, a measure presented in Cramér (1946) is used, often referred to as *Cramér V* after its creator:

[6]Yates' correction for continuity amounts to reducing each cell value by 0.5 before squaring it to correct for the risk of false positive results with small samples, cf. Hinton (2004, 258).

$$V = \sqrt{\frac{\chi_u^2}{n \times (min\{I, J\} - 1)}}. \tag{4.6}$$

The relationship with the ϕ coefficient is obvious. The main difference is that the number of observations n in the denominator is multiplied by the smaller of either the number of rows or columns (I, J) minus 1. Although this means that the measure can be applied to larger tables, the value is no longer strictly bounded by the range 0 – 1, although the approximation is good enough for most practical purposes.

In both cases, the denominator of the fraction corrects for the inflation bias of the Pearson chi-square test, a bias which increases the risk of spuriously low p-values simply due to a large sample size, as discussed above.

Gries (2005) illustrates the value of effect size measures. The paper is a follow-up on a comment from Kilgarriff (2005) that null hypothesis testing on random words will give significant p-values (thus showing the practical problems inherent to such an approach). However, while this is true to some extent, Gries shows that when effect size measures and other corrective steps are taken, the problem is limited. Specifically, Gries tested word frequencies from ten files in the British National Corpus (BNC) and then compared them with each other, which amounted to some 1.2 million comparisons of word pairs. Although around 20% of these were significant at $p < 0.05$, applying the Cramér V effect size measure showed that only 5% of the total had an effect size of 0.005 or higher (Gries, 2005, 282). In other words, random word frequencies might easily reach significance due to the inflation effect discussed above, but effect size measures at least in part mitigate this.

Regarding the size of an effect from a chi-square test, Cohen (1988, 222–227) classifies ϕ as follows:

- Small = 0.10
- Medium = 0.30
- Large = 0.50

Values for V will partly depend on the size of the contingency table, so that V decreases as the size of the smaller of rows and columns ($min\{I, J\}$) increases (Cohen, 1988, 224). However, the values are of a comparable size. For comparative values with different matrix sizes, see table 7.2.3 of Cohen (1988, 222).

An example

Table 4.3 is re-created from McEnery and Wilson (2001, 84), and shows frequencies of the use of the 3rd person singular present and perfect forms of the Latin verb *dico* ("say") in two of the Gospels. For ease of interpretation, I have added the expected frequencies next to the observed frequencies.

TABLE 4.3: *The number of present and perfect forms of the Latin verb "say" in two Gospels, from McEnery and Wilson (2001, 84), with added expected frequencies computed in R.*

	dicit	Exp	*dixit*	Exp
Matthew	46	(64.3)	107	(88.7)
John	118	(99.7)	119	(137.3)

As the table shows, Matthew uses *dicit* less and *dixit* more than expected, whereas for John we find the opposite pattern. McEnery and Wilson (2001, 85), having observed that the table yields a p-value which is significant at the 5% level ($\chi^2_{df(1)} = 14.0$, $p = 0.00017$),[7] conclude that there is a real variation in the use of these two verb forms in the texts.[8] However, the authors do not touch upon whether the difference is large, i.e. whether we should attach any great practical importance to this difference.

The ϕ value for table 4.3 is 0.195, or 0.2 for convenience. This corresponds to a small association between rows and columns, i.e. between the Gospels and the two verb forms. Although Matthew seems to use *dicit* a little less than expected, and John seems to use *dixit* a little less than expected, this does not amount to a great difference. (In fairness, it should be added that McEnery and Wilson (2001) do not claim to give an introduction to statistics. However, the book is so well known that the example used in it makes for a suitable illustration.)

A convenient way of visualizing this is with a Cohen-Friendly association plot, cf. Cohen (1980) and Friendly (1995), which shows the contributions to the chi-square value for each cell in the table. Such a plot for table 4.3 is shown in figure 4.1. Here, each of the cells are represented by a rectangle which is placed either above or below a dotted line. The dotted line represents the expected values, while the direction of the rectangle shows whether the observed values are larger or smaller than the expected

[7] In subsequent chapters, p-values are normally given as limit values, i.e. above or below the specified alpha.

[8] I leave aside here the methodological point regarding null hypothesis tests of something which can be considered the complete population.

values. The total area covered by the rectangle is a function of the size of the discrepancy between observed and expected values, that is, a small rectangle means that the cell contributes little to the overall significance of the table.

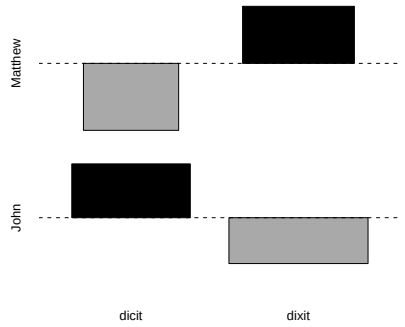


FIGURE 4.1: *Cohen-Friendly plot for table 4.3. The dotted lines represent expected frequencies. It is clear that the four cells' contribution to the chi-square value are more or less the same.*

As the plot in figure 4.1 shows, the cells' contributions to the chi-square value appear to be more or less of the same magnitude. Thus, the graphical display complements the effect size measure in the sense that the weak association appears to be a general property of the whole table, rather than being caused by a single cell.

Thus, effect size measures and Cohen-Friendly plots are easy and efficient ways of obtaining a more nuanced picture than what can be gleaned from a simple p-value produced by a Pearson chi-square test. The statement in McEnery and Wilson (2001, 84) that discrepancies in sample size are unimportant for the chi-square p-value, while technically true, is wrong in the wider sense that a larger corpus usually means higher observed frequencies. The main methodological point made here is that, given a large enough corpus, it is a trivial task to obtain significant p-values. A proper interpretation of such a value requires additional information.

4.4.2 The odds ratio

The odds ratio (*OR*) is a convenient measure of effect size for 2×2 tables, and it is the recommended effect size measure for Fisher's exact test. See Fleiss et al. (2003,

102) for a more thorough introduction to the odds ratio. Simply put, the odds ratio is the ratio between two odds. As shown in the formula below, it is computed by dividing the odds for A with those of $\neg A$. The subscripted ns in the formula below refer to the numbering of cells in a 2×2 contingency table as illustrated in table 4.2.

$$OR = \frac{O_A}{O_{\neg A}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} \tag{4.7}$$

The *odds* of something happening can be thought of as the probability of the desired outcome divided by the probability of the undesired outcome, or $p/(1-p)$. The odds of tossing a fair coin and getting a head are thus $0.5/0.5$, i.e. 1. The *ratio* between two odds is an estimate of an event (say, tossing a head with a fair coin) occurring in one group versus it occurring in another group. Consequently, an odds ratio of 1 indicates equal odds, that is, a probability of 0.5, while an odds ratio of 0.33 would indicate that the odds are greater in the second group than in the first group (when the odds ratio is larger than 1, this indicates that the odds are greater in the first group).

Although the odds ratio may not be immediately intuitive, it is important since the outcome of a logistic regression (see below) is reported in log odds ratios. Essentially, the odds ratio can be represented as a common fraction: how many expected occurrences of A do we find for each occurrence of $\neg A$? As such, the odds-ratio is useful in situations where we want to conceptualize the occurrence of some category over a pool of other categories.

4.5 Conditional probability

Although perhaps less often encountered in linguistic research than the previously discussed null-hypothesis tests, conditional probabilities can be useful tools in language related research, as demonstrated in e.g. Jurafsky, Bell, Gregory, and Raymond (2001). Conditional probabilities are often, although not necessarily, connected with a Bayesian approach to statistics.

4.5.1 Basic conditional probability

Conditional probabilities are useful tools for quantifying the relative frequency of some events (or in a Bayesian conceptualization, the belief in some outcome), without imposing the conditions required for null-hypothesis testing. Thus, they can be considered a relatively assumption free non-parametric statistics.

Jurafsky, Bell, Gregory, and Raymond (2001) discuss probabilistic relations between words and give a good outline of methods to measure them with examples, see

also e.g. Bilisoly (2008, 116–117); Bod (2003, 15–17). The probability of an event E is simply its occurrence expressed as a fraction of all possible occurrences. Intuitively, it is clear that the probability of, say, a word w_i occurring in a text is the frequency $n(w_i)$ with which this word appears in the text divided by the total sample size, or text size N. This is the *marginal probability*, expressed in 4.8.

$$P(w_i) = \frac{n(w_i)}{N} \tag{4.8}$$

Similarly, the probability of two events occurring together is the intersection of the two events divided by the total. If we have two words, then their combined frequency of co-occurrence divided by the total sample size gives us their *joint probability*, as shown in 4.9:

$$P(w_i \cap w_j) = \frac{n(w_i \cap w_j)}{N}. \tag{4.9}$$

With this we can go on to calculate the *conditional probability* of one event E given another event F, usually written $P(E \mid F)$; e.g. the probability of one word given the other. That is, having observed one word, what is the probability of seeing the other? It turns out that this is simply the joint probability of words i and j divided by the probability of the conditioning word.[9] The conditional probability of word i given word j is thus

$$P(w_i \mid w_j) = \frac{n(w_i \cap w_j)}{n(w_j)}. \tag{4.10}$$

Note that in this last step, order matters, as pointed out in Bilisoly (2008, 117), even if the mathematical probabilities are not intrinsically related to the order of real world events and occurrences. The numerator of the calculation will stay the same, since $n(w_i \cap w_j) = n(w_j \cap w_i)$. However, the denominator changes depending on the conditioning frequency. Effectively, the denominator represents a "sample space" where the events in the numerator may co-occur and changing the sample size changes the conditional probabilities. These frequencies will tend to be different, unless $n(w_i) = n(w_j)$. Thus, given two events E and F, $P(E \mid F) \neq P(F \mid E)$.

[9] Since dividing both numerator and denominator by the same number does not change the fraction, we can simply use the observed frequencies rather than the probabilities as in the examples above.

4.6 Linear models

In the subsequent chapters, a number of techniques will be employed to test the hypotheses outlined in chapter 2. Although some use will be made of classical null-hypothesis testing and relatively intuitive measures such as conditional probability, a major methodological point of this dissertation is to explore more advanced techniques. The previous sections discussed the well-known problems associated with many classical null-hypothesis testing techniques which make their use or interpretation in linguistics difficult.

In the present section, I will discuss another approach, viz. *linear models*. I take this to include (in the broad sense) both regression models and correspondence analysis techniques. Although they are distinctively different, as Greenacre (2007, 47) remarks, they do share certain characteristics which make it reasonable to group them together for now: both sets of models use mathematical models to fit lines to data points in multidimensional spaces, and both methods make use of least squares fitting (see below).

4.6.1 Regression

Regression modeling is an application of linear models to data analysis. Regression models have been used in linguistics for some time, especially as logistic regression in the so-called *variable rule analysis* in sociolinguistics, cf. e.g. Cedergren and Sankoff (1974); Sankoff and Labov (1979); Labov (2008); Bayley (2002). Regression modeling has been shown to be useful in corpus linguistics, cf. Biber (1992), and it has also been used in historical corpus studies Kroch (1989).

In the following sections the logic behind regression modeling is outlined briefly. The aim is to present the reasoning which underlies the use of regression modeling, not an in-depth exploration of the mathematical foundations.[10] The focus will be on the necessary background to understand and interpret regression models used in subsequent chapters.

Classical linear regression is a suitable introduction to the ideas behind all types of regression modeling. In the sections below, the reasoning will be expanded to other models which are based on the same basic idea, but make use of different assumptions. However, the fundamental logic is the same as for linear regression. Consequently, although these other models are more powerful than classical regression, they are conceptually more demanding and correspondingly less suited as an introduction to the motivations behind this type of data analysis. The outline on linear regression is primarily based on Faraway (2005, 11–13).

[10] A good introduction to the mathematical concepts behind calculating classical linear regression parameters can be found in Gill (2006, 155–157).

Linear functions

Data invariably comes with variation, whereas we as researchers often are interested in any underlying trends that might be present in the data. The goal of statistics is typically to capture such trends with numbers. Finding numeric properties of some variables can be thought of in terms of a mathematical *function*. A function in mathematical terms is a "mapping" or set of instructions "which gives a correspondence from one measure onto exactly one other for that value" (Gill, 2006, 19). This means that given the correct function, we can take a set of input values and (via the function) find the corresponding output values, cf. (Gill, 2006, 20).

A typical technique in many sciences is to assume that this function takes the form of a straight line. This is known as a *linear model*. The straight line model is attractive because it is a simple function with known properties.

A straight line can be described in the form

$$y = \beta_0 + \beta x \tag{4.11}$$

where β (the *slope*) is the steepness of the line, and β_0 (the *intercept*) is the point where the line meets the y-axis of a Cartesian coordinate plane (Gill, 2006, 26). This can be visualized as in figure 4.2 on page 93. If the slope is positive, as in 4.2, there is an increase in y for each increase in x. In the case of a negative slope, there is a decrease in y.

By taking a line as a model for the data, it is possible to abstract away some of the variation and get a clearer picture of the trend (if it is present). Since lines have known properties, it becomes possible to estimate (or map) from the observed cases to unobserved ones.

Linear regression

At its most basic level, a regression model is an attempt to capture a *response*, typically in the form of measurements on interval level data. The goal of regression modeling is to choose a systematic structure and the correct random variation that best matches the data we have observed (Faraway, 2005, 13):

(1) Data = Systematic structure + Random variation

The systematic components of the model should be chosen so that the systematic part explains the most (and the random part the least) of the model.

This is achieved by modeling the systematic component with a linear function, thus approximating the response y in the form of a line: the *regression line*. The regression model is then defined in terms of one or more *predictors* X and an error term ε. This

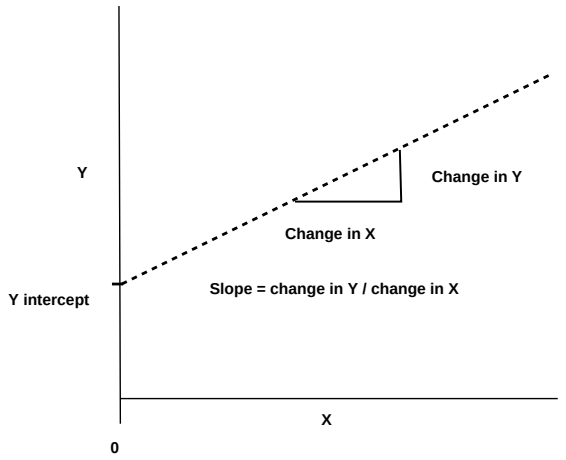


FIGURE 4.2: *An illustration of the intercept and slope of a linear function.*

entails finding a β, so that βX (the linear function) approximates the response y as much as possible: the estimated β, $\hat{\beta}$ ("beta hat"), is the best possible estimation of β within the space of the model (Faraway, 2005, 12).

$$\hat{y} = \hat{\beta}X + \hat{\varepsilon} \tag{4.12}$$

The simplest interpretation of the $\hat{\beta}$ is that changing a unit in any of the input variables (the X-part of the equation in (4.12)) will lead to a change in y (the response) of a magnitude of the *coefficient* $\hat{\beta}$. To take an example from Faraway (2005, 44): if we think of the response y as annual income, and X as representing the number of years of education, then $\hat{\beta}$ might (if the model is correct!) represent the predicted change in annual income for an individual with one extra year of education.

The difference between the response in the data, y, and the predicted response $\hat{y}$, is called *residual error*, or *residuals*, denoted $\hat{\varepsilon}$. We would like the $\hat{\varepsilon}$ part of the model to be as random as possible, that is, it should follow a (random) normal distribution, cf. Hinton (2004, 328).[11]

Now we can model y with e.g. two predictors using the general equation in (4.13), cf. Faraway (2005, 11):

[11] Additional requirements to the residuals are that they should have a mean of zero, add to zero, and be independent, cf. Hinton (2004, 328).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4.13}$$

In the equation above, $x_i,\ i = 1, 2$ are predictors, and β_0 is the intercept. $\beta_i,\ i = 1, 2$ are the unknown parameters of the model, i.e. what we are trying to estimate. ε is the error term or the residual error in the model, i.e. the part which cannot be explained by the predictors.

The input data typically comes with a number n of cases or observations, with measurements on the response y and the predictors x. In a tabular form, the data might look something like the example below, from Faraway (2005, 12):

$$\begin{array}{lll} y_1 & x_{11} & x_{12} \\ y_2 & x_{21} & x_{22} \\ \ldots & \ldots & \ldots \\ y_n & x_{n1} & x_{n2} \end{array}$$

There are as many rows as there are cases, and in the example above there are two predictors. Generally, we want more cases than predictors in the dataset to ensure correct estimation of the parameters (or *coefficients*) of the models. These coefficients represent the weighting of the predictors in estimating the response.

The model including the predictor data points can then be written as in equation (4.14).

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \; i = 1, \ldots, n \tag{4.14}$$

In geometrical terms, a regression model finds the line that best fits the input points. The linear model can be fitted to data points located in the space denoted by the x and y coordinates, by using the *residual values*, i.e. the differences (or distances) between the line drawn by the linear model and the data points – or to put it differently, the difference between the observed values and those predicted by the model. The best fit to the data is achieved by finding the model with the smallest summed squared residuals. This *least squares method* is fundamental to all the linear models considered here, including correspondence analysis (see below), cf. Faraway (2006, 117); Greenacre (2007, 47).

In the present study the regression models will be presented in the form of R code, in order to relate the predictors and the response to the data being analyzed in a more intuitive way. The R notation is introduced in section 4.6.2 below.

This exposition of linear regression has focussed on the conceptual aspects, rather than model checking and applications, since classical linear regression is not appropriate for most of the data under consideration in the present work. For this reason,

the technicalities of motivating and checking linear regression models will be passed over in silence. For more on linear regression, see Crawley (2005, 125–154); Faraway (2005) or Baayen (2008a, 165–240).

Generalized linear models

A *Generalized Linear Model* (GLM) is a generalization of a linear regression model to handle a wider set of problems.[12] While linear regression is suitable for continuous data, the GLM family can be used on continuous, count, and proportion data. The list below, taken from Crawley (2005, 113), summarizes some of the cases where GLMs can be used:

- count data expressed as proportions (e.g. logistic regressions),
- count data that are not proportions (e.g. log linear models of counts),
- binary response variables (e.g. dead or alive) or
- data on time-to-death where the variance increases faster than linearly with the mean (e.g. time data with gamma errors).

Thus, to define a GLM, it is necessary to specify an *exponential family* distribution for the response (y), and a *link* function which describes "how the mean of the response and a linear combination of the predictors are related" (Faraway, 2006, 115). Different exponential families are used for the response variable depending on the type of data. Typically, the *normal distribution* is used for continuous data, the *Poisson* (or in some cases the *negative binomial*) distribution for count data, and the *binomial* distribution for proportion data.

The underlying logic is still the same as for the ordinary least squares linear model: we assume that there is some shared, underlying probability of successes and failures in the material which constitutes a systematic structure, together with some random variation. The goal is then to specify the correct structure for the systematic part of the shared underlying probabilities using predictors, so that the remaining part of the probabilities are randomly distributed.

The logistic GLM

The most commonly used model in the subsequent chapters is the *logistic* or *logit* model, so this will serve as an example for the present discussion. A logistic regression

[12] Technically, the ordinary least squares linear regression described in the previous section is a special case of the generalized linear model.

model is defined in the same way as the linear model above, but with some further specifications:

(i) The response is a binary variable, usually referred to as "successes" and "failures";

(ii) the response has a binomial error distribution;

(iii) the response has a constant probability of success or failure;

(iv) the link function is the logit.

Some of these assumptions deserve further comments.

A binary variable can, as its name implies, take on one of two values, the classic example being coin flipping which can result in heads or tails. One such coin flipping – known as a "trial" – is a *Bernoulli outcome*, where one outcome is designated as success (conventionally labeled 1) and the other as failure (conventionally labeled 0), cf. Gill (2006, 339). Although the vocabulary, which has long historical traditions, may seem somewhat unfamiliar, this is in fact a very useful conceptualization of real world events, as noted by Gill (2006, 339–340), since many events may be classified as binary. Examples include but are not limited to: whether a bill is passed in a legislative assembly or not; whether a given district votes one way or another in a political system dominated by two political parties; whether the participants of an experiment make the correct or incorrect classification in a lexical decision task; etc. As the examples suggest, logistic regression is used extensively in political science, but also in linguistics and biology, for examples see Gelman and Hill (2007); Baayen (2008a); Crawley (2005).

A *binomial* probability function is the extension of the single Bernoulli case to multiple trials, cf. Gill (2006, 340). The binomial distribution is a model which can be used to analyze multiple Bernoulli experiments (Gill, 2006, 340). Using this model allows us to assign probabilities of success and failure to discrete nondeterministic events, i.e. events where the outcome is not predetermined.

Probabilities are somewhat problematic to work with directly, as pointed out by Baayen (2008a, 196): first, although probabilities are strictly bounded by 0 and 1, the functions used in regression modeling cannot incorporate this knowledge and may make estimates below or above these boundaries. Second, the variance of probabilities increases with the mean, but the functions used in regression assume a constant variance. Third, it is desirable to weight the result for the number of observations supporting the outcome. All of this can be achieved by using a *link function* which describes how the mean of the response is related to a linear combination of the predictors (Faraway, 2006, 115–116). For logistic regression, the link function is the *logit*, which

can be interpreted as log odds ratios. Formally, $\text{logit}(x) = \ln(x/(1-x))$ is a function which maps the restricted range [0, 1] to the continuous range [$-\infty$,∞] (Gelman and Hill, 2007, 80). For this reason, it is necessary to *back-transform* the results from a logistic regression to get them back on the probability scale from 0 to 1 again. The advantage of this approach is that it gives a continuous response which can take any real value, but the probabilities associated with the log odds ratio are restricted to their proper range, cf. Everitt and Hothorn (2006, 93). This satisfies the requirements of the GLM, while at the same time making it possible to model binary data.

In a binomial regression there are a number of link functions available, cf. Gelman and Hill (2007, 124–125), and according to Faraway (2006, 36) it is usually not possible to choose an appropriate link function based only on the data.[13] For the present study, only the default link function for the binomial family, the logit, has been considered. The reason is simple: only the logit coefficients have a reasonably intuitive interpretation (as log odds ratios), which is not the case for the other link functions with the binomial family, cf. Faraway (2006, 32). Thus, while some of the models might have been further optimized by exploring other link functions, the focus has instead been kept on ease of interpretation.

The numerical aspects of logistic regression will not be dealt with further here, see instead subsequent chapters on analysis of the corpus data. Instead, a dummy example will be described to give an outline of the basic idea.[14] Consider a situation where we have measured the habits of the people in an area regarding how they travel to the supermarket. The two options are to either drive or not to drive, i.e. a binary variable. We suspect that distance to the supermarket may be a contributing factor in their choice.

We can now think of the response as probabilities of either driving (success, or 1) or not driving (failure, 0) and fit a logistic regression model where the probability of driving is the response, and distance in meters to the supermarket is the predictor. Using the `glm()` function in R, this could look as follows:

(2) `glm(drive ~ distance, family = binomial)`

Figure 4.3 shows a plot of the inverse logit curve for this hypothetical example, i.e. the values have been back transformed to probabilities. Note how the logit line in the plot goes asymptotically towards 0 and 1 at low and high distances, respectively. The plot shows that people with short distances to the supermarket are likely to use other means of transportation, but as the distance in meters increases, driving becomes the predominant mode of transportation. We can see that somewhere between 500 and 600

[13] See the R help pages for *family* objects for details about which links are available.

[14] The data are based on an example found in Faraway (2006, 28), but heavily altered to fit the present purposes.

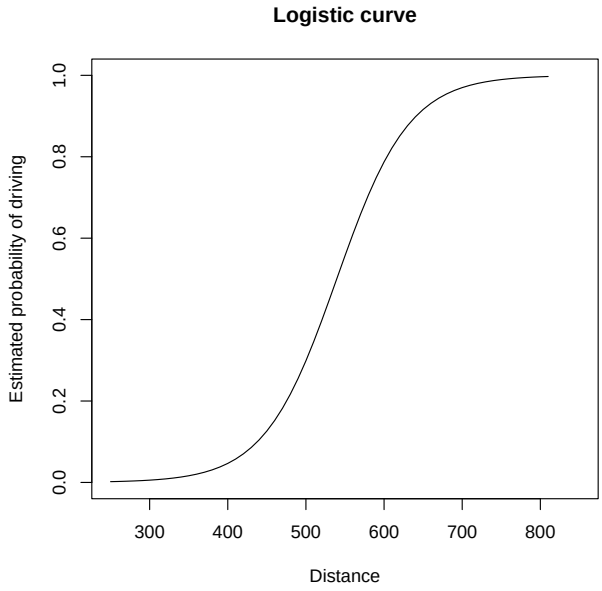


FIGURE 4.3: *A fitted logistic curve showing the probabilities of switching from walking to driving to the supermarket in a fictional example. The y-axis shows probability of driving to the supermarket, while the x-axis shows distance from home to the supermarket in meters.*

meters the probability is 0.5, that is, both modes of transportation are equally likely. At this point, we can estimate the *maximum likelihood* of switching from not driving to driving. Because of the nonlinear shape of the curve, the coefficients cannot be interpreted generally as probabilities. Instead, we must choose some specific point of the curve to interpret the model coefficients on a probability scale (Gelman and Hill, 2007, 81). For this reason, care needs to be taken when interpreting the coefficients from a logistic regression. Gelman and Hill (2007, 82) present a convenience rule of *divide by 4*: we can estimate the upper bound of the predicted difference in probabilities around the center of the curve by simply dividing the coefficients by 4. Subsequent chapters will deal with the analysis of logistic regression in more detail, including model checking and the interpretation of coefficients.

Using a binomial (logistic) regression on proportion data requires some assumptions to be fulfilled, cf. Crawley (2005, 117) and Faraway (2006, 26). Specifically, it

is assumed that the data can be expressed as proportions between 0 and 1; and that the probability of success for each trial is independent and constant. These assumptions are potentially problematic. Take the assumption of strictly bounded proportions between 0 and 1. This entails that we know both how many times an event occurred and how many times it did not occur (as opposed to a Poisson GLM where we only know how many times an event occurred). However, we only know this relative to the sampling frame. For example, the number of observations of *there* expressed as a proportion of all observations (which must also be defined somehow, i.e. a non-technical research problem) can only form the basis of a generalization from a binomial-logistic GLM if we are convinced that no systematic bias has influenced the proportions in the sample. The assumptions that the variance increases with the mean and is independent and non-constant are linked to the use of the binomial distribution as a model for the response. Since the model assumes that the probability of success is both independent and constant, a violation of either of these conditions will lead to a lack of fit.

Mixed effects models

Classical regression and GLMs specify a single model to fit the whole dataset. The reasoning behind this, as explained above, is to abstract away variation to study the main tendencies. However, the variation in the dataset is often systematic depending on membership in some group. For instance, the test score of 10 year old pupils in a single school district might vary systematically by school, so that one school has a higher mean score than the others, etc. Classical regression and GLMs would pool all the responses (i.e. the test results) in the school district and fit a single model. Sometimes this is fine, but we might be interested in the group level variation for its own sake. A *mixed effects model* is able to accommodate the group level variation and the individual level responses at the same time, that is, it can provide the same information as a GLM, but with added information about systematic group variation in the data.

Because of the ability to incorporate group level information, mixed effects models are also known as "multilevel models", or "hierarchical models". Mixed effects modeling has been used for some time with other types of tests, see e.g. the discussion in Cohen (1976). According to Gorard (2003, 215), the major change over the past 15 years has been in the advances of available software, rather than statistical breakthroughs. Unlike GLMs, mixed effects models make a distinction between *fixed* and *random effects* (hence yet another name these models are known by: "random effects models"). Faraway (2006, 153) defines a fixed effect as "an unknown constant we are trying to estimate from the data". He defines a random effect as a random variable such that we "try to estimate the parameters that describe this random effect".

Essentially, the fixed effects of the model are the predictors discussed above. The

random effects are group level constants, which are assumed to come from a normally distributed population of such constants. To stick with the school example, the parents' level of education might be a good predictor (i.e. fixed effect) for pupils' test results in all schools, but switching from one school to another might systematically increase (or decrease) the mean of the group.

Gelman and Hill (2007, 246) discuss three scenarios where multilevel modeling is particularly useful:

- accounting for individual and group-level variation in the same model;
- accounting for variation among individual-level regression coefficients;
- estimating regression coefficients for specific groups, even with small sample sizes.

The strength of multilevel model approaches lies in their ability to handle complex correlations and non-independent data, i.e. pseudoreplication. For example, all kinds of data pooling constitute an intended loss of information, in order to get a clearer picture of the data. A classical regression analysis attempts to fit the same model for the whole data set, which may cause problems. Hinton (2004, 327), in his presentation of GLMs, points out that models need to fit the data all the way, i.e., the regression line should be an equally fair model of the data anywhere along line. If this is not case, the result will be a poor model fit. Multilevel models can account for the pooled data (means, group level trends) while at the same time including variation at the individual level.

Assumptions regarding data and error distributions are mostly as for GLMs in general. They can be fitted as multilevel models, i.e. as *Generalized Linear Mixed Models* (GLMMs). Diagnostic checks for goodness of fit are also, mostly, as for GLMs (see below): the model assumes constant residual variance.

The great advantage of multilevel models for the current research project is that they can be used to neutralize the effect of non-independent data. It is assumed in GLMs that observations are independent, and if this assumption is violated the model might run into problems. Examples could be spatial non-independence (often an issue in biology), where observations occur in geographical clusters. Perhaps more pertinent in historical linguistics is the temporal non-independence of the observations: we know that a manuscript composed in, say, the mid 11^{th} century was written prior to one composed in the mid 12^{th} century, something which violates the classical assumptions of random sampling. In a GLM, the likely result is overdispersion, or other indications of a poor fit, due to pseudoreplication. In a multilevel model, however, the temporal non-independence can be included in the model by including a group level effect for (in this case) the time period. The specifics regarding the interpretation of multilevel models will be returned to in the data analysis in the following chapters.

Criticisms

Gorard (2003, 216–219) warns against the uncritical use of mixed models. Specifically, he challenges the real advantages of including group level variables uncritically, since the true value of mixed models are only apparent when the auto-correlation is cluster-random. This implies a situation with clearly defined clusters from which cases are sampled at random, but excludes convenience sampling and opportunity clustering, i.e. clustering and selecting cases because it is a convenient way of collecting and organizing data. Furthermore, he warns that mixed models for some seem to represent an alternative to good research design:

> There is a danger ... that [mixed effects modeling] is seen as a kind of 'magic bullet', uncovering causation, overcoming poor design and allowing researchers to draw robust conclusions from poor datasets. It is nothing of the sort ... It is simply a useful technique for specific situations.
>
> Gorard (2003, 217)

The simple point advocated by Gorard is compelling. Rather than relying on ever bigger and more advanced statistical hammers to hit our problems with, the real advances in data analysis are bound to lie in increased attention to measurement variables, better sampling techniques, and combinations of observational and experimental data.

However, this desired attention to design of studies does not invalidate the value that mixed effects modeling can bring. All statistical modeling, even simple chi-square tests of independence, require some simplifying assumptions to work. It is not clear why the – in principle – comparable simplifying assumptions of mixed effects models are less desirable than those of more standard null hypothesis tests. The *practical* difference is of course as, Gorard (2003, 222) points out, that the increased complexity of mixed effects models might mislead both researchers and readers. The researcher, as pointed out above, might fall into the trap of mistaking mixed effects models for a cure-all for all manner of flaws regarding the design of the study, while at the same time providing a causal explanation. Conversely, the reader might easily be baffled by technical jargon into thinking that "anything so 'clever' must be OK" (Gorard, 2003, 222). Presenting a parsimonious summary of mixed effects models without lapsing into merely listing coefficients and other technicalities requires a fine balance between interpretation and giving the reader enough information to judge the results.

Nevertheless, the real value of mixed effects models should not be underestimated, *pace* Gorard. In practice, a mixed effect model will often give an acceptable fit to the

data in cases where ordinary GLMs suffer from overdispersion or other symptoms indicating a lack of fit to the data. Since the interpretation of a regression model hinges on how well it fits the data, this is a real advantage, even if the aim is purely descriptive. Furthermore, if the aim is to use statistical "mumbo-jumbo" as a rhetorical device, this can be achieved equally effectively through null hypothesis tests and p-values. The potential for misleading use is in other words in no way tied exclusively to mixed effects models, but is a *general* problem which always must be kept in mind when dealing with statistics. Similarly, the temptation to use statistics as a quick fix for what are really measurement and sampling problems is not exclusively tied to mixed effects models. Using a simple chi-square test on a large table with unclear measurement variables and bad sampling is essentially doing the same thing. Finally, when it comes to causality, this is again not a specific problem with mixed effects models. The problem of mistaking correlation for causation is a general caveat for statistical approaches, cf. Gorard (2003, 155–159); Goldthorpe (2001).

As long as the researcher has a clear understanding of the limitations of a given statistical method and is able to provide parsimonious interpretations of the results, there are no special problems connected with mixed effects models compared with other types of statistical testing. Proper sampling and data collection, as well as explications of the research questions in relation to the measurement variables are *always* the responsibilities of the researcher. Even if the mixed effects models are conceptually challenging, the same problems manifest themselves with a chi-square test if the researcher has not properly grasped the logic behind, and limitations of, the test (the calculations are in either case done by computers). There is no principled difference between a researcher who attempts to fix design flaws with a mixed effect model and one who tries to fix them with a chi-square test. In either case the result is bad quantitative analyses, even if the technicalities of the tests might be correct. Thus, the criticisms presented in Gorard (2003) are not really good counterarguments against the use of mixed effects models. Rather, they apply to any statistical test or model, and the central message – the importance of good research design and sampling – applies universally, irrespective of the choice of test.

4.6.2 A note on R

This section briefly deals with some R-specific issues: the choice of function for the multilevel models and the procedure for model fitting. Although these are software-related issues, they are nevertheless of interest as they cannot be entirely separated from the output of the models and the interpretation of them. Linear regression and GLM models can be fitted in R without adding further libraries. However, the most widely used library for fitting multilevel models is probably the *lme4* library from

Bates, Maechler, and Dai (2008). For the present study, the `lmer()` function from *lme4* is the preferred function for fitting multilevel models.

Notation

The following symbols are used when presenting various model formulas from R:

- The *tilde* symbol: `y ~ x` reads "model y as a function of x"
- The *pipe* symbol: `x | z` reads "group x by z"
- The *colon*: `x : z` reads "let x and z interact"
- The *plus sign*: `y ~ x + z` reads "model y as a function of x and z"
- The *asterisk*: `y ~ x * z` reads "model y as a function of x and z and the interaction between x and z" (shorthand for `y ~ x + z + x : z`)

Model fitting

Crawley (2005, 234–237) recommends fitting a saturated model, i.e. a model with as many parameters as there are response values, and then gradually remove the non-significant interactions to reach the minimal model. A maximal, saturated model involves all two-way interactions between predictors, all three-way interactions, and so on until all *n*-way interactions have been fitted. Crawley's point is well argued; however, with large data sets this quickly becomes a problem due to R's memory limitations. According to R's help files, the memory limit of R is set to 2 Gb on most 32-bit Windows systems, and it cannot exceed 3 Gb. This proved to be a problem when attempting to fit some saturated models for the present project, since some of them would lead to models taking up 7 Gb or more of memory, according to R's error messages.

Fortunately, this is not the only opinion on the topic. Gelman and Hill (2007, 547) recommend starting with a simple model and expanding it. Ultimately, this is a question of preferences combined with such real-life constraints as computer memory and data; however, the ethical aspects of the discussion apply in both top-down and bottom-up modeling: fitting data requires care so that potentially misleading results are avoided.

4.6.3 Model diagnostics

The following sections provide some considerations on choosing and assessing models with GLMs and GLMMs. The treatment here is not exhaustive, nor is it a full introduction. Instead, two approaches that will be used later are exemplified and discussed.

For a very readable general introduction to model diagnostics with linear models see chapter four of Faraway (2005).

Numeric measures

We can broadly distinguish between two types of goodness of fit, *calibration* and *discrimination*, cf. Harrell, Lee, and Mark (1996). The former refers to the bias in the model, while the latter refers to accuracy. To take an example from Harrell et al. (1996, 366), a forecast which predicts a mean chance of rain of 0.15 for a given area may be well calibrated if the annual number of days with rainfall is 55 on average. In other words, calibration is an expression of the model's ability to distinguish between different responses. However, a daily forecast which predicts a 0.15 chance of rain every day is obviously completely uninformative. Discrimination refers to the ability to accurately predict *a given response*, i.e. predicting rain only or mostly on the days when it actually does rain.

A widely used measure of calibration is R-squared or R^2. This measure, also known as the *coefficient of determination*, is most intuitively related to simple least squares regression models. The basic R^2 is a coefficient bounded by 0 and 1, which is typically reported as a goodness of fit statistic for ordinary least squares regression models. It measures the sum squared differences between observed and predicted responses y, divided by the sum of the squared differences between the observed responses and their mean value. This measure can be interpreted in several ways. One interpretation is that the coefficient gives the total amount of variability in the data explained by the model. Another one is that R^2 measures the improvement of the regression model over a *null model*, a model which only guesses the mean value.

However, unlike least squares regression, model fitting with GLMs such as logistic regression is done in an iterative manner to produce a maximum likelihood, not to reduce variance. For this reason, the normal least squares R^2 is not applicable to GLMs or GLMMs. Instead, a number of *pseudo* R^2 measures have been developed. Here the case of logistic regression will be considered, in the form of *Nagelkerke's* R^2. A good basic introduction to such measures can be found on-line,[15] but they are also briefly discussed in Long and Freese (2001, 84–86). A technical definition of Nagelkerke's R^2 is given in Nagelkerke (1991).

As discussed in Nagelkerke (1991), this R^2 has many of the same properties as an ordinary R^2: it is independent of the original scale of measurement, it is approximately (not exactly, as in the least squares case) bounded between 0 and 1, and it is

[15] See e.g. UCLA Academic Technology Services' FAQ on pseudo R^2 measures, http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm.

interpretable as the improvement of the fitted model over a null model which only gives the most frequent outcome.[16]

R^2 measures are surrounded by a certain amount of scepticism. Long and Freese (2001, 119) explicitly state that they do not take a high pseudo R^2 to be evidence of the "best" model. More generally, they discuss a wider class of numeric measures of fit[17] Long and Freese (2001, 83–87), taking care to point out that although such measures can be useful

> there is no convincing evidence that selecting a model that maximizes the value of a given measure results in a model that is optimal in any sense other than the model having a larger (or, in some instances, smaller) value of that measure.
>
> Long and Freese (2001, 80)

Thus, although such measures will be referred to in the subsequent analysis, they will not be used exclusively. The form of the model will be considered against the research questions to be answered and, crucially, extensive use will be made of graphics for model checking and fitting. (This topic is dealt with in the following section.)

As indicated above, it is also necessary to take into account model calibration. A simple way of doing this is taking the squared correlation between the response and the predicted values, cf. Johnson (2008, 237–240). A more principled approach is advocated in Harrell, Lee, and Mark (1996) and Baayen (2008a, 204). This approach amounts to doing a rank based test (Kendall's τ) on all pairs of responses and predicted values. Two such widely used (and related) measures are the *C* index and *Somer's* D_{xy}. In the former case, 1 denotes perfect prediction and 0.5 denotes randomness, while in the latter case 1 denotes perfect prediction and 0 denotes randomness. See Baayen (2008a, 204), Harrell, Lee, and Mark (1996), and Newson (2002) for further technical details on these measures.[18] The important point defended in Harrell, Lee, and Mark (1996) is that while a model might perform well at the group level in predicting different responses, it is also important to know the accuracy of the model, i.e. whether cases predicted to have property A in fact do so. In many ways, such a calibration measure serves as a corrective for discrimination measures, cf. Harrell et al. (1996, 366–367).

[16] An R script for calculating Nagelkerke's R^2 for mixed logistic models written by T. Florian Jaeger is available from http://hlplab.wordpress.com/2009/08/29/nagelkerke-and-coxsnell-pseudo-r2-for-mixed-logit-models/.

[17] Such as log-likelihood measures, and information measures such as AIC and BIC.

[18] Calculation of these measures is done in R with the `Hmisc` library, cf. Harrell (2009).

Plots

An important diagnostic tool for regression model-checking is making plots of the key assumptions the models are based on: constant variance and normal residuals, in addition to checking for outliers and overly influential observations. A visual approach to model-checking is recommended by e.g. Cameron and Travin (1998) and Faraway (2005), and while looking at plots of residuals can be uninformative when one is unaccustomed to them, they quickly gain in usefulness when enough plots have been examined. This is a more empirical approach to model-checking than simply relying on a statistical test which "may have a reassuring aura of exactitude about it, but ... may be powerless to detect problems of an unsuspected nature" (Faraway, 2005, 58). Faraway's views are consistent with those of Cameron and Travin (1998, 140) who write:

> Residual analysis, particularly visual analysis, can potentially indicate the nature of misspecification and ways that it may be corrected, as well as provide a feel for the magnitude of the effect of the misspecification. By contrast, formal statistical tests of model misspecification can be black boxes, producing only a single number that is then compared to a critical value. Moreover, if one tests at the same significance level (usually 5%) without regard to sample size, any model using real data will be rejected with a sufficiently large sample even if it does fit the data well.

The problem is not only tied to the nature and interpretation of misspecifications, but also to the detection itself, as Faraway (2005, 60) points out:

> After all, with a large dataset, even mild deviations from nonnormality may be detected, but there would be little reason to abandon least squares because the effects of nonnormality are mitigated by large sample sizes. For smaller sample sizes, formal tests lack power.

Thus, relying on plots to check the model fit achieves two goals: first, it avoids any spurious sense of statistical precision regarding the model's fit. Second, it avoids the problem of significance (i.e. deviance from the theoretical model) based purely on large amounts of data; which also opens up for checking model assumptions in large data sets: a standard, recommended test for normality such as the Shapiro-Wilks test cannot handle more than 5 000 data points in R.

However, reading such graphs can be a bit of a challenge before one is used to them, and Faraway (2005) recommends generating plots from randomly generated data with known properties as a form of calibration. Figure 4.4 on page 108 shows examples of randomly generated data from four different distributions following Faraway (2005,

59), plotted with a quantile-quantile (Q-Q) plot[19] function in R. Note that the function produces a linearly transformed fit to the Normal distribution, instead of the more well-known bell-shaped version.[20] The four distributions are:

i) Normal

ii) Lognormal: A skewed distribution

iii) Cauchy: A long-tailed distribution

iv) Uniform: A short-tailed distribution

Only normal-shaped residuals indicate a good fit, however, the severity of the problems with the other three varies, cf. Faraway (2005, 59–60). With a uniform, or short-tailed distribution the nonnormality can safely be disregarded since this is not a particularly serious deviation. For skewed errors the problem needs to be addressed, but it can often be solved by some kind of transformation of the response variable, typically a logarithmic or square root transformation, although other transformations are also possible cf. Gelman and Hill (2007, 65–66). With long-tailed errors, a number of techniques such as bootstrap or permutation tests could be used, however, a good alternative according to Faraway (2005, 60) is simply to use "robust methods, which give less weight to outlying observations".

Authors differ in their opinions about the severity of non-normal error distributions. Gelman and Hill (2007, 46) state that normality of errors is probably the least important assumption, and they specifically "do *not* recommend diagnostics of the normality of regression residuals" (original emphasis). Both Crawley (2005) and Faraway (2005) are more positive to the use of normality checks of errors.

Ultimately, there are no definitive answers to interpreting a quantile-quantile plot, and the final judgment must rest on a combination of experience (seeing a lot of plots), sound knowledge of the data and a clear understanding of what the model is intended to be a model of.

For GLMs with non-normal responses the errors are not expected to be linear anyway, but, as noted by Faraway (2006, 129), it is still of interest to look for influential observations and outliers.[21] Such diagnostics of GLMs can be carried out

[19] A Q-Q plot is used to compare two random distributions by dividing them into quantiles, i.e. equally sized portions, and plotting them against each other. Similarity will manifest itself as an approximately straight line.

[20] However, as Gelman and Hill (2007, 14) point out: "Linearly transformed normal distributions are still normal".

[21] An *outlier* is an observations which does not fit the model well, whereas an unusual observation is a point which "changes the fit of the model in a substantive manner" (Faraway, 2005, 64).

with a half-normal plot. In the present work, half-normal plots are generated with the `halfnorm()` function from the *faraway* library, cf. Faraway (2009). The half-normal plot, as its names suggests, is related to the normal plot, but employs only the positive half of the normal curve. Atkinson (1981, 15) suggests that half-norm plots are more effective in displaying large outliers than normal plots.

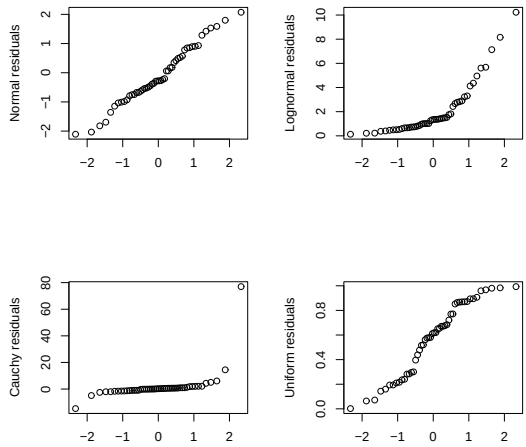


FIGURE 4.4: *Randomly generated data from four different distributions: Normal, Lognormal, Cauchy, and Uniform. These are examples of possible shapes of residuals from a regression model. Only an approximately normal distribution indicates a good fit, the other three indicate some kind of problem with the model fit, such as outliers and extreme cases.*

When it comes to equal variance, this assumption can be checked by plotting residuals against fitted values. Anything but a constant variance indicates a problem with the model, but the problems can be of different types and magnitudes. Typical problems involve some degree of nonconstant variance (heteroscedasticity), which can be dealt with using some kind of transformation of the response, or nonlinearity, which indicates that some kind of structural change in model is required, cf. Faraway (2005, 53–54). Following a procedure from Faraway (2005, 55), figure 4.5 on page 109 shows randomly generated residual vs. fitted plots with examples of constant variance, strong and mild heteroscedasticity and nonlinearity.

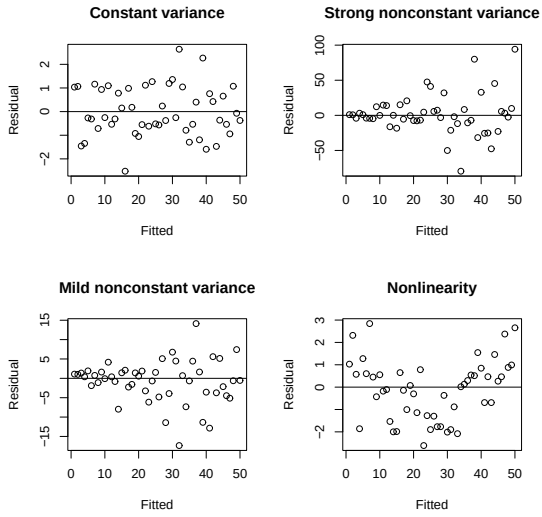


FIGURE 4.5: *Randomly generated data illustrating four different scenarios: Constant variance, strong nonconstant variance, mild nonconstant variance and nonlinearity. Only constant variance indicates a good model fit to the data.*

As mentioned above, the residuals are the difference between the response in the data (i.e. the y-part of the regression equation) and the *estimated* response from the model. The fitted values are the estimated values.

In the case of a diagonal streak in the plot, this is usually caused by a large number of zero values in the response variable. This can safely be ignored, cf. Faraway (2005, 156).

4.6.4 Summary: regression analysis

Regression modeling is a powerful tool for data analysis, and the default option in many observational sciences where some kind of explanation of observed data are sought. The strengths of regression models, especially with GLMs and multilevel models, lie in their ability to handle a whole range of data types and their flexibility in modeling the contributions of – and interactions between – a number of different predictors. As with other statistical models, however, there are pitfalls. Regression models rely on

specific assumptions about error distributions and independent data just as much as other statistical methods. However, these problems are, at least in principle, solvable: input variables and responses can be transformed to a different scale to meet the assumptions of the model if one of the many distributions available for GLMs are not appropriate, and introducing group level effects in multilevel models is a reasonably simple workaround fix for the independence issue. There is of course the inevitable danger of interpreting association and causation, but as with all statistical testing this is really the responsibility of the researcher, not the test.

The main practical *disadvantage* with regression modeling is perhaps the seemingly daunting effort required to set up, evaluate, and interpret the models. However, I prefer to look at this differently, seeing the effort required to set up, evaluate, and interpret the models as the main *advantage* of regression models. As pointed out in the previously cited quote from Fleiss et al. (2003, 50–51) and highlighted in the preceding discussion, the seductively simple use and output from a Pearson chi-square test is far more complicated and problematic than it might appear at first glance. Thus, there is a double advantage in regression modeling: in addition to the more detailed and useful information it can provide, the model specification also keeps the researcher acutely aware of the assumptions of the tests and the conditions for interpreting and reporting the results. Once the full implications of this for data analysis are appreciated, the effort of getting to know regression modeling seems like a small investment compared to the rich rewards it can yield.

4.6.5 Correspondence analysis

Before leaving the topic of linear models, a short introduction to correspondence analysis is in order. Often presented as a non-parametric statistical technique which is suitable for data exploration without making too many assumptions about the data, correspondence analysis is a useful addition to regression analysis.

Correspondence analysis (CA) first found its use within the French sociology tradition, following initial work by linguist and data-analyst Jean-Paul Benzécri who began working on the method in an effort to solve philological problems, cf. Apollon (1990, 195–197) and Nenadic and Greenacre (2007). The method is widespread in many disciplines dealing with categorical data, and Baayen (2008a, 128–136) illustrates its usefulness in linguistics. The brief description here is intended as a background to analyses found in subsequent chapters. For a more in-depth introduction to the technical aspects of CA, see Greenacre (2007). There are a number of ways to do a CA analysis in R, see e.g. Baayen (2008a, 128–136) and Faraway (2006, 75–79). In the present dissertation, the CA analyses have been performed with the *ca* library, cf. Greenacre and Nenadic (2007).

CA in brief

As mentioned previously, there are certain similarities between CA and regression in that both types of techniques employ least squares methods to fit lines to data points. However, as Greenacre (2007, 47) explains, there is a fundamental difference in that regression analysis treats one variable as a response, whereas in CA there is no response. Instead, the model fit in CA is done by reducing all variables to lower linear dimensions and minimizing the variation perpendicular to the dimension (or direction) being fitted, so that the first dimension accounts for the maximum variation in one direction, the second dimension accounts for the maximum variation in another direction and so on.

CA is part of a larger class of methods for matrix analysis, including factor analysis, principal component analysis, multidimensional scaling and others, cf. Greenacre (2007, 201–203); Baayen (2008a, 118–138). The specific techniques are suited for different data levels, and CA is the method developed for counts of categorical (i.e. nominal) data. The aim in CA is to *reduce* variation without too much loss of information, which is very useful, since the result can be presented in the form of a plot, and CA is primarily a *visual* method for data analysis. Simply put, CA is "a method of data analysis for representing tabulated data graphically" (Greenacre, 2007, 1).

The output from a CA analysis can be represented as a plot in two dimensions (or sometimes three dimensions), where the data are represented as points relative to two (or three, in the case of a three-dimensional plot) axes or dimensions. Such a plot creates a "spatial map of the data", (Greenacre, 2007, 1), which gives us the *approximated* distances between the row and column variables in the data. Thus, CA is an excellent tool for exploratory data analysis. What sets CA apart from a simple scatter plot of the data matrix, is that CA reduces the matrix to a subspace of the original space but with a lower number of dimensions, whereas a scatter plot represents all variability in the data.

A detailed discussion of the mathematics of this method falls outside the scope of the present dissertation. However, a brief consideration is useful to appreciate the output of such an analysis. All the matrix analysis methods mentioned above rely on a classical technique in matrix theory: *singular value decomposition* (SVD) (Greenacre, 2007, 201). The goal of CA is to use SVD to reduce the original matrix to a subspace (that is, fewer dimensions) that can best represent the original matrix. The maximum number of dimensions required to represent all the variation in the matrix equals the number of columns. CA is based on the χ^2 statistic for the matrix, but instead of using this statistic to measure dependence (as in the classical Pearson chi-square test), it is used to measure *distance* from a given cell to its row mean (Greenacre, 2007, 27). By modifying the χ^2, cf. Greenacre (2007, 25–32), the χ^2 can be used to measure the dis-

tance between rows and columns of the matrix. Thus, row and column associations can be expressed as distances: the closer the row and column points are to their averages, the lower the association between rows and columns.

A CA example: hair and eye color

To give an example of CA, the data data collected by Snee on the correlation between hair and eye color, presented in Faraway (2006, 75), are useful. The original data are presented in table 4.4 on page 112, giving frequencies of the eye colors green, hazel, blue, and brown for people with black, brown, red, and blond hair. A test of significance (like a Pearson chi-square) on this data set would merely indicate that some eye and hair colors are more frequently encountered than others, and as pointed out by Faraway (2006, 76), this is something we already know. What we would like to know is the relationship, or association, between *specific* eye and *specific* hair colors.

TABLE 4.4: *Frequencies of hair and eye color, from Faraway (2009).*

	green	hazel	blue	brown
BLACK	5	15	20	68
BROWN	29	54	84	119
RED	14	14	17	26
BLOND	16	10	94	7

Figure 4.6 on page 113 shows a plot based on a CA analysis of the data in table 4.4. To interpret the plot, we start with the x-axis and look at the distance between the points and the center of the plot (i.e. the origin). Points near the center are more expected (and tend to have less explanatory value) than the points near the edges of the plot, since the plot is created by allowing the points to contribute unequally to the axes. This procedure is then repeated for the y-axis (and for any further axes that need to be examined). In the hair/eye color plot shown in figure 4.6, the x-axis is defined mainly by the opposition between black and blond hair, whereas brown hair is near the origin and contributes much less to the dimensions of the plot.

The CA plot provides us with information about which hair and eye colors are associated with each other, and also which are negatively associated. On the horizontal (x) axis, "blond" and "BLUE" are far from the origin and close to each other. Thus, there appears to be a positive correlation between these hair and eye colors. If a row and column point are situated diametrically apart from each other on either side of the

origin, the points are negatively correlated (Faraway, 2006, 78). There appears to be a negative association between blond hair and hazel eyes.

The plot in 4.6 is a two-dimensional projection of the original subspace analysis. How much of the variation in the full matrix does this this two-dimensional plot represent? The horizontal axis in this case represents 89.4% of the variation, whereas the vertical axis represents 9.5%. Thus, a two-dimensional plot in this case gives a very good picture of the variation in the data, by explaining 98.9% of the variation. This can be compared with the theoretical maximum number of dimension required to represent the matrix, which would equal the number of columns i.e. four. Obviously, it is not possible to represent four dimensions in a graphical display, so a reduction to two dimensions greatly assists us in the interpretation of the plot.

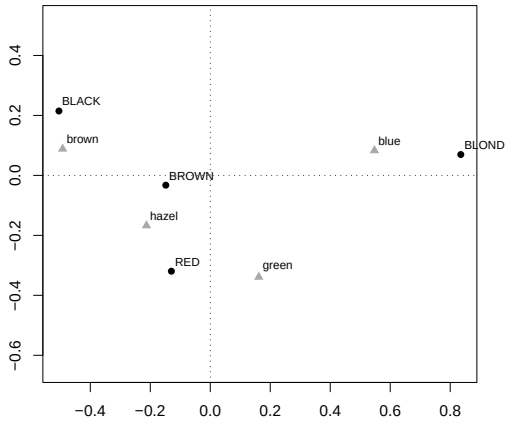


FIGURE 4.6: *CA plot of the correlation between hair color and eye color, reproduced from Faraway (2006, 78). Atypical observations in the two represented dimensions lie far from the origin. The x axis represents 89.4% of the variation in the matrix, the y axis represents 9.5%. The cumulative variation accounted for by the two dimensions in the plot is thus 98.9, or virtually all the variation.*

This is not the full story of CA, however. A number of pieces of information need to be assessed, such as the total inertia of the matrix (that is, the overall association between rows and columns), the contributions to the plot made by each point, and

questions regarding scaling the plot. These aspects of CA will be further discussed in conjunction with the analyses presented in later chapters, but see e.g. Greenacre (2007); Greenacre (2006); Greenacre (1994); Blasius (1994). For now it will suffice to state that although CA is a primarily visual analysis technique, there is an important numerical aspect to it; that is, CA is more than merely "looking at plots", cf. also the comments in Blasius (1994, 35).

CA: summary

One of the strengths of CA is that it is more robust to differences in sample size than e.g. a Pearson chi-square test. Furthermore, a CA analysis can be used in situations where the assumptions of the chi-square test are not fulfilled, as is the case when the observations are not independent of each other and we have a pseudo-contingency table, cf. Greenacre (2007, 77–78). CA proves useful as an alternative to regression models in cases where we are interested in associations between multiple categories which cannot easily be fit to the "response – predictor" formula of regression models.

4.7 Summary

In this chapter I have reviewed some common statistical tests used in linguistics and discussed their strengths and weaknesses. Necessary amendments, such as the use of effect size measures, have been illustrated. The importance of taking the background assumptions of the tests seriously, even for ordinary linguistic consumers of statistical tests, has been stressed. Some advanced models, notably GLMs and multilevel regression models have been discussed, together with correspondence analysis. It is argued that these approaches are superior to many previously employed tests and that their increased demand for conceptual understanding of the models is a benefit, which ensures the correct use of these models.

Chapter 5

Data

The primary aim of this project is to investigate syntactic change in early English. Much of the relevant material is available in tagged and parsed form, in three manually annotated treebanks. The Old English material taken from the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE), cf. Taylor, Warner, Pintzuk, and Beths (2003). The Middle English material is taken from the *Penn-Helsinki Parsed Corpus of Middle English, second edition* (PPME2), cf. Kroch and Taylor (2000), and the Early Modern English material from the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPEME), cf. Kroch, Santorini, and Delfs (2004).

5.1 The treebanks

Below follows a short technical summary of the three treebanks used for the study. For a more thorough description, see the respective accompanying official websites. For each treebank there is also a url where the processed datasets which form the basis of the current investigation can be downloaded from.

5.1.1 YCOE

Size: 1 449 722 words. The prose part of YCOE consists of the entire set of known Old English prose texts. The processed dataset is available from
http://home.hib.no/ansatte/gbj/data/oe.txt.

5.1.2 PPME2

Size: 1 155 965 words. Comprises 55 prose text samples, based on the diachronic part of the Helsinki corpus, but with "considerably larger" text samples, cf. Kroch and Taylor (2000).
The processed dataset is available from http://home.hib.no/ansatte/gbj/data/me.txt.

5.1.3 PPEME

Size: 1 794 010 words. Consists of the prose part of the Early Modern English part of the Helsinki corpus and two supplements. The processed dataset is available from http://home.hib.no/ansatte/gbj/data/eme.txt.

5.2 The structure of the data

All the three historical corpora are tagged according to the same criteria, with minor differences. Some of these differences are discussed below. For a full overview of the differences, see the corpus documentation or http://www.ling.upenn.edu/histcorpora/annotation/differences.htm. The annotation scheme is loosely based on a flat-structure Government and Binding approach to syntax.

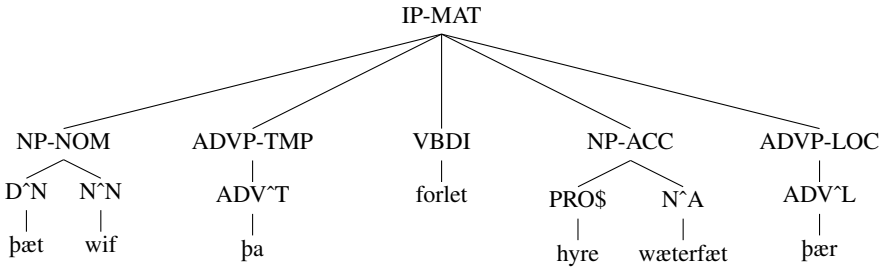
An example from YCOE is given in (1) below:

(1) þæt wif þa forlet hyre wæterfæt þær
"That woman then left her water-pot there'"

The corpus representation of tokens is in the form of a phrase structure grammar (PSG) format. Words and phrases constitute nodes in a tree structure, as illustrated by the corpus format of (1):

```
( (CODE <T03400002100,64>)
  (IP-MAT(NP-NOM (D^N +T+at) (N^N wif))
            (ADVP-TMP (ADV^T +ta))
            (VBDI forlet)
            (NP-ACC (PRO$ hyre) (N^A w+aterf+at))
            (ADVP-LOC (ADV^L +t+ar))
  (. ,))
(ID coaelhom,+AHom_5:64.722))
```

The tagged corpus example above corresponds to a straightforward shallow syntactic tree as illustrated below:



In the tree, the corpus codes for OE characters have been replaced with the actual characters: "+a" corresponds to æ and "+t" corresponds to þ (similarly, "+d" corresponds to ð and "+g" corresponds to ȝ).

5.3 Extracting data

No current off-the-shelf solutions currently exist for automatic extraction from these corpora. Although free software to search for entire tokens that match a given query exists, this represents a different search philosophy and is not easily compatible with a quantitative analysis of large corpora. Consequently, the problem of extracting the data and outputting it in a suitable format is a non-trivial one. Extracting and handling the data for the current study was done using three different programs, CorpusSearch 2.0, Perl and R.

5.3.1 CorpusSearch 2.0

Despite the recency of YCOE, PPME2, and PPEME corpora (all three were released between 2000 and 2004), their format are unfortunately not well suited for automated extraction of data. Instead, they are designed to be used with CorpusSearch 2.0, a command line based Java program written by Beth Randall to search these corpora.

The philosophy behind CorpusSearch (hereafter "CS") is to search all trees in the selected treebank files that match a given query. An example query is given below, searching all IP nodes for the existence of all forms of the verb *be*:

(2)
```
node:  IP*
query:  (BE* exists)
```

Such query files must be given as input to CS, which, having executed it, will close

and needs to be restarted for subsequent queries. The output is in the form of a text file containing all trees that match this query. In other words, CS provides a subset of trees, but presents them in their original format. This is a search-philosophy which lends itself well to example-driven studies where the corpus is primarily a source of examples used to back up the linguistic argumentation. It is also faithful to the original data format, which in some respects is a good thing. However, it is unsuited for extracting and processing large amounts of corpus tokens and preparing them for statistical analysis. For such an approach a different search philosophy is needed: instead of searching for a subset of trees, each tree needs to be parsed, analyzed and transformed into a format suitable for processing in a statistics program.

Although the output from CS is a file containing those tokens that matched the search terms, e.g. all trees with a locative adverb, some qualification is necessary here due to the way CS carries out a search. A CS search file might look as below (which is the one used to search for locative adverbs in YCOE):

(3)
```
node:  IP*
query:  (ADV^L exists)
```

The code in (3) tells CS to look for the tag "ADV" in any "IP" node, where "*" is a wildcard in the CS regular expression language. This means that CS will look for all "ADV" tags contained within IPs. As explained in the CS Lite documentation, this is mainly an issue when making use of the summary statistics produced by CS, which is not the case for this project. However, a query like the one below does produce a subtly different output:

(4)
```
node:  ADV^L
query:  (ADV^L exists)
```

With node set to `ADV^L`, CS finds 9213 tokens, whereas with node set to IP it finds 9203 tokens. The difference is accounted for as follows: With node set to ADV^L, CS finds all tokens that contain this tag, whether the token itself is tagged as an IP or not. The ten tokens that constitute the difference between the two searches are all tagged as "X" in YCOE, that is, they are considered empty or problematic and not conventional IPs. In the present study only tokens tagged as IPs have been included.

According to the documentation, CS can produce output which can be read by programs such as SPSS, Varbrul and others. I was not able to successfully make use of this feature, and the procedure only produced a blank output. Similar problems have been reported by others, e.g. Roy (2006, 6). Instead, Perl was used for formatting the

tokens to a form suitable for statistical analysis. The Perl scripts are found in appendix C.

5.3.2 Perl

Perl is a freely available scripting language, specially designed for text processing. See Hammond (2003) and Danielsson (2004) for introductions to Perl aimed specifically at linguists. Lemay and Colburn (2002) is a good general introduction to Perl, while Bilisoly (2008) shows how Perl can be used in text mining and related fields, including some corpus linguistic applications.

My primary reason for using Perl in the current project was that no existing text mining or concordance software packages could provide data in the format I needed to read the data into the statistics application R. With Perl, on the other hand, I could write scripts to extract exactly the information I needed from the syntactic trees, and organize it in the form most convenient for my needs. This came at a cost, since separate scripts had to be developed for each treebank due to various annotation differences. Furthermore, inconsistencies and occasional errors in the annotation led to numerous rounds of debugging and modification of the Perl scripts. Despite this, I chose Perl for the current study since the syntax is very close to other familiar programming languages such as Java, and for its ease in handling regular expressions for text processing.

5.3.3 R

Some extraction and processing tasks were also carried out in R. Notably, certain operations on large tables are more conveniently handled in R, and in many cases R served in a support role to CorpusSearch and Perl. I frequently relied on R to split a large dataframe into columns for further processing with Perl, before loading the processed data back into R where it was merged with other dataframes to create the datasets described below. Thus, the different software systems employed were used together in a way which allowed them to do what they do best, and to be augmented by other systems in areas where they are less efficient.

5.3.4 Alternatives

Extracting large amounts of data from these treebanks is not a new problem. Mitchener (2006) faced the same problem and solved it by writing a set of search scripts in the programming language Haskell, available as *Crouton*.[1] Roy (2006) also opted for Perl

[1] Freely available from http://crouton.sourceforge.net/.

scripts to extract relevant data.

Although Perl scripts are easy to write and run, they are not without difficulties. Perl is not particularly suited for balanced text, i.e. sequences of opening and closing brackets. This put some rather heavy constraints on what could be extracted from the treebanks (finding the start of a phrase is for instance very easy, but finding the end of it is quite challenging). This is actually a parsing problem, where it is necessary to keep track of a number of nested opening and closing brackets. Instead of trying to write a parser in Perl I ended up with trying to push the standard regular expression pattern matching as far as I could. Although not entirely satisfactory, it was still sufficient to extract a good deal of relevant information about the tokens in the treebanks, but at the cost of very complex scripts: as the regular expression patterns become more complex they also become exceedingly difficult to debug.

Alternatives to Perl include parsing with Python, which is a scripting language like Perl. Python includes add-on parsing modules such as *PLY* and *Pyparsing* (the latter seems especially user friendly and appears to have some advantages over the *PLY*/*YACC*/*LEX* approach). Another option would be to use *ANTLER*[2] to generate parsers in languages such as Python, Java, C, etc.[3]

Whatever the technical solution, it is going to require a fair amount of programming or scripting skills. This is unfortunate since this is currently not a standard requirement in most linguistics programs. Thus, the most useful electronic resource for studying historical English is severely handicapped by its lack of supported user friendly software for automatic searching, retrieval and parsing.[4]

In summary, performing a large-scale quantitative analysis on historical English data is in principle an achievable task. However, with the lack of appropriate software for automatic extraction of tokens, this remains achievable mainly in principle. Although regular expression searches with Perl can perform remarkably well given its limitations, the task of writing and debugging such scripts quickly becomes unmanageable for complex searches. It seems beyond questioning that some kind of software to search these treebanks automatically and produce a structured output suitable for software programs is required. Although the development of such software was beyond the scope of the current project it seems a natural and obvious next step in making these treebanks even more useful for the research community.

[2] See http://www.antlr.org/.

[3] According to the *ANTLER* web pages an early prototype version of Perl output for *ANTLER* is also under development.

[4] This also applies to example-based searching, to some extent. CS is not a particularly user friendly program, as the description in section 5.3.1 above attests to.

5.4 The data frame format

A data frame in R is a table where each row corresponds to exactly one token or observation (Crawley, 2005, 15). This can be compared with the summary statistics from CorpusSearch which produces the number of hits per file. In such a format, every row corresponds to a different number of hits. Such a format is not suitable for most statistical testing purposes in R. Below is a short description of the data frame format used for the present study. The format below is modeled on the format of some of the data sets distributed with the *languageR* package in R, Baayen (2008b), notably "lexdec","oldFrench", "oldFrenchMeta", and "verbs". The descriptions follow a conceptual division between meta data and the data set proper with linguistic information on each token. For the analysis the two were treated as a single dataset. Such sets were created for all three treebanks and the categories were kept as similar as possible. Where there are differences, this is commented on.

5.4.1 Short description: Data

The present section gives an overview of the datasets used for the current study. A close scrutiny of the datasets will reveal more variables than those discussed below. These are for the most part "ad-hoc" convenience variables created to make certain processing tasks easier. For instance, to quickly exclude the cases where no year was given in the corpus documentation, a `YearNA` variable was created, setting as TRUE those cases that lack a year. Similarly, information from several variables below was combined, so that the `InitialThere` variable below is simply constructed from cases with *there* as adverb and where the adverb is in initial position. Only the most central of these convenience variables are described below.

ID

a factor specifying which text the token comes from, with the individual corpus ID-tag codes as levels.

Text

a factor specifying the text (or filename) from which the token comes, with the text names as levels.

ADV

a factor with the attested locative adverbs as levels.

Context

a factor with ADV's immediately following element(s) as levels (unprocessed).

ConTag

a factor with only the tag of Context as levels.

ConFiller

a factor with only the lexical filler of the Context element as levels.

Context2

a factor with Context's immediately following element(s) as levels (unprocessed). This constitutes the second (linear) left context element of ADV.

ConTag2

a factor with only the tag of Context2 as levels (not present for the PPEME data).

ConFiller2

a factor with only the lexical filler of the Context2 element as levels (not present for the PPEME data).

AdvLength

a numeric vector with the length of the adverb in letter characters (excluding coding characters like + and $).

BeContext

a factor coding whether the context-factor contains a form of to *be*, with F (FALSE) and T (TRUE) as levels.

Initial

a factor coding whether the sentence contains an initial locative ADV, with F (FALSE) and T (TRUE) as levels.

ConjClause

a factor coding whether the clauses has a conjunction in initial position, with F (FALSE) and T (TRUE) as levels. Bech (2008) shows a tendency for existential verbs in OE to be over-represented in coordinated clauses. This hypothesis was not pursued further in the current project due to time constraints.

Transitive

a factor coding whether the sentence contains an NP in the accusative, with F (FALSE) and T (TRUE) as levels.

There

a factor coding whether the adverb is a form of *there* (in YCOE this would include the forms þar, þær, þæra, þara, þare, þære, ðær, ðar, þer, þere, ðer, þear, ðaer), with F (FALSE) and T (TRUE) as levels.

Here

a factor coding whether the adverb is a form of *here*, with F (FALSE) and T (TRUE) as levels.

THTclause

a factor coding wether the token contains a *that*-clause, with F (FALSE) and T (TRUE) as levels. Only present in the OE dataset.

EmptyExpl

a factor coding wether an empty expletive subject is present, with F (FALSE) and T (TRUE) as levels.

Pro

a factor coding whether any other empty subjects are present, with F (FALSE) and T (TRUE) as levels.

Nodes

a numeric vector with the count of the total number of nodes in each sentence.

LogNodes

a numeric vector with the log-transformed count of the total number of nodes in each sentence.

IP

a numeric vector with the count of the total number of IPs in each sentence, including the matrix IP.

LogIP

a numeric vector with the log-transformed count of the total number of IPs in each sentence, including the matrix IP.

NP

a numeric vector with the count of the total number of NPs in each sentence. In some tokens no NP is present. In those cases, the NP count has been set to 0.1. While the basic assumption for the complexity measure used is that complexity is a product of verbs and NPs, it is unreasonable to assume that clauses without NPs have no syntactic complexity whatsoever. Furthermore, it is impossible to do a logarithmic transformation of 0. Thus, both theoretical and practical considerations led to this decision.

LogNP

a numeric vector with the log-transformed count of the total number of NPs in each sentence.

Locative

a numeric vector with the count of the total number of locative elements in each sentence, based on elements tagged with -LOC. Only used for YCOE.

LocTag

a factor specifying whether the tag of the adverb is locative or not, with F (FALSE) and T (TRUE) as levels. Only used for PPME2 and PPEME.

ExTag

a factor specifying whether the tag of the adverb is existential or not, with F (FALSE) and T (TRUE) as levels. Only used for PPME2 and PPEME.

PredExTag

a factor specifying whether the predicted tag of the adverb is existential or not, with F (FALSE), T (TRUE), and "NA" as levels. Only used for YCOE.

Complexity

a numeric vector with the SCR of the sentence, calculated from raw counts as $NP \times IP/\sqrt{Nodes}$. See chapter 6 for a further discussion.

LogComplexity

a numeric vector with the log-transformed SCR of the sentence, treated as interval data.

InitialThere

a factor coding whether the adverb *there* is present in initial position in the token, with T (true) and F (false) as levels. Initial conjunctions are ignored.

WordOrder

a factor coding the (approximate) word order of the token, with "VX" (verb initial), "OV" (object initial), "SV" (subject initial) and "XV" (other elements) as levels. For all word orders, initial conjunctions are ignored. For the VX pattern, enclitic forms of *be* (e.g. *nis* or *nas*) are counted as verb initial. The SV pattern comprises all tokens with an initial nominative NP. The OV pattern comprises all tokens with an initial NP in accusative, dative, genitive or a bare NP (in some cases including question words in initial position). The XV pattern comprises all other initial elements, such as negation, PPs, adverbs, or clauses.

Object

a factor specifying whether the sentence token contains an object or not, with F (FALSE) and T (TRUE) as levels. For YCOE this is mostly NPs tagged as either NP-DAT or NP-ACC. For PPME2 and PPEME, they are tagged as NP-OBJ.

Verb

a factor giving the (first finite) verb for each token, with the (raw) corpus verb forms as levels. There were difficulties with this semi-automatic procedure in all three treebanks. First, verbs were automatically extracted, then they were lemmatized and semantically classified manually, before they were automatically paired with the corresponding corpus token in the dataset. The task of lemmatizing and classifying the verbs caused some problems. Below, some examples of these difficulties with examples from PPME2 are discussed briefly.

In PPME2 semantic classification is, as in the case of YCOE and PPEME, done manually and exclusively on semantic grounds, although the verb classes in Levin (1993) for Present-day English have been used as a guiding principle. The semantic classification is primarily based on Stratmann (1940), but also to some extent Burrow and Turville-Petre (2005). Notably, I have not followed Stratmann (1940) in consistently using the oldest attested forms (i.e. those nearest to Old English) as the lemma form. More recent forms have for instance been chosen in cases where this seems agreeable with the corpus material. One example of this is "bury", listed as *bürȝen* in Stratmann (1940, 99) whereas I have chosen the more modern lemma form *birien*. Other examples include "build" listed as *bülden*, (Stratmann, 1940, 97), whereas I have chosen *bilden*; "die" listed as *dēȝen*, (Stratmann, 1940, 156), lemmatized as *deien*; and "kill" listed as *cüllen*, (Stratmann, 1940, 143), lemmatized as *killen*.

As in the other datasets, the semantic classification is fairly coarse-grained in that it is based on summary decontextualized occurrences. This means that there will be some misrepresentation of certain verbs with multiple meanings. Furthermore, accents are not annotated in the treebanks, which makes it impossible to distinguish between forms such as *witen* – primarily "know, be certain of", but see Stratmann (1940, 689), and *wīten* – primarily "see, keep", but see Stratmann (1940, 689). A similar problem occurs with "call, command" *hāten* and "hate" *hátien*, where the present tense third person singular forms are respectively *hateð* and *háteð* – obviously indistinguishable without accents and context. In all these cases what seemed like the most central or common meaning has been chosen. Other difficult cases include *seggen* "say" and *seon* "see", which share several forms. The precision with which some of these forms have been classified is not impressive, although for both verbs there are also several unambiguous forms. Although this is far from satisfactory, with the limited time and resources available and the lack of lemmatization in the corpus, this was deemed the only practical solution. However, these deficiencies should not obscure the fact that for many, perhaps most, cases it was nevertheless possible to identify the lemma form, even without context.

VerbTag

a factor with the syntactic label of the first verb for each sentence as levels.

VerbType

a factor giving the type of verb, with BE, Do, Have, Modal, and LexV (lexical verb) as levels.

Lemma

a factor giving the proposed lemma form of the token's verb, with the lemma forms as levels.

Modern

a factor giving the proposed modern/Present-day English translation of the token's verb, with the modern verbs as levels.

VerbFreq

a numeric vector giving the number of occurrences for the given (lemmatized) form in the extracted data.

SemDensity

a numeric vector giving the number of verbs (cf. "Verb" above) per semantic class (cf. "SemClass" below), as an approximation to the semantic "density" of the given verb class.

SemClass

a factor coding the semantic class of the token's verb with the semantic classes as levels. Semantic classes are primarily based on Levin (1993), but with some minor adaptations where necessary. Unlike Levin's verb classes which reflect both semantics and syntax, the present classification is based exclusively on semantics. See chapter 7 for a further discussion.

PossibleDynVerb

a factor coding whether the verb is a possible dynamic verb, with F (FALSE) and T (TRUE) as levels.

PossibleTrVerb

a factor coding whether the verb is a possible transitive verb, with F (FALSE) and T (TRUE) as levels.

ExTag

a factor coding whether the tag of the adverb is existential (EX) or not, with F (FALSE) and T (TRUE) as levels.

LocTag

a factor coding whether the tag of the adverb is locative (LOC) or not, F (FALSE) and T (TRUE) as levels.

5.4.2 Short description: Meta

Text

a factor with text codes as levels.

Year

a factor specifying the (approximate) year of composition for each text where this is known, with years, ranges of years, or in the case of missing information 'NA', as levels. For most OE texts the exact year of composition is not known. Ker (1957) gives approximate periods of composition for most of the texts, and these are provided in the YCOE documentation. YCOE follows the notation in Ker (1957), where centuries are given in small roman numerals, followed by "in", "med" or superscript arabic numerals indicating quartiles. Mostly, Ker (1957, xx) gives approximate quarter centuries, noting that exact numbers might give a false sense of precision. For the present study this has been modified somewhat. The "Year" factor is given as a number with arabic numerals, corresponding roughly to Ker's quarter centuries. Thus, "xi^{1}" (the middle of the first quarter of the eleventh century) is represented as "1025" which is merely a convenient representation format for the purposes of the statistical analysis. In some cases more

or less arbitrary decisions need to be made in order to assign a single year to a text, but in those cases where Ker and YCOE give long time periods I have gone with the lower estimate for the shorter spans, or some kind of middle ground for the longer spans. Hopefully, this necessary categorization should will not introduce too much noise in the data. Further examples are given in table 5.1.

TABLE 5.1: *Examples of how the approximate date of composition for OE texts from Ker (1957) are represented in the oeMeta data frame.*

Ker/YCOE		oeMeta
x/xi, xi.in	→	1000
xi^1	→	1025
xi^1 – xi^2, xi.med	→	1050
xi^2, xi 3rd quarter	→	1075
ix/x – x^2	→	925
Various	→	NA

For the PPME2 and PPEME treebanks, this 25-year interval has been coded as `Year25`, so that `Year` in those corpora corresponds to the actual date supplied in the corpus documentation.

Dialect

a factor giving information about the dialect(s) used in the text, with the dialects as factors. In YCOE, the dialects used are ws (West Saxon), k (Kentish), a (Anglian) or combinations of these as recorded in the philological information in the corpus, such as "ws.k.a" or "k.a".

For PPME2, the dialects are "EastMidlands", "Kentish", "Northern", "Southern", and "WestMidlands". No dialects are included for PPEME.

Translation

a factor specifying whether the text is a translation from Latin, with F (FALSE), T (TRUE) and "NA" (no information or uncertain) as levels. In PPME2 and PPEME this variable simply codes whether the text is translated or not, since source language varies.

Original

a factor specifying the source language of translated texts, with the various sources languages as levels, in addition to "NA" for the non-translated texts. Only used in PPME2 and PPEME.

Genre

a factor specifying the genre of the text, with the individual genres as levels.

WordCount

a numeric vector giving the total word count for each text, as supplied by the corpus documentation.

TokensPerText

a numeric vector giving the total number of sentence-tokens for each text, as calculated in the CS output files.

ObsPerYear

a numeric vector giving the total number of observations (in the dataset) per 25-year interval.

Author

a factor specifying the name of the author where known. Not used for the YCOE material. PPME2 and PPEME give author information. "NA" where not known.

AuthorGender

a factor specifying the gender of the author where known, with levels "female", "male" and "NA".

File

a factor specifying the corpus file the token is taken from, with the respective corpus file names as levels.

Chapter 6

Units of measurement: A new measure of syntactic complexity

6.1 Introduction

A measure of syntactic complexity which integrates memory cost and integration cost is proposed. The measure is motivated on psychological grounds, but is submitted primarily as a useful heuristic tool in the present corpus study. The psychological reality of the measure is an empirical matter more properly dealt with in an experimental design and consequently not covered here.

6.2 Units of measurement

The units of measurement for most variables in the present study is counts, that is, most variables are nominal (counts of words, counts of characters, counts of phrases etc.). Some units are logical (true/false) binary variables (such as whether a token contains a form of *there* or not). The time period is for convenience measured in 25 year intervals, thus, the "year" variable can arguably be classified as interval data. Most of these units of measurement are fairly uncontroversial (some are already given, either in the corpus documentation regarding e.g. whether a text is classified as a translation or not, or in the corpus annotation). However, one variable stands out, and deserves a more thorough treatment: syntactic complexity.

A potential predictor of the use of *there* as subject is the complexity of the sentence.

Lakoff (1987, 572) gives the following examples, and notes that (2) sounds better than (1), due to the greater complexity of (2).

(1) There bled a hemophiliac.

(2) For two hours there had been bleeding on the emergency room floor a poor hemophiliac who had fainted before he could sign his Blue Cross form.

Whether complexity in fact is a good predictor of the use or non-use of *there* in early English is of course an empirical question to be answered in subsequent chapters. However, before undertaking such an investigation it is necessary to establish a measure of complexity. Szmrecsányi (2004) discusses syntactic complexity in terms of the number of nodes in the syntactic tree, the number of words, or an Index of Syntactic Complexity (ISC) which is based on the number of embedded clauses. See Frazier (1985) for a further discussion on some grammatically motivated measures of complexity. Since my data comes from treebanks, it is possible to make use of syntactic information in measuring the complexity of the tokens (matrix IPs, or sentences). The question is rather which measure to use. Rather than using one of the measures discussed by Szmrecsányi (2004), I propose a new simplified measure of syntactic complexity which could be argued to potentially have some psychological relevance.

6.3 A new measure of complexity

Why a new measure? Gibson (2000) argues that a psychologically realistic measure of complexity should include both storage (i.e. a memory cost) and integration (i.e. syntactic processing). Warren and Gibson (2002) argue that integration of the NP material with the verbs of the sentence is the task most prone to cause intuitive complexity, cf. also Hawkins (2004, 262). See Damasio and Tranel (1993) for results suggesting that nouns and verbs operate from different neural substrates and pose different difficulties in syntactic processing. I therefore suggest a simplified measure of complexity, the sentence complexity ratio, or SCR. SCR is based on the number of NPs and the number of embedded IPs (i.e. the material to be processed), and the number of nodes (i.e. the "distance" over which this material must be processed or integrated). The number of IPs corresponds to the number of finite verbs, verbs since all IPs contain a finite verb, while maintaining a distinction between auxiliary and lexical verbs.

Experiments by Just and Carpenter (1992) seem to support the idea that humans use one verbal working memory which both stores words and meanings and carries out syntactic processing. Lieberman (2000, 73–74) discusses the findings of Just and Carpenter (1992) and **?**, arguing that the latter's support for modular syntactic process-

ing is merely a result of low-span working memory. Test results appear to be highly dependent upon individual differences of processing capacity, which would contradict a hypothesis of a truly universal, or innate, syntactic processing module. At same time this strengthens the case for syntactic complexity as a phenomenon dependent on working memory. A full presentation of the hypotheses on how working memory is organized falls outside the scope of the current study. Suffice it to say memory appears to be divided into two main components, short-term and long-term memory. Furthermore, short-term memory can be subdivided into several subcomponents (**?**). An influential assumption in studies on working memory is that these subcomponents contribute to syntactic and semantic processing in different ways and degrees, cf. **?**; **?**. For an overview and further discussion of the role of working memory in language processing see **?**.

In summary, there seem to be a number of good reasons why verbs and nouns should both play a role in the perception of syntactic complexity. The number of nodes in a syntactic tree has no *direct* psychological counterpart, but serves as a useful approximation to the limits within which syntactic structure must be integrated. It is likely that working memory is involved in processing syntactic complexity, although the details of the processing system are still being investigated (**?**).

6.4 Defining the SCR

For the purposes of the current definition, a *node* is a tree element in a corpus token from one of the three treebanks YCOE, PPME2, and PPEME. The tree in (3) below has three nodes, A, B, and C. A is the top (or root) node:

(3)

A *phrase* is defined as the root node and any nodes below it in the tree. Thus, the total number of nodes is a measure of the total size of the corpus token. An *IP* ("Inflectional Phrase") is a phrase with a finite verb as its head (i.e. a clause). All tokens extracted for the present study have an IP as their root node (i.e. the top node of a corpus token), as explained in section 5.3.1. An *NP* ("Noun Phrase") is a phrase with a noun as its head, together with all nodes dominated by the NP node itself. Through recursion, IPs can contain other IPs and NPS can contain other NPs.

Let ND be the total number of nodes dominated by the sentence, minus the sentence node itself; let NP be the total number of noun phrases (including those embedded in PPs); and let IP be the total number of IPs including the sentence IP. This has the

advantage of keeping complexity low for matrix clauses (IP-MAT) without embedded subordinate clauses, while the presence of subordinate clauses (IP-SUB) will cause the complexity to grow. A single IP with many nodes and many NPs but no embedded IPs (list-like constructions) will grow in complexity more slowly.

The formula in (6.1) defines SCR:

$$\text{SCR} = \frac{\text{NP}}{\text{ND}} \times \text{IP} \times \sqrt{\text{ND}} \tag{6.1}$$

The ratio of NPs to nodes is used as a complexity measure, with added complexity being introduced by the number of IPs. The IP factor will only make a difference when IP > 1, i.e., when there are embedded sentences in the tree. Multiplying with the square root of the number of nodes will compensate for the increased memory cost of keeping long (but not necessarily complex) structures in working memory. Additionally, this step will guard against situations where a longer sentence is simply twice the number of NPs and nodes of a shorter one. Multiplying with the square root of nodes will ensure that the longer sentence will receive more memory complexity than the shorter one. The formula above can be written in a simplified form as in (6.2) below:

$$\text{SCR} = \frac{\text{NP}}{\text{ND}} \times \text{IP} \times \sqrt{\text{ND}} \tag{6.2a}$$

$$= \frac{\text{NP} \times \sqrt{\text{ND}}}{\text{ND}} \times \text{IP} \tag{6.2b}$$

$$= \frac{\text{NP} \times \sqrt{\text{ND}}}{\sqrt{\text{ND}} \times \sqrt{\text{ND}}} \times \text{IP} \tag{6.2c}$$

$$= \frac{\text{NP} \times \text{IP}}{\sqrt{\text{ND}}} \tag{6.2d}$$

Thus, for each sentence the SCR is the product of NPs and IPs per square root tree node. Syntactic complexity is expressed as a ratio of two approximate entities: how much material needs to be integrated (NPs and embedded IPs), and what is the "distance" (i.e. the memory cost, expressed as the number of nodes) that this operation needs to be carried out over. It is assumed that a low integration cost (few NPs and IPs) together with a low memory cost (few nodes) amount to low complexity in a sentence.

Consider the following examples from the YCOE (the numbers in parenthesis are the number of IPs, NPs and nodes):

(4) a) and þar restað haligra manna saula að domesdæig (**1**, **3**, **13**)
b) þa Iudeiscan gesawon swutele tacna on þam wodan men þe ðær wæs gehæled (**2**, **4**, **22**)

c) Se godspellere Iohannes sæde on þysum godspelle þæt Crist ure Hælend, þa þa he her on life wæs, come on sumne sæl to Samarian byrig, to ðæs heahfæderes wurþige þe wæs gehaten Iacob (**3**, **14**, **65**)
d) Him andwyrde þæt wif, Hlaford, syle me of þysum liflican wætere, þæt me heonon forð ne þyrste, ne ic her ne þurfe hladan (**2**, **8**, **44**)
e) & þær gefongen wæs (**1**, **1**, **6**)

The complexity measures for the sentences in (4) are given in table 1, with corresponding ordinal ranks. All the calculations have been carried out using R. If Warren and Gibson (2002) are correct in assuming that syntactic processing and memory storage and retrieval are using some of the same resources (which would also be reasonable assumption under a cognitive-functional framework), d) should be more complex than b); c) should be the most complex; while a) and e) should have the lowest complexity.

TABLE 6.1: *Syntactic complexity ratios for the example sentences, with corresponding ranks sorted in descending order from the lowest* SCR *to the highest.*

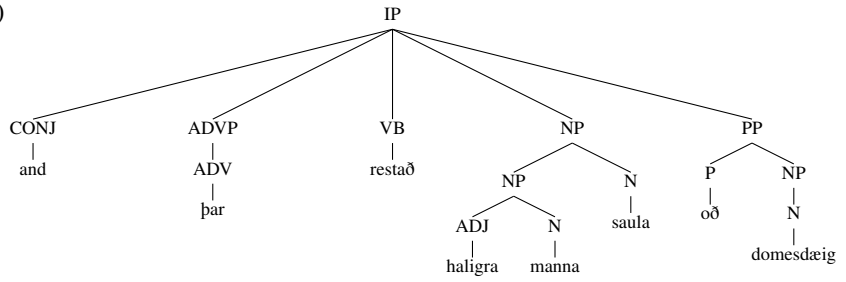
Sentence	SCR	Rank
a	0.83	2
b	1.71	3
c	5.21	5
d	2.41	4
e	0.41	1

The measure gives more complexity to sentence a) than sentence e), as expected (see (5) for a syntactic representation). The interpretation is relatively straightforward, since the NP elements are primarily responsible for introducing new discourse elements. In cases like a) and b), where there is only 1 IP, the SCR is simply the number of NPs divided by the square root of the number of nodes to account for working memory load. Table 6.1 shows that in terms of SCR, a) is less complex than b), and that c) is the most complex and that e) is the least complex. This corresponds to a fairly intuitive complexity ranking of these sentences based on the number of nodes, NPs and IPs. Furthermore, we can note the relation between b) and d), where the node and NP counts in d) are twice those of b). As expected, d) is assigned a higher complexity ratio score than b). This illustrates the motivation for dividing by the square root of the node count, since simply dividing by the raw node count would result in identical scores for

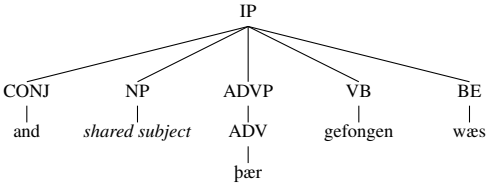
b) and d).

Two of the simpler sentences are given below with the tree analysis used in YCOE. The numbering corresponds to (4).

(5) a)



e)



6.5 Properties

Figure 6.1 shows the SCR computed for all tokens in the datasets extracted from the three corpora; Old English (OE), Middle English (ME), and Early Modern English (EME). The figure shows the distributions as density plots and Q-Q plots. Evidently, the complexity ratio scores are far from normally distributed in any of the three datasets. However, there are ways of transforming a numeric variable to achieve better fit to the normal distribution. One way of transforming a variable for normality is to take the logarithm of the variable (**?**, 317–324). Figure 6.2 shows the same data as in figure 6.1 transformed by taking the natural logarithm ($\ln$ or log_e) of the distributions. As the plots show, the (natural) log transformed SCR (log SCR) in all three datasets achieves a fairly good fit to the normal distribution. This offers an advantage for statistical testing, since many tests assume normally distributed data (cf. chapter 4).

The shape of the SCR / log SCR distribution appears to be fairly constant for all three datasets. As table 6.2 shows, there is an increase in the complexity ratio from the Old English to the Early Modern English data. The increase appears to be uninterrupted and gradual from the earliest to the later data. The datasets from YCOE and PPEME (i.e. the endpoints) are of approximately equal size ($\sim$ 9 000 observations), while the Middle English dataset from PPME2 consists of about 5 500 observations. Thus, the

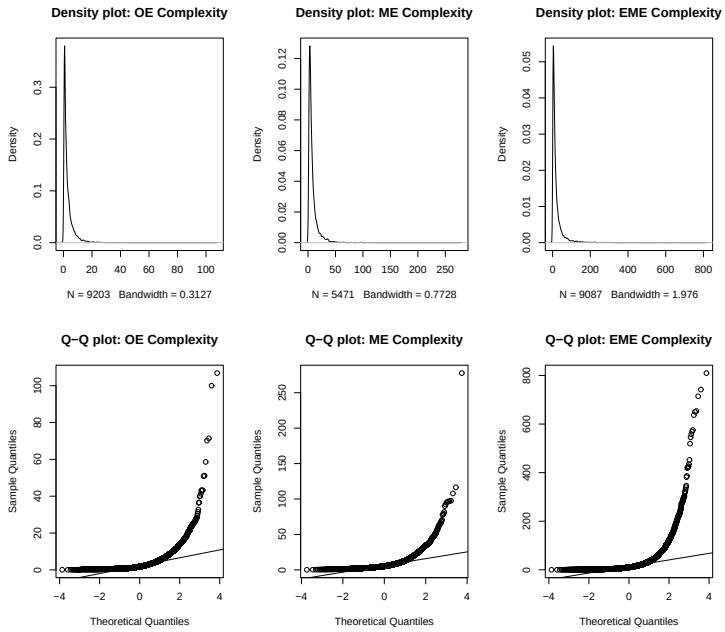


FIGURE 6.1: *Density and Q-Q plots for* SCR *values from OE, ME, and EME. As the plots show, the distribution of the* SCR *is far from normally distributed.*

size of the datasets does not appear to be responsible for this increase, although the genre composition of the corpora might play a role.

Tables 6.3, 6.4, 6.5, and 6.6 show some examples of the SCR and the log SCR for different values of IP, NP, and ND. These tables are intended for illustration and the numbers do not always translate into realistic sentences.

6.6 Discussion

6.6.1 Psychological validity

The validity of the SCR in a psychological sense is of course an empirical question, but such an evaluation falls outside the scope of the current investigation. There are two obvious reasons which favor the SCR, and justifies its use as a heuristic tool in corpus

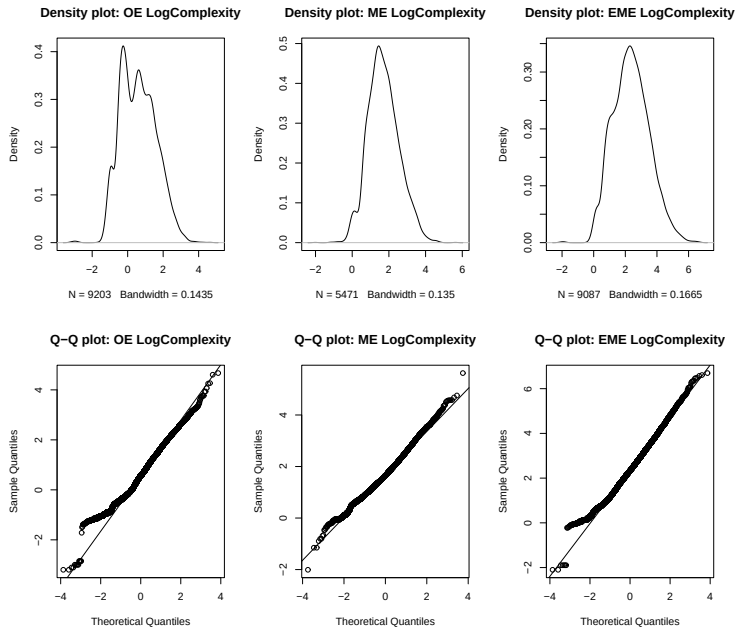


FIGURE 6.2: *Density and Q-Q plots for the log transformed* SCR *from OE, ME, and EME. The plots show that a (natural) logarithmic transformation brings the variable reasonably close to the normal distribution.*

analysis, even when its psychological validity remains untested: the first reason has to do with granularity, the second with interpretation.

Despite this, there are admittedly some validity problems with the proposed measure. However, the measure ought to be approximately proportional to a product of memory and integration cost. Although the exact numbers remain in doubt, it seems plausible that the relative order of complexity is adequately handled in most cases.

6.6.2 Nodes vs. SCR

Why not simply use the number of nodes in the tree as a measure of complexity? As a measure of complexity the number of nodes is relatively crude. Furthermore, it disguises a lot of variation.

As an example, I looked at node count vs. SCR in the Old English selection of

TABLE 6.2: *Summary statistics for the* SCR *of the three datasets from* YCOE *(OE),* PPME2 *(ME), and* PPEME *(EME).*

Corpus	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
OE	0.04	0.83	1.75	3.06	3.72	106.90
ME	0.13	3.10	5.20	8.18	9.53	277.70
EME	0.12	4.65	10.29	21.73	22.86	809.80

TABLE 6.3: *Constructed data showing the increase of* SCR *and log* SCR *as the number of IPs, NPs and nodes increase.*

IP	NP	Nodes	SCR	log SCR
1.00	3.00	7.00	1.13	0.13
2.00	4.00	8.00	2.83	1.04
3.00	5.00	9.00	5.00	1.61
4.00	6.00	10.00	7.59	2.03
5.00	7.00	11.00	10.55	2.36
6.00	8.00	12.00	13.86	2.63
7.00	9.00	13.00	17.47	2.86
8.00	10.00	14.00	21.38	3.06
9.00	11.00	15.00	25.56	3.24
10.00	12.00	16.00	30.00	3.40

tokens with locative adverbs (see chapter 8). The median node count in the whole selection is 25. I then looked at all tokens with 25 nodes, 176 tokens in all. Interestingly, the mean SCR for those 176 tokens was 1.8, with a maximum of 5.0 and a minimum of 0.6. There were six tokens with the lowest SCR, all of them consisting of one IP with 3 NPs and 25 nodes. Only one token had the highest SCR score of 5.0, consisting of 5 IPs, with 5 NPs and 25 nodes.

SCR is also a better indicator of intuitive complexity than the number of nodes for sentences with center embedding.[1] See e.g. **?** for a discussion on the difficulties that center embedded clauses pose for impaired processing.

[1] I am grateful to Christer Johansson for bringing this point to my attention.

TABLE 6.4: *Constructed data showing the increase of* SCR *and log* SCR *as the number of IPs, NPs and nodes increase. Note how the* SCR *scores are lower than in table 6.3 due to the higher number of nodes.*

IP	NP	Nodes	SCR	log SCR
1.00	3.00	12.00	0.87	-0.14
2.00	4.00	13.00	2.22	0.80
3.00	5.00	14.00	4.01	1.39
4.00	6.00	15.00	6.20	1.82
5.00	7.00	16.00	8.75	2.17
6.00	8.00	17.00	11.64	2.45
7.00	9.00	18.00	14.85	2.70
8.00	10.00	19.00	18.35	2.91
9.00	11.00	20.00	22.14	3.10
10.00	12.00	21.00	26.19	3.27

Take the two constructed sentences below, where (6) has a structure with center embedding, whereas (7) does not:

(6) The man the boy hit ran

(7) The man who was hit by the boy ran

Both sentences contain two NPs and two IPs. Although the total node count is fairly similar (10 in the case of (6) and 14 in the case of (7)) we would expect (6) to be more complex to process because of the center embedding. Based on the number of nodes, however, (7) would be considered more complex. Calculating the SCR for the two sentences above is straightforward:

$$\frac{2 \times 2}{\sqrt{10}} = 1.26 \tag{6.3}$$

$$\frac{2 \times 2}{\sqrt{14}} = 1.07 \tag{6.4}$$

As the calculations above show, the sentence with center embedding emerges with the highest SCR. The sentence in (7) receives a lower SCR score precisely *because* it has more nodes and hence a greater distance to distribute the IPs and NPs over.

TABLE 6.5: *Keeping the number of IPs constant causes the* SCR *to grow more slowly.*

IP	NP	Nodes	SCR	log SCR
1.00	3.00	7.00	1.13	0.13
1.00	4.00	8.00	1.41	0.35
1.00	5.00	9.00	1.67	0.51
1.00	6.00	10.00	1.90	0.64
1.00	7.00	11.00	2.11	0.75
1.00	8.00	12.00	2.31	0.84
1.00	9.00	13.00	2.50	0.91
1.00	10.00	14.00	2.67	0.98
1.00	11.00	15.00	2.84	1.04
1.00	12.00	16.00	3.00	1.10

It is clear from these brief examples that the SCR score gives a far more fine-grained and intuitive measure of syntactic complexity than simply counting nodes.

6.6.3 Using log counts

One of the useful properties of the SCR is that it is close to a normal distribution when log transformed, as shown in figure 6.2. A possible alternative to the scr would be to simply use the (natural) log count of IPs, NPs and nodes. Figure 6.3 shows Q-Q plots for these quantities in Old, Middle and Early Modern English. A normal distribution is as above indicated by the thin straight line, and if the observations are close to this line, the fit to the normal distribution is good.

The plots in figure 6.3 show that the log counts of IPs seriously deviate from a normal distribution in all three periods. The log counts of NPs have a less bad fit, but there are evident problems in the lower tail. Only for the log counts of nodes does the normal distribution appear to be a good fit.

One possibility, then, would have been to use only the log count of nodes. However, as the previous section showed, the count of nodes hides much variation in terms of the number of NPs and IPs. Thus, although working with the log counts of IPs, NPs and nodes would give a more intuitive interpretation, it would create problems for statistical tests that assume a normal distribution, and (in the case of working with nodes only), disregard information about the sentence such as the number of embedded clauses. In this perspective, I consider the properties of the SCR when log transformed to be

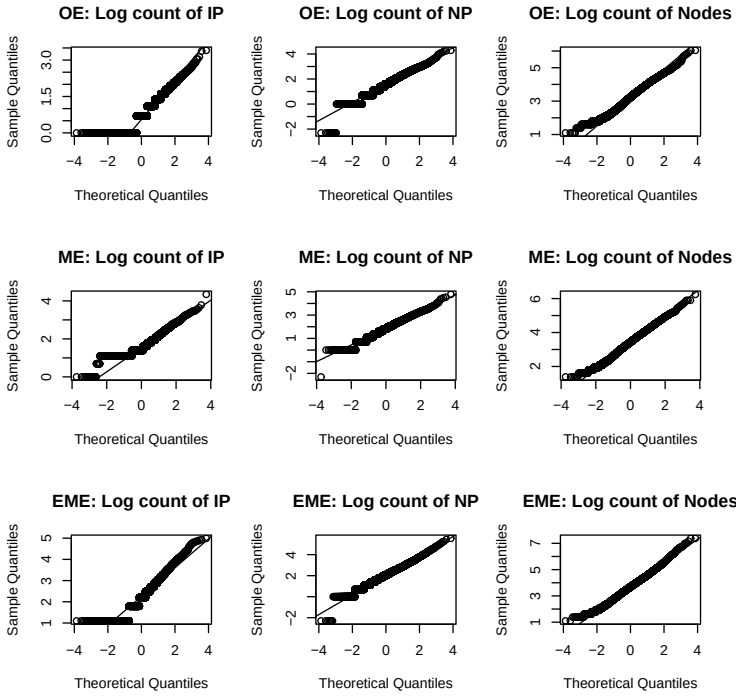


FIGURE 6.3: *Quantile-Quantile plots for the log count of IPs, NPs and nodes in OE, ME and EME. Only the log count of nodes has a good fit to the normal distribution in all three periods.*

TABLE 6.6: *Keeping the number of NPs constant and increasing the number of IPs causes more rapid growth, with the final score being close to the number of IPs.*

IP	NP	Nodes	SCR	log SCR
1.00	3.00	7.00	1.13	0.13
2.00	3.00	8.00	2.12	0.75
3.00	3.00	9.00	3.00	1.10
4.00	3.00	10.00	3.79	1.33
5.00	3.00	11.00	4.52	1.51
6.00	3.00	12.00	5.20	1.65
7.00	3.00	13.00	5.82	1.76
8.00	3.00	14.00	6.41	1.86
9.00	3.00	15.00	6.97	1.94
10.00	3.00	16.00	7.50	2.01

a reasonable trade off between desirable properties and interpretability. Although the SCR does not *directly* represent something in the corpora, it still expresses *indirectly* for each token a relationship between the number of IPs, NPs and their "density".

6.6.4 Interpretation

As noted above, the SCR has been justified in explicitly psychological terms, with reference to working memory capacity. Gibson (2000, 104–105) defines a *simplified discourse processing cost* in term of *energy units* which refer to the *discourse processing cost* (whether a unit contains new information or not) and the *structural integration cost* (the complexity of integrating a new structure into the structure already active in working memory). However, the interpretation of these energy units remains vague. The strength of the SCR is that it is motivated by the directly observable relationship between clause elements, and that keeping many of these elements active in memory at the same time will put more pressure on working memory capacity than keeping fewer of them active.

Whether a high SCR score *always* translates into a higher subjective impression of complexity is outside the scope of the present study, although it appears to be the case for some common structures. It is perhaps unfortunate to call the measure a *complexity* ratio, as this might suggest that the SCR measures something else than it actually does. It must again be stressed that the *motivation* for the SCR score lies in a psycho-

logical phenomenon and previous research on processing, but this does not entail that it directly measures such processing. Consequently, the term *complexity* as employed in the subsequent chapters is used in a strictly technical sense to denote what the SCR refers to; not to some psychological phenomenon. The SCR is used as a heuristic, and a tool to summarize a technical, observable property of the corpus annotation, namely the number of nodes, IPs and NPs, in a single number. As the plots in figure 6.2 showed, the natural logarithm of the SCR is close to a normal distribution in all three corpora, and for convenience the subsequent chapters will treat the log SCR as a continuous normally distributed interval scale variable.

6.7 Summary

The proposed measure of complexity, the Sentence Complexity Ratio (SCR) seems to compare favorably with other ways of measuring complexity. In some cases, it appears to be superior to a simple count of syntactic nodes. Its added value lies in its ease of use and potential psychological relevance. Whether the psychological interpretation is correct is ultimately an empirical question which should be tested experimentally. In the present study, the rationale for the measure is submitted, and its usefulness as a heuristic tool in cognitive corpus research is defended.

Chapter 7

Semantic verb classes

7.1 Introduction

The semantic verb classes presented in the present chapter are based on the classification by Levin (1993). For applications and evaluations of Levin's classes see e.g. **?**; **?**; **?**. Unlike Levin's classes which are syntactically and semantically motivated, the classes constructed for the present study are based on semantics only. The main practical motivation for the classes is to reduce the large number of verbs to a more restricted set of classes. With fewer levels than the number of individual verbs, the semantic verb classes become a more manageable input variable for statistical analysis. However, a more theoretical motivation can also be found. Breivik (1990, 164–168) discusses the importance of the verb for ECs explicitly in terms of broader classes such as "existence", "appearance", etc.

7.2 Semantic classes

As mentioned in chapter 5, the verb classes are constructed from decontextualized occurrences. This was necessary to make the work feasible, since the combined datasets comprise close to 24 000 verb tokens from YCOE, PPME2 and PPEME. As a result, the classification of individual verbs must be taken with a pinch of salt, since some are difficult to lemmatize and semantically classify out of context, cf. the discussion in chapter 5. Particularly difficult cases have been labeled with a question mark. However, many verbs were fairly clear cut cases, without (or with only related) alternative meanings. In the cases where it was completely impossible to either lemmatize or find

the meaning of a verb, semantic class was set to "NA" (this applies to a total of 162 tokens in the three datasets). Unless it is explicitly stated otherwise, the classes are attested in all three datasets. Most classes are based on Levin's verb classes. In cases where I have created them, or modified Levin's classes substantially, this is commented upon.

7.2.1 Overview

The classification resulted in a total of 48 semantic verb classes from the three corpora. Note that not all semantic classes are attested in each corpus. Table 7.1 gives an overview of the semantic classes constructed for the verbs in the datasets from YCOE, PPME2 and PPEME, together with their frequencies.

TABLE 7.1: *Semantic verb classes with combined frequencies of occurrence in the datasets from* YCOE, PPME2 *and* PPEME. *Note that not all semantic classes are attested in all three datasets.*

No.	Semantic class	Frequency
1	Ability	536
2	Activity	205
3	Appearance	1036
4	Aspect	404
5	Assessment	33
6	BodyProcess	121
7	ChangeOfPossession	943
8	ChangeOfState	389
9	Coloring	1
10	Combining	74
11	Communication	2082
12	Concealment	35
13	Contact	94
14	Contain	725
15	Creation	321
16	Cutting	38
17	Decision	11
18	Destruction	43
19	Disappearance	25
20	Emission	41

Continued on next page

No.	Semantic class	Frequency
21	Engender	10
22	Existence	8776
23	Grooming	1
24	HoldKeep	241
25	ImageCreation	1
26	Ingesting	99
27	Intention	754
28	Judgment	99
29	Killing	30
30	Learning	26
31	Lodge	19
32	Marking	112
33	Motion	1556
34	Obligation	86
35	Occurrence	159
36	Perception	546
37	Poke	20
38	Prevent	5
39	PsychologicalState	1292
40	PushPull	26
41	Putting	407
42	Removing	59
43	Searching	50
44	SendCarry	201
45	SocialInteraction	1145
46	SpatialConfiguration	652
47	Throwing	68
48	Weather	2

7.2.2 Short description of verb classes

The following sections give a list of verbs (from all three datasets) which constitute the semantic classes. The frequencies in table 7.1 attests to the uneven rate at which the classes are found in the datasets. As the frequency of verbs within each class is also uneven, all verbs are presented with their total frequency in the datasets in parenthesis.

Ability

can (108), cunnan (7), cunnen (6), magan (128), may (122), motan (14), mowen (151)

Comments This class was created for the present study to accommodate verbs of ability, such as modal verbs (and their precursors).

Activity

ʒedon (2), anchor (3), bitaken (1), bysien (1), commit (1), crossen (1), do (126), don (37), executen (1), laborin (1), plesen (1), practice (2), read (11), reden (1), reherce (2), toil (1), translacioun (2), treten (3), weiten (2), werken (6)

Comments This class was created for the present study to accommodate various verbs that did not seem to fit any of the other classes. Typically, the verbs describe some kind of general activity. Only found in the ME and EME datasets.

Appearance

ðyncan (3), appear (26), apperen (3), arrive (11), awakans (1), aweccan (2), becuman (31), bicomen (4), bode (3), come (338), cuman (374), cumen (208), discover (1), enter (19), entrin (3), issue (2), lenden (2), manifest (1), present (1), represent (3)

Comments This class is based on Levin's subclass of Appearance verbs.

Aspect

abisgian (4), ablinnan (1), aginnen (1), begin (22), beginnan (3), biginnen (73), blinnan (2), brucan (2), cause (12), causen (1), cease (1), cessen (2), change (3), conclude (3), continue (19), continuin (1), don (93), effect (2), end (5), enden (14), endian (42), endure (1), establish (1), finish (1), haunt (2), kepen (7), læsten (1), leven (23), onginnan (19), reducen (1), secure (1), stablen (1), sterten (2), stop (1), stoppin (1), storen (1), strive (3), stutten (1), undernimen (1), use (16), usen (3), varien (1), wake (1), wakien (2), wardien (2), witan (4), wrixlan (1)

Comments This class is based on Levin's Aspectual verbs, and denote beginning of, end of, and ongoing activity.

Assessment

bicomen (2), compare (1), consideren (2), examine (3), examinen (1), exceed (2), except (5), liken (1), measure (1), meten (6), prove (6), proven (1), suffice (1), suffise (1)

Comments This class is based on Levin' Assessment verbs, and reflect assessment of something without considering the outcome of the assessment.

BodyProcess

a-drenchen (1), adrincan (4), afæstan (1), aspiwan (2), asweltan (1), bite (1), biwepen (2), blawan (1), bleden (1), brethin (1), decease (2), deien (33), die (15), druncnen (1), fast (1), gape (2), laugh (1), orðian (4), pinien (3), slæpan (6), slæpen (2), sleep (3), smilin (1), spiwen (1), stupian (1), sucan (1), sweat (1), sweltan (18), tremblen (2), wepan (5), wepen (3)

Comments This class was created for the present study. It is a combination of several of Levin's verbs involving the body. The class includes both processes proper and physical states.

ChangeofPossession

ȝelden (2), ȝeven (24), þigcan (2), afford (1), afindan (2), alætan (1), aprouen (1), atteigne (1), befon (2), begietan (12), beniman (4), biggen (3), biniman (1), bitaken (1), borrow (1), bycgan (3), ceosan (8), cepan (1), contribute (1), deal (2), entitle (1), fangen (3), find (53), findan (46), finden (69), fiscian (1), fon (133), forgiefan (2), forgieldan (2), forhergian (7), forlætan (26), forleosan (6), forleosen (5), forlorian (1), forspendan (1), fortune (4), furnish (5), gearwian (5), get (7), geten (4), giefan (16), gieldan (2), gietan (1), give (26), læccan (3), leosen (2), lose (5), niman (132), nimen (6), noten (3), obtain (2), ofearnen (1), offrian (11), pay (5), procure (1), profit (1), prosper (1), provide (31), purchase (1), reafian (1), reaven (1), receive (19), ripan (1), seisen (1), seize (2), sell (1), sellan (53), spend (3), spoil (2), squander (1), steal (2), stelan (2), strienan (3), succeed (2), tacan (1), take (68), taken (62), tiðian (4), unarmen (3), unnan (1), victual (1), withholden (2), wreþian (1), yield (2)

Comments This class is based on Levin's verbs of change of possession.

ChangeOfState

æberstan (6), ætmtian (1), ætþringan (1), ðryccan (1), þennan (1), a-druʒien (2), aðeostrian (31), aðieadan (1), aðindan (1), abrecan (8), adrencan (9), adrugian (1), adwæscan (3), afylan? (1), agitan (1), aheardian (1), aidlian (1), alefian (1), aliesan (1), alihten (5), amenden (1), apparel (3), arasian (6), astreccan (7), attain (1), avoiden (1), aweaxan (2), awierdan (1), beclysan (2), become (3), befæstan (1), befealdan (3), beliðan (1), belucan (2), beorgan (1), berstan (1), bescitan (1), besmitan (1), biegan (1), biernan (6), blenden (1), brædan (1), break (8), brecan (10), breken (3), brennen (5), briwan (1), burn (2), burst (1), bærnan (4), close (1), closen (2), clothe (1), conforten (1), cool (1), dælan (3), dælen (3), delight (2), depart (4), derian (8), devise (1), distill (2), distribute (1), divert (1), divide (3), dressen (1), drigen (1), eacan (4), ealdian (1), efenlæcan (1), empty (1), encresen (1), enlighten (1), exhaust (1), extend (2), fæstan (4), fæstnian (6), faʒien (2), fain (1), fastnien (1), fatten (1), feolan (4), forbeornan (1), fordon (5), formeltan (1), forspildan (1), fortendan (1), gearcian (1), gierwan (1), gladien (1), grow (17), growan (1), growen (4), hard (11), hierdan (4), hladan (1), hwemman (2), increase (1), infect (1), inform (3), kimen (3), lacnian (1), leohtnen (1), light (5), lihtan (3), lock (2), masen (1), mawan (1), menden (2), misbeodan (1), nivian (1), ontendan (1), parten (2), perfect (1), petrify (1), redeem (1), reflect (1), refreschen (1), refresh (6), renew (1), reparin (1), rescoue (3), resolve (4), restore (2), rotian (1), round (1), rusten (1), scafan (2), scearpian (1), scortian (1), separate (2), softin (1), spildan (1), sprecchen (1), staðolian (3), streitin (2), strengthen (1), stretch (1), swamian (1), tenden (1), tendren (1), teoðian (1), tilien (1), to-draʒen (1), trymian (3), unfealdan (1), unfealdan (1), unlock (1), upwaxen (1), vary (1), wallen (1), wax (1), waxen (4), weaxan (10), werodian (2), wierdan (4), worðen (3), wright (1), wyrcan (1), wyrsian (1)

Comments This class is based on Levin's class of change of state verbs. It includes all kinds of changes, primarily material ones, to an entity.

Coloring

stain (1)

Comments This class is based on Levin's class of coloring. It involves changing the color of an entity somehow. Only attested in the EME dataset.

Combining

ðeodan (1), þeodan (3), add (7), adjoin (1), allay (1), assemble (4), bind (3), bindan (5), concur (1), fegan (2), gæderien (5), gadrian (15), gather (3), gebindan (1), inoculate (1), join (4), mix (2), nail (1), recollect (1), samnian (11), tiegan (1), weld (1)

Comments This class is based on Levin's class of combining and attaching. The class includes all forms of combining, also more metaphorical uses.

Communication

ætiewan (28), ætsacan (1), ðancian (1), ðingian (12), ȝecyden (1), ȝelpen (1), abiddan (2), acorden (1), address (1), admit (1), advertise (1), advise (4), affirm (1), agree (5), allege (3), ameldian (2), andswarian (25), andwyrdan (1), answer (19), answeren (9), argue (1), ascian (28), ask (24), avise (3), avouch (1), axen (26), beden (2), befrinan (4), beg (1), behatan (7), beseech (3), betellan (1), biddan (64), bidden (15), biecnan (1), bihoten (5), bodian (27), bodien (1), call (31), callen (6), certify (1), ciegan (1), claimen (1), clepien (11), clipian (12), complain (5), compleinen (2), conseilere (2), contract (1), convey (1), cry (5), cweðan (313), cweðen (10), cyðan (41), debate (1), declare (7), declaren (1), demand (5), deniien (1), denounce (1), deny (4), design (1), direct (3), disputen (1), exclaim (1), explain (1), expound (1), expresse (1), forename (1), foresay (2), forsake (1), freinen (1), frignan (9), galan (1), gieddian? (1), grant (3), granten (2), halsian (1), hwistlian (1), informin (1), inquire (3), insist (1), læran (34), læren (1), light (1), mærsian (1), maintain (1), manian (3), mention (5), nicken (1), oðiewan (8), persuade (2), post (1), pray (33), preach (5), prechen (7), preien (16), preserven (1), promise (1), promisen (2), prophecien (1), propose (2), protest (4), purpose (1), purposin (2), queath (31), rædan (24), ræden (12), reason (1), recite (1), reckon (1), recommend (1), relate (2), repeat (2), reply (6), request (1), require (2), say (253), scheawen (24), scyan (1), secgan (85), seggen (301), seofian (1), show (19), siken (1), sing (3), singan (11), singen (4), speak (27), speken (28), sprecan (35), suggest (1), swaren (2), swear (10), sweotolian (18), swerian (4), swerien (9), swigian (2), tæcan (8), talk (11), talken (5), teach (5), techen (17), tell (79), tellen (42), thank (13), treowsian (1), tyn (2), warien (1), warnian (5), wið-cweðen (1), wið-seggen (1), witegian (2)

Comments This class is based on Levin's class of communication. It includes all forms of communication and transfer of messages.

Concealment

þeccan (1), ahwylfan (1), ahydan (1), bediglian (1), bedydrian (1), behydan (1), beteldan (4), bewreon (1), bi-byrien (2), bihengen (1), birien (7), bury (1), cover (1), forswigian (1), happen2 (1), helen (1), hude (2), huden (1), hulen (2), hydan (2), lurk (1), lutian (1)

Comments This class is based on Levin's class of concealment. It includes verbs that somehow conceals something for someone.

Contact

þerscan (2), þydan (2), aslean (1), bat (1), beatan (2), club (1), cnokien (2), grope (1), hit (1), hrepian (1), lettan (5), punchin (1), qveisen (1), ram (1), slean (50), spurnan (1), stempan (1), striken (1), swingan (2), touch (16), wundien (1)

Comments This class is based on Levin's verbs of contact and verbs of impact. As used here, contact also includes verbs of impact.

Contain

contain (3), habben (205), harbor (2), have (513), hereberȝen (1), stuff (1)

Comments This class was constructed for the present study to accommodate verbs that denote or entail something being contained inside something else, also metaphorically. Only attested in the ME and EME datasets.

Creation

bedician (3), bilden (11), bredan (1), build (15), byldan (1), byrgan (17), create (1), dreogan (2), fashion (1), figuren (1), found (5), fremman (2), ground (1), macian (2), make (79), makien (75), produce (3), schapien (1), scieppan (2), stælan (1), stalian (1), swincan (2), timbran (29), wall (1), werche (2), winnan (14), wyrcan (48)

Comments This class is based on Levin's class of creation and transformation. It primarily refers to creating something.

Cutting

aslitan (1), besceafan (1), bescreadian (1), ceorfan (1), cut (5), fellen (12), forceorfan (4), forslean (2), gymmian (1), heawan (3), keorven (5), scafan (1), sniðan (1)

Comments This class is based on Levin's class of cutting. It primarily involves verbs to do with cutting and carving.

Decision

bilæfen (1), cheosen (2), choose (1), differen (1), fasten (2), forbugen (1), forgan (1), forlæten (1), forsaken (1)

Comments This class was created for the present study to accommodate verbs that involve choosing between two or more options. Only attested in the ME and EME datasets.

Destruction

astyfecian (1), awestan (2), destruien (2), for-dilgian (1), forbærnan (6), hergian (29), spill (2)

Comments This class is based on Levin's class of destroy verbs. It involves both physical and more metaphorical destruction.

Disappearance

aswinden (1), forweorðan (21), losian (3)

Comments This class is based on Levin's subclass of disappearance verbs. It involves disappearance from the scene, also metaphorically. Only attested in the OE and ME datasets.

Emission

bimænen (1), blast (1), crawen (2), crien (3), cronen (2), din (5), glimmer (1), hlydan (1), hrieman (2), neȝen (1), recan (3), schinen (1), scinan (15), shine (2), sparkin (1)

Comments This class is based on Levin's class of emission verbs. It denotes emission of something from an entity.

Engender

biʒeten (5), breed (2), conceive (1), raisen (2)

Comments This class is based on Levin's engender verbs. It involves one agent being brought into existence by another.

Existence

þeon (1), abidan (10), abide (20), abiden (27), alibban (1), anbidian (2), appertain (1), awunian (2), be (3566), ben (1884), beon (2812), bidan (9), biden (3), bogian (3), consist (2), dwell (25), dwellen (57), eradian (12), geanbidian (1), hanten (1), libban (34), lifgan (3), linger (1), live (11), livien (21), onbidan (3), remain (24), reside (1), sælan (1), stay (49), tarry (11), wait (4), wician (13), wonen (12), wunian (149)

Comments This class is based on Levin's subclass of existence verbs. It involves some kind of existence at some kind of location. Only attested in the ME and EME datasets.

Grooming

cemban (1)

Comments This class is based on Levin's class of grooming and bodily care. Only attested in the OE dataset.

HoldKeep

behealdan (4), belong (8), bilongen (1), distrain (1), goldhordian (2), gripen (1), habban (122), halden (5), healdan (42), helden (27), hold (14), horden (1), hordian (1), keep (8), ongietan (1), overtake (2), reach (1)

Comments This class is based on Levin's class of hold and keep verbs. It involves longer contact between entities than hitting or other forms of impact.

ImageCreation

paint (1)

Comments This class is based on Levin's class of image creation verbs. Only attested in the EME dataset.

Ingesting

afedan (5), beginan (3), biten (1), browse (1), cheowen (1), dine (25), drincan (2), drink (8), drinken (4), eat (2), etan (12), eten (9), fedan (9), forswelgan (1), freten (1), gnaȝen (1), reordian (2), sup (10), swelȝen (1), water (1)

Comments This class is based on Levin's class of ingesting verbs. It covers ingestion of both food and drink.

Intention

schulen (273), sculan (86), shall (237), sierwan (1), will (111), willen (46)

Comments This class was created for the present study to accommodate verbs that involve the expression of intention of something.

Judgment

ðreagan (1), þreagan (1), accuse (1), adeadian (1), allouen (1), ateorian (1), avail (1), availen (1), belæfan (11), blame (6), chastien (1), condemn (1), convicten (1), dafenian (2), deman (1), demen (6), deserven (2), dugan (7), fail (4), failen (3), favorin (1), forbeodan (3), forbeoden (1), forbid (2), fordeman (2), fordemen (1), forwregan (1), fremian (3), gereccan (1), heȝen (3), herian (2), herien (2), judge (1), justify (1), lasten (1), leanian (3), pardon (1), prescribe (1), prosecute (1), punish (1), pursuin (1), scrifan (2), tælan (1), trukien (3), unnen (1), utlagian (3), warnien (1), wreȝen (1), wuldrian (1)

Comments This class is based on Levin's class of judgment verbs. It involves judgments and opinions.

Killing

a-deaden (1), acwellan (7), behead (1), cwellan (2), killen (3), slay (2), slean (13), venimin (1)

Comments This class is based on Levin's class of killing verbs.

Learning

learn (6), leornian (10), leornien (10)

Comments This class is based on Levin's class of learn verbs.

Lodge

inn (2), lodge (13), loggen (4)

Comments This class is based on Levin's class of lodge verbs. The verbs involved describe living situations.

Marking

account (1), awritan (6), bewritan (1), biseggen (1), bitacnen (3), ciegan (3), count (2), date (1), descriven (1), diten (1), dub (1), dubbian (1), eahtian (1), haten (3), mark (5), mearcian (1), mearkien (1), multiplie (1), name (9), nemnan (11), number (1), print (1), register (2), rekenen (2), signify (3), surname (1), tacnen (1), tacnian (9), talian (1), tellan (2), writan (15), write (12), writen (8)

Comments This class was constructed for the present study to accommodate verbs that involve classifying and naming (partially based on Levin's class of appoint verbs).

Motion

ærnan (2), ætwindan (1), þræstan (1), a-hon (1), afaran (1), afeallan (1), aflieman (4), agan (8), approach (2), arise (10), arisen (3), ascend (5), aspringan (6), astandan (3), astigan (1), aswæman (1), ateon (2), awendan (1), awindan (1), becierran (1), becreopan (1), befaran (5), beferan (1), began (6), behweorfan (1), beridan (4), besincan (1), bestelan (10), beteon (10), bewindan (2), bigan (3), bregdan (1), caccen (3), catch (1), chase (1), cierran (9), convey (1), creep (2), creopan (1), creopen (1), cross (1), cwellen

(2), dance (1), depart (9), descend (5), discourse (2), dræfan (3), draȝen (4), drifan (1), drifen (2), drive (4), drop (1), efestan (2), ensue (1), eoden (8), erect (1), escape (2), escapen (2), exalt (1), fall (30), fallen (5), faran (229), fare (2), faren (3), feallan (30), feran (99), flee (4), fleoȝen (5), fleogan (9), fleotan (1), flieman (7), fling (1), flowan (6), flowen (1), fly (2), folgian (3), follow (28), folwen (6), forferan (8), fylgan (6), gan (157), gengan (1), glidan (2), go (210), haste (4), hleapan (2), hreosan (1), hunt (2), hwearfian (19), ieran (9), iernan (12), issen (8), læfan (1), lætan (25), læven (1), land (3), lead (5), leapen (1), march (2), mousteren (2), move (3), moven (2), muster (2), nealæcan (11), oðberstan (1), oðfleogan (1), onettan (1), ongan (11), part (2), pass (63), passen (8), plegan (2), proceed (3), ræsan (4), remyen (1), rennen (5), repass (1), resort (2), resorten (1), retire (1), return (20), ridan (12), ride (29), riden (12), risan (2), rise (7), risen (4), roil (1), row (1), rowan (1), run (13), ruten (1), sail (9), scent (1), schipen (1), seilien (1), shake (1), shoot (2), shuffle (1), siðian (4), siȝen (1), siglan (3), sihen (1), sincan (1), skip (1), sniken (1), sojournen (2), spring (2), springan (1), springen (1), spyrian (1), step (1), stiȝen (4), stiellan (1), stigan (1), stirien (2), surmounten (1), swican (2), swifan (1), swimman (1), travel (25), traverse (1), tread (1), tredan (1), trend (1), turn (2), turnen (10), walk (46), wander (2), wealwian (1), wendan (5), wenden (52), wið-teon (1), windan (3), winden (2), wrecan (1), wriðan (1), ymbsellan (1)

Comments This class is based on Levin's class of motion verbs. The class involves all kinds of motion.

Obligation

awiht (12), mot (2), must (66), ought (6)

Comments This class was constructed for the present study to accommodate verbs that involve an obligation to do something. Only attested in the ME and EME datasets.

Occurrence

befall (1), belimpan (4), bifallen (10), bitimen (1), byrian (6), chance (9), happen (27), let (42), limpan (58), limpen (1)

Comments This class is based on Levin's subclass of occurrence verbs. It involves occurrence of an event.

Perception

behold (18), beseon (6), biholden (9), biloken (2), forseon (2), forseon? (1), gate (4), gaze (1), hark (1), hawian (1), hear (33), herken (1), hieran (42), locian (4), lokien (18), look (19), luren (1), note (10), observe (13), onbyrgan (1), overhear (1), regard (1), sceawian (14), see (138), seon (197), smell (1), stincan (1), stinken (2), toten (2), view (1), witnessen (1)

Comments This class is based on Levin's class of perception verbs. It includes both active and passive perception.

Poke

adelfan (2), bedelfan (2), delfan (2), diggin (1), picchen (3), prick (2), stician (2), stick (3), sting (1), stingan (2)

Comments This class is based on Levin's subclass of rummage verbs. It involves searching by means of piercing some surface.

Prevent

merren (2), tarien (3)

Comments This class was created for the present study to accommodate verbs that involve preventing something from happening. Only attested in the ME dataset.

PsychologicalState

æfestian (1), ðolian (3), ðrowian (5), þencan (36), þenchen (34), þolian (3), þolien (4), þrowian (31), þurfan (7), acknowledge (2), admire (1), adore (1), afæran (1), agrymetian (1), aleogan (1), andgiet (10), angsumian (1), arweorðian (1), asmeagan (1), awakien (1), beðurfan (1), behofian (2), believe (14), belyfan (8), besierwan (1), beware (3), biþenchen (6), bihofen (10), bileven (4), blissian (5), care (1), cennan (1), cheer (1), cnawan (3), cnawelechi (1), cnawen (20), cnodan (1), comprehend (1), conject (1), conjecture (1), consent (1), consider (9), covet (1), crave (1), cunnen (8), cwylmian (1), deputen (1), desire (17), desiren (4), disdain (1), doubt (13), dræden (3), dreccan (1), durran (13), duten (5), dyrstlæcan (1), eaðmedan (3), eglan (8), eilen (2), encourage (1), endure (1), ensure (3), envy (1), excite (1), expect (3), fægnian (1), fancy

(4), fear (9), feel (3), feinin (1), felen (10), feogan (1), forʒeteh (2), forgieman (3), forgyltan (1), forhealdan (3), forhogian (1), forseon (2), freosen (1), fright (2), fundian (1), giernan (1), gremian (2), hate (1), hatien (5), heofian (2), hope (14), hungren (1), hurten (1), hycgan (1), iersian (1), imaginen (1), intend (6), know (55), laðien (1), lack (1), laken (2), lament (1), langen (1), leoʒen (1), leven_2 (4), lician (4), lie (31), like (3), liken (2), long (3), love (4), lufian (5), lusten (1), lustfullian (1), luvien (6), lystan (2), mænan (4), mænen (1), marvel (5), mean (7), mervailen (1), miltsian (3), mind (3), mistake (1), misunderstanden (1), munan (3), myndgian (6), mynegen (5), myntan (2), neden (4), need (10), offeren (3), oncnawan (4), ondrædan (9), perceive (9), reason (1), recall (2), rejoice (1), remember (14), remembren (1), respect (1), sarettan (1), scamian (2), scarnen (2), scendan (2), scyldigian (1), seem (24), smeagan (5), stomach (1), styrian (1), suffer (4), suffren (5), suppose (17), supposen (3), suspect (1), swencan (1), syngian (3), tenten (2), think (55), tilian (8), tintregian (1), treowan (1), treowen (7), trow (3), trust (9), trusten (1), tweon (4), understand (7), understanden (12), urge (2), wærcan (1), wacian (4), want (11), wanten (2), weallan (1), wenan (64), weornian (1), will (108), willan (85), willen (47), willian (1), wilnian (8), wish (12), wist (2), witan (59), witen (43), wonder (3), wrænsian (1), wregan (2), wundrian (7), wynsumian (1), wyscan (1)

Comments This class is based on Levin's verbs of psychological state. Here it refers to bringing about or being in some psychological state.

PushPull

bescufan (1), draw (8), haul (2), pluck (1), pull (1), pullen (1), teon (10), thrust (2)

Comments This class is based on Levin's class of exerting force verbs. It involves some entity being pushed or pulled somehow.

Putting

afyllan (4), ageotan (1), alecgan (4), areccan (1), asettan (13), besawan (1), besettan (1), charge (4), clæman (4), depose (1), fill (5), fullen (3), fyllan (10), geotan (9), hebban (1), hellen (1), lay (65), leccan (2), lecgan (23), leggen (16), logian (4), place (2), plant (3), put (25), raise (4), ramify (1), sædian (70), scatter (1), seowen (2), set (23), settan (44), setten (27), smierwan (21), smitan (1), spræden (1), spread (3), stregdan (3), strewian (1), teldian (1), widmærsian (1)

Comments This class is based on Levin's class of putting verbs. It involves putting or placing something somewhere.

Removing

ðwean (1), þwean (2), aðwean (2), adræfan (3), adrifan (10), afierran (1), baðian (1), clænsian (7), fordrifan (12), omit (2), purgen (2), remove (4), rub (4), substract (1), subtract (1), swilian (1), waschen (4), wash (1)

Comments This class is based on Levin's class of removing verbs. It involves removing something from a location, also metaphorically.

Searching

acunnian (1), assaien (1), attempt (1), cunnian (1), endeavor (2), neosan (4), presume (2), search (2), secan (21), sechen (10), seek (4), try (1)

Comments This class is based on Levin's class of searching verbs. It includes metaphorical searching such as attempting something.

SendCarry

aberan? (2), asendan (9), bear (8), beran (13), beren (9), betæcan (1), bring (16), bringan (15), bringen (19), carry (5), deliver (6), delivren (5), fetian (1), forsendan (1), recan (1), send (29), sendan (40), senden (18), tieman (2), weȝen (1)

Comments This class is based on Levin's class of sending and delivering verbs. It involves sending, carrying, and delivering something.

SocialInteraction

ðafian (5), ðeowian (1), ȝeeode (1), þegnian (1), þeowian (4), þingan (5), þreaten (1), þwærian (2), a-cursien (2), abate (1), accompany (2), accord (4), acquaint (2), affect (1), alædan (1), amansumian (3), andettan (7), andetten (1), anointin (2), appoint (8), aspanan (1), assailen (1), assert (1), assure (5), attend (3), awerian (7), baptize (1), befeohtan (1), behave (1), beodan (46), besiege (1), beswican (7), betan (4), bettan (2), bicker (1), bid (11), bisechen (1), bisegen (1), bismerian? (1), biswiken (1), bless (4), bletsian (3), bletsien (4), brute (1), campian (1), ceapian (1), check (1), command (4), commaund (9), compass (1), concern (7), confer (1), confess (11), confirm (1),

confusen (1), conjure (1), conquere (3), conspire (1), contend (1), converten (1), costian (2), cursien (3), dally (2), dare (5), defend (1), defende (1), dihtan (2), disallow (1), disappoint (1), discordin (1), discredit (1), dispose (1), distourben (1), drohtian (5), durren (5), embrace (1), employ (1), enforce (2), enteren (2), entreat (2), espy (2), exilin (1), fehten (10), fenden (1), feohtan (187), feormian (2), flitan (3), forðteon (1), forʒifen (4), forgive (1), forstandan (1), freogan (3), fulwen (4), fulwian (7), fylcian (1), fyrdian (2), gemungian (2), gieman (3), gislen (1), gree (1), greet (1), gretan? (2), greten (1), griðian (2), hadian (12), halʒien (1), halgian (13), hatan (120), help (2), helpan (2), helpen (4), here (15), heren (1), hiersumian (2), hire (2), honor (1), intermeddle (1), interrupt (1), justen (1), kiss (2), lædan (75), læden (13), læstan (4), læten (38), laðian (2), leave (26), loose (1), mæssian (1), manage (1), marry (2), me_ten (18), meddle (1), meet (43), metan (53), minister (1), misdon (1), miss (2), mittan (4), niedan (4), offend (3), offer (14), offrien (4), onliesan (1), oppose (1), oppress (1), ordenen (11), order (2), play (4), pleʒen (1), please (11), praise (1), pretend (1), prevent (3), profess (2), provoke (1), refer (1), refuse (4), regnen (13), rehearse (1), release (1), resignen (1), restan (7), restrain (1), retain (1), ricsian (19), rikien (1), rule (1), sacren (1), sacrifice (1), salfen (2), salute (1), satisfy (2), save (5), semen (7), serve (4), servien (5), sisen (1), skirmish (1), spien (3), spouse (2), spowan (1), spusen (1), spy (3), stead (1), stihtan (1), swiken (1), symblian (1), travail (2), treat (1), trouble (1), underðeodan (1), visit (7), vow (1), wealdan (1), wedden (2), werrien (5), wieldan (1), win (2), winnan (1), winnen (1), worship (1), wreken (1)

Comments This class is based on Levin's class of social interaction verbs. This is understood broadly as interaction proper, but also as actions that follow from (or constitute) social conventions.

SpatialConfiguration

ætsittan (3), ætstandan (2), ahangian (4), ahebban (1), ahon (4), aræran (18), belutian (1), besittan (17), bestandan (1), betynan (2), bi-clyppen (1), buʒen (1), bugan (3), cneolien (3), cneowian (1), cruchen (1), forridan (2), forsætian (1), hang (2), hangian (3), hengen (7), hieldan (1), kneel (2), licgan (75), lie (28), liggen (38), luften (1), ontynan (1), open (2), openian (1), pitch (1), rest (9), restan (14), resten (14), ring (1), settle (2), sit (38), sittan (155), sitten (17), stand (61), standan (88), standen (25)

Comments This class is based on Levin's subclass of verbs of spatial configuration. It involves the placement or position of persons or objects in a space.

Throwing

asceotan (1), aweorpan (11), aworpen (2), beweorpan (1), cast (7), casten (5), puten (13), sceotan (6), smiten (6), throw (1), weorpan (13), werpen (2)

Comments This class is based on Levin's class of throwing verbs. It includes throwing projectiles.

Weather

blawen (1), reinen (1)

Comments This class is based on Levin's class of weather verbs. Only attested in the ME dataset.

Chapter 8

The Old English EC

8.1 Introduction

The present chapter will present an overview of the data from the Old English treebank along with investigations of some of the hypotheses proposed previously. Specifically, I will present empirical evidence regarding the following hypotheses:

(i) *there* is more frequent than other locative adverbs in Old and Middle English ECs

(ii) *there* is more frequent than temporal adverbs in Old and Middle English ECs

(iii) the proportion of adverbs in initial position increased in the Old English period

(iv) sentences with *there* in initial position had a lower syntactic complexity than other sentences

Some practical interpretation must be given to these hypotheses. As far as the two first are concerned I will establish whether *there* is used overall more than other locative adverbs or temporal adverbs. The association with the existential construction (EC) cannot be established directly, but will be approximated through measuring associations with the verb *be*. As *be* is taken to be the prototypical existential verb, it is expected that any association between *there* and the EC should be measurable through *be*. As far as what counts as initial position, this will be further defined in section 8.4 below. The question of whether initial adverbs grew more frequent in Old English is recast as the probability of having an initial locative adverb in a token from the York-Toronto-Helsinki Corpus of Old English, discussed further in section 8.8. Additionally, questions such as corpus homogeneity and size will addressed.

8.2 Data

The prose part of the York-Toronto-Helsinki Corpus of Old English (YCOE) treebank consists of 1 449 722 words, in about 110 136 main clauses (mostly tagged as IP-MAT, i.e. a matrix Inflection Phrase). The main clauses are the basic units of the corpus and are referred to as *tokens*. These tokens are syntactically annotated and the words tagged for parts of speech.

8.2.1 Collecting the data

Based on the hypotheses, the aim was to collect all instances of locative adverbs in YCOE and the syntactic tree they occur in. This initial data collection step was carried out by running the following query with Corpus Search:

(1)
```
node:  IP*
query:  (ADVP-LOC exists)
```

The next step consisted of processing the extracted tokens with the scripting language Perl to fit a dataframe format which can be read by R. Table 8.1 shows five columns of the first six rows of the data frame. The full dataframe has 9 203 rows and 62 columns. This processing step made use of Perl's regular expression capabilities, see sample code in appendix C. For a good introduction to Perl see e.g. Lemay and Colburn (2002). For an example of the use of Perl in corpus linguistics see e.g. Danielsson (2004).

TABLE 8.1: *An excerpt from the* OE *dataframe, showing six columns from the first six rows. See chapter 5 for a description of the measurement variables.*

	Adverb	Text	There	LogComplexity	Year
1	+tar	coadrian	TRUE	-0.18	1050
2	ufan	coadrian	FALSE	0.56	1050
3	+t+ar	coadrian	TRUE	0.63	1050
4	+t+ar	coadrian	TRUE	1.14	1050
5	her	coaelhom	FALSE	1.63	NA
6	her	coaelhom	FALSE	1.54	NA

Since the YCOE material is not tagged for existential *þær*, the search in (1) which returns all adverbs tagged as locative adverbs,[1] was considered an acceptable approximation. Keep in mind that one of the questions to be answered is what made *þær* special enough to be singled out as an existential subject. To properly assess this, we need to compare the characteristics of *þær* with those of other adverbs. This is not a "sample" in the sense of a randomly selected sample; in fact, it could be considered a full population from a statistical point of view (all cases of locative adverbs in YCOE). For this reason, I use the neutral term "selection" to refer to this and other collected data.

YCOE itself is not a random sample of Old English, but since the corpus was not created with the study of existentials in mind, no particular bias is expected regarding the present object of study (beyond any coincidental bias due to a non-random selection of texts). See also the discussion on data in chapter 1.

As shown in figure 8.1, much of the material is translated. This might be considered problematic, but for the purposes of the current research a pragmatic approach is followed. Initially, all tokens whether translated or not are considered to belong to the same population of Old English utterances.

A quick approximation to this question can be gained by simply looking at the frequencies of *þær* in translated vs. non-translated text in YCOE, as shown in table 8.2. The frequencies appear to be very similar for both the translated and the non-translated texts and the deviation from the expected values are small. Figure 8.2 shows a Cohen-Friendly plot for the table, with the relative size of each cell's contribution to the chi-square value. A Pearson chi-square test for 8.2 shows that translation and use of *þær* are not independent of each other under the null-hypothesis of a uniform distribution ($\chi^2_{df(1)} = 27.93$, $p < 0.01$, $\phi = 0.07$). However, the association as measured with the ϕ coefficient shows that the association between rows and columns is extremely low, and that the significant p-value is probably a result of the large number of observations (6 491 – the total number of tokens in the selection for which translation status is known). Thus, for all practical purposes the use of *þær* in YCOE seems unrelated to whether the text is a translation or not.

8.2.2 An overview of adverbs

There is clearly the most frequently occurring locative adverb in Old English. Table 8.3 lists the 12 most frequent locative adverbs in YCOE. Out of 9 203 corpus tokens with at least one locative adverb, 5 356 contain *there*. In other words, *there* is present in

[1] With the additional restriction that they should occur within an IP. Some tokens in YCOE are not tagged as IPs, but these fragments are few in number.

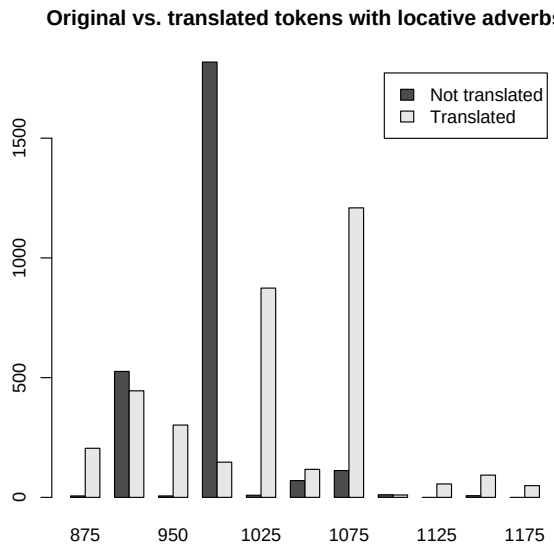


FIGURE 8.1: *The distribution of translated and original sentence tokens with locative adverbs in OE, by period. Total number of tokens included is 6 491, tokens with an uncertain status are not shown.*

about 58% of the tokens with locative adverbs in the material, while the other 42% of the tokens are shared by at least 100 other locative adverbs (of which *here* is the most frequent).[2]

Table 8.4 shows raw frequencies by 25-year interval for *þær* vs. other locative adverbs in YCOE.

[2]There is considerable spelling variation in the material, so this remains a tentative figure as no serious attempt has been made to identify the exact frequencies for low-frequent adverbs that anyway are beyond the scope of this investigation.

TABLE 8.2: *Frequencies of* there *vs. all other locative adverbs for translated and non-translated texts in OE. The observations in the table comprise the 6 491 tokens for which translation status is known. The columns labeled* Exp *give the expected frequencies, which are close to the observed data. The association between* there *and translation is very weak* ($\chi^2_{df(1)} = 27.93$, $p < 0.01$, $\phi = 0.07$).

	¬there	Exp	*there*	Exp
Non-translated	1 195	(1 092)	1789	(1 892)
Translated	1 181	(1 284)	2 326	(2 223)

8.3 Sampling and representativity

Although the basic assumption in the present study is that there is no specific bias in the material, the lack of random sampling might still be of some concern. If the collection of tokens with locative adverbs obtained from YCOE is overly influenced by the size and composition of the corpus, it will be difficult to make comparisons with later linguistic periods. Even worse: if there are systematic selection mechanisms creating biases in the material (over- or underrepresentation), it is likely that the common assumptions of most statistical tests (random sampling) will be severely violated. It is thus important to give some thought to this issue before proceeding with the analysis of the corpus.

8.3.1 Representativity

One way to approach this problem is to look at the distribution of locative adverbs in the material. If the material is a reasonably random sample, then we would expect their frequency to follow a mathematical distribution. Figure 8.3 on p. 169 plots a *frequency spectrum* for the locative adverbs. A frequency spectrum is the number of types per frequency class – i.e. how many types (in this case adverbs) occur once, twice, and so on, cf. Baayen (2001, 8). Using the terminology in Baayen (2001), m denotes the frequency class, i.e. the number of token occurrences. $V(N)$ is the total type frequency in a sample of N tokens, and $V(m, N)$ is the number of types occurring exactly m times, so that $V(1, N)$ is the number of types occurring only once, $V(2, N)$ is the number of types occurring twice, etc. (Baayen, 2001, 8). As the plot shows, there are 79 adverbs that only occur once (hapax legomena),[3] 21 that occur twice (dis legomena), and 14 that occur three times, etc. The curve drops sharply and ends with a

[3]From Greek *hapax* "once" and *legomenon* "read" (Baayen, 2001, 8).

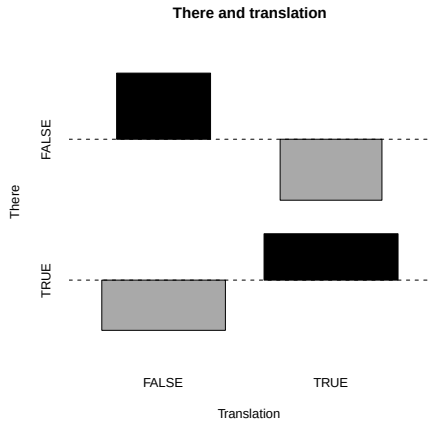


FIGURE 8.2: *Cohen-Friendly plot showing the contributions to the chi-square value for each cell in table 8.2. As the figure shows, there is an overrepresentation of* there *in translated material, but all the contributions to the chi-square value are of approximately the same size.*

long right-tail, suggesting that we have a scenario where a few events occur frequently while a large number of events are infrequent.

An initial reaction might be to try to model this as a *Zipf distribution*, based on the relation between decreasing rank order and frequency of tokens with this rank order, cf. Baayen (2001, 13–15). However, as discussed in Baayen (2001, 17–34) and Baayen (2008a, 222–236), the Zipf distribution is overly sensitive to differences in sample size. Essentially, the number of unseen tokens in small and/or uneven text samples will cause an overestimation of the observed types, cf. also Gale and Sampson (1995) and Gale (1994). Put differently, a Zipf-model will not be able to inform us about how much the proportion of, say, *there* is overestimated in the current data collection compared with a hypothetical full population of utterances.

An alternative, as discussed in Baayen (2001, 51–57) and Baayen (2008a, 230–236), is to use a *Large Number of Rare Events* (LNRE) model, which attempts to estimate the missing or unseen types. Such a model is implemented in the *zipfR* library in R, cf. Evert and Baroni (2008).

Finding the right distribution to fit the model to is a question of trial and error: a *Generalized Inverse Gauss-Poisson* (GIGP) model in *zipfR* gives poor results (GIGP: $\chi^2_{df(3)} = 28.79$, $p < 0.001$), which suggests that this is the wrong model. A *finite Zipf-Mandelbrot* (fZM) model performs much better (fZM: $\chi^2_{df(3)} = 7.15$, $p = 0.067$), as

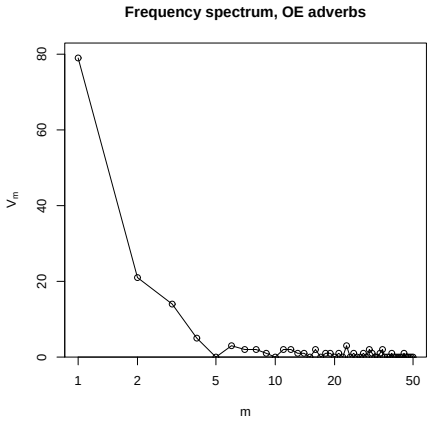


FIGURE 8.3: *Frequency spectrum plot for the first 50 types of locative adverbs. The x-axis shows token frequencies (m), whereas the y-axis shows the number of types V that occur m times. A large number of adverbs have very low frequencies (79 hapax legomena, 21 dis legomena), while a few types contribute most of the tokens.*

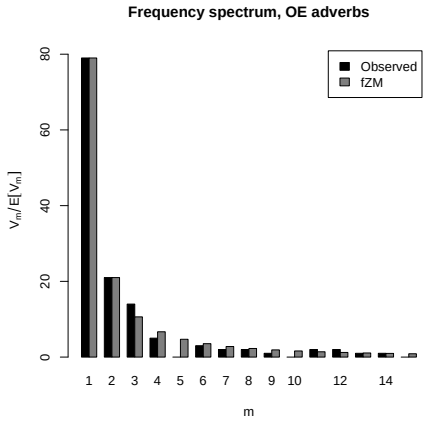


FIGURE 8.4: *Frequency spectrum for OE locative adverbs alongside predictions of the finite Zipf-Mandelbrot LNRE model. The y-axis shows the observed and expected number of types V that occur exactly m times. The plot shows the 15 types with the lowest frequencies.*

TABLE 8.3: *The twelve most frequent locative adverbs in* YCOE*, with the cutoff point set to 50 observations. In the table spelling is normalized for items marked with an asterisk.*

OE locative adverbs (freq. $>$ 50)	
Adverb	Frequency
þær∗	5 356
her∗	2 313
utan	129
neah	124
feor	100
innan	99
inne	79
gehwær	63
feorr	59
ufan	58
gehende	58
ute	56

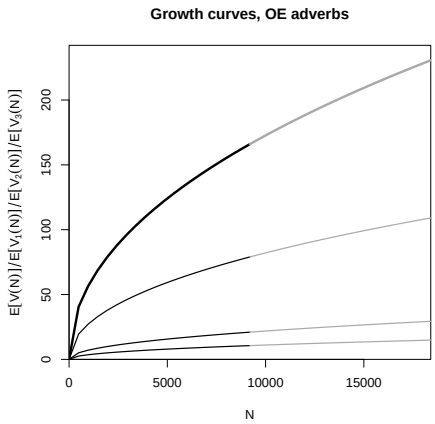


FIGURE 8.5: *Growth curves for OE locative adverbs. The thick upper line represents all adverbs, while the three thinner lines represent (from top to bottom), hapax, dis, and tris legomena. Black lines represent interpolation to smaller sample sizes and gray lines are extrapolation to twice the observed sample size. The y-axis represents type frequency and the x-axis sample size.*

TABLE 8.4: *Frequencies of* þær *vs. all other locative adverbs in* YCOE *by 25-year intervals.*

Year25	There	Other
850	29	22
875	92	119
925	526	445
950	148	160
975	189	123
1000	1 386	614
1025	550	333
1050	693	846
1075	942	401
1100	11	10
1125	417	508
1150	101	47
1175	56	16
Total	5 140	3 644

indicated by the low χ^2 value and relatively high p above the conventional cut-off point of 0.05, cf. Baayen (2008a, 233). Figure 8.4 on page 169 shows the observed frequency spectrum and the fitted fZM predictions. The plot supports the multivariate chi-square statistics, showing a very good fit. There are evidently some problems, especially at five and ten occurrences, possibly due to the irregularity introduced by the same type being represented by different spelling variations in the data. However, the overall fit appears to be good.[4]

The point here is to model the reliability of the material, as discussed in Baayen (2001, 199–203). If the sampling error when compiling the corpus was high, it might be the case that *there* is over- or underrepresented in the material. Figure 8.5 on p. 170 plots growth curves for the adverbs in the material. The thick, upper line plots the overall curve for the adverbs, while the thinner lower curves plot (respectively) curves for hapax, dis, and tris legomena, with black curves representing observed interpolation to lower sample size and gray curves extrapolation to larger samples. What this

[4]The fZM, as indicated by its name, assumes a population of finite size. A non-finite Zipf-Mandelbrot (ZM) model typically gives a poor fit to the data according to Baroni and Evert (2006, 20), although it may be useful for estimation purposes.

shows is that with a data set of 9 203 adverbs, we have probably captured a fairly good range of the variation in the (hypothetical) population of utterances, although for hapax legomena the sample still appears to be in the LNRE zone where adding more material will give more types, cf. Baayen (2001, 56–57). As the interpolated curves to smaller sample sizes show, a smaller sample would risk missing a good part of the types, with an estimate of only 80 adverb types in a sample of 2 000 tokens and only about 57 types in a sample of 1 000 tokens. Even 5 000 tokens would only capture around 80 types. Thus, going from 5 000 to 10 000 tokens gives about twice as many types, while adding another 5 000 would, at an estimated sample size of 15 000, give an estimated 210 types. Note that already with tris legomena there is little gain from increasing sample size. This means that for high frequency phenomena, the sample should be adequate.

Productivity

It is worth asking what practical difference increasing sample size would make to the current project, since it is the high-frequent types that are of interest here: would the adverb types that might emerge at higher sample sizes be relevant? It would be interesting to measure the degree of productivity for adverbs as a whole to assess this, since adverbs as a class probably would be less productive than other classes. It should be noted that "productivity" is a concept with many uses in linguistics as documented in Barðdal (2008, 10–19). Here the term is used in a way related to Baayen (2001, 154) and Baayen (2003, 234-242), that is, in an explicitly probabilistic sense. One such measure of productivity is $\mathscr{P}$, proposed as a measure of morphological productivity in Baayen (1992) and Baayen (2001, 154–158). $\mathscr{P}$ measures productivity by dividing the number of hapax legomena by the size of the sample. In the case of Old English adverbs $\mathscr{P} = 0.01$, which is a low number compared with the results reported in e.g. Lüdeling and Evert (2005, 360–361). However, as they point out, $\mathscr{P}$ is still dependent on sample size, and it should be augmented by further measures, such as the fZM model parameter α. Lüdeling and Evert (2005, 362) suggest that when $\alpha \approx 0.5$, a process is moderately productive, on a scale from 0 to 1 where 0 means no productivity at all. For the Old English locative adverbs fZM model $\alpha = 0.48$. As discussed in Baayen (2001, 206–208) there tends to be an overestimation bias in growth curves, and although it seems likely that more types of adverbs would emerge if more Old English text became available, the weight of the evidence is against a large change in proportions for the high-frequency phenomena under investigation here.

To sum up this section, it appears that the current data of about 9 000 locative adverbs constitute a fair model of the range of adverb types in Old English. A fZM model provided a good fit to the data, suggesting that the data drawn from a period spanning

over 300 years as a whole are reasonably homogenous, and that under- or overrepresentation of the feature under investigation (*there*) should not be a major concern.

8.3.2 Dispersion

Related to the previous section is the question of whether *there* and initial *there* are used very unevenly throughout the Old English period, that is, how well dispersed is the phenomenon under investigation over the categories or classes under consideration? In the previous section the fZM model suggested that the locative adverbs are fairly evenly distributed. However, the data were treated as coming from a homogenous population, but what if the proportion of *there* changed radically during the Old English period? In this case we would expect to find occurrences of *there* lumped together in some parts of the corpus, rather than being evenly dispersed throughout the all the parts.

A way of testing this is using a *dispersion measure*, cf. Gries (2008). The deviation of proportions or *DP* measure, introduced in Gries (2008, 414), measures how much the observed proportions deviate from the theoretically expected proportions[5] if the proportions were distributed perfectly even.

The DP is defined in formula 8.1, cf. Gries (2008, 415), where *EP* and *OP* stand for "expected proportion" and "observed proportion", respectively.

$$DP = \sum_{i=1}^{n} \frac{\mid EP_i - OP_i \mid}{2} \tag{8.1}$$

Thus, DP is the sum of the absolute value of the expected proportions minus the absolute value of the observed proportions, divided by two. DP can range from approximately zero to approximately one, where a value close to zero means that the word is distributed over all corpus parts, whereas a value close to 1 means that it is lumped together.

Computationally, this is done by taking the observed cases of *there* versus other locative adverbs and dividing by the total number of corpus tokens per 25-year interval in the corpus. From these proportions the expected proportions are computed. Applying the formula in 8.1 gives a DP of 0.11, i.e. fairly close to zero. That is, the proportion of *there* is more or less similar across the whole Old English period, when measured against other locative adverbs by total sample size.

[5] Given the assumption of a discrete uniform distribution, that is, every value is equally likely, the expected proportions are computed as the outer product of the row sums and column sums of a $n \times m$ matrix divided by the sum of the matrix.

8.4 Syntactic complexity

One of the hypotheses laid out earlier was that we would expect tokens with existential *þær* to have a lower syntactic complexity than tokens with the locative use. Remember that the log syntactic complexity ratio of a token is expressed as

$$log\ SCR = \ln\left(\frac{NP \times IP}{\sqrt{Nodes}}\right) \tag{8.2}$$

As chapter 6 showed, the log transformed complexity ratio is reasonably close to a normal distribution. A normal distribution of data is a key assumption for using a t-test. The assumption of equal variance in the classical *Student*'s test can be met by using a *Welch* t-test which compensates for unequal variances. Chapter 6 discussed possible validity problems with the log SCR which might make it difficult to argue that the data are really on an interval scale. However, it was concluded that the advantages of this assumption were greater than the disadvantages.

The mean of `LogComplexity` for the YCOE selection data is 0.60, with a standard deviation of 0.99. Thus, we would expect 95% of the observations to lie within two standard deviations from the mean, or the range [-1.38, 2.58]. The mean log complexity for tokens with *þær* is 0.64, while it is 0.55 for tokens with another locative adverb. In other words, tokens with *þær* appear to have a *higher* complexity, but the differences seem small and well within the expected range.

The difference can also be subjected to a formal statistical test. With a normally distributed variable, a t-test is an obvious choice for testing the difference between two groups. A nondirectional two sample Welch t-test[6] reveals that the difference in means is statistically different ($t_{df(8059.61)} = -3.98$, $p < 0.01$). The estimated 95% confidence interval for the mean difference is [0.04, 0.12], in other words, the true mean difference between the two groups is most likely with this range, which means that the true difference is between $\frac{1}{23}$ and $\frac{1}{8}$ of one standard deviation from the mean. This does not seem like a great difference, and an effect size measure can be used to get a better impression of the magnitude of the difference.

The size of the effect can be evaluated with Cohen's d. Cohen's d is defined as mean A minus mean B divided by the standard deviation, cf. Cohen (1988, 20):

$$d = \frac{\mu_A - \mu_B}{\sigma} \tag{8.3}$$

where μ_A and μ_B are the two (population) means, and σ is the standard deviation. Since we have two standard deviations, the pooled standard deviation is used, defined

[6] A Welch t-test compensates for unequal variances as mentioned above, hence the decimal degrees of freedom.

in Cohen (1988, 44) as the square root of the summed squared standard deviations divided by two:

$$\sigma\prime = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}} \tag{8.4}$$

Cohen's d can thus be calculated as follows:

$$d = \frac{0.55 - 0.64}{\sqrt{\frac{1.01^2 + 0.97^2}{2}}} = -0.08 \tag{8.5}$$

The interpretation of d should (like with all other effect size measures) ideally be made with respect to specific expected values. Cohen nevertheless offers some guiding principles for the interpretation, stating that a d of 0.2 to 0.3 corresponds to a small effect, 0.5 corresponds to a medium effect, and 0.8 and up corresponds to a large effect (Cohen, 1988, 24–27). The effect size as measured by Cohen's d for the log-transformed complexity in (8.5) shows that this is a very weak effect ($d = 0.08$,[7] which must be considered trivial for a data set with more than 9 000 observations).

However, this includes all cases of *þær*, including locative uses, which would not be expected to have lower syntactic complexity. There are 781 tokens in the selection with *þær* in initial position and theses cases would seem like a much better approximation of the existential use. The mean log complexity ratio for tokens with *þær* in initial position is 0.12, while it is 0.65 for other tokens, so the difference looks substantial. A nondirectional two sample Welch t-test shows a significant difference in log complexity between tokens with *þær* in initial position and all other tokens ($t_{df(994.10)} = 16.56$, $p < 0.01$, $d = 0.58$). Note that the effect size as measured with Cohen's d is much higher, showing a moderate effect. The 95% confidence interval for the mean difference between groups is [0.46, 0.59], or around half a standard deviation.

Nevertheless, this only shows that tokens with *þær* in initial position have a lower syntactic complexity than other tokens; it does not in itself offer an explanation of why this is the case. In particular, it is necessary to consider whether this is a peculiar effect for *þær*, or whether it is related to the clause position, i.e. whether all tokens with initial locative adverbs have a lower log complexity. The mean log complexity for tokens with any locative adverb in initial position is 0.04, whereas for tokens without an initial locative adverb it is 0.79. The mean for tokens with initial locative adverbs is at the lower boundary of the 95% confidence interval for `LogComplexity`, so the difference appears to be substantial. A nondirectional two sample Welch t-test confirms this ($t_{df(5256.18)} = 38.27$, $p < 0.01$, $d = 0.85$) and also shows a strong effect size. The

[7] The absolute value of d indicates a two-directional test, cf. Cohen (1988, 20).

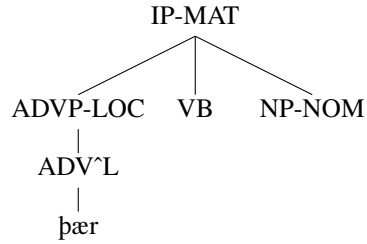
estimated 95% confidence interval for the differences in means in this case is [0.71, 0.79]; an interval which is small and close to one standard deviation.

In summary, there is no real difference in the mean log complexity ratio for tokens with and without *þær*. A medium sized difference is found between tokens with *þær* in initial position and other tokens, but the strongest difference in log complexity ratio is between tokens with an adverb in initial position and tokens with the adverb in other positions. This suggests that there is no specific effect for *þær*, but a position effect, where tokens with a locative adverb in initial position tend to have lower log complexity than tokens with locative adverbs in other positions.

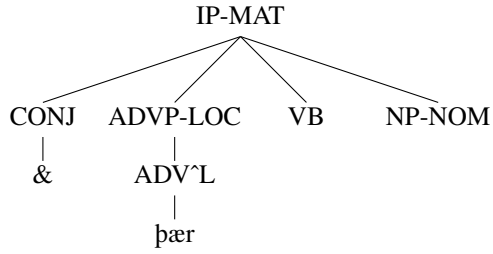
8.5 Initial position

In view of the above findings, a closer inspection of the properties of locative adverbs in initial position seems in order. In the previous section, "initial position" was not precisely defined. What counts as initial position is the first linear order position of the clause, excluding any initial conjunctions. Figures (2) and (3) show schematically cases where *þær* is the first (linear) element of the matrix clause (IP-MAT). The present definition excludes any initial conjunctions if they are present present, so that the adverb in figure (3) is still considered to be initial.

(2)

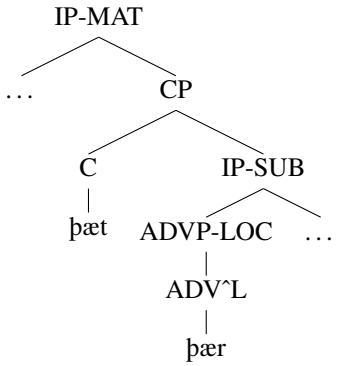


(3)



A different situation is illustrated in figure (4). In this case *þær* is the first element of a subordinate IP, that is, it occurs initially in an embedded clause:

(4)



Such a pattern occurs e.g. when the adverb is preceded by a *that*-complementizer.

Examples of *þær* and *her* in clause-initial position are given below:[8]

(5) Þær wæron eac untrume oðre twægen cnapan,
"there were also ill two other children/youths"
(coaelive,+ALS_[Sebastian]:144.1295)

(6) & ðær is ece gryre;
"and there is also horror/terror;"
(cowulf,WHom_7:122.467)

(7) Her ys se yrfeweard;
"Here is the cattle-keeper"
(cowsgosp,Lk_[WSCp]:20.14.5287)

(8) and her nis nan þearfa.
"and here is not one poor."
(coaelive,+ALS_[Martin]:925.6562)

As examples (6) and (8) illustrate, adverbs that occur with an initial conjunction are still considered to be in clause-initial position as explained above.

There are 2 338 initial locative adverbs in the YCOE material, which means that 25% of the extracted tokens have a locative adverb in initial position. As mentioned above, 781 of these initial adverbs are realized by *þær*. Thus, *þær* constitutes some 33% of all initial adverbs, and tokens with initial *þær* make up about 8% of the selection from YCOE.

[8]The texts are *Ælfric's Lives of Saints* (coaelive), *Wulfstan's Homilies* (cowulf), and *The West-Saxon Gospels* (cowsgosp).

FIGURE 8.6: *Cohen-Friendly plot for table 8.5, giving relative contributions to the chi-square value for the four cells. The underrepresentation of* þær *– and corresponding overrepresentation of other locative adverbs – in initial position is a major contributor to the overall significance of the table.*

A quick way of checking whether *þær* is overrepresented in initial position is to make a contingency table listing the frequencies *þær* vs. other locative adverbs in initial and non-initial position, as shown in table 8.5.

TABLE 8.5: *Frequencies of* there *and other locative adverbs in initial and non-initial position, with expected frequencies in parentheses.*

	¬Initial position	Exp	Initial position	Exp
¬There	2 290	(2 870)	1 557	(977)
There	4 575	(3 995)	781	(1 361)

A Pearson chi-square test of independence is not very informative in this situation ($\chi^2_{df(1)} = 790.62$, $p < 0.01$, $\phi = 0.29$). As expected, the result shows a low p-value formally rejecting a null-hypothesis of independence between *þær* and initial position, but the effect size as measured with the ϕ coefficient is low. As the expected values in table 8.5 show, a major contribution to the statistical significance of the table is that

þær is *underrepresented* in initial position compared with the expected value. This is illustrated by the Cohen-Friendly plot in figure 8.6. Although *þær* is the most frequent locative adverb, this is not entirely unexpected given that the table compares *þær* with all other locative adverbs.

Instead, since table 8.5 is a 2×2 table, we can calculate the *odds ratio* for the table. This gives a ratio between the odds for one outcome compared with the odds of another outcome. Although the odds ratio can easily be calculated by hand,[9] a better option is to use the function for Fisher's exact test (`fisher.test()`) in R. In addition to calculating the p-value for a Fisher exact test (in this case, $p < 0.01$), this function provides an estimated odds ratio for the table, as well as an estimated 95% confidence interval for the odds ratio (i.e. we get an estimate of the precision of the odds ratio). For table 8.5, the estimated odds ratio is 0.25, i.e. for every 1 occurrence of *þær* in initial position, we would expect 4 other locative adverbs. The estimates for the 95% confidence interval of the odds ratio range from 0.23 to 0.28; i.e. a ratio of *þær* to other locative adverbs starting from around $\frac{1}{4}$ approaching, but not quite as much as, $\frac{1}{3}$. Another way of putting this is that out of 100 tokens with initial locative adverbs, we would typically expect the number of *þær* in initial position to vary from 23 to 27 for every hundred tokens. Since *þær* in initial position has an odds ratio of 0.25, i.e. right in the middle of a narrow confidence interval of [0.23, 0.27], the estimate is precise. There seems to be a small but consistent overrepresentation of *þær* in initial position.

The interpretation of an odds ratio is not always intuitive. However, it has an advantage over measures of association such as ϕ in that the odds ratio can be tied to specific outcomes in the contingency table. This is especially useful when the categories are constructed as in table 8.5 with *þær* vs. all other locative adverbs.

8.5.1 Position and semantic verb class

An interesting follow-up question here is whether the position of the locative adverb is somehow associated with certain semantic verb classes. The semantic verb classes are described in more detail in chapter 7. If there are specific constructional links between verbs and the position of certain arguments, this might manifest itself in verb class preferences for adverb placement. We would expect the more "existential" (*Existence*, *Appearance* etc.) verb classes to have a preference for placing the adverb in initial position. For such exploratory analyses of large tables, correspondence analysis (CA) is a very useful alternative to a Pearson chi-square test, as will be demonstrated below. Keep in mind that p-values and association measures such as ϕ and *Cramér V* are

[9] See e.g. Fleiss et al. (2003, 100–125) for an introduction to the odds ratio. See also the discussion in chapter 4.

global measures which pertain to the entire table, whereas here we are interested in specific associations between many categories.

To investigate the association between adverb position and semantic verb class, a table of frequencies for each semantic class by adverb position was constructed using R, where position made up the rows and the 36 verb classes the columns. Adverb position is here defined rather loosely with respect to the matrix clause (IP-MAT) of the corpus annotation: as previously, an adverb is initial if it occurs first in the linear order of the Inflection Phrase (IP), preceding coordinating conjunctions being ignored. The final position is the final linear order position of the IP (marked by a comma or a semi-colon in the corpus). The mid position is any position which is not initial or final, and consequently not a true position at all. Obviously, this is a coarse distinction which disregards much structural linguistic information.

Although this tripartite division is primarily done for pragmatic reasons of coding, I nevertheless believe it has some theoretical justification based on the so-called *serial-position effect*, which states that in a list structure the initial and final elements tend to be remembered more easily than the middle elements, cf. Murdock (1960). Especially the initial position has been found to be associated with distinctiveness (Neath, 1993a,b).[10]

The resulting 3×36 table of position versus semantic class is so large that its size makes it difficult to interpret both in terms of raw figures and percentages. The size also makes it difficult to interpret the outcome of a Pearson chi-square test. The result of such a test is statistically significant ($\chi^2_{df(72)} = 618.81$, $p < 0.01$, *Cramér V* $= 0.184$). However, the extremely low p is not surprising given the number of observations, and the low value for the association measure V with 72 degrees of freedom leaves serious doubts about the *practical* usefulness of the Pearson chi-square test in this case.

As mentioned above, we can get more specific information from such a table using CA. If we take adverb position as rows and semantic class as columns, we can look at the variation in the range of semantic verb classes for the three positions. We would a priori expect the range of semantic verb classes to be more or less equally distributed over the positions, given that all the adverbs in question are locative adverbs.

CA results

The results of the CA are summed up in table 8.6 on p. 182 and table 8.7 on p. 183, the former gives the results of the overall fit, while the latter sums up the results for the rows – i.e. the adverb positions. The numerical results for the columns – the semantic classes – will be discussed in passing and not presented in their entirety here

[10]The full implications of this view will not be pursued at the moment; instead they will be more thoroughly discussed later in the general summary of the corpus findings.

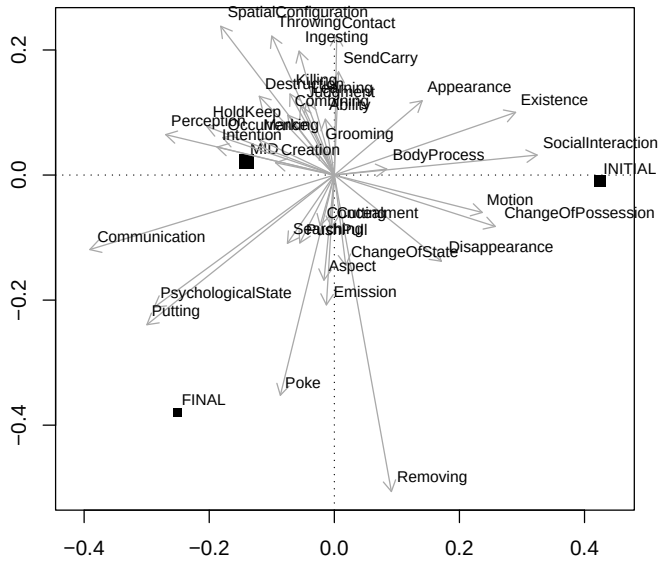


FIGURE 8.7: *Standard CA biplot showing the adverb-position/semantic class data. The horizontal axis accounts for 92.0% of the variance in the data, the vertical axis accounts for an additional 8.0%. Total inertia is approximately 0.07, i.e. fairly low. The rows (adverb position) are in principal coordinates, while columns (semantic class) are in standard coordinates times the square root of the mass. Row point sizes are plotted proportionally to their relative frequency.*

TABLE 8.6: *Summary of the fit statistics for the CA of the adverb-position/semantic class data. The high proportion of explained variation is a sign that the quality of representation in the biplot is good, and that first (horizontal) dimension accounts for nearly all variance in the data. Inertia is low indicating that there is only a small association between rows and columns.*

Dimension (axis)	Inertia	Explained variance (%)
1 (horizontal)	0.062	92.0
2 (vertical)	0.005	8.0
Total	0.067	100.0

for reasons of space. Instead, the full R output is printed in section A.1 of appendix A. See Nenadic and Greenacre (2007) for documentation regarding the output format, see Blasius (1994) for a good example-based introduction to the interpretation of the numerical output.

The quality of the two-dimensional representation is excellent, in that the first dimension (the x axis) accounts for 92% of the variance in the data. What is *represented* by these percentages is the *total inertia* (see chapter 4), which in this case is 0.07. As pointed out by Blasius (1994, 24), the total inertia can be used as a crude measure of total association between rows and columns (adverb position and semantic class, in this case), not unlike the ϕ or *Cramér V* coefficients. Unlike these coefficients, the theoretical maximum of the total inertia is not 1, but the minimum of either rows or columns (whichever is smaller) minus 1; i.e. the *rank* of the matrix. In the present case, the smaller of rows and columns is the number of rows (3 adverb positions), which gives a total theoretical maximum inertia of 2. Thus, in this case the total inertia of 0.07 is around one thirtieth ($\frac{1}{29}$) of its theoretical maximum. This suggests that overall the verb classes are in fact used quite similarly across the adverb positions. However, there is obviously some variation, and this is presented in table 8.7.

Turning to the biplot in figure 8.7, we can interpret the graph as a map of the data. The interesting factor is the position of the adverb in the clause, i.e. rows, so the plot shows the rows scaled in principal coordinates, cf. Greenacre (2007, 68). This implies that the we get a very good picture of the distances between row points (i.e. adverb position) plotted as squares, relative to the first dimension. The semantic classes (represented as arrows) are plotted in standard coordinates, that is, they are adjusted to have a mean of 0 and a variance of 1. The scaling adopted in this case multiplies the

TABLE 8.7: *Summary of the rows in the adverb-position/semantic class data. "Initial" has a small to medium sized relative frequency (mass) and high inertia, indicating high explanatory value.*

Adverb position	Absolute frequency	Relative frequency (%)	Inertia (%)
Final	329	0.035	0.107
Initial	2391	0.256	0.681
Mid	6483	0.709	0.212

columns with the square root of their relative frequency,[11] cf. Nenadic and Greenacre (2007, 9). Note that with this scaling no distance-based interpretation of the column points to the row points is possible, cf. Greenacre (2007, 103). Instead, we can look at the *direction* which the arrows (i.e. the semantic classes) take, relative to the adverb positions. An arrow pointing towards a square indicates that this semantic class is associated with this adverb position (see further details below).

Looking at the adverb positions, we see that they are placed relatively far apart. This is partially an effect of the scaling chosen for the plot. The row profiles are, as mentioned above, in fact very similar. Choosing another scaling would lead to the rows being clustered together in one lump, which would make the interpretation of what patterning is actually present very difficult. As the plot shows, both INITIAL and FINAL are farther from to the origin than MID on the horizontal axis, which suggests that the initial and final positions deviate more from the row average than the mid position.

The adverb positions can be interpreted relative to the directions of the arrows representing the semantic verb classes in the following manner: follow the arrow from its point to the origin and draw an imaginary line from the origin to the adverb position which is to be evaluated. In cases where the angle between the line from the square and the arrow is smaller than $90°$ the adverb position and the semantic classes are positively associated. In cases where the angle is $90°$ the adverb position and the semantic classes do not interact. In the cases where the angle is between $90°$ and $180°$, the adverb positions and semantic classes are negatively associated. Thus, the fact that the endpoint of, say, the *Existence* arrow is closer to INITIAL than *Appearance* has no interpretation here. The interpretation is rather that the former has a smaller angle to

[11] The scaling chosen here – in R code: `map = "rowgreen"` – is a purely pragmatic decision for optimal display of the patterning in the data. This does not affect the interpretation of the plot in any substantial way, save for the interpretation guidelines discussed in the text.

the adverb position point than the latter.

A striking pattern emerges from this interpretation of the plot. Six semantic verb classes appear to be positively associated with locative adverbs in initial position: *Appearance*, *Change of possession*, *Disappearance*, *Existence*, *Motion*, and *Social interaction*. For each of these classes, the arrow representing it has an angle to the INITIAL point in figure 8.7 which is smaller than $90°$. However, turning Firth's well known maxim of knowing words by the company they keep on its head, it is just as interesting to look at the *negative association* in the plot. For instance, verbs like *sit*, *stand*, and *lie* grouped as *Spatial configuration* are located in the upper left panel to the left of the y axis running through the origin. The angle between INITIAL and the *Spatial configuration* arrow is clearly greater than $90°$, which points to a negative association between this row and this column. Other verb classes, such as *Send Carry* and *Killing* simply do not interact with INITIAL since the angle appears to be very close to $90°$.

The numeric output, presented in appendix A, shows that *Appearance* and *Disappearance* contribute very little to the map in figure 8.7. *Motion* and *Change of possession* contribute somewhat more, but that *Existence* and *Social interaction* contribute more to the total inertia, with *Social interaction* being more strongly correlated with initial position than *Existence*.

To sum up this section, it was shown that a cross-table of locative adverb position in the clause against semantic verb classes could provide some indications of construction use. Although the overall explanatory potential of such a table was found to be small, breaking the effects down by adverb position and semantic class showed that most of the observed effect as in fact attributable to the initial position. Six semantic verb classes, including *Existence*, *Appearance*, and *Disappearance* could be shown to be positively associated with the initial position, albeit weakly. Additionally, a number of other verb classes could be shown to either not interact with or be negatively associated with the initial position. Two points thus emerge from this, one relating to the hypotheses being tested, the other methodological. First, relating to the hypotheses, there is a weak association between the choice of verb class and the positioning of the locative adverb. The methodological point pertains to the use of non-directional omnibus tests like the Pearson chi-square test on large tables. While the p-value and any association measure obtained for such a table are certainly valid, their practical usefulness is limited by the large size of the table, which makes it difficult to interpret the result. For such tables, CA is a good alternative providing detailed information about the relationship between individual categories.

8.5.2 Is *þær* a subject?

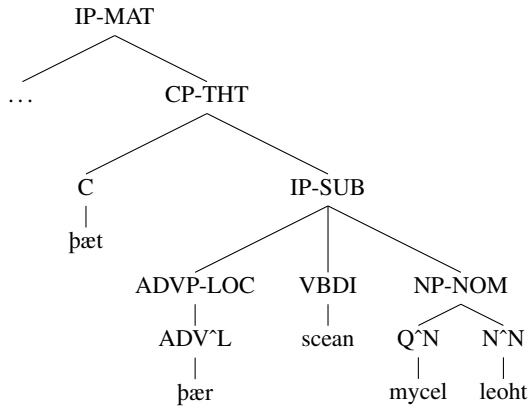
Despite a whole range of attested combinations of the subject, verb and object or complement elements in Old English, the canonical subject position was the initial position: "The prose ... display[s] a considerable tendency towards the order S V O/C in non-dependent clauses" (Quirk and Wrenn, 1957, 92). Quirk and Wrenn (1957, 87) remark that despite the attested variation in Old English word order, there is "considerable conformity to describable patterns" which "to a great extent coincide with present-day usage".

Pfenninger (2009, 13) suggests, in accordance with the data in Breivik (1990), that Old English had an expletive *þær* to fill a required subject position in finite subordinate clauses. An example is given below in (9), with a partial syntactic structure presented in (10).

(9) Þa foresceawode Godes gifu þæt þær scean mycel leoht
"Then provided God's grace that there shone much light"
(coaelive,+ALS_[Julian_and_Basilissa]:213.1068)

As the example illustrates, *þær* occurs initially in a finite subordinate *that*-clause. Although *þær* might refer to a location – light shone "at that place'" – it is not implausible to interpret this as an existential use of *þær*.

(10)



Although other types of finite subclauses would be equally relevant, e.g. clauses with *þonne* (Pfenninger, 2009, 13), I have limited myself to *that*-clauses for reasons of space.

There is another issue at stake here, namely whether a distinction can be drawn between a locative and an existential *þær* in Old English regarding subordinate clauses.

Lakoff (1987, 469) argues that existential *there* constructions in Present-day English can be freely embedded in subordinate clauses, whereas locative *there* constructions are generally restricted from this, with some exceptions. The following examples from Lakoff (1987, 469) illustrate this:

(11) a) I doubt that there's anyone in the kitchen
b) *I doubt that there's Harry in the kitchen

The second (starred) sentence tries to combine *that*-clause embedding with a locative interpretation of *there* ("Harry is in the kitchen, over there"). Moving *there* would make the sentence acceptable, along the lines of *I doubt that Harry's over there, in the kitchen*. However, this argumentation relies on the judgments of native speakers, an option which is obviously not present in the study of Old English. Instead, distributional approaches may provide a clue.

To fully appreciate the role of *þær* in these subclauses, we may consider what is the canonical subject of a *that*-clause in Old English. A search of YCOE revealed that there are 18 746 *that*-clauses in the treebank, distributed over 14 407 corpus tokens (i.e. matrix IPs). That is, 13% of the tokens in the treebank contain a *that*-clause. Of these, at least 12 567 cases of *þæt* are followed by a nominative NP, whereas at least 223 are followed by an accusative NP, and at least 236 by a dative NP. Looking at cases with *þær*, we find that the combination *þæt þær* (with spelling variations) occurs 178 times.

Using conditional probabilities (see chapter 4), we can quantify the association between *that* and these subject-slot fillers. In this case, we are looking at a level below the token (or IP) level, so instead of counting occurrences of IPs, I have counted occurrences of words (or, in the case of NPs, phrases which can be justified by the fact that each NP will have a nominative element as head). Consequently, the denominator in the fractions for calculating independence below have the total number of words in YCOE, not the total number of sentence-tokens.

As explained in section 4.5, the conditional probability of a word given another word is

$$P(w_i \mid w_j) = \frac{n(w_i \cap w_j)}{n(w_j)}. \tag{8.6}$$

The baseline, i.e. words used independently of each other is the product of the marginal probabilities:

$$P(w_i) \times P(w_j) = \frac{n(w_i)}{N} \times \frac{n(w_j)}{N}. \tag{8.7}$$

Following Bilisoly (2008, 116), the conditional probability should as a rule of thumb be at least 8 times higher than the product of the marginal probabilities. With these formulas, we get the following calculations:

First, the most frequent category, viz. NP-NOM:

$$P(NP\text{-}NOM \mid that) = \frac{12567}{18413} = 0.683 \tag{8.8}$$

To see if this indicates an association, we take the marginal probabilities (i.e. the independent probability of NP-NOM and *þæt*) and multiply them:

$$P(NP\text{-}NOM) \times P(that) = \frac{192890}{1449722} \times \frac{18413}{1449722} = 0.002 \tag{8.9}$$

As it turns out, the conditional probability of NP-NOM given *that* is around 400 times higher than the marginal probabilities. Next, accusative NPs:

$$P(NP\text{-}ACC \mid that) = \frac{223}{18413} = 0.012 \tag{8.10}$$

$$P(NP\text{-}ACC) \times P(that) = \frac{71745}{1449722} \times \frac{18413}{1449722} = 6.29\,e-4 \tag{8.11}$$

In this case, the conditional probability is around 20 times higher than the product of the marginal probabilities. That is, there is an association between *þæt* and NP-ACC, but not as strong an association as with NP-NOM. The same holds for dative NPs:

$$P(NP\text{-}DAT \mid that) = \frac{236}{18413} = 0.013 \tag{8.12}$$

$$P(NP\text{-}DAT) \times P(that) = \frac{76783}{1449722} \times \frac{18413}{1449722} = 6.73\,e-4 \tag{8.13}$$

There is an association between *þæt* and NP-DAT since the conditional probability here too is around 20 times higher than the product of the marginal probabilities. As with NP-ACC, however, the association is less pronounced than with NP-NOM. Next, let us consider *there* as a subject-position filler in *that*-clauses:

$$P(there \mid that) = \frac{178}{18413} = 0.010 \tag{8.14}$$

$$P(there) \times P(that) = \frac{5626}{1449722} \times \frac{18413}{1449722} = 4.93\,e-5 \tag{8.15}$$

Here we find that the conditional probability of *there* given *that* is around 200 times higher than the multiplied marginal probabilities. Thus, the two are clearly correlated, and the effect is stronger than for accusative and dative NPs.

This can be compared with *here*, for which there are four occurrences as possible subjects of Old English *þæt*-clauses.

$$P(here \mid that) = \frac{4}{18413} = 2.2\,e-4 \tag{8.16}$$

$$P(here) \times P(that) = \frac{2313}{1449722} \times \frac{18413}{1449722} = 2.0\,e-5 \tag{8.17}$$

As the calculations above show, the conditional probability of *her* given *þæt* is minute, which is not surprising given the frequencies. Although the conditional probability is about 10 times higher than the product of the marginal probabilities, the size of the frequencies involved and the size of the probability itself must also be taken into account. In this case it seems unlikely that any real association exists between *her* and *þæt*.

In view of the above, I would suggest that the category of subject in Old English *þæt*-clauses has a radial structure, with nominative NPs as the prototypical subject. This interpretation is based on Geeraerts (1997, 21). The category ThatClauseSbj (i.e. subject of a *that*-construction) is taken to be composed of members with unequal structural or cognitive salience and overlapping properties. "Salience" can, for the purposes of the present study, be operationally defined conditional probability. The categories are overlapping, or non-discrete, in the sense that both nominative and accusative NPs belong to the category NP, and even *there* can be described as having NP-like properties (Breivik, 1997). Based on the investigation above, nominative NPs are thus taken to be the central members of this category. *There*, accusative, and dative NPs are taken to be non-prototypical *that*-construction subjects due to their lower conditional probabilities of appearing in this position. There is no evidence based on this that *her* plays any role as a possible subject of *þæt*-clauses.

The analysis presented above is only a rudimentary one, since only *that*-subordinate clauses have been considered. Nevertheless, it offers a tantalizing piece of circumstantial evidence for the subject-status of *there* in Old English.

8.6 Associations with *be*

It seems likely that there is a special relationship between *there* and *be* already in Old English. In the present section, this relationship will be described in further depth.

8.6.1 *There* and *be*

It was shown in section 8.4 that there is a tendency towards a lower syntactic complexity score for *there* in initial position. In light of this, it is interesting to note that there is no similarly strong relationship between *be* and syntactic complexity. Mean log complexity for tokens with and without *be* immediately following the adverb is 0.40 and 0.63, respectively. The difference in means is (as expected, due to a large number of observations) statistically significant in a nondirectional two sample Welch t-test, but shows only a small effect size ($t_{df(1602.23)} = 7.69$, $p < 0.01$, $d = 0.24$). An estimated 95% confidence interval for the difference covers the range [0.17, 0.29], which amounts to between approximately $\frac{1}{6}$ and $\frac{2}{7}$ of a standard deviation from the mean.

Based on the investigations carried out so far, it is reasonable to expect some kind of relationship between *þær* in initial position and *be*. A simple way of testing this is to set up a contingency table listing the possible combinations and carrying out a chi-square test. Such a table is shown in table 8.8. A Pearson chi-square test with Yates' correction for continuity comes out as as significant, i.e. initial *þær* and *be* are not independent of each other ($\chi^2_{df(1)} = 1430.26$, $p < 0.01$, $\phi = 0.39$). However, association as measured with the ϕ coefficient is medium sized.

TABLE 8.8: *Initial* þær *and other locative adverbs (including non-initial* þær*) vs.* beon *as right context and other right contexts in* YCOE*. Association as measured with the ϕ coefficient is low to moderate. Numbers in parentheses are expected frequencies.*

	¬*Be*	Exp	*Be*	Exp
¬InitialThere	7 664	(7 323)	758	(1 099)
InitialThere	338	(679)	443	(102)

As noted previously, a notorious problem with contingency tables is that it is difficult to see *where* in the table an effect arises, since the χ^2 value is computed based on the whole table, a problem which also affects ϕ. As above, we can get more information by looking at expected values and by looking at the contributions to the χ^2 value in a Cohen-Friendly plot.

Looking at the expected frequencies in table 8.8, we find that *þær* initial position is overrepresented both in contexts with (and underrepresented in contexts without) *beon*. Figure 8.8 on page 190 shows a Cohen-Friendly plot for table 8.8. The plot confirms that *þær* in initial position immediately followed by *beon* is overrepresented and probably has the greatest contribution to the overall significance.

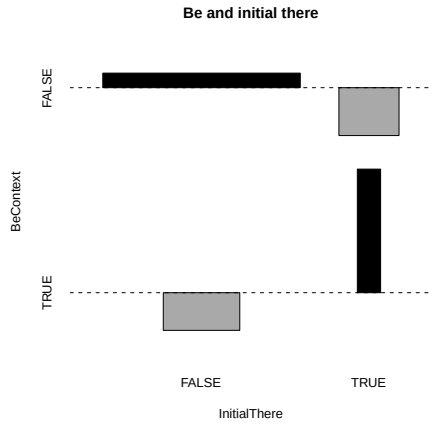


FIGURE 8.8: *Cohen-Friendly plot for table 8.8,* there *in initial position vs.* be *immediately following. The dotted lines represent expected values, whereas the size and direction of the bars represent deviations from expected values.*

Although these methods are useful, they compare *þær* in initial position and *beon* as the right context with everything else, which results in a rather crude picture of the situation.

Another way of looking at the relationship between *þær* and *beon* is in terms of chances of occurrence using conditional probability, rather than formal null-hypothesis testing.

With conditional probabilities, we can estimate the chances of seeing *there* or *be*, given information about one of them. The conditional probability of *be* given *there* is thus calculated as follows:

$$P(Be \mid There) = \frac{n(There \cap Be)}{n(There)} = \frac{846}{5356} = 0.146 \tag{8.18}$$

In other words, if we have observed *there* in the sample, the probability of observing a form of *be* immediately after it is about 14.6%. Conversely, we can calculate the conditional probability of *there* given *be* as follows:

$$P(There \mid Be) = \frac{n(There \cap Be)}{n(Be)} = \frac{846}{39350} = 0.021 \tag{8.19}$$

This suggests that the probability of observing *there* in front of a form of *be* is about 2%. Formally $P(Be \mid There) \neq P(There \mid Be)$, since the order matters when dealing with conditional probabilities, cf. Bilisoly (2008, 117). This might seem counterintuitive at first: how can the conditional probability of *there* given *be* be different from *be* given *there*? The answer lies in the denominator of the fraction. Since the conditioning word (on the right hand side of the vertical bar) is used in the denominator, it constrains the sample space of possible events to occurrences where this word is present. Put differently, although *there* might co-occur with *be*, *be* might co-occur a lot more with something else than *there*.

However, what does this conditional probability really amount to? Clearly, some kind of baseline is needed to compare these probabilities. Bilisoly (2008, 117–118) explains that conditional probabilities can be used as a test for independence of event A and event B: if $P(A \mid B)$ is close to $P(A) \times P(B)$, then A and B are independent. The conditional probabilities above can thus be evaluated by multiplying the marginal probabilities. The marginal probabilities multiplied gives the following result:

$$P(There \times Be) = \frac{5356}{110136} \times \frac{39350}{110136} = 0.017 \tag{8.20}$$

Since $P(There \mid Be) \sim P(There) \times P(Be)$, the relationship between *there* and *be* appears to be one-directional. The probability of *there* given *be* is only around 1.2 times higher than independence, which for all practical purposes means that there is no association. For *be* given *there*, on the other hand, the conditional probability is about 8.4 times higher than the product of the marginal probabilities. Bilisoly (2008, 118) suggests as a rule of thumb that a factor of at least 8 indicates an association. Although I will follow this convention, any such cut-off point must of course be treated with caution. Based on this rule of thumb, there is a higher-than-chance probability of encountering *beon* after seeing *þær* in YCOE.

For *þær* in initial position and *beon*, the calculations are equally straightforward.

$$P(Be \mid Init.there) = \frac{n(Init.there \cap Be)}{n(Init.there)} = \frac{443}{781} = 0.567 \tag{8.21}$$

$$P(Init.there \mid Be) = \frac{n(Init.there \cap Be)}{n(Be)} = \frac{443}{39350} = 0.011 \tag{8.22}$$

$$P(Init.there) \times P(Be) = \frac{781}{110136} \times \frac{39350}{110136} = 0.003 \quad (8.23)$$

In other words, there is a strong effect for *beon* given *þær* in initial position which has a conditional probability 224 times higher than independence, whereas the converse probability is only 4.4 times higher and thus so close to independence that no association seems to be present. Since a strong result was found for *beon* given *þær* in initial position, and the result for *beon* given *þær* in general was less strong, we might suspect that most of the effect is caused by *þær* in initial position. To check this, we can remove those cases of *þær* that occur in initial position and calculate the conditional probability of *beon* given the remaining cases of *þær*:

$$P(Be \mid \neg Init.there) = \frac{n(\neg Init.there \cap Be)}{n(\neg Init.there)} = \frac{403}{4872} = 0.08 \quad (8.24)$$

And the check for independence:

$$P(\neg Init.there \times P(Be) = \frac{4872}{110136} \times \frac{39350}{110136} = 0.016 \quad (8.25)$$

The conditional probability of *beon* given non-initial *þær* is only 5 times higher than independence, which suggests that the association – if any – is very weak. It seems warranted to conclude that the effect which was observed for all cases of *þær* above was mainly caused by *þær* in initial position.

To summarize this, we can plot the co-occurrences of *þær* and other elements occurring in its left contexts. Figure 8.9 shows Cohen-Frienldy plots for the possible combinations of *þær* with *beon* and nominative NPs. As the plots show, *beon* is overrepresented in both the first and second right context of *þær*. Nominative NPs, on the other hand are underrepresented, compared to an expected equal distribution of proportions, for both right contexts. Although the differences are significant, the effect sizes are low, and it is reasonable to conclude that there is only a weak association between *þær* in general and *beon* and nominative NPs.

However, if we turn to *þær* in initial position we get a different result, as shown in figure 8.10. Here it is clear that there is an overrepresentation of the patterns *þær* + *beon* and *þær* + . . . + nominative NP. The differences are significant with an alpha level of 0.01, and the effect sizes are close to moderate. Conversely, the differences for the patterns *þær* + nominative NP and *þær* + . . . + *beon* are not significant with an alpha set to 0.01 (in fact, it is not even significant at the 0.05 level) and effect sizes are negligible. Thus, for *þær* in initial position we find a moderate association with *beon* as the first right context and nominative NPs as the second right context.

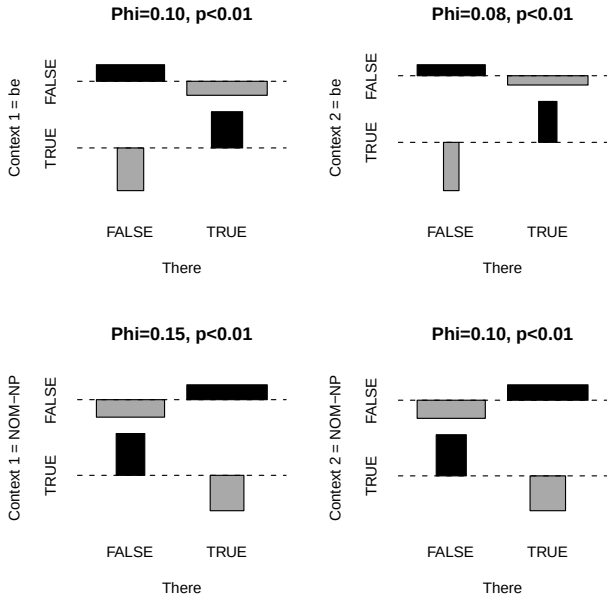


FIGURE 8.9: *Cohen-Friendly plots for associations between* there*,* be*, and nominative NPs. The first and second right context of* there *is coded as either* be *or non-*be*, or nominative NP or not nominative NP. All differences are statistically significant at the 1% level, but effect sizes are small. There does not appear to be a strong association between* there*,* be *and nominative NP.*

8.6.2 Other contexts of *be*

To put the observations above in context, it is useful to look at what other elements can occur with *be* in YCOE. This was done through extracting all IP-tokens with *be* and running a Perl script which carried out a simple KWIC (Key-Word In Context) search and returned a trigram with a left and a right context for *beon*.[12] As stated above, there are 39 350 tokens containing a form of *beon* in YCOE, with a total of 50 266 instances of *beon*.

The most frequent left context for *be* are trace positions of NPs, with at least 1 684 occurrences. The most frequent non-empty left context for *be* in YCOE is the pronoun *he* in the nominative with 1 828 occurrences. Although there is much variation, pro-

[12]The Perl code can be found in appendix C.

nouns stand out: there are 6 041 pronouns occurring as the left context of *be*, most of which are in the nominative (638 or 10.6% of the total pronoun count are datives, no other case markings were found with frequencies above 50). Thus, 12.0% of all the 50 266 instances of *be* in YCOE have a pronoun as left context. If we add traces, empty expletives and other nominal elements, these make up at least 8 778, or about 22%, of the total left contexts for *be*. The temporal adverbs *þa* and *þonne* occur at least 1 240 and 609 times, respectively. *þær* occurs 846 times, and *here* 285 times.

Another high-frequency item is the negative particle *ne* which is found 1 220 times, but the *Large Number of Rare Events* effects seems clear: with the exception of a few highly frequent categories, most elements occurring as the left context of *be* have low frequencies. PPs, for instance, occur only 99 times, the modals *sceal* and *mæg* occur only 110 and 65 times, respectively. It would seem that the left context of *be* displays a wide variation, but with nominal elements in the nominative and especially pronouns being the dominant category.

Turning next to the right context, the impression is that the choice of categories is more restricted, that the categories on the whole have higher frequencies, and that the transition from high to low frequencies is smoother. In the right context, PP is the most frequent single category with at least 8 205 occurrences (approximately 21% of all right contexts – if 1 293 single prepositions are added to this, the percentage increases to 24%). The second most frequent category is the end of the token (as attested by its ID tag turning up in the trigram), with 6 797 occurrences (17%). Next we find verbal nouns, complementizers, adverb phrases and conjunction phrases before the frequencies drop below 2 000 (the respective frequencies are 5 497 or 14%, 3 939 or 10%, 2 307 or 6%, and 2 234 or 6%). A number of NPs (1 324), conjunctions (1 142) and adverbs (1 098) are also found, each of which amounts to 3% with rounding to whole percentages.

For the sake of comparison with the sections above, it is interesting to look at the conditional probability of finding a form of *beon* after a nominative NP, i.e. more or less a SUBJECT-BE pattern. There are some 8 373 occurrences of nominative NPs and nouns preceding *beon*. Including pronouns in the nominative raises the count to 14 503, which gives the following calculations:

$$P(be \mid NP\text{-}NOM = \frac{14503}{200930} = 0.072 \tag{8.26}$$

$$P(be) \times P(NP\text{-}NOM = \frac{50266}{1449722} \times \frac{200930}{1449722} = 0.005 \tag{8.27}$$

The conditional probability of finding a form of *beon* following a nominative NP is around 0.07. Although not very high, this is nevertheless 15 times higher than the

product of the marginal probabilities, so it seems likely that there is some association between the two categories. However, compare this with the numbers for initial *þær* and *beon* discusses above: a conditional probability of over 0.5, which was more than 200 times higher than we would expect if they were used independently. There is in other words a much closer affinity between *þær* and *beon* than between the general subject category of nominative NPs and the same verb.

In summary, both left and right contexts of *be* display abundant variation regarding the attested arguments. However, for the left context the frequencies drop much more quickly than for the right context. Pronouns constitute the single largest group of elements in the left context, whereas for the right context the arguments are much less constrained. This does not in any way constitute an exhaustive study of the argument structure of *be* in YCOE, and is intended only as a background to the other analyses. Nevertheless, it is instructive in that it illustrates the "ecological niche", to use a biological metaphor, which these contexts represent.

8.7 Locative vs. temporal adverbs

So far only locative adverbs have been considered, for the simple reason that the various forms of *there* are tagged as locative adverbs in YCOE. However, there are other interesting Old English adverbs, such as the temporal adverbs *þa* and *þonne* (both meaning "then", see also the discussion in section 2.9.1). For this purpose all 35 409 tokens in YCOE with temporal adverbs were collected and processed with Perl.

8.7.1 Temporal adverbs

As the plots in figures 8.11 and 8.12 on p. 205 show, temporal adverbs display a much more varied behavior in terms of frequencies than locative adverbs. The plots show frequencies of temporal adverbs in initial position and those occurring in initial position immediately followed by *be*. Like with the locative adverbs, a few adverbs stand out. However, there are more high-frequency adverbs among the temporal adverbs: in addition to the different spelling variations of *þa* and *þonne*, it is possible to see adverbs[13] such as *nu* ("now"), *eft* ("afterwards"), and to barely make out *sona* ("immediately"), in figures 8.11 and 8.12. The two figures show plots based on frequencies of initial temporal adverbs and initial temporal adverbs followed by *beon*, respectively. *Þa* and *þonne* are obviously the dominant temporal adverbs in initial position. Note that the

[13] Several of these adverbs have corresponding conjunctions that are identical in form, cf. Baker (2007, 97). The material in the present work has been extracted on the basis of YCOE parts of speech tags, so that only adverbial uses have been included.

relative positions of words are independent of whether they are followed by *beon* or not.

8.7.2 Temporal adverbs in context

Figures 8.13 and 8.14 show the right context of temporal adverbs occurring in all position, and in initial position, respectively. As the plots show, right contexts for temporal adverbs are much more varied than for the locative adverbs.

Nevertheless, when the two sets of plots for initial position – temporal adverbs and their right-contexts – are interpreted together, it appears that an extremely frequent pattern is the adverb *þa* followed by *cwæð*, the past tense indicative of the verb *cweðan* "speak, say". Thus, the pattern "then spoke" seems to be a highly prominent one when we consider temporal adverbs in initial position.

8.7.3 Why not existential *þa*?

If frequencies alone are considered, it would seem that *þa*, or perhaps *þonne*, would be a much stronger candidate for the job of formal subject than *þær*. Breivik (1990, 99–100) notes that *þær* and *þa* are interchangeable in some contexts. Suggestively, Enkvist (1972, 95) states that *þa* "could be described as an overt link [to the narrative structure], 'detachable' because its removal does not seem to destroy the basic grammatical well-formedness of the sentence". Thus, it would seem that *þa* and *þær* do share many features. Furthermore, based on the plots in figures 8.13 and 8.14 on page 205, it looks as if *be* is a very common right context for both temporal adverbs in general and in initial position, although it does not seem to have quite the same position as with *there*. To test this relationship, conditional probabilities are again employed. First, a non-position-specific test:

$$P(Be \mid Tha) = \frac{Be \cap Tha}{Tha} = \frac{1174}{16830} = 0.070 \tag{8.28}$$

Next, the test for independence, viz. the multiplied simple probabilities:

$$P(Be) \times P(Tha) = \frac{39350}{110136} \times \frac{16830}{110136} = 0.055 \tag{8.29}$$

Thus, the conditional probability of *be* given *þa* is only larger than independence by a factor of about 1.3, which clearly indicates that the two are independent. However, for *there*, the effect was much stronger when position was taken into consideration. For *þa* this works out as follows:

$$P(Be \mid Init.tha) = \frac{770}{16830} = 0.046 \tag{8.30}$$

$$P(Be) \times P(Init.tha) = \frac{39350}{110136} \times \frac{8434}{110136} = 0.027 \tag{8.31}$$

As the equations above show, the relationship between *þa* and *be* appears to be much weaker than that between *there* and *be*. The conditional probability of *be* given *þa* in initial position is small, being larger than the multiplied simple probabilities by only a factor of about 1.7. Again, this suggests that the two events are independent.

Given the goals of this dissertation, viz. charting the development which led to the prototypical construction of *there* + *be* in Present-day English, it would seem that while temporal adverbs are not uninteresting, it might be permissible to disregard them for the present discussion, due to limits of time and space. However, looking at the case of *þa* and *þonne* is instructive in that it suggests certain constraints operating on adverbs that might become formal subjects: high frequency is not sufficient. In addition, it would seem that a certain restrictedness of context is desired, and since *þær* is more selective in that it is more closely linked with *be* it comes out as a stronger candidate for subject status than *þa* and *þonne*. Enkvist (1972, 93) notes "one of the functions of adverbial *þā* is to mark actions and sequences of actions." It thus needs to be considered whether this role was also compatible with one marking existence.

As pointed out above, the pattern *þa cwæð* is the most frequent one for temporal adverbs in initial position. Compared with the locative counterpart, the highly frequent combination of *þær wæs*, it is not surprising that *þær* had better chances than *þa* of becoming a formal subject in existential sentences, despite the higher overall frequency of the latter.

8.8 The likelihood of initial adverbs

Although the following paragraphs deal with locative adverbs only, I consider this an acceptable limitation of the hypothesis outlined above that there was an increase of adverbs occurring clause-initially during the Old English period. If *all* adverbs show an increasing tendency towards occurring in initial position, then the locative adverbs should also increase, and locative adverbs are anyway the most relevant type of adverbs for the present research project.

If there was an increasing proportion of initial locative adverbs in Old English it is certainly not evident from the raw frequencies in table 8.9 on p. 198. It appears that the wildly different numbers of corpus tokens per 25-year interval obscures any trend.

TABLE 8.9: *Frequencies for initial and non-initial locative adverbs in* YCOE *from texts which are dated in the corpus documentation. The tokens with initial adverbs constitute about 26% of the total number of tokens in the table (8 784). 419 tokens from undated texts are not included in the table.*

Year25	InitialAdv	NonInitialAdv
850	4	47
875	9	202
925	409	562
950	18	290
975	82	230
1000	273	1727
1025	113	770
1050	769	770
1075	103	1240
1100	2	19
1125	470	455
1150	16	132
1175	3	69
Total	2 271	6 513

A plot gives a more intuitive impression of how size influences the numbers. Figure 8.15 on p. 206 shows proportions of initial and non-initial locative adverbs in YCOE. The frequencies have been recast as proportions of all corpus tokens for a given 25-year interval. The scale of the y-axis is occurrences per 1 000 tokens.

As with the raw frequencies, it is unclear whether there actually is a diachronic change in the proportions of initial locative adverbs; if there is one, it does not appear very pronounced. The pattern is clearly influenced by fluctuations in corpus size, which causes a pattern of "peaks" and "valleys". This makes it very difficult to directly assess the development (let alone test hypotheses) by simply inspecting the plot. Clearly, some formalized procedure is needed to ascertain whether a change in proportions takes place. Such a procedure must not only take into account the frequencies of initial locative adverbs, but also the fluctuations in sample size.

Table 8.10 on p. 199 gives the total number of corpus token per 25-year interval. The plot also illustrates the importance of taking the entire corpus into account. In

isolation, the proportions in any smaller diachronic window spanning one "valley" followed by a "peak" might easily give the impression of a strong, increasing trend. The plot in figure 8.15 clearly shows the effect of differences in corpus size.

TABLE 8.10: *The total number of corpus tokens per 25-year interval in* YCOE. *The table excludes 6 129 tokens from texts which are not dated in the corpus documentation.*

Year25	#TokensPerYear
850	660
875	3 556
925	5 157
950	7 242
975	3 021
1000	33 789
1025	19 682
1050	8 667
1075	14 039
1100	361
1125	4 467
1150	2 507
1175	859
Total	104 007

8.8.1 Modeling initial adverbs

If there was a diachronic change in Old English, it should be possible to detect it using information about the time variable `Year25` which codes the 25-year intervals in the selected data. Instead of measuring the corpus proportions of initial and non-initial adverbs directly, we can reformulate the hypothesis of increasing numbers of adverbs in initial position as an increasing *probability* of adverbs in initial position. That is, if there really is a trend of increasing initial adverbs, it is expected to cause an overall increasing likelihood that any clause with a locative adverb has its adverb in initial position. With this operational definition of the hypothesis, the problem can be approached through regression modeling. The following paragraphs will employ different models (using R code rather than algebraic notation) to test whether the 25-year intervals can be used to

predict an increasing likelihood of initial locative adverbs. If this is the case, we should find that increasing the time variable also increases the likelihood of having a locative adverb in initial position.

Since `InitialAdv` is a binary variable, logistic regression is a natural choice. The data are not temporally independent, so it seems unlikely that an ordinary logistic regression fitted with the `glm()` function in R will be very successful. However, the results form an instructive background to later analyses. The model fitted with `glm()` looks like this:

(12) `glm(InitialAdv ~ Year25, family = binomial)`

This means that the binary response `InitialAdv`, coded as `TRUE`/`FALSE`, is modeled as a response of the time intervals in `Year25`. The error distribution is `binomial` and the link function is `logit`, but since `logit` is the default for a binomial logistic regression it is not necessary to specify this.

However, the best guess is that a mixed effects model is needed. Using the `lmer()` function, a binomial logistic GLMM (mixed effects model) can be fitted. The simplest model uses only a random intercept for the time intervals in `Year25`, that is, the intercept is allowed to vary by time interval:

(13) `lmer(InitialAdv ~ (1 | Year25), family = binomial)`

A more elaborate model includes some of the candidate predictors such as log SCR and the semantic class of the verb:

(14)
```
lmer(InitialAdv ~ LogComplexity + SemClass + (1 | Year25),
family = binomial)
```

Finally, it might be interesting to see whether *beon* is the right context of the adverb or not affects the estimated probability of a token having an initial locative adverb:

(15)
```
lmer(InitialAdv ~ LogComplexity + SemClass + BeContext
+ (1 | Year25), family = binomial)
```

8.8.2 Model evaluation

As explained in a previous chapter, the most important diagnostic tool with such models is the residuals vs. fitted plot. This plot shows whether the model actually fits the data, and in case of a bad fit they give an indication of where the problems lie. Such plots for the four models above are shown in figure 8.17. As the upper left plot in figure 8.17 shows, the ordinary GLM model is a terrible fit: only one point is within the 95%

confidence intervals, and the model only guesses the most frequent outcome, i.e. non-initial adverbs. The upper right plot for the mixed intercept model in (13) fares a little better, in that the points are inside the 95% confidence interval, but the model still only guesses the most frequent outcome. The plot in the lower right corner for the model in (14) looks better, although possible non-linearities can be detected. The model includes both high and low probabilities of `InitialAdv` and there are no problems with non-constant variance. A few points lie outside the 95% confidence interval, but the model is clearly much better than the other two. In the lower left corner we see the plot for the model in (15). Two things are worth noticing here: first, the number of points outside the 95% confidence interval is the same (i.e. no improvement) as for the model in (14), and second, there is a strong non-linearity in the shape of an inverse "V" in the middle of the plot, indicating a structural problem with the model. Thus, based on the plots, the model in (14) seems to be the best one. The specific models discussed above will not be explored in further detail, since their intended goal was to investigate *whether* a diachronic variable is a valuable indicator of initial adverbs in YCOE, not to identify which predictors that describe this phenomenon well.

We can quantify the improvements that these models make, using Nagelkerke's pseudo R-squared (R^2) measure of goodness of fit for logistic models. This model quantifies the approximate degree of improvement over other models (on an *approximate* scale from 0 to 1). The model in (13) gives a Nagelkerke R^2 of 0.27 when evaluated against an ordinary intercept model such as (12). The full model in (14) gives an R^2 of 0.40 against (12), but only 0.17 against (13). That is, the full mixed model in (14) is much better than an ordinary GLM model, but the improvement over a mixed intercept-only model is more modest. The model in (15) with `BeContext` added has a somewhat higher R^2 over the mixed model (0.24), but the difference is marginal from (14), and as the plot in figure 8.17 showed, the model in (15) was not a good fit.

8.8.3 Interpretation

With the output from the model in (14), it is possible to quantify the variance between the 25-year intervals. The within-interval standard deviation is 0.95 on the log odds-ratio scale, and dividing it by 4 is a quick way to get an upper bound of the difference between the categories as probabilities (Gelman and Hill, 2007, 82). This works out as follows: 0.95/4 = 0.24, which shows that the estimated mean probability of `InitialAdv:TRUE` (i.e. having a locative adverb in initial position), differs by $\pm 24\%$ *within* each 25-year interval. This is a fairly large difference, and we would expect to find it in the form of an upwards trend for the probability of initial adverbs. However, inspecting the year-effects for the model shown in figure 8.16 on p. 206 leads

to the same conclusion as the plot in figure 8.15 on p. 206, namely that the early and late parts of the corpus contain less initial adverbs (probably due to missing data), while the well-represented intervals remain stable.

The GLM model with `Year25` as predictor in (12) was clearly not a good fit. The model in (14), which includes semantic and syntactic information, was shown to be the best of the models explored here, although better models could doubtless be found.[14] The point has been to show that within YCOE semantic and syntactic factors take precedence over diachronic ones when it comes to explaining the probability of locative adverbs in initial position.

Since the uncertainties with respect to initial adverbs for all practical purposes are as great within the diachronic units as between them, it seems that there is no real change taking place in YCOE regarding the probability of a locative adverb in initial position. In other words, the situation in YCOE with respect to initial adverbs is characterized by great variation, not change. The model presented above represents a reasonable estimate for the effects of the diachronic variable on the probabilities of initial adverbs in Old English. Thus, the picture that best represents the probability of initial adverbs in Old English is one of stability.

8.9 Summary

In the present chapter it was shown that

(i) forms of *þær* are the most frequent locative adverb in YCOE;

(ii) forms of *þær* are somewhat overrepresented in initial position;

(iii) forms of *þær* are strongly associated with *beon* when the adverb occurs in initial position;

(iv) no such association can be found for the temporal adverb *þa*;

(v) there is circumstantial evidence which suggests that *þær* could be used as a subject in Old English;

(vi) no verb classes have a strong preference for adverbs in initial position, but a weak effect was found for `SocialInteraction` and `Existence`;

[14] Including `Adverb`, or `There` – the former contains the adverb of the token and the latter is a binary variable coding whether a form of *þær* is present or not – led to severe non-linearities in the residual vs. fitted plots, i.e. the model had an incorrect structure. The reason for this is probably that these predictors are not linearly independent of the response variable.

(vii) finally, no firm evidence was found supporting the view that the number of adverbs in initial position in YCOE increases throughout the Old English period.

Forms of *there* are by far the most frequently occurring locative adverb in YCOE, providing initial support for the view that frequencies are somehow involved in the evolution of existential *there*. It was also found that *þær* tends to occur in initial position and that it often co-occurs with *be*. An examination of *that*-clauses provided circumstantial evidence for the status of *þær* as a non-prototypical subject. Perhaps the most important finding is that the probability of having an initial adverb does not increase during the Old English period. This suggests that there is no clear diachronic trend related to initial locative adverbs. Consequently, there was no "piggy-back ride" situation whereby *þær* was carried into initial position and subsequent subject status through some general wave of adverbs in initial position. Instead, syntactic and semantic factors related specifically to *þær* seem much more promising when it comes to explanatory potential. Interestingly, the picture emerging from the YCOE data is decidedly one of stability. Little diachronic change on the subject of *þær* and locative adverbs can be detected within the Old English period. It seems that much of the development in question must have taken place during the Middle English period; a conjecture to be tested in the next chapter.

Thus, some of the original hypotheses have been strengthened whereas other have been rejected. The hypothesis that *þær* is the most frequent locative adverb in Old English still stands, as does the hypothesis that *þær* is more frequent than temporal adverbs in the left context of *beon*, and that there is a strong association between *þær* and the prototypical existential verb *beon* when the adverb occurs initially. The hypothesis that *þær* should be tied to particular, identifiable contexts was also strengthened.

Other hypotheses were rejected. It was found that tokens with *þær* do have a low syntactic complexity ratio, but this was shown to be tied to the initial position itself, not specifically to the use of *þær*. Furthermore, the hypothesis that the likelihood of *þær* in initial position should increase during the course of the Old English period found no support in the material from YCOE.

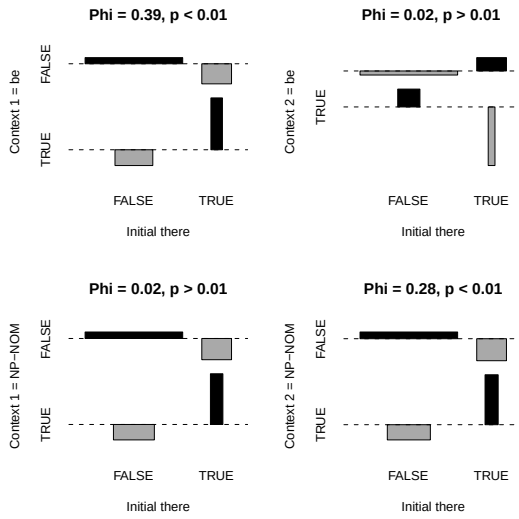


FIGURE 8.10: *Cohen-Friendly plots for associations between* there *in initial position,* be *and nominative NP in first and second right contexts. Note that the pattern* initial there + be *and* initial there + . . . + *nominative NP are significant at the 1% level, with reasonable effect sizes. The opposite patterns* initial there + *nominative NP;* initial there + . . . + be *are not significant, even at the 5% level, and effect sizes are negligible.*

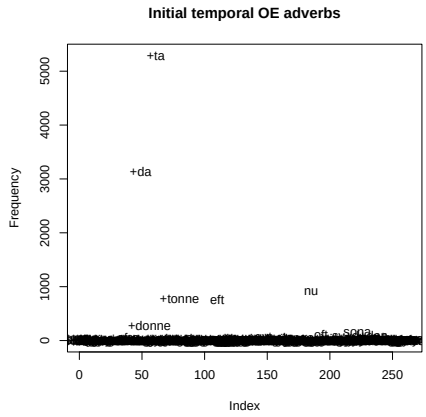


FIGURE 8.11: *Frequencies of temporal adverbs in clause-initial position. Total number of occurrences is 12 098.*

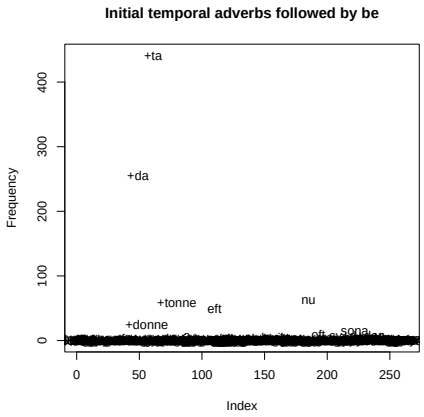


FIGURE 8.12: *Frequencies of temporal adverbs in clause-initial position that are immediately followed by* be*. Total number of occurrences is 1 306.*

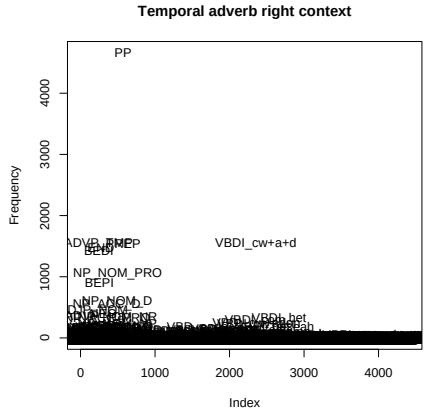


FIGURE 8.13: *Frequencies of right contexts for temporal adverbs in Old English.*

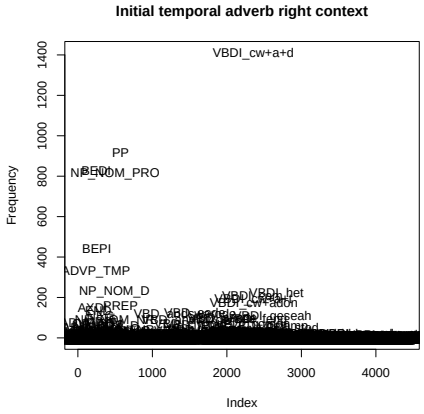


FIGURE 8.14: *Frequencies of right contexts for temporal adverbs occurring in initial position.*

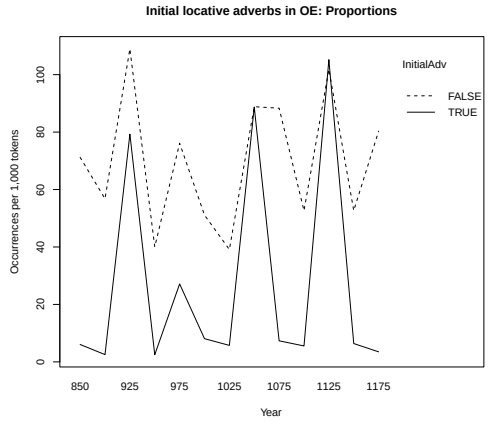


FIGURE 8.15: *Proportions of initial and non-initial locative adverbs in* YCOE*, scaled to occurrences per 1 000 corpus tokens. The proportions show massive fluctuations due to large differences in texts available for the different 25-year intervals. It is difficult to spot an obvious diachronic trend.*

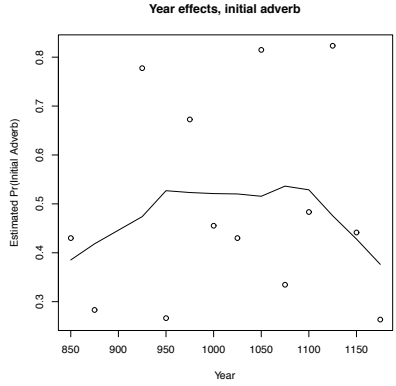


FIGURE 8.16: *Estimated probabilities per 25-year interval for the model in (14), with a non-parametric smoothing regression line. No clear diachronic trend can be detected, and the main difference appears to be the one between intervals with little at the start and end and intervals with much data in the middle.*

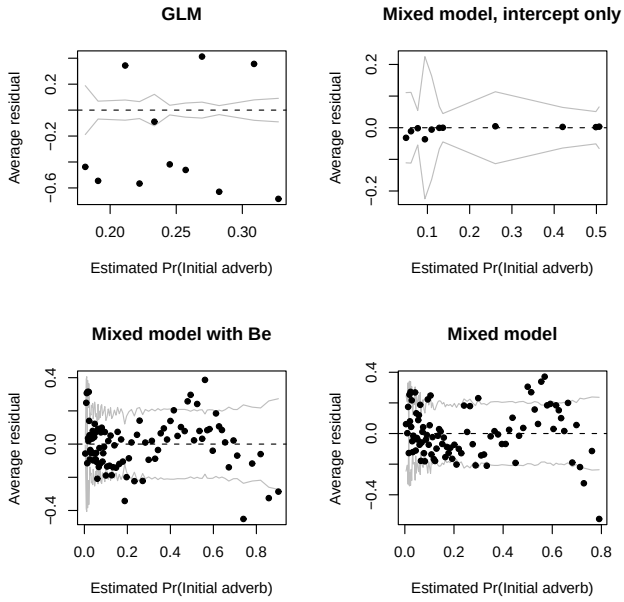


FIGURE 8.17: *Binned residual vs. fitted plots for four logistic-binomial models of* `InitialAdv` *in the selected tokens from* YCOE. *The plots show, clockwise from top left, the GLM in (12), the GLMM with random intercept only in (13), the GLMM with* `LogComplexity` *and* `SemClass` *as fixed effects and a random intercept in (14), and finally the GLMM with* `BeContext`, `LogComplexity` *and* `SemClass` *as fixed effects and a random intercept in (15). The model in the lower right corner appears to be the best of the four.*

Chapter 9

The Middle English EC

9.1 Introduction

This chapter provides an overview of the uses of existential and locative *there* in the Middle English treebank PPME2. I will discuss some characteristics of the Middle English EC and present a formal statistical model describing the relationship between existential *there* and these characteristics. As part of the presentation of this model, the chapter illustrates the process of model fitting and model checking with generalized linear models with mixed effects (GLMMs). Through these steps, I will address some of the hypotheses presented earlier:

(i) *there* is more frequent than other locative adverbs in Middle English;

(ii) the proportion of adverbs in initial position increased in Middle English;

(iii) sentences with *there* in initial position had a lower syntactic complexity than other sentences.

Additionally, the relationship between *there* and *be* is investigated further, alongside explorations of diachronic effects and the influence of authors. The results of the present investigation will in chapter 11 be used to estimate the distribution of existential *there* in Old English, and contribute to a theoretical interpretation of the statistical models.

9.2 Data

The prose part of the Middle English treebank, comprising 1 155 965 words in some 84 685 tokens, follows a slightly different annotation scheme where *there* is tagged with "EX" for the existential uses and "ADV" for the locative uses. For consistency with the Old English material, both existential and locative uses were included in the analysis.

As the examples below from respectively *The Brut*, *The Parson's Tale* and *Boethius* illustrate, both existential and locative uses of *there* are attested in PPME2:

(1) And so there Brut quelled his fader.
"And so there Brut killed his father." (*locative*)
(CMBRUT3,6.122)

(2) For he shal nat taken kep who sit there, but in whos place that he sitteth.
"For he shall not take notice who sits there, but in whose place that he sits." (*locative*)
(CMCTPARS,324.C1.1535)

(3) and thow hast schewyd me wel that over thilke good ther nys no thyng more to ben desired.
"and you have shown me that over this good there is no other thing to desire." (*existential*)
(CMBOETH,430.C2.105)

(4) But certes by nature ther nys but o God
"But certainly by nature there is but one God" (*existential*)
(CMBOETH,433.C1.183)

As in Old English, the classification of what is existential and what is locative use of *there* in Middle English is not always straightforward. Consequently, relying on the annotation in PPME2 has two great advantages. First, it makes the study easier to replicate; and second, I avoid adding my own bias to the data, since the corpus was not specifically annotated with studies of *there* in mind.

9.2.1 Collecting data

The CorpusSearch query in (5) below yields all tokens containing either an existential *there* (EX) or a locative adverbial (referred to as *target words* in the following text).

(5)
```
node: IP*
query: (EX exists) OR (ADVP-LOC exists)
```

The query yielded 5 889 tokens, one of which is a false hit: CMMANDEV,123.2989, from *Mandeville's Travels*, has a Latin quote where the Latin "EX" is treated as an expletive tag by CorpusSearch. This token was manually removed before the subsequent analysis of the remaining 5 888 tokens. However, it was necessary to further edit the files, since some tokens were missing an ID-tag. Furthermore, 413 tokens from *Ormulum* were removed. This text is included in the prose part of PPME2, but is marked as poetry. This leaves a final selection of 5 471 tokens. These tokens were further processed with Perl to fit a data frame format which can be read into R. As with YCOE, developing the script to account for all the variation in the corpus required and iterative process of testing and re-testing, which, although time consuming, also gave the added benefit higher familiarity with the data through having to scrutinize selected parts of the corpus closely.

Table 9.1 shows an excerpt from the dataframe. In table 9.1, six columns from the first six rows are shown. The full dataframe contains rows 5 471 and 71 columns, i.e. one row per token, with 71 fields of information about that token.

TABLE 9.1: *An excerpt from the* ME *dataframe, showing six columns from the first six rows.*

	Adverb	AdverbTag	Text	InitialAdv	AdverbPosition	Year25
1	heore	ADV	CMAELR3	FALSE	Mid	1400
2	here	ADV	CMAELR3	FALSE	Mid	1400
3	+ter	ADV	CMAELR3	TRUE	Initial	1400
4	+ter	EX	CMAELR3	FALSE	Mid	1400
5	*t*	Trace	CMAELR3	FALSE	Mid	1400
6	+ter	EX	CMAELR3	TRUE	Initial	1400

9.2.2 The adverbs and existential *there*

The selection of 5 471 contains some 250 locative adverb types, as well as existential *there* (referred to as "target words"),[1] however, just like in Old English the spelling variation is substantial. Due to a smaller sample size the total number of types is somewhat lower, i.e. more unseen events. However, as the data from YCOE indicated, this should only affect adverbs with very low frequencies. *There* is the most common target word with 3 309 occurrences, closely followed by *here* with 831 occurrences. This

[1] In addition, a number of traces of existential *there* or locative adverbs were encountered. This feature has not been analyzed further, since RCG does not recognize empty traces.

means that *there* alone accounts for some 60% of the selection from Middle English, whereas *here* only constitutes about 15%, a decrease from Old English where *here* made up about 25% of the tokens.[2]

The third most frequent category is "trace" (*T*), with 647 occurrences. The trace category will not be dealt with further here for two reasons: scope and comparison. First, the emphasis of the study is on the evolution of *there*. Second, since the dataset from *ycoe* was extracted based on the presence of locative adverbs, the trace category is infrequent in that dataset. Thus, for comparison with the YCOE dataset it makes sense to focus on *there*.

Interestingly, few medium frequency items were found. The next adverb attested is *where* with some 105 occurrences for all the spelling variations; after this the frequencies drop sharply and fall below 50 occurrences. Thus, *there* and *here* completely dominate the selection for Middle English, as opposed to the variation found in Old English. This might be a result of a smaller corpus size for Middle English, or it might reflect some real diachronic change. However, a detailed study of the reasons for this falls outside the scope of the present study. What seems clear is that only *here* and *there* look like plausible candidates for existential subjects, and consequently only they will be dealt with further.

As mentioned above, PPME2 tokens are tagged for the presence of existential *there*. A total of 1 386 tokens with existential *there* was found,[3] which amounts to 25% of the selection, and 42% of all instances of *there*. Figure 9.1 shows frequencies of *there* vs. other target words. The relationship between locative and existential *there* is dealt with further in the following section. The distribution of existential *there* vs. all locative adverbs is visualized in figure 9.2.

A look at the right context of the target words reveals that for the immediate right context, forms of *be* is the most common element, with 1 214 occurrences, which is about 22% of the total. Other verbal elements occur in a further 886 tokens, bringing the total number of verbal elements in the immediate right context up to 2 100 or about 38% of the cases. There are 1 547 tokens with nominal elements (nouns and NPs, including pronouns) occurring in the immediate right context, equal to about 28% of the total, which means that verbs and nominal elements together make up 3 647 or about 66% of the total number of right contexts for the target words.

In the second right context (i.e. the second element following the target word), nominal elements dominate with 2 173 occurrences, which corresponds to about 40% of the total number of tokens. Verbs are less frequently encountered in the second right context, being attested in 989 of the cases, i.e. some 18% of the tokens. Forms of *be*

[2]The effect size for the difference in proportions as measured with Cohen's h is 0.25, i.e. a fairly small effect.

[3]But see examples (6) and (7).

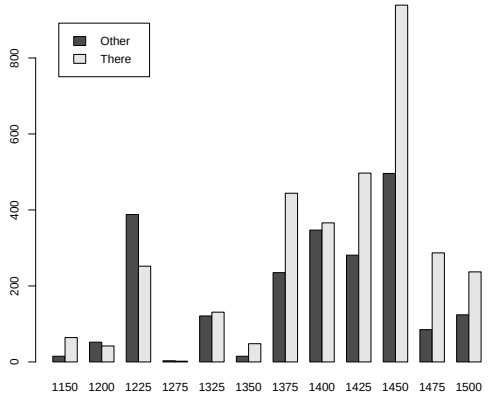


FIGURE 9.1: *Barplot showing frequencies of* there *vs. other target word realizations in ME in 25-year intervals.*

are comparatively infrequent here, occurring only in some 294 tokens, or 5% of the total. Verbs and nominal elements together occur in 3 162 or 58% of the tokens.

There is obviously much variation in the right context of the target words, but verbs and nominal elements are the most frequently attested elements. It is worth noting that verbs are found more often than nominal elements in the first right context, while the situation is reversed for the second right context. Forms of *be* are the most frequently encountered realization of the first right context.

Turning to the cases where the target word (i.e. existential *there* or locative adverb) is existential *there*, the first right context is a form of *be* in 888 tokens or 64% of the total number of tokens with existential *there*. The second right context is a nominal element in 924 tokens, i.e. 67% of the cases with existential *there*. The intersection of the two, i.e. cases of existential *there* where the first right context is a form of *be* and the second right context is a nominal element, consists of 660 tokens, or some 48% of all instances of tokens with existential *there*. In other words, in almost half the cases where existential *there* is present, we find the linear order THERE + BE + NP.

Interestingly, among the collected tokens we find one instance annotated as an existential use of *here*, from *Sawles Warde*, a text from around 1225 belonging to *The*

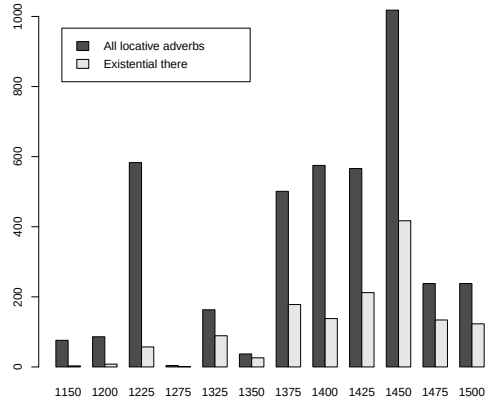


FIGURE 9.2: *Barplot showing frequencies of existential* there *vs. all locative adverbs.*

Katherine Group:[4]

(6) nis hare nan þe ne feareð ofte untoheliche.
NEG=be.PRES here none COMP NEG go.3SG.PRES often rudely
"here is none who does not often behave rudely"
(CMSAWLES,166.17)

One existential construction without *there* or *here* is also found (annotated as *exp* in PPME2), in *Mandeville's Travels* from the early 15th century:

(7) And abouen the Gerneres withouten *exp* ben many scriptures of
and about the granaries outside EXPL be.PRES many scriptures of
dyuerse langages.
diverse langauges
"And around the pyramids outside [there] are many inscriptions in various languages."
(CMMANDEV,34.848)

Mandeville's Travels is translated from French and one might consider this a trans-

[4]See Jenset (2005, 97–104) for a discussion on an example of a possibly existential use of *here* in the works of Chaucer.

lation error, but the documentation for PPME2 states that although the translator often misunderstands the French original, the English translation is of good quality.[5] For technical reasons these two tokens are for the most part included in the total count for existential *there*.

In summary, it would appear that the Middle English EC is quite predictable in its composition. The most frequently attested pattern is THERE + BE + NP, although much variation can be found. Some extremely rare uses are found, such as one case of existential *here* and one case with an empty existential subject. Overall, *there* is the most frequent target word realization and existential *there* has a substantial frequency among the selected target words.

9.2.3 Syntactic complexity

According to one of the hypotheses outlined previously, we might expect tokens with existential *there* to have a lower syntactic complexity than clauses without. As it turns out, the mean log complexity ratio of the selected tokens from PPME2 is 1.71, with a standard deviation of 0.84. For tokens with existential *there*, the mean log complexity ratio is 1.83, while it is 1.67 for tokens without existential *there*. The difference does not appear to be great, but the impression can formally be tested with a t-test, since the log SCR is approximately normally distributed, as shown in chapter 6. The result of a nondirectional Welch two sample t-test reveals a significant difference ($t_{df(2887.48)} = -6.67$, $p < 0.01$, $d = 0.20$), but the effect size as measured with Cohen's d is low. The estimated 95% confidence interval for the difference in means is [0.11, 0.20] or somewhere between $\frac{1}{8}$ and $\frac{1}{4}$ of a standard deviation. In other words, although significant, the difference in syntactic complexity ratio between tokens with and without existential *there* is negligible in practical terms.

Perhaps the syntactic complexity ratio plays no role at all regarding adverbs and existential *there*? A quick look at target words (locative adverbs and existential *there*) occurring in initial position shows that this is not the case. The tokens with initial target words have a mean syntactic complexity ratio of 1.42, whereas the tokens with non-initial ones have a mean log syntactic complexity ratio of 1.83. The difference is statistically significant when tested with a nondirectional two sample Welch t-test ($t_{df(2984.46)} = 17.21$, $p < 0.01$, $d = 0.51$). The estimated 95% confidence interval for the difference in means is [0.36, 0.46], in other words an estimated true difference which is close to one half of a standard deviation. As the effect size measure d indicates, this is a medium sized effect.

[5] See Faverty (1928, 98) for a note on how the author of the *Mandeville's Travels* manuscript considers the Egyptian pyramids to be granaries.

Looking only at target words in initial position, the mean log SCR is 1.61 for tokens with existential *there*, whereas it is 1.35 for other tokens, i.e. somewhat *lower*. The difference is statistically significant, and the effect size is small to medium sized ($t_{df(817.07)} = -6.57$, $p < 0.01$, $d = 0.37$), with a 95% confidence interval for the difference in means of [0.19, 0.35] or between 0.22 and 0.41 of a standard deviation from the mean of log SCR.

It would appear that syntactic complexity hardly plays any role in describing the distribution of existential *there*, whereas there are real differences in syntactic complexity between the tokens with either existential *there* or a locative adverb in initial position, and those with the target word in other positions. The hypotheses that tokens with existential *there* would have a lower syntactic complexity score is thus rejected.

9.3 Adverb position

The position of the adverb is hypothesized to play a role in the evolution of existential *there* as discussed initially in this chapter and in section 8.8. In PPME2 1 498, or 27.3%, of the tokens have a target word in initial position. 397, or 26.5%, of those 1 498 initial target words are tagged as existential *there*. This means that in 28.6% of the tokens with existential *there* the morpheme occurs initially. Thus, it would seem that the proportion of existential *there* perfectly mirrors the overall distribution of target words in initial position.

As mentioned above, Butler (1980, 279) suggests that adverb fronting played a role in the history of *there*. If *there* is attracted to the EC via initial position, it would not be unreasonable to expect some kind of association between target words in initial position and verbs of existence, and possibly also verbs of occurrence and appearance. Such a possible association can be investigated using correspondence analysis (CA), as in Old English. First, a 3×45 matrix was created, with the sentence positions INITIAL, MID and FINAL as rows and with the 45 verb classes attested in the material as columns. A CA biplot of the analysis is shown in figure 9.3.

The total inertia in figure 9.3 is not more than 0.06 out of a theoretical maximum of 2, but the quality of representation is good, with the x-axis accounting for 74.2% of the total inertia and the y-axis for 25.8%, i.e. 100% in two dimensions. Note that this is of the same order of magnitude as the inertia of 0.07 for the similar plot in chapter 8. Judging by the plot in figure 9.3 and the numeric output (found in appendix A.2), there is little or no association between initial position and typically "existential" verb classes such as `Existence`, `Appearance`, and `Occurrence`. The plot shows that rather than being positively correlated with *initial* position, these classes are negatively correlated with *final* position, an impression which is confirmed by the numeric output.

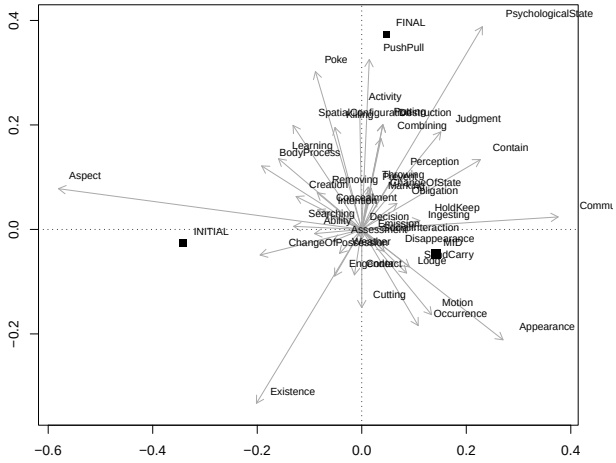


FIGURE 9.3: *Standard CA biplot showing the adverb-position/semantic class data. The horizontal axis accounts for 74.2% of the variance in the data, the vertical axis accounts for an additional 25.8%. Total inertia is approximately 0.06 out of a maximum of 2, i.e. fairly low. The rows (adverb position) are in principal coordinates, while columns are in standard coordinates times the square root of the mass. Row point sizes are plotted proportionally to their relative frequency.*

Initial and mid position together draw up the x-axis, while final position is responsible for the y-axis. The "existential" semantic classes are correlated with the y-axis, but not with the x-axis.

Based on this, it seems unlikely that there is a strong correlation between existential *there*, target word position in the sentence, and the semantic class of the verb. The low inertia of the matrix with target word position vs. semantic class suggests that the semantic classes are used fairly equally over the three positions, although verbs of existence, appearance and occurrence tend to disfavor target words in final position.

Whether there is an increase in initial adverbs in Middle English can be tested using the same approach as in the previous chapter. As in the previous chapter, a simple model using only a random intercept does not give good results, whereas a model incorporating the semantic class of the verb and the syntactic complexity ratio provides a good fit to the data judging by the residuals vs. fitted plot. The models are

shown below:

(8) `lmer(InitialAdv ~ (1 | Year25), family = binomial)`

(9) `lmer(InitialAdv ~ LogComplexity + SemClass + (1 | Year25), family = binomial)`

It is difficult to know whether the mean likelihood of initial locative adverbs increases, thereby causing existential *there* to increase, or whether existential *there* increases thereby causing the likelihood of initial adverbs to increase (remember that the variable `InitialAdv` does not distinguish between locative and existential target words). To correct for this, we rerun the model in (10), but this time using only sentence-tokens without existential *there* as input. The results can be seen in figures 9.4 and 9.5. In both cases, there is a clear upwards trend showing a gradual increase of target words in initial position.

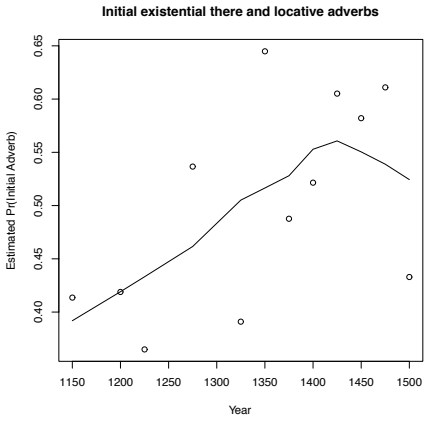


FIGURE 9.4: *Year-effects for all locative adverbs and existential* there *in* PPME2. *Note the increase in the estimated mean probability of initial position, as indicated by the smoothed nonparametric regression line.*

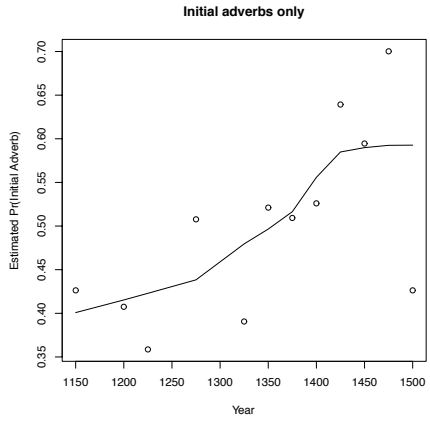


FIGURE 9.5: *Year-effects for locative adverbs only. The increase is still present, as indicated by the smoothed nonparametric regression line, suggesting that the effect is not caused by the presence of existential* there *in the material.*

There is reason to suspect that the amount of data available is involved here, though. The previous chapter showed that the situation in Old English was one of stability, with

dips in the mean probabilities in the early and late period, corresponding to less well-represented periods in the data. Figure 9.6 shows raw frequencies of initial target words in PPME2. However, viewing the frequencies scaled as proportions of the number of tokens per 25-year interval shows that while there might be a slight increase, the picture is far from clear due to fluctuations in the amount of material, cf. figure 9.7.

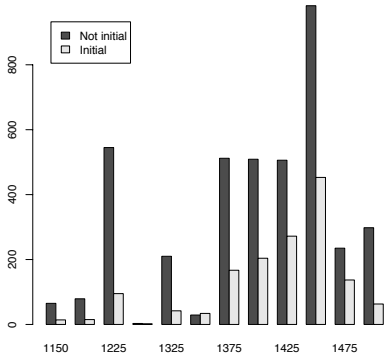


FIGURE 9.6: *Raw frequencies of initial locative adverbs and existential* there *by 25-year interval.*

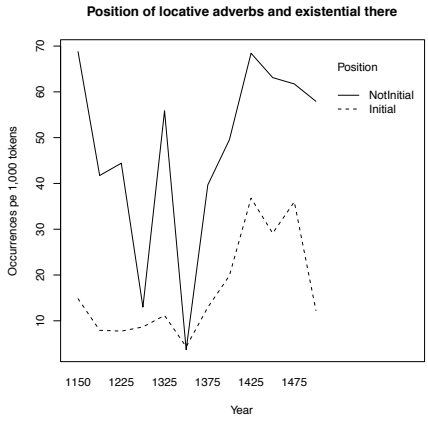


FIGURE 9.7: *Initial locative adverbs and existential* there *by 25-year interval as proportions of corpus size for the interval. The scale is occurrences per 1 000 corpus tokens.*

To check whether there is an increase, the model was modified by including the variable `Obs100`, which is the number of observations per 25-year interval in the corpus, divided by 100 to reflect increases of 100 sentence tokens.

(10)
```
lmer(InitialAdv ~ LogComplexity + SemClass + Obs100 +
(1 | Year25), family = binomial)
```

As before, the residual vs. fitted plot gives no reason for concern, cf. figure 9.8. Inspecting the coefficient for the observations-variable shows that it has a log odds ratio of 0.01, corresponding to a probability of 0.50. That is, increasing the number of observations per 25-year interval results in a 50% increase in the estimated probability of finding a target word in initial position. However, the standard error of the coefficient is 0.04, i.e. four times as large as the coefficient. This suggests that the uncertainties

of the estimates cancel out much of the effect (the corresponding probabilities for the standard error is 0.51). In other words, while increasing the amount of data leads to higher estimated probabilities of initial target words, the uncertainty is as large as the effect itself.

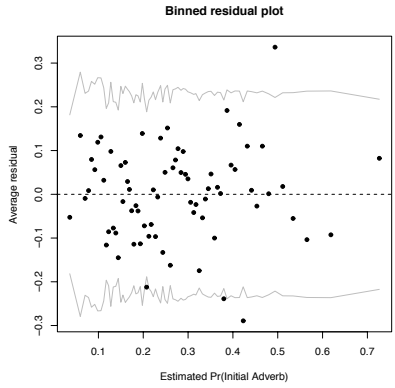


FIGURE 9.8: *Residual vs. fitted plot for the model in (10). There are no obvious problems with the fit to the data.*

The tentative conclusion to this section then, is that there is an increase in initial target words during the late Middle English period. The availability of more material towards the end of the period might account for some of this effect, but the estimates are uncertain, making it difficult to draw firm conclusions.

9.4 An overview of *there*

Is the use of *there* increasing in Middle English? In answering this question it is vital to look at the relevant numbers in their proper context. Put bluntly, raw frequencies and percentages of raw frequencies are more misleading than informative when dealing with data collections of unequal sizes.

Consider figure 9.9, which shows frequencies of existential and locative *there* in Middle English, that is, the figure represents all and only uses of *there*. It is tempting to look at the solid line representing existential *there* and conclude that there was a substantial increase in existential *there* from the end of the 14th century. However, the figure also shows frequencies for locative *there*, represented by the dashed line. The two follow each other closely, and this makes it worthwhile to ask if sample size could

play a role. In figure 9.10, the frequencies in figure 9.9 have been rescaled as proportions based on the total number of corpus tokens available in PPME2 for a given 25-year interval. The scale chosen for the y-axis is number of occurrences per 1 000 tokens, to avoid fractions of occurrences. The plot in figure 9.10 illustrates that although an upward trend can be detected, there seem to be only only small changes taking place regarding the use of existential *there* when we take sample size into account.

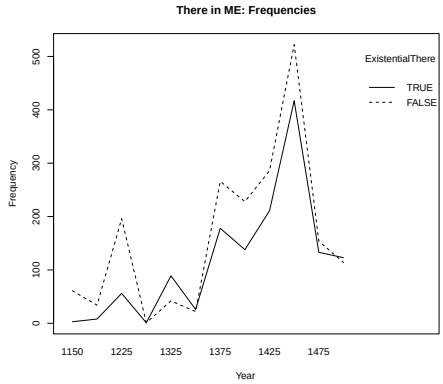


FIGURE 9.9: *Raw frequencies of locative and existential* there *by 25-year interval.*

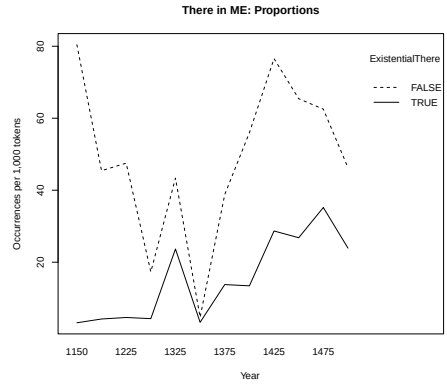


FIGURE 9.10: *Locative and existential* there *by 25-year interval as proportions of corpus size for the corresponding interval. The scale is occurrences per 1 000 corpus tokens.*

For 1275, i.e. the interval from 1275 to 1300 in figure 9.10, there is a dramatic peak which needs to be explained. The total number of corpus tokens available for this interval in PPME2 is very small compared to the rest of the Middle English period, only 231 tokens. Additionally, the number of existential and locative instances of *there* are extremely low. Table 9.2 gives an overview of the number of corpus tokens from PPME2 per 25-year interval. These numbers are taken from the CorpusSearch output file (which shows how many tokens were searched from each corpus file) and subsequently summed based on which 25-year interval they occur in. Table also 9.2 shows frequencies of locative and existential *there* by 25-year interval.

Table 9.2 also reveals that 1150, the third quarter of the 12$^{\text{th}}$ century, has a comparably low number of tokens in the corpus, which explains the seemingly sudden drop in proportions for 1150 in plot. In other words, judging from the plot in figure 9.10, some change is taking place in Middle English regarding the uses of *there*, although

we have reason to believe that sample size might be involved in this. The existential and locative use follow each other closely, with the locative use being somewhat more frequent. This is reflected in the overall proportions of uses of *there* in Middle English presented in section 9.2.2, with 42% being existential and 58% locative. Dispersion, as measured by the DP index, is 0.06 which is sufficiently close to zero to conclude that both locative and existential *there* are evenly distributed over the 25-year intervals in PPME2.

TABLE 9.2: *Frequencies of locative and existential use of* there *in* PPME2*. The two tokens in (6) and (7) are not included in the column for existential* there*. Note the extremely low frequencies in the early part of the period.*

Year25	ExThere	LocThere	# corpus tokens
1150	3	61	945
1200	8	34	1 893
1225	56	196	12 269
1275	1	1	231
1325	89	42	3 759
1350	26	22	7 869
1375	178	266	12 907
1400	138	228	10 271
1425	211	286	7 393
1450	417	522	15 559
1475	134	154	3 807
1500	123	114	5 144
Total	1 384	1 926	82 047

9.5 *There* by author

In Middle English, 2 378 or about 43% of the selected tokens come from a text with a known author. As table 9.3 shows, most of the known authors come from the later part of the Middle English period, with a concentration around the mid 15th century.

As figure 9.11 shows, it is clear that some authors use existential *ther* more often than others. However, no clear detailed trend can be seen in the diachronic develop-

TABLE 9.3: *Overview of ME authors by time period included in the selected material.*

Year25	Author(s)
1325	Rolle
1375	Chaucer, Purvey, Trevisa
1425	Gaytryge, Thornton
1450	Capgrave, Hilton, Julian of Norwich, Kempe, Malory
1475	Caxton, Fitzjames
1500	Mirk

ment. For instance, Walter Hilton clearly uses more existential than locative *ther*, but a later author such as John Mirk use the two in more or less equal proportions.

Since more than half the tokens in the selection from PPME2 *lack* an identifiable author, including author in some of the models discussed above would exclude substantial amounts of data. It is possible to get an impression of the impact of authors by using a logistic GLMM as shown below:

(11)
```
lmer(ExTag ~ LogComplexity + BeContext + Author +
(1|Year25), family = binomial
```

The model above estimates the probability of finding existential *there* for a given token based on the log scr of the token, the presence or absence of *be* in the immediate right context, and the author. A random effect is included for the 25-year intervals. To check the assumptions of the model a residual vs. fitted plot was made (not included) indicating that the variance was reasonably constant. The plot showed signs of mild nonlinearities for probabilities of existential *there* greater than 0.2, suggesting that there were some structural problems with the model. This is not unexpected, since `Author` and `Year25` are not independent of each other, quite the contrary, they are highly correlated. However, the fit seemed reasonable enough for the present purpose of getting an idea of the effects of authors. Table 9.4 gives the coefficients and standard errors for authors, *be*, and log syntactic complexity ratio. As the numbers in the table clearly show, the effects for the authors differ considerably, as does the variation. Some authors have standard errors almost as large as the coefficient, probably caused by too little data for accurate estimates. However, compared with the large effect of having *be* appearing in the right context of the target word, the effects for the authors are small. (But see the effects for Geoffrey Chaucer and Walter Hilton, where Chaucer stands out

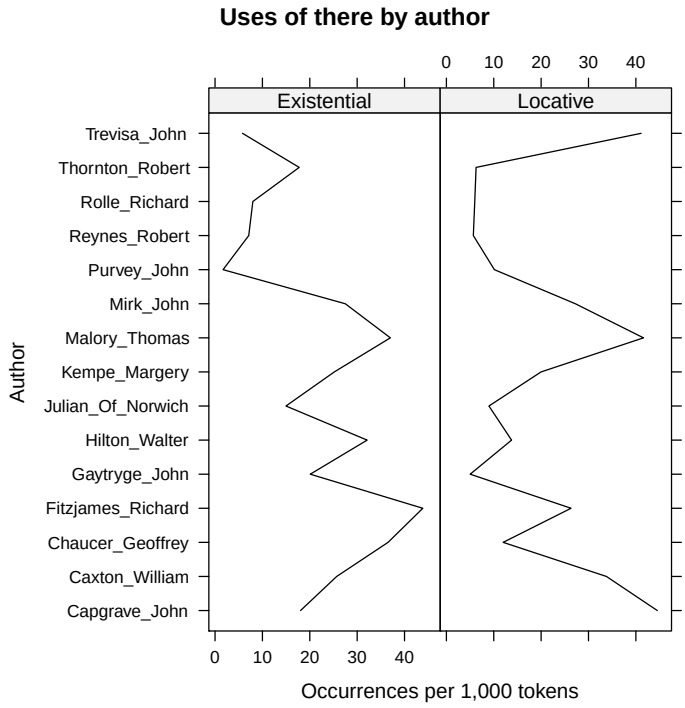


FIGURE 9.11: *Proportions of existential and locative* there *by author in ME, scaled to occurrences per 1 000 corpus tokens.*

with a unique combination of a high estimated probability of existential *there* and a fairly low standard error).

Thus, while individual authors clearly do play a role regarding the use of existential *there*, there are clearly other (linguistic) factors that play a greater role. There are also data-analytical considerations to make. Including `Author` would necessitate leaving out much data, and even then we have to deal with great uncertainty for some authors due to little data for those particular authors. In summary, while not unimportant, the author appears to be a relatively minor factor in estimating existential *there*. The benefits of including the author in the analysis are clearly smaller than the expenditure, and I will disregard author as a predictor in subsequent analyses.

TABLE 9.4: *Output from the GLMM in (11). Total number of observations is 2 378, between-group variance (standard deviation of the error term) is 0. The estimated differences between authors are small compared with the effect for* `BeContext`*. A binned residual vs. fitted plot of the model showed mild nonlinearities for estimated Pr(Existential there) > 0.2. The reference level (intercept) is* `AuthorCapgrave_John`*.*

Fixed effects	Est. coef	Std. Error	Pr(Existential there)
(Intercept)	-2.54	0.19	0.07
LogComplexity	0.14	0.07	0.08
BeContextTRUE	2.65	0.13	0.53
AuthorCaxton_William	0.48	0.34	0.11
AuthorChaucer_Geoffrey	1.81	0.22	0.32
AuthorFitzjames_Richard	1.16	0.49	0.20
AuthorGaytryge_John	0.96	0.70	0.17
AuthorHilton_Walter	1.90	0.75	0.34
AuthorJulian_Of_Norwich	0.25	0.67	0.09
AuthorKempe_Margery	0.78	0.21	0.15
AuthorMalory_Thomas	1.06	0.18	0.19
AuthorMirk_John	1.14	0.21	0.20
AuthorPurvey_John	-1.30	0.56	0.02
AuthorReynes_Robert	0.11	0.61	0.08
AuthorRolle_Richard	-0.35	0.47	0.05
AuthorThornton_Robert	0.44	0.38	0.11
AuthorTrevisa_John	-0.70	0.29	0.04

9.6 Associations with *be*

Following the procedure laid out in the previous chapter, it is possible to describe the relationship between the various target words and *be* using conditional probability. Keep in mind that

$$P(w_i \mid w_j) = \frac{n(w_i \cap w_j)}{n(w_j)}. \tag{9.1}$$

9.6.1 *There* and *be*

First, let us consider the case of *there* in initial position:

$$P(be \mid Init.there) = \frac{425}{1201} = 0.354 \tag{9.2}$$

This is about 65 times higher than independence, i.e. the product of the marginal probabilities:

$$P(be) \times P(Init.there) = \frac{1201}{84685} \times \frac{32273}{84685} = 0.005. \tag{9.3}$$

For non-initial *there*, we get:

$$P(be \mid \neg Init.there) = \frac{656}{2108} = 0.311 \tag{9.4}$$

As in Old English, the association is weaker in non-initial position; in this case the conditional probability of *be* given *there* in non-initial position is about 33 times higher than independence:

$$P(be) \times P(\neg Init.there) = \frac{2108}{84685} \times \frac{32273}{84685} = 0.009. \tag{9.5}$$

9.6.2 *Here* and *be*

Next, we turn to the relationship between *here* and *be*. There are only 58 cases in the selection where a form of *be* is immediately preceded by *here*. The conditional probability of *here* in initial position is much lower than for *there*

$$P(be \mid Init.here) = \frac{28}{233} = 0.120. \tag{9.6}$$

However, the deviation from independence is larger, around 115 times higher than independence:

$$P(be) \times P(Init.here) = \frac{233}{84685} \times \frac{32273}{84685} = 0.001. \tag{9.7}$$

For non-initial *here*, the conditional probability of *be* is much lower:

$$P(be \mid \neg Init.here) = \frac{30}{598} = 0.05. \tag{9.8}$$

As with *there* we find a deviation from independence also for the non-initial positions, in this case around 19 times higher:

$$P(be) \times P(\neg Init.here) = \frac{598}{84685} \times \frac{32273}{84685} = 0.003. \tag{9.9}$$

Thus, there seems to be some evidence suggesting a relationship between *here* in initial position and *be*, but the actual frequencies compared with those of *there* suggest that this is a marginal phenomenon.

9.6.3 Existential *there* and *be*

Finally, the association between existential *there* and *be*.

$$P(be \mid Ex.there) = \frac{888}{1386} = 0.641 \tag{9.10}$$

The conditional probability of *be* given existential *there* turns out to be approximately 103 times higher than the multiplied marginal probabilities:

$$P(be) \times P(Ex.there) = \frac{1386}{84685} \times \frac{32273}{84685} = 0.006. \tag{9.11}$$

We can make this further comparable with the previous results by looking at cases where existential *there* occurs in initial position:

$$P(be \mid Init.ex.there) = \frac{303}{397} = 0.763 \tag{9.12}$$

Not surprisingly, we get an even stronger effect for existential *there* in initial position, where the conditional probability of *be* given existential *there* is almost 430 times higher than than the product of the marginal probabilities:

$$P(be) \times P(Init.ex.there) = \frac{397}{84685} \times \frac{32273}{84685} = 0.002. \tag{9.13}$$

The magnitude of the factors (103 and 430 times) separating random use from co-occurrence in the case of all cases and initial cases of existential *there* suggests that there is a strong association between existential *there* and *be* in PPME2. This can be compared with the results from YCOE, where $P(be \mid there)$ was 0.15, or 8.4 times higher than chance, while $P(be \mid Initial.there)$ was 0.57, or 224 times higher than chance. Thus, it seems that existential *there* in Middle English is more strongly associated with initial position than *there* was in Old English.

9.6.4 Other verbs

As mentioned above, forms of *ben* ("be") are by far the most frequent verb in the target word's first right context with 1 214 occurrences. The second most frequent verb is *comen* ("come"), with 133 occurrences; other verbs are very infrequent. If the scope

is restricted to cases with existential *there*, we find 118 tokens with forms of *comen* as the first right context. Below are examples of existential *there* used with *comen*:[6]

(12) and ofte þer comeþ / greate ziknesses.
"and often there comes great sicknesses."
(CMAYENBI,53.944)

(13) and þere come wolfes,
"and there came wolves,"
(CMBRUT3,15.418)

(14) Þer come a angele
"There came an angel"
(CMSIEGE,85.459)

A quick search of the PPME2 shows that there are 4 056 instances of *comen* in the corpus. If we make the simplifying assumption that there is one token for each of the cases of *comen*, the conditional probabilities can be calculated as follows:

$$P(come \mid Ex.there) = \frac{133}{1386} = 0.09 \tag{9.14}$$

$$P(come) \times P(Ex.there) = \frac{1386}{84685} \times \frac{4056}{84685} = 7.8\,e-4. \tag{9.15}$$

As this shows, the conditional probability of *comen* given existential *there* is more than 100 times higher than independence. Although there seems to be an association, the actual frequencies involved are low, and *comen* is by far the most frequent among the non-*be* verbs occurring in the first right context.

This suggests that existential there in Middle English is not found exclusively with *be*, but that the range of verbs associated with the prototypical core of the EC is nevertheless restricted.

9.6.5 Interim summary

In the present section, it was shown that like in Old English, there is a *position-effect* for the conditional probability of co-occurrence with *be*, where items in initial position tend to have a higher conditional probability of occurring with *be*. It was shown that the conditional probability of *be* given existential *there* is very high, regardless of position, but that existential *there* also exhibits a position-effect. Apart from existential *there*,

[6]The examples are taken from *Ayenbite of Inwyt* ("CMAYENBI"), *The Brut* ("CMBRUT"), and *The Siege of Jerusalem* ("CMSIEGE").

the strongest effect was found for initial *here*. However, with only 28 instances of *here* in initial position followed by *be* (as compared with 303 for existential *there*), this appears to be a more marginal phenomenon than in YCOE where raw frequencies were much higher and close to those of *there*. An association was also found for existential *there* and *come*. Together with the results from section 9.3, this can be interpreted to mean that the effects are stronger for individual verbs than for broader semantic verb classes. The results for the co-occurrence of *there*, *be*, and nominative NPs is presented in figure 9.12. A strong effect is found for the pattern *there* + *be*, whereas a medium effect is found for the pattern *be* + … + NP.

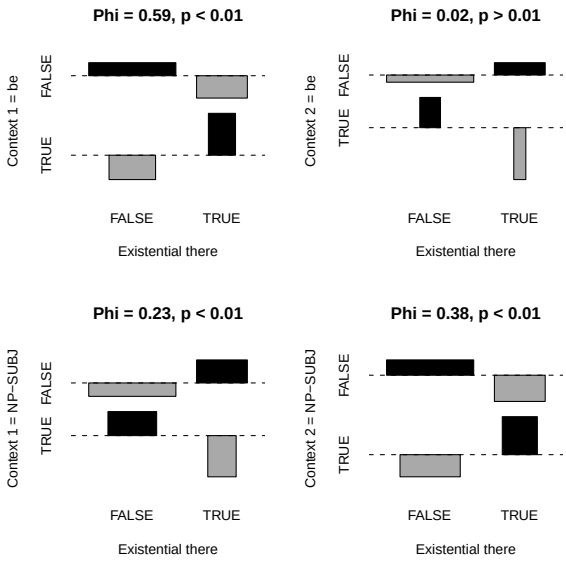


FIGURE 9.12: *Cohen-Friendly plots showing combinations of* there, be*, and nominative/subject NPs. For each plot the p-value from a Pearson chi-square test is presented alongside the ϕ effect size coefficient. A strong effect is found for* there *followed by* be *(upper left), whereas a medium effect is found for* there *followed by an NP in second position, i.e. with some other element between the two (lower right).*

9.7 A model of existential *there*

Since existential *there* can be coded as a binary variable (existential *there* vs. all other cases, i.e. all locative adverbs), it is possible to use a logistic regression model to examine which factors are associated with existential *there*. The observations are not temporally independent, so a random effect for the time-variable `Year` is called for. Using the `lmer()` function, the following logistic GLMM was found to be a reasonably good fit to the data:

(15)
```
ExThere ~ LogComplexity * BeContext + SemClass + NomNP
+ InitialAdv + (1|Year), family = binomial
```

In the model in (15), the binary response `ExThere` (coded as `ExTag` in the dataset) has the levels `TRUE` and `FALSE`, where `TRUE` represents all and only instances of existential *there* as coded in the PPME2 corpus, while `FALSE` represents all other locative adverbs, including locative uses of *there*. For simplicity, the neutral term "target word" will be used when referring to occurrences in the selection of either existential *there* or a locative adverb. `LogComplexity` is the syntactic complexity of the corpus token. `BeContext` is a binary variable with levels `TRUE` and `FALSE`, where `TRUE` indicates that the immediate right context of the target word is a form of *be*, while `FALSE` is any other right context. `SemClass` gives the semantic class of the first verb of the clause (i.e. not necessarily the main verb of the main clause, but in most cases they coincide), a total of 45 semantic classes were assigned for the Middle English data selection. Finally, `NomNP` is a binary variable with the levels `TRUE` and `FALSE`, where `TRUE` means that the second right context of the target word is a noun phrase coded as either `NP-SUBJ` or `NP` in the corpus, which would be indicative of a post-verbal NP in an EC.

9.7.1 Model evaluation

In the model fitting phase, a number of predictors were tried, including but not restricted to whether the token comes from a translated text, whether the token contains an object NP, the number of tokens from that particular 25-year interval, the position of the target word, and the number of verbs making up the semantic class of the token's first verb. These predictors proved either to lead to a worse fit (as judged by looking at the residuals vs. fitted plots), or to have no predictive power as measured by looking at coefficients, pseudo-R^2, and log-likelihood tests of the number of predictors. The interaction between `LogComplexity` and `BeContext` is a significant improvement over a model with no interaction as measured with a log-likelihood test ($\chi^2_{(1)} = 46.88$,

$p < 0.01$). However, the interaction term does not amount to a great improvement for the full model over the mixed intercept model as measured with Nagelkerke's R^2. Nagelkerke R^2 is 0.48 for the model with interaction and 0.47 for the model with no interaction. Thus, the simpler model displayed in (16) could also be argued to be a reasonable fit to the data.

(16)
```
ExThere ~ LogComplexity + BeContext + SemClass
+ NomNP + (1|Year), family = binomial
```

Looking at the residuals vs. fitted plots for (15) and (16), shown in figures 9.13 and 9.14 respectively, it seems that removing the interaction effect as in (16) creates a slightly better fit. As figure 9.13 illustrates, there are outliers and what appears to be a mild non-constant variance effect at higher probabilities. Conversely, the plot in figure 9.14 displays fewer outliers and a somewhat more constant variance.

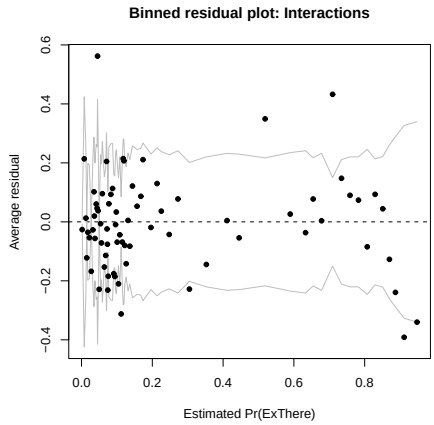


FIGURE 9.13: *Binned residuals vs. fitted plot for the model in (15) with an interaction effect between* `LogComplexity` *and* `BeContext`*. The x-axis shows the estimated probability of existential* there.

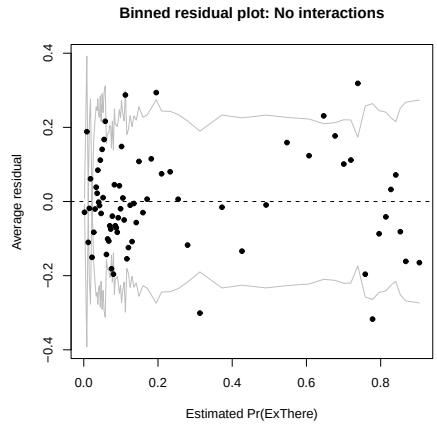


FIGURE 9.14: *Binned residuals vs. fitted plot for the model in (16) with no interaction between* `LogComplexity` *and* `BeContext`*. The x-axis shows the estimated probability of existential* there.

Note that the y-axis scale is a little different for the two plots, and that the plot in figure 9.14 actually has a smaller 95% confidence interval than the plot in figure 9.13. Thus, since the difference in improvement for the two models over the mixed intercept-only model as measured with Nakelkerke's R^2 was very small, and since the

plot in figure 9.14 appears to be a slightly better model, it seems warranted to choose the simpler model in (16). The input for the `LogComplexity` variable is centered around the mean value to make it easier to interpret the coefficient. This simply amount to subtracting the mean of the input variable from the input variable itself, cf. Gelman and Hill (2007, 93). The modified model is displayed in (17):

(17)
```
ExThere ~ c.LogComplexity + BeContext + SemClass
+ NomNP + (1|Year), family = binomial
```

Centering `LogComplexity` around the mean does not affect the model as such, so it is not necessary to re-check the model. For reasons of space, table 9.5 only shows the coefficients for `c.LogComplexity`, `BeContext`, and `NomNP`, in addition to the intercept. The full output from the model in (17) including the effects for `SemClass` is found in appendix B.

TABLE 9.5: *Fixed effects for the model in (17). The reference level is* `SemClass:Ability`. *Note the large effect of co-occurrence with* be.

Fixed effects	Est. coef.	Std. Error	z value	Pr(>|z|)	Pr(ExThere)
(Intercept)	-2.0	0.29	-6.86	6.71e-12	0.12
c.LogComplexity	0.27	0.05	5.29	1.20e-07	0.15
BeContextTRUE	2.80	0.10	26.93	$<$ 2e-16	0.69
NomNPTRUE	0.92	0.09	10.47	$<$ 2e-16	0.25

If we look at the values covered by two standard deviations, i.e. the uncertainty estimates, around the coefficients, we find the following: for `c.LogComplexity`: [0.14, 0.16]; co-occurrence with *be* (i.e. `BeContext:TRUE`): [0.64, 0.73]; nominative NP (`NomNP:TRUE`): [0.22, 0.29].

The standard deviation for the random effect (`Year`) is 0.64, and we can quickly find the mean difference between the 25-year intervals on the probability scale by dividing by four, cf. Gelman and Hill (2007, 82). As it turns out, the 25-year intervals in Middle English differ in their estimated mean probability of existential *there* by $\pm$16%. Thus, for the intercept with `BeContext:FALSE` and `NomNP:FALSE` (i.e. no co-occurrence with *be* and no nominative NPs) we would expect mean probabilities anywhere in the range of [0.04, 0.33].

The effects for log SCR and nominative NP alone are within this range, whereas co-occurrence with *be* clearly stands out. Although `c.LogComplexity` and `NomNP:TRUE`

have higher estimated mean probabilities than the intercept, they are still not entirely unexpected given the wide standard errors for the 25-year interval groups. Co-occurrence with *be* on the other hand has a very strong effect and even the lower bound estimate for its two standard deviations confidence interval is much higher than the upper bound estimate for the groups.

Given this, it is worth asking whether only `BeContext` should be kept as a predictor in the model. Consider the four models below:

(18)
```
ExThere ~ c.LogComplexity + BeContext + SemClass
+ (1|Year), family = binomial
```

(19)
```
ExThere ~ c.LogComplexity + BeContext + NomNP
+ (1|Year), family = binomial
```

(20)
```
ExThere ~ c.LogComplexity + BeContext
+ (1|Year), family = binomial
```

(21)
```
ExThere ~ BeContext + (1|Year), family = binomial
```

Removing `NomNP` as in (18) reduces Nagelkerke's R^2 to 0.45, while removing `SemClass` as in (19) gives a Nagelkerke R^2 of 0.42. Removing both of them at same time gives a Nagelkerke R^2 of 0.39. Keeping only `BeContext` as a predictor in the model, shown in (21) also gives a Nakelkerke R^2 for the full model over the mixed intercept model of 0.39.

Removing `LogComplexity` only does not affect Nagelkerke's R^2 which remains at 0.47 (with two decimals), but causes other problems. The `lmer()` function gives an output which appears reasonable enough with only `BeContext`, `SemClass`, and `NomNP` as predictors. However, attempting to produce a residuals vs. fitted plot for this figure results in an error message, saying that infinite values are passed to to the plotting function in R. Possible causes of this could be trying to perform an undefined operation, such as dividing by zero or taking the logarithm of zero. Thus, as long as `BeContext`, `SemClass`, and `NomNP` are all kept in the model, it makes sense to keep `LogComplexity` too, even if removing it does not change the R^2 value.

Based on the R^2 values it seems warranted to keep `SemClass` as a predictor in the model, while the case for `NomNP` is more dubious. To assess these models further, residuals vs. fitted plots are presented in figure 9.15. The four plots show the cases outlined above, with `NomNP` removed, with `SemClass` removed, with both of them removed, and with only `BeContext` as a predictor. As explained above, it was not possible to produce such a plot for the case where only `LogComplexity` was removed.

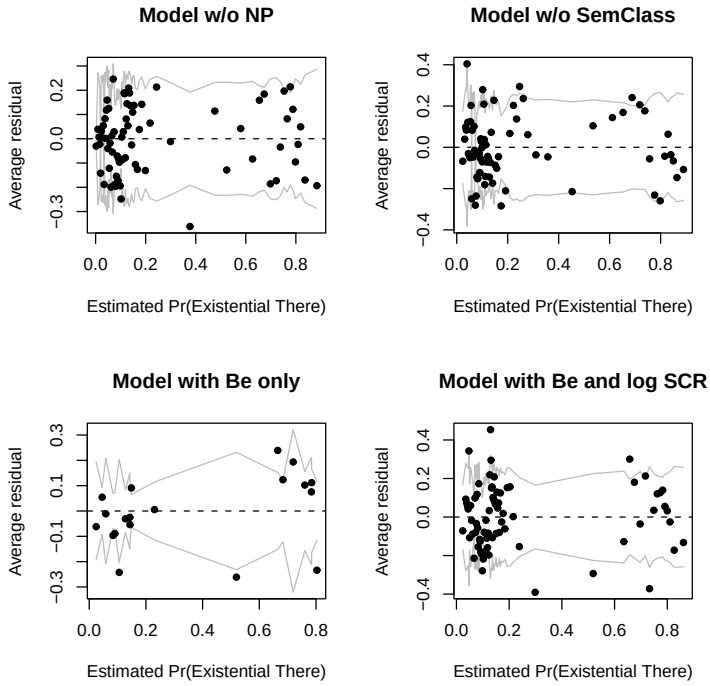


FIGURE 9.15: *Binned residual vs. fitted plots for models (18), (19), (20), and (21), shown clockwise from top left. Compare with figure 9.14 on page 231.*

For all the plots in figure 9.15, the 95% confidence intervals are smaller than or equal to the interval in figure 9.14, although the differences are not great. The best plot appears to be the one in the upper left corner, for the model in (18) where `NomNP` has been removed, but the other predictors are kept. Here, the 95% confidence interval is about the same size as in figure 9.14, but there is only one clear outlier.

It is not immediately obvious which model to choose in this case, but the best candidates seem to be either (17), (18) or (20). While the residuals vs. fitted plot for (18) is a little better than for (17), the former model has a lower R^2 than the latter (0.45 vs. 0.47). The lower right plot in figure 9.15 for model (20) also has an acceptable fit, although it does seem inferior to previous two. This illustrates some of the dilemmas in model fitting with linear models: which criteria should be used to choose between

models and how should they be evaluated against each other? Further criteria, such as *AIC*, *BIC* and log-likelihood have not been included in the discussion since they do not have as intuitive interpretations as the residual vs. fitted plots and Nagelkerke's R^2. A good rule of thumb is to choose the simpler one when two models are otherwise similar, cf. Faraway (2005, 25).

However, in addition to *describing* the data in Middle English, the current model will also be used to *estimate* the distribution of existential *there* in Old English. Hence, a criterion for choosing between the models should also be how well they perform in such as task. To test this, the rank measures C and D_{xy} can be used, as discussed in chapter 4. The results are instructive: the full model which includes log SCR, whether *be* is the context or not, semantic class and presence of a nominative NP has a C of 0.68 and a D_{xy} of 0.37. Recall that a C of 1 indicates perfect prediction of the response while 0.5 indicates a model which performs no better than random; the corresponding values for D_{xy} are 1 and 0. Baayen (2008a, 204) states that for a model to have some predictive capacity, C should be at least 0.8, which gives reason to doubt the usefulness of this model. The model in which `NomNP` has been removed performs worse ($C = 0.60$, $D_{xy} = 0.20$). The best model in this respect is in fact the simplest one in (20) which seems to have some real predictive capacity ($C = 0.84$, $D_{xy} = 0.68$). In other words, extending the coverage or calibration of the model (R^2) comes at the expense of discrimination (C/D_{xy}).

The final decision as to which model to use for estimating existential *there* in YCOE will be made in a subsequent chapter. Suffice it to say that we have valid reasons to suspect that the simpler model will perform better. However, for exploratory data analysis, it might also be worthwhile to consider larger models, as illustrated in table 9.5 which shows that having a nominative NP as the last element in the 3-gram gives a substantial effect size.

Regarding the diachronic picture, it is worth noting that both the more complex model and the simpler one give very close estimates for between-year variation. (17) has a standard deviation for the random effect of 0.64 on the logit scale which through the divide by 4 rule, cf. Gelman and Hill (2007, 82), gives an estimated difference in probability of existential *there* between the 25-year intervals of $\pm 16.00\%$ above the differences explained by linguistic factors. The corresponding value for the simpler model in (20) is 0.69, which yields a difference of $\pm 17.25\%$. In other words, most of the variation is accounted for by the linguistic factors, not a diachronic trend.

9.8 Summary

In the present chapter, the investigation of PPME2 has showed that *there* is the most frequent locative adverb. Existential *there* is also frequent, particularly in the later part of the corpus, although this might be due to the larger sample size for the later material. When *there* is used existentially, the most commonly attested pattern is the canonical THERE + BE + NP pattern also found in Present-day English.

Furthermore, existential *there* was found to have a high conditional probability of co-occurring with *be* (and to some lesser extent with *come*). Interestingly, the relatively high conditional probability of initial *here* followed by *be* attested in YCOE was not found in PPME2.

The hypothesis that tokens with existential *there* would have a lower syntactic complexity ratio was rejected. Instead, it was found that in initial position, tokens with existential *there* have a significantly higher mean log complexity ratio, with a small to medium effect size. As in YCOE, target word position showed only a weak association with the semantic class of the verb.

A series of logistic generalized mixed effects models were fitted, and two competing models were closely examined in terms of goodness of fit, calibration and discrimination. The evidence favored the simpler model, where the occurrence of existential *there* could be predicted using only the occurrence of *be* in the right context and the log SCR as fixed effects.

Thus, it would seem that the selectional gap between *there* and *here* has widened from Old English to Middle English, and that the presence of *be* is the single strongest indicator of whether *there* is used as an existential subject. Plots showing the proportions of locative and existential *there* by 25-year interval points to an increase in the latter towards the end of the period. However, the DP index shows that the proportions of locative and existential *there* are evenly distributed. Thus, it is difficult to decide whether the status of *there* in Middle English is changing or not, although the evidence seems to point in the direction of stability.

Chapter 10

The Early Modern English EC

10.1 Introduction

The current chapter will present some key data from the Early Modern English period, represented by the Penn-Helsinki Parsed Corpus of Early Modern English (PPEME), cf. Kroch, Santorini, and Delfs (2004). The emphasis will be on exploration rather than hypothesis testing, since the hypotheses discussed in earlier chapter have been directed at Old and Middle English. However, it is important to analyze the Early Modern period for the following reasons:

(i) it provides a contrast for the analyses of the previous corpora,

(ii) it provides a background for explicit and implicit comparisons between the older periods and present-day English.

Given that the hypotheses tested earlier actually hold, we would expect to see little development regarding existential *there* in Early Modern English.

10.2 Data

The prose part of the Early Modern English treebank PPEME consists of 1 794 010 words in some 106 302 tokens. As in Middle English (and present-day English), we find both the existential and the locative use of *there*.

10.2.1 Collecting data

The CorpusSearch query in (1), identical to the query employed for Middle English, yields all tokens containing either an existential *there* (EX) or a locative adverbial:

(1)
```
node: IP*
query: (EX exists) OR (ADVP-LOC exists)
```

As with the previous dataset, these tokens were further processed with Perl to fit a dataframe format which can be read into R. Table 10.1 shows six of the columns for the first six rows in the dataframe. The full dataframe has 9 087 rows and 62 columns. Among the information included is which target word, i.e. locative adverb or existential *there*, was found (for convenience labeled "Adverb"), complemented by the corpus tag which distinguishes between the locative and existential uses.

TABLE 10.1: *An excerpt from the* EME *dataframe, showing six columns from the first six rows.*

	Adverb	AdverbTag	ConTag	Text	BeContext	Year25
1	were	WADV	END	ALHATTON-E3	FALSE	1675
2	ther	EX	BEP	ALHATTON-E3	TRUE	1675
3	there	EX	BEP	ALHATTON-E3	TRUE	1675
4	ther	EX	BED	ALHATTON-E3	TRUE	1675
5	there	ADV	END	ALHATTON-E3	FALSE	1675
6	here	ADV	CP-THT	ALHATTON-E3	FALSE	1675

10.2.2 The adverbs and existential *there*

In Early Modern English, the selection of 9 087 tokens contains around 130 types, including forms of existential *there* (the figure is not exact, since there are spelling variations). This implies a reduction in the number of types compared with the approximately 250 types found in PPME2, but not so far from the approximately 150–160 types seen in YCOE. It is not immediately obvious why the smallest of the three corpora should have the largest variation in types, although one possible explanation could simply be more spelling variation, since none of the corpora are lemmatized.

Figure 10.1 shows frequencies of *there* vs. all other target words (locative adverbs) in PPEME. Unsurprisingly, *there* still dominates, with 5 380 tokens. In other words, tokens with *there* account for some 59% of the selection, while the other types are distributed over the remaining 3 707 tokens.

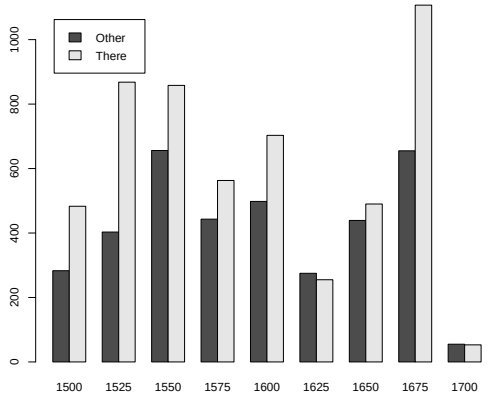


FIGURE 10.1: *Barplot showing frequencies of* there *vs. other target word realizations in EME in 25-year intervals.*

Figure 10.2 plots the frequencies for the use of *there*. As the plot shows, the locative use of *there* still dominates in the first quarter of the 16th century, but later the existential use is more frequent. The low number of occurrences in both plots for the first quarter of the 18th is a result of the corpus design. Very few texts in PPEME extend beyond 1700; the few that do cause a sharp drop in frequencies which is merely an artifact of the corpus design.

Turning to the immediate right context of the target word, we find that forms of *be* dominate, with at 2 810 occurrences, or 31% of all tokens. The second most frequent category is composed of nouns and NPs (primarily tagged as subjects), with 2 260 occurrences, or 25%. This is largely correlated with whether the target word is existential *there* or not.

Table 10.2 shows frequencies and percentages for the use of forms of *be* and subject NPs in first and second context position for existential and locative *there*. Clearly, the use of *be* as the first context word and subject NPs as the second context word is quite distinctive for existential *there*, whereas for locative *there* the most characteristic feature is the use of subject NPs in second context position.

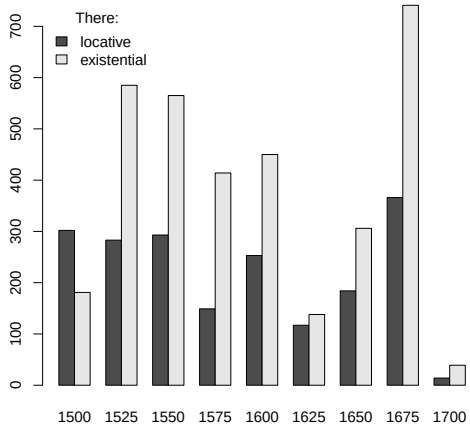


FIGURE 10.2: *Barplot showing frequencies of existential* there *vs. locative* there *in EME in 25-year intervals.*

TABLE 10.2: *The use of* be *and subject NPs with existential and locative* there *in EME, for first and second context position of the target word.*

There	1st context		2nd context	
	Be	NP	Be	NP
Existential	2469 (72%)	98 (3%)	228 (7%)	2161 (63%)
Locative	58 (3%)	465 (24%)	99 (5%)	249 (13%)

10.2.3 Syntactic complexity

Based on the findings regarding syntactic complexity in Middle English, we would expect the same pattern in Early Modern English, i.e. a somewhat higher mean complexity ratio for tokens with existential *there*. In the selected tokens from PPEME, the mean log complexity is 2.37, with a standard deviation of 1.15. Tokens with existential *there* have a mean log complexity of 2.45, whereas tokens without have a mean of 2.32. As in PPME2, the difference in mean log complexity is significant when measured with a nondirectional two sample Welch t-test, but the effect size as measured with Cohen's d is low ($t_{df(8118.48)} = 5.33$, $p < 0.01$, $d = 0.11$). The estimated 95% standard deviation for the difference in means is [0.08, 0.17], which corresponds to between approximately $\frac{1}{15}$ to $\frac{1}{7}$ of one standard deviation from the mean. Turning to tokens with a target word in initial position, we find that the mean log complexity is 1.80, whereas for tokens with the target word in non-initial position the mean log complexity is 2.52. Again, this was tested with a nondirection two sample Welch t-test. The difference in means is statistically significant, and in this case the effect size is moderate to strong ($t_{df(3615.48)} = 28.55$, $p < 0.01$, $d = 0.69$). The estimated 95% confidence interval of the difference in means is [0.68, 0.78], that is, consistently close to 0.5 standard deviations from the overall mean.

10.3 Adverb position

As in the previous two chapters, a CA analysis was carried out on the position of target words and the semantic classes of verbs. The numeric output can be found in appendix A.3, while the CA biplot is shown in figure 10.3 on p. 242. As in the other CA analyses, the representation is good, with the first dimension accounting for 91.2% of the variation, while the inertia is fairly low (approximately 0.08 out of a maximum of 2, i.e. of the same magnitude as in previous chapters).

Although the overall association between target word position and semantic class is still low, it is evident from the map in figure 10.3 that the association between initial position and verbs of existence has become clearer. This is the only semantic class in the PPEME material which is clearly associated with target words in initial position; all other classes are either unrelated or negatively associated with initial target words. Thus, the association between initial target words and verbs of existence seems to be fairly well established in Early Middle English.

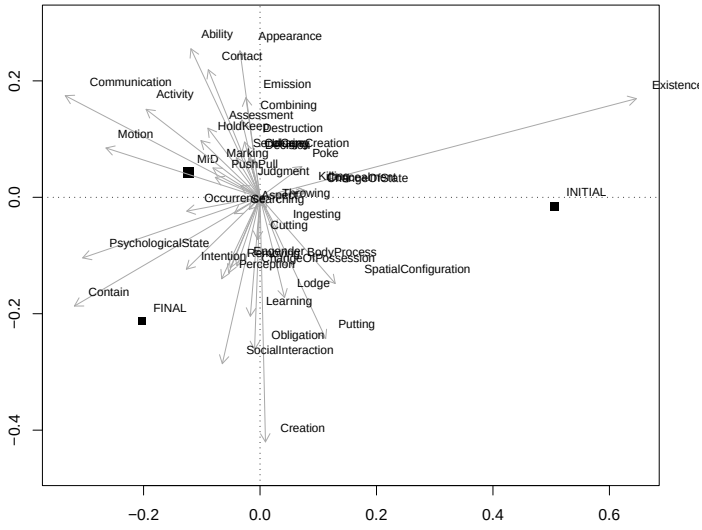


FIGURE 10.3: *x: 91.2%, y: 8.8%. Scaling:* `map = "rowgreen"`*. Total inertia is 0.08 of a maximum of 2, i.e. fairly low, but marginally higher than in previous chapters. The rows (adverb position) are in principal coordinates, while columns are in standard coordinates times the square root of the mass. Row point sizes are plotted proportionally to their relative frequency.*

10.4 An overview of *there*

Table 10.3 presents frequencies and proportions of existential and locative uses of *there* in nine 25-year intervals for the PPEME data. With the exception of the interval from 1525–1550, it is clear that the existential use is more frequent than the locative use, unlike in PPME2.

TABLE 10.3: *Frequencies and proportions of existential and locative uses of* there*, by 25-year intervals from* PPEME*.*

Year25	Ex. there	%	Loc. there	%	Total there
1500	181	0.37	302	0.63	483
1525	585	0.67	283	0.33	868
1550	565	0.66	293	0.34	858
1575	414	0.74	149	0.26	563
1600	450	0.64	253	0.36	703
1625	138	0.54	117	0.46	255
1650	306	0.62	184	0.38	490
1675	741	0.67	366	0.33	1107
1700	39	0.74	14	0.26	53
Total	3419		1961		5380

The distribution of existential and locative *there* in PPEME is homogenous, as indicated by a DP of 0.05. As explained in chapter 8, a DP value close to zero indicates that the differences between expected and observed proportions of existential and locative *there* for the 25-year intervals are very small. Table 10.4 gives the total number of corpus tokens per 25-year interval, and although there are clear differences in sample size, the coverage is nevertheless better than for PPME2, with no interval represented by less than 2 000 tokens.

Figures 10.4 and 10.5 plots (respectively) the raw frequencies and proportions of locative and existential uses of *there* in PPEME. There are two things to note here, first the sharp drop for the 1700 mark. As explained above, this is merely a result of the corpus design and says very little about the development of existential *there*. More interestingly, around the 1525 mark we can see the cross-effect from the numbers in table 10.3. At this time existential *there* seems to stabilize itself as more frequent than the locative use. In other words, in the first part of the Early Modern English period we

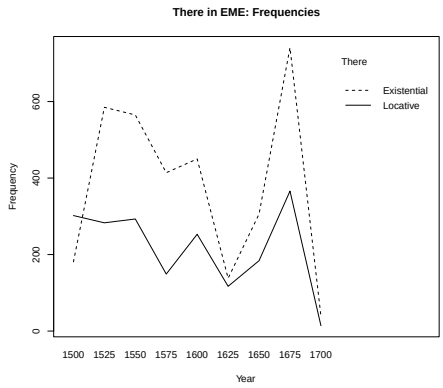


FIGURE 10.4: *Frequencies of locative and existential uses of* there *in* PPEME*, by 25-year interval.*

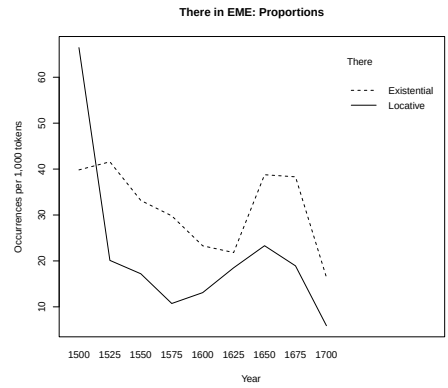


FIGURE 10.5: *Proportions of locative and existential uses of* there *in* PPEME*, by 25-year interval. The scale on the* y*-axis is occurrences per 1 000 corpus tokens.*

find a qualitative change in the quantitative distribution between the two uses of *there*, whereby the existential use becomes the dominant one.

10.5 Associations with *be*

As in the previous corpora, there seems to be a clear association between existential *there* and *be*. Rather than calculating conditional probabilities, I will here simply report the results of four Pearson chi-square tests of independence, as was done in the previous chapter to supplement the calculations of conditional probability.

Table 10.6 shows Cohen-Friendly association plots for the four cases tested:

(2) a) *There* + *be*
b) *There* + . . . + *be*
c) *There* + NP
d) *There* + . . . + NP

where the ellipses represent an unknown element. The patterns are to some extent overlapping, but they are nevertheless interesting since they indicate which elements tend to co-occur with *there* in different positions. All p-values are significant at the 1%

TABLE 10.4: *Total number of corpus tokens from* PPEME *by 25-year interval.*

Year25	No. of tokens
1500	4545
1525	14068
1550	17034
1575	13887
1600	19320
1625	6324
1650	7893
1675	19341
1700	2376

level, but the effect sizes vary considerably.

As the plots in figure 10.6 clearly demonstrate, the strongest effect is found for the case in a), i.e. *there* immediately followed by *be*. The pattern *there* followed by something and then followed by *be* (i.e. the case in b)), is significant but has a trivial effect size, which is not surprising, since we would not expect to find a strong association here anyway. Turning to the bottom row, we see that the pattern in c), *there* followed immediately by an NP is significant, but that the effect size is smaller than for *there* and *be*. Finally, the pattern in d), *there* followed by something (e.g. *be*) followed by an NP is significant and has a medium effect size. In other words, there is a tendency for *there* to occur with *be* immediately following, and also for *there* to occur with an NP in the second context. A secondary pattern is evident as *there* followed by an NP, although the effect size is smaller than for the two previous patterns. The pattern where some other element occurs between *there* and *be* has such a small effect size that it can be disregarded. Thus, strong associations exist for *there* with *be* and NPs, but as expected these effects are position-specific.

10.6 A model of existential *there*

Based on the results from the previous chapter, the following logistic GLMM models seem sensible to try out:

(3) `ExThere ~ LogComplexity + BeContext + SemClass + NomNP`

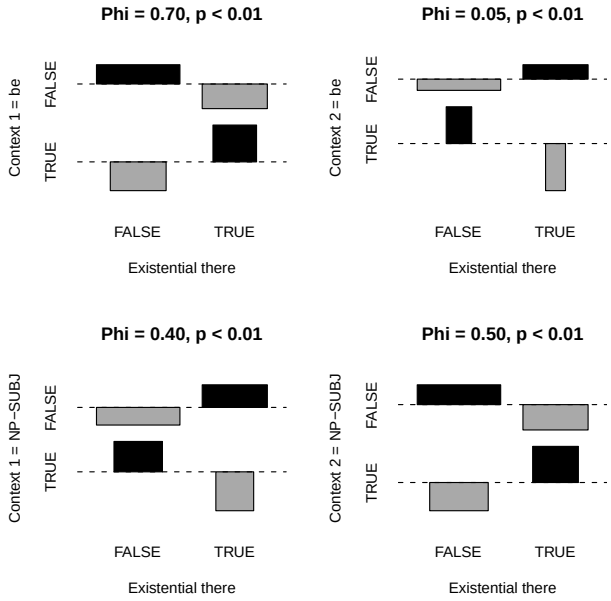


FIGURE 10.6: *Four Cohen-Friendly plots illustrating the association between existential* there*,* be *and NPs. The plots compare existential* there *with* be *and NPs in the first and second position of the linear order following* there*. For each plot, the* p*-value from a Pearson chi-square test of independence and the* ϕ *effect size coefficient is reported. The biggest effects are found for* there *followed by* be *(top left), and* there *followed by an NP in the third position, i.e.* there . . . *NP (bottom right).*

```
    + (1|Year), family = binomial
```

(4)
```
    ExThere ~ LogComplexity + BeContext + (1|Year), family
    = binomial
```

As figures 10.7 and 10.8 show, both models fulfill the assumptions of constant variance, and no suspicious patterns are evident. The two models have fairly similar Nagelkerke R^2 values: 0.60 for the larger model and 0.55 for the smaller model, when measuring improvement over a mixed-intercept-only model. That is, both model are able to account for a fairly large amount of variation. However, the precision in the two models is different. The smaller model has a C of 0.86 and a D_{xy} of 0.73, both of which

indicate a good fit. The corresponding figures for the larger models are $C = 0.59$ and $D_{xy} = 0.19$, i.e. not very much better than chance. The situation is thus similar to what was found in Middle English, viz. that the larger model gains a higher R^2 simply by predicting more of the background noise, not the signal (existential *there*). Again, the smaller model seems the superior one, because it avoids overfitting (i.e. modeling the noise).

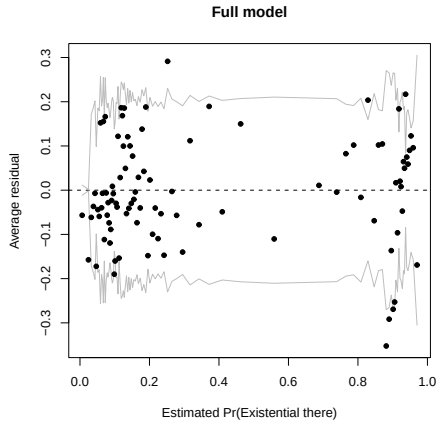


FIGURE 10.7: *Binned residuals vs. fitted plot for the model in (3). No particular problems are evident.*

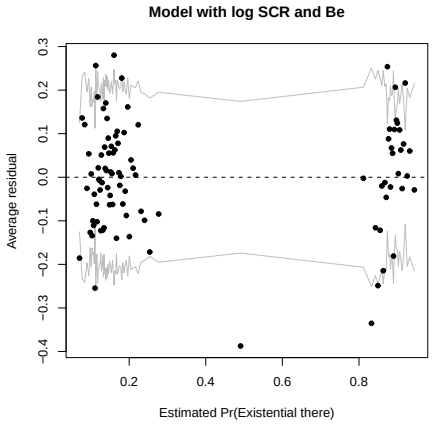


FIGURE 10.8: *Binned residuals vs. fitted plot for the model in (4). No particular problems are evident.*

Table 10.5 shows the fixed effects from model (4). The standard deviation of the random effect (year) is 0.30, which gives an estimated maximum between-year difference for the probability of existential *there* of 7.5%. In other words, the average difference between the 25-year intervals is very small.

As the results in table 10.5 show, the standard errors for the fixed effects are small, indicating that there is enough data to estimate them properly. The mean estimated probability of existential *there* with a *be* context is 0.81, but only 0.09 for a non-*be* context. Thus, the model accounts for the data in an adequate way and gives a strong indication that the single best predictor for existential use of *there* is co-occurrence with *be*.

TABLE 10.5: *Fixed effects from the logistic GLMM model in (4), with a random effect for the* `Year25` *variable. The standard deviation of the random effect is 0.30.*

Fixed effect	Est. coef.	Std. Error	Pr(Existential there)
(Intercept)	-2.32	0.13	0.09
LogComplexity	0.21	0.03	0.11
BeContextTRUE	3.80	0.07	0.81

10.6.1 Which measure?

In chapter 6, the (log) syntactic complexity ratio was presented as a heuristic measure of syntactic complexity. However, above we could see the effect of this measure in table 10.5, that is, increasing the log SCR by 1 increases the probability of existential *there* by 0.11. As I mentioned in chapter 6, it is not entirely clear what this measure represents. For the purposes of interpretation it is necessary to ask the question of whether other ways of approximating the size of the sentence token works equally well. The obvious thing to do, as discussed in chapter 6, is to take the components of the SCR itself – the number of IPs, NPs and nodes – and transform them to a natural log scale.

The four logistic GLMMs below try these three options, substituting log SCR with the count of respectively the number of nodes, NPs and IPs. The fourth model uses only *be* as a predictor for reference.

(5)
```
ExThere ~ Nodes3 + BeContext
+ (1|Year25), family = binomial
```

(6)
```
ExThere ~ NP + BeContext
+ (1|Year25), family = binomial
```

(7)
```
ExThere ~ IP + BeContext
+ (1|Year25), family = binomial
```

(8)
```
ExThere ~ BeContext + (1|Year25), family = binomial
```

Figure 10.9 shows the residuals vs. fitted plots for the four models. Using only *be* as a predictor seems to result in a mild nonconstant variance, while the other three models display a fairly constant variance, although outliers are a problem.[1]

[1] A log transformation of the counts makes the results somewhat better for nodes, but does not solve all the interpretation issues.

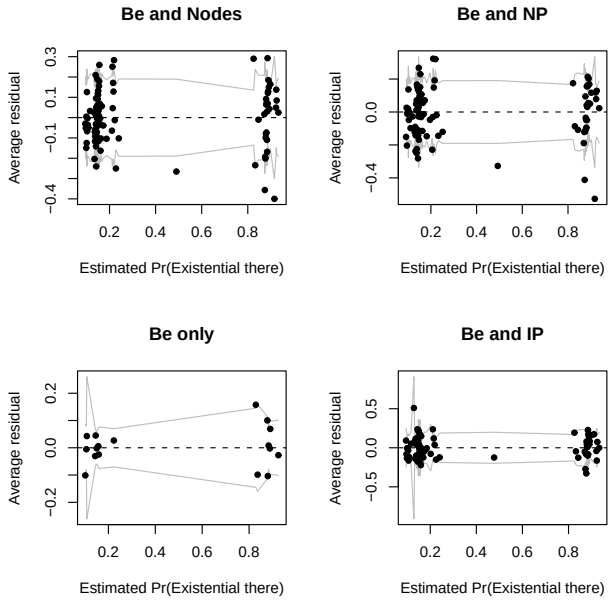


FIGURE 10.9: *Binned residual vs. fitted plots for the models in (5), (6), (7), and (8). There appears to be some problems with all of them, although the severity of the problems varies.*

Interestingly, substituting log SCR with one of the predictors outlined above changes very little. The coefficient for be is stable around 3.78 with an estimated probability of around 0.86, and the between-period variation remains constant. The effects of nodes, NPs and IPs are rather small. See appendix B.2 for the full output. Centered around its mean value, the effect of adding one NP translates into a en estimated probability of existential *there* of 0.13. The same probability is found for adding one IP, or adding three nodes (adding a single node makes little sense, since even a single word will consist of one phrasal node and one leaf node). Centering the input variables makes very little difference. The three models in (5), (6), and (7) all show a good fit as measured with C and D_{xy} (for all three, the C value is around 0.86 and the D_{xy} around 0.72). However, there is reason to suspect that this is mainly due to the influence of *be*.

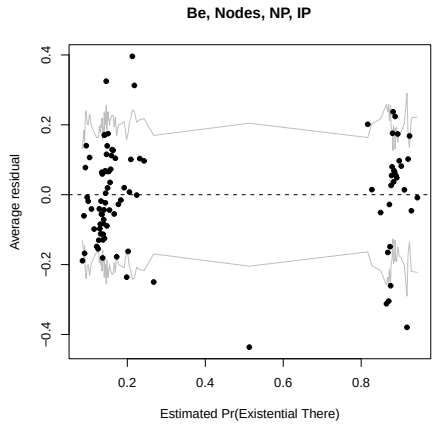


FIGURE 10.10: *Binned residual vs. fitted plot for a logistic GLMM including raw counts of nodes, NPs and IPs as predictors alongside* be*, with a random effect for year.*

10.6.2 Concluding remarks on SCR

The model in (8) with only *be* as a predictor has a C of 0.85 and a D_{xy} of 0.71. The plot in figure 10.10 shows a residuals vs. fitted plot for a model which combines counts of nodes, NPs and IPs with *be* in one model. The estimated probabilities of *there* does not change in any substantial way from the ones already reported. As the plot shows, the model is not entirely unacceptable, but there are more points falling outside the confidence intervals at the highest and lowest probabilities. Thus, it seems that a useful way – for modeling purposes – of combining the information in node, NP and IP counts is to use the proposed log SCR measure, even if it makes interpretation more complex. As I have shown, the main effect is anyway accounted for by *be*, and the most important function of the numeric predictor is to provide a better model fit.

In summary, it was shown that for some interpretation purposes, it is certainly advantageous to use raw counts as predictors. However, this would measure something else than the SCR, which was proposed as a measure which integrates memory and processing costs. The best fit is achieved with the log SCR, and for modeling purposes it seems useful to opt for the solution which provides the best fit.

10.7 Summary

Three points have been made in the present chapter. First, the summaries show that *there* behaves exactly as expected given the data and results presented in chapter 9. Second, that a shift in frequency takes place around 1525–1550, where existential *there* becomes more common than the locative use. Finally, it was shown that the log complexity ratio measure, introduced in chapter 6, is difficult to interpret in a GLMM. Instead, the measure could be substituted by raw counts of nodes, NPs and IPs. The result of this operation was to show that, individually, these counts have effect size of the same magnitude as the log *scr*, i.e. fairly low. Furthermore, the fit of the model (judging by the residual vs. fitted plots) seemed to be better with the log SCR. The subsequent chapter will employ the results from the current chapter and the previous one to estimate the probability of existential *there* in Old English, and then to tie the various results together.

Chapter 11

Discussion: The early English EC

> The hallmark of good science is that it uses models and "theory" but never believes them.
>
> Martin Wilk

11.1 Introduction

In the previous chapters it was shown that variants of *there* constitute the most frequent locative adverb. Furthermore, it was shown that there is a strong tendency for *be* to correlate with *there* in the Old English period. In Middle and Early Modern English it was not surprisingly found that existential *there* also is strongly associated with *be*. The present chapter will deal with some of the results and observations from the separate periods dealt with in previous chapters. Specifically, the logistic model from Middle English will be used to estimate the distribution of existential *there* in Old English. Based on this, some further analyses of existential *there* for the whole early English period (i.e. all three corpora) will be attempted. Subsequently, some views on how this fits into a RCG model of the early English EC will be presented.

11.2 The status of *there* in OE

YCOE does not distinguish between existential and locative uses of *þær*. However, as pointed out above there are good reasons to believe that *þær* was used as an existential subject in Old English. The problem to be solved here is two-fold: first, the information from the analyses in Middle and Early Modern English concerning which factors predict an existential use of *there* should be employed to make backward-predictions for the Old English data. Second, this should be done automatically for all 9 203 cases of locative tokens extracted from YCOE. This amounts to a classification problem which can be solved using a *Classification and Regression Tree* (CART) analysis with functions from the *rpart* library in R, cf. Therneau and Atkinson (2009). Such an analysis divides the initial training data into a series of non-overlapping subsets, based on the most useful predictors (Baayen, 2008a, 149). The model from the training data can then be used to make predictions for the test data. In this case, Middle and Early Modern English can be considered training sets, since we know what the result should be. The model can be extended to make predictions for Old English.

11.2.1 CART analysis

The `rpart()` function performs what is variously known as a *decision tree* or *recursive partitioning analysis* of the data. The algorithm creates decision rules for partitioning the response, based on the properties of the predictor variables. The basic idea here is to first give `rpart()` a regression model and a dataset with a known response (existential *there* in Middle English). This provides a baseline for evaluating the performance of the model. Then the model can be extended to a dataset with an unknown response (existential *þær* in Old English) and let the classification tree decide the cases which qualify as existential, based on the decision rules from Middle English applied to the properties of the Old English tokens.

The model

The best fit to the data in the previous chapters was achieved with the following logistic GLMM model, using `lmer()`. However, simpler models were also discussed, and in the following paragraphs, the best model for the estimation will be chosen. We start with the more complex model:

(1)
```
ExThere ~ LogComplexity + BeContext + SemClass + NomNP
+ (1 | Year)
```

The focus here is on the predictors (i.e. the fixed effects), leading to a simplified model as shown below:

(2) `ExThere ~ LogComplexity + BeContext + SemClass + NomNP`

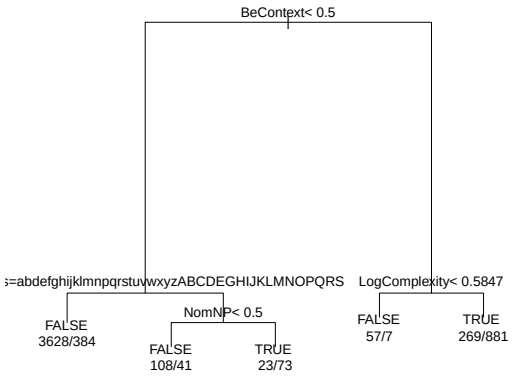


FIGURE 11.1: *Unpruned* CART *tree for ME existential* there*. The nodes give the decision rules for choosing between existential* there *(*`TRUE`*) or not existential* there *(*`FALSE`*). The numbers under the leaf nodes show how many cases that support / go against the given rule.*

Running the CART analysis above gives an initial, unpruned CART tree as shown in figure 11.1. However, as explained in Baayen (2008a, 150–151), there is a risk that this tree overfits the data. That is, the tree has too many splits or branches with no real predictive value for new data. The length of the branches is an indication of their explanatory value. Thus, we can see in figure 11.1 that *be* has a strong effect, as expected, and that the decision rule is based on a probability of *be* above or below 0.5. In the former case, a decision rule based on syntactic complexity comes into play. In the latter case, a rule based on semantic verb classes is invoked.[1] Finally, a rule based on the presence of a nominative NP is employed for some of the semantic classes. All branches after the first split look suspiciously short, the numbers under

[1] The letters are short reference codes for the semantic classes, to save space.

the leaf nodes (cases that support / go against the rule) show that not all the rules are that well supported, and it is possible that the tree is overfitting the data. To avoid this, the tree can be pruned by removing branches without predictive value. This is achieved through *cost-complexity pruning*, an algorithm which compares the size of the tree with its success in reducing *impurity*. Node purity here means that the ratio of existential *there* to non-existential *there* in a daughter node should be more extreme (or closer to 1 or 0) compared with the mother node (Baayen, 2008a, 150–151). Figure 11.2 shows a cost-complexity cross-validation plot for the tree in figure 11.1. The horizontal axis

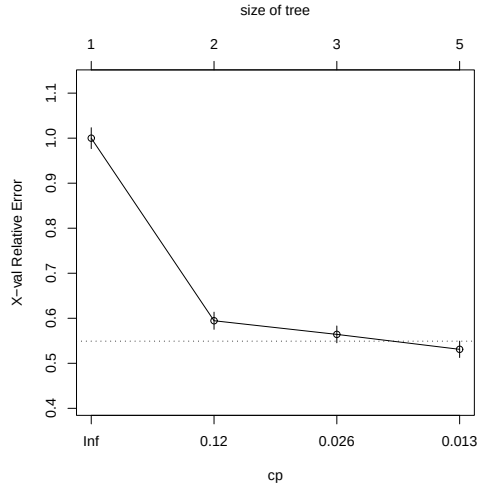


FIGURE 11.2: *Cost-complexity cross-validation plot for the unpruned* CART *tree (figure 11.1) for ME existential* there.

displays the cost-complexity parameter with the corresponding size of the tree. The dotted horizontal line represents one standard error above the mean of the lowest point, and according to the help files for the *rpart* library, a good cost-complexity (`cp`) value for pruning is the leftmost value with a mean above this line. This `cp` value is 0.026, and the pruned tree is displayed in figure 11.3. As the figure shows, only two rules from the initial tree have been kept, whether *be* is the immediately following word, and a rule based on the syntactic complexity of the clause. The final model is displayed in (3) below.

(3) `ExThere ~ LogComplexity + BeContext`

The single best predictor for the presence of existential *there* is clearly having a form of *be* as the immediate right context, whereas logcomplexity only makes a small (and less accurate) contribution. This supports the view in previous chapters that the larger model was overfitting the data.

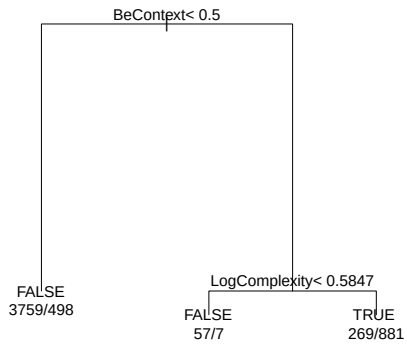


FIGURE 11.3: *Cost-complexity pruned* CART *tree for ME existential* there*. The decision tree is notably smaller than the one presented in figure 11.1. As the rightmost node shows, 881 cases are correctly classified as existential* there*, while 269 cases are wrongly classified as existential* there*.*

Before this classifier can be applied to the Old English data, we need some information about its accuracy. To check this, the model is used to predict the distribution of existential *there* in Middle English. The results are shown in table 11.1 on page 258.

As the table shows, 774 tokens are misclassified, which gives an error rate of 14%. Although this might seem high, it is a noticeable improvement from a baseline model which *only* guesses the most frequent category ($\neg$ existential *there*) and hence would be mistaken in all the 1 386 cases with existential *there*, i.e. 25.3% of the time. It is possible to quantify the model's performance further by measuring the *precision* and *recall* of the model (**?**, 81). The precision is the number of correctly classified existential *there* out of all predicted existential *there*, i.e. $881/1150$ which is 0.77. The recall is the number of correctly classified existential *there* out all cases of existential

TABLE 11.1: *Counts of observed and predicted existential* there *in ME, based on the CART tree in figure 11.3. 774 tokens are misclassified, which gives an error rate of 14%.*

	Observed	
Predicted	¬ ExThere	ExThere
¬ ExThere	3816	505
ExThere	269	881

TABLE 11.2: *Summary of observed and predicted existential* there *in ME, based on the CART tree in figure 11.3. Precision is 0.77 and recall is 0.64, which gives an F-score of 0.69.*

	Observed	Predicted	Correctly pred.
ExThere	1386	1150	881
¬ ExThere	4085	4321	3816

there in PPME2, i.e. $881/1386$ which is 0.64. In this case, both precision and recall are quite good. Table 11.2 provides a summary of this information by showing the number of correctly predicted occurrences out of all predicted occurrences compared with the observed frequencies from PPME2. Precision and recall can be further combined into a single measure of performance, the so-called *F-score*:

$$F\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \tag{11.1}$$

Applying the formula above results in an F-score of 0.69, which gives a balanced summary of the model's completeness and reliability.

It should furthermore be emphasized that the bias in the model is conservative with respect to the hypotheses under investigation: 505 (9%) instances of existential *there* are wrongly classified as ¬ existential *there*, whereas 269 (5%) cases of ¬ existential *there* are wrongly classified as being existential. In summary, the CART tree in figure 11.3 is deemed as an acceptable tool for estimating the distribution of existential *þær* in Old English.

Using the `predict()` function from *rpart*, the model can then be used to predict occurrences of existential *þær* in Old English given the input variables from YCOE.

The model predicts a total of 483 occurrences of existential *þær* in YCOE. Based on the evaluation of the model, this is probably too low. Given the error rate and the same distribution between over- and underestimation as in the Middle English data, it is possible to estimate an upper limit for existential *there* in YCOE. The model mistakenly classified 505 tokens with existential *there* as locative for the PPME2 data, i.e. $505/1386 = 0.36$. If we assume the exact same misclassification rate for the YCOE data, around 277 cases of existential *there* should have been left out, bringing the total to 760 tokens.[2]

The potential magnitude of the effects of any misclassifications can be judged by looking at the frequencies as proportions. Table 11.3 compares the observed proportion of existential *there* in PPME2 and the estimated minimum (i.e. 483 occurrences) and maximum (i.e. 760 occurrences) proportions of existential *there* from YCOE. The table gives proportions of all tokens with *there* and of the whole dataset. As the table shows, a noticeable increase has in proportions have taken place from YCOE to PPME2. It seems that a substantial change in proportions has taken place, irrespective of whether the higher or the more conservative estimate of existential *there* in YCOE is chosen.

TABLE 11.3: *Proportions of existential* there *in ME, and estimated maximum and minimum proportions of existential* there *in OE. Proportions are shown out of all tokens with* there*, and all tokens in the respective datasets.*

	Proportion ExThere	
	All there	Whole dataset
ME	0.42	0.25
OE $\text{est}_{min.}$	0.09	0.05
OE $\text{est}_{max.}$	0.14	0.08

Since the evaluations on the Middle English data showed that the model tends to underestimate the number of existential *there*, it seems reasonable to have more faith in a higher than in a lower number of occurrences. Thus, the number 483 is taken as a conservative estimate and it seems likely that the real number is higher. Since the subsequent analysis is based on the more conservative estimate, there is less risk of false positive results ("type I errors"). At the same time, the difference in proportions from YCOE to PPME2 is substantial no matter which of the two estimates is chosen. With

[2]Calculated as follows: $277/(483 + 277) = 0.36$.

robust models, it should be possible to get an accurate picture of the main tendencies based on the more conservative estimate.

11.2.2 Interim Summary

The present section has shown that by using CART trees, an attempt could be made at estimating the distribution of existential *there* in Old English. Starting with the larger model discussed in chapter 9 suspected of overfitting the data, it was shown that a smaller model provided the best results. Although Everitt and Hothorn (2006, 142) warn that CART models are simple, and can never be more than rough approximations, they can nevertheless be a useful tool for estimating the occurrences of existential *there*. The CART model above is a very good improvement over a baseline model (which only predicts locative uses of *there*), with an error rate of 14% and an F-score of 0.69. The estimates are conservative in the sense that more instances of existential *there* are erroneously classified as locative than vice versa. The resulting data will in the subsequent sections be used as a reasonable approximation to the lower bound of occurrences of existential *there* in Old English.

11.3 A diachronic picture

With reasonable estimates for the number of occurrences of existential *þær* in Old English, a diachronic picture starts to emerge. By adding the CART estimates to the data discussed in chapter 8, the estimates can be used to trace the evolution of *there* throughout the three corpora.

If we consider the proportions of all locative adverbs to the size of the corpora in Early English, displayed in figure 11.4, we find that the proportion is more or less constant, but with a small receding trend from the late Middle English period. Turning to locative uses of *there*, in figure 11.5 on page 261, it is clear that the proportion of locative *there* is falling throughout the period. Finally, the proportion of existential uses of *there*, as seen in figure 11.6 on page 262, increases sharply from the late Middle English period.

In other words, it does not seem reasonable to ascribe the observed differences in proportions between locative and existential *there* to some *general* trend pertaining to all locative adverbs. Rather, it seems more likely that this is a phenomenon which is isolated to *there*.

For all three plots, a *lowess* (locally weighted scatter plot smoothing) curve has been fitted, to better highlight the trend in the plot (in this case using the `lowess()`

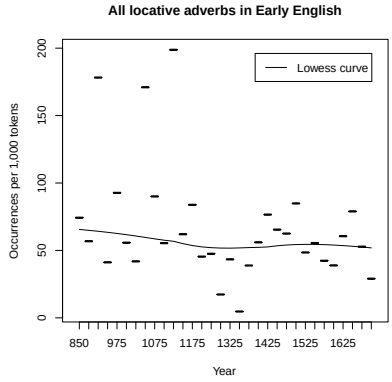


FIGURE 11.4: *Proportions of locative adverbs in Early English, with added smoothing lowess curve. The scale is number of occurrences per 1 000 corpus tokens.*

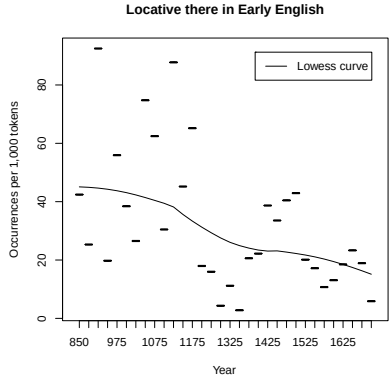


FIGURE 11.5: *Proportions of locative* there *in Early English, with added smoothing lowess curve. The scale is number of occurrences per 1 000 corpus tokens.*

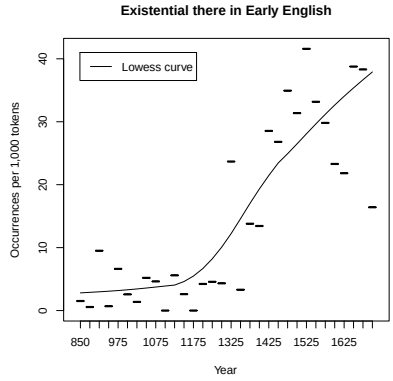


FIGURE 11.6: *Proportions of existential* there *in Early English, with added smoothing lowess curve. The scale is number of occurrences per 1 000 corpus tokens.*

function in R). This is a form of nonparametric regression which fits local polynomial models, an approach which is robust to outliers, cf. Faraway (2006, 221).

Figure 11.7 on page 263 shows all cases of *there* in Early English, divided up between existential and locative *there*. As in the other plots, the scale on the y-axis is the normalized number of occurrences per 1 000 tokens, in order to avoid fractions of occurrences.

As the figure shows, there appears to be a clear cross-over effect, where locative *there* is initially the most frequent type of *there*, with a minority of existential uses. In Middle English, there is less data, but it looks like the start of the Middle English period has more in common with Old English, whereas in late Middle English the gap between the two categories is diminishing. A major transition takes place in Early Modern English, where existential *there* becomes more frequent than locative *there*. Interestingly, this takes place *after* the transition from the PPME2 to the PPEME corpus. Keeping in mind that the years in the graph represent 25-year intervals, it is worth noting that "1500" (representing the first quarter of the 16$^{\text{th}}$ century) is an overlap between the two corpora. However, in both corpora the frequencies and proportions are very similar for the two categories of *there* in this period. The cross-over effect is found in "1525", that is the second quarter of the 16$^{\text{th}}$ century or 1525–1550, when existential *there* stabilizes at higher frequencies than locative *there* for the first time.

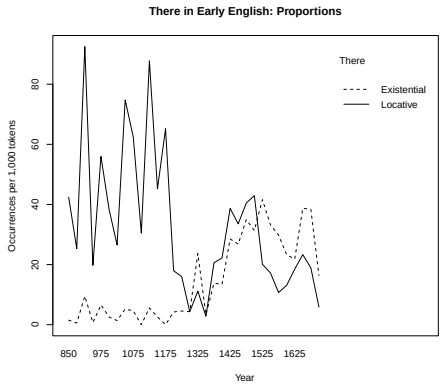


FIGURE 11.7: *Proportions of all occurrences of* there *in Early English, plotting existential uses vs. locative uses. The scale is number of occurrences per 1 000 corpus tokens.*

To put this into perspective, it is worthwhile to briefly consider the situation in Present-day English. The British part of the *International Corpus of English* (ICE-GB) is a suitable Present-day English comparison for the Early English corpora. Like the historical corpora it is a syntactically annotated phrase structure grammar corpus. It also has the added benefit of being of a comparable size: with 1 061 263 words and 83 394 parse trees it is about the same size as the PPME2, and only a little smaller than the other two historical treebanks.

Searching ICE-GB for *there* and existential *there* gives suggestive results: there are 4 457 syntactic trees containing the word *there*, 3 286 of which are annotated as existential *there*. In other words, 74% of all instances of *there* in ICE-GB are of the existential type. This can be compared with the estimated 9.0% existential *there* in YCOE, 41.8% in PPME2, and 63.6% existential uses of *there* in PPEME.

Thus, the data from ICE-GB point to a situation where the Early Modern English shift in the proportions between locative and existential *there* has stabilized. This raises a further question, namely how big is the actual change in proportions from the estimated situation in YCOE with a very small proportion of existential *þær* to a much higher proportion in PPEME. Or to put it differently, since PPEME seems to be close to the proportions in ICE-GB, does the observed shift from YCOE to PPEME constitute a small or a large change?

To measure the degree of diachronic similarity across the three corpora, we can again turn to the DP for a simple summary. The DP for proportions of locative and

existential *there* in all the 25-year intervals of the three corpora combined is 0.37. This number reflects what appears to be a gradual change in proportions between locative and existential uses of *there*. There are clearly differences in the material, but the difference is moderate, on a scale from 0 to 1. This points to a situation with only small changes from each corpus to the next, i.e. a gradual change. For a more detailed picture of the diachronic development, GLMMs are useful, as shown in the following sections.

11.4 Tying up loose ends

Two variables that have not been discussed much previously are dealt with below, namely genre and translation status. Although they might appear to be excellent candidates for predicting semantic and syntactic phenomena, they did not prove to be valuable predictors during the model fitting process in the previous chapters. The sections below show that, for the whole Early English period, diachronic factors take precedence over genre, and that translation status does not affect the use of existential *there* in the three diachronic corpora.

11.4.1 Genre

None of the models presented so far have dealt with genre, which may seem surprising. In fact, I attempted to include genre at many stages of the model fitting, but due to the number of levels (from 28 to 36 depending on the corpus), this proved too computationally demanding for such large datasets with other predictors and random effects. However, the genre composition of the three treebanks is not identical, and the question must be asked whether the differences regarding existential *there* is not so much about time periods as about genres. To test this, the following assumptions were made: if genre is a good predictor for existential *there*, then it should be able to do better in predicting the response (existential *there*) than the mean (i.e. the intercept) of a mixed effects model, with a random effect for year. The models in (4) and (5) below were fitted for the full Early English dataset of 23 761 observations, using `lmer()`.

The baseline model uses only a random effect for the `Year`-variable (i.e. the intercept):

(4) `ExThere ~ (1 | Year), family = binomial`

This can be compared with the full model which includes `Genre`:

(5) `ExThere ~ Genre + (1 | Year), family = binomial`

Figures 11.8 and 11.9 show residual vs. fitted plots for the two models. As the plots show, there is more unexplained variation in the plot for (5), but this model predicts existential *there* better, whereas the plot for (4) shows that this model does not not go beyond a probability for existential *there* of 0.4 to 0.5.

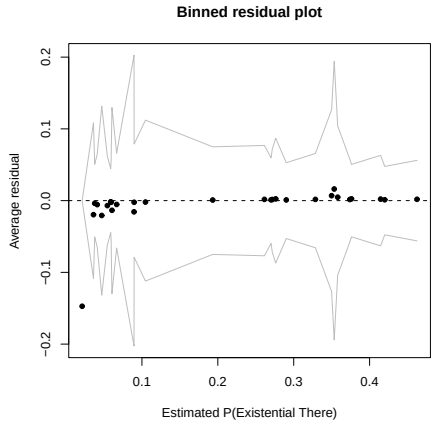


FIGURE 11.8: *Binned residuals vs. fitted plot for the model in (4). The x-axis shows the estimated probability of existential* there.

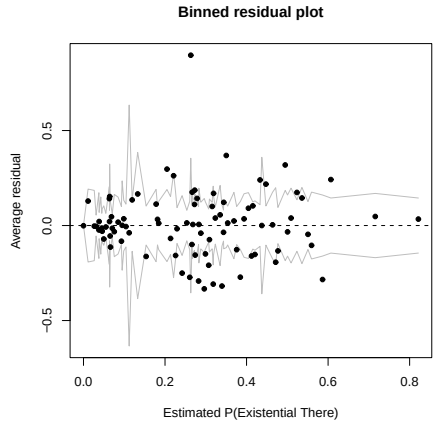


FIGURE 11.9: *Binned residuals vs. fitted plot for the model in (5) which includes genre. The x-axis shows the estimated probability of existential* there.

How much of an improvement does this amount to? A log-likelihood test of the two models shows that the model with genres in (5) is a significant improvement over the model in (4) with only a random effect for the intercept ($\chi^2_{(34)} = 726.43$, $p < 0.01$). However, the size of the overall improvement effect as measured with the Nagelkerke R^2 is only 0.05, i.e. a very small effect for the improvement of the model. A look at the size of the coefficients from the full model in (5) shows that the effects are small. Table 11.4 shows the coefficients that were significant at the 5% level (see appendix for the full output).

The standard error for the intercept is 1.27 on the logit scale. Dividing this number by 4 gives the estimated upper bound for the predictive difference of a one-unit change in `Year`, near the midpoint of the logistic curve, cf. Gelman and Hill (2007, 82). Dividing 1.27/4 gives 0.32. This indicates that the 25-year intervals differ by approximately $\pm$32% on the probability scale with respect to the estimated mean probability of existential *there*, which corresponds to an expected variation of the probability of

TABLE 11.4: *The probabilities (Pr(ExThere)) give estimated mean probability of existential* there *for the reference level (*`Genre:Apocrypha`*), and for the genres shown. Only predictors with* $p < 0.05$ *are included. The standard error for the* `Year`*-variable is 1.27 on the logit scale.*

Fixed effects	Est. coef.	Std. Error	z value	Pr(>|z|)	Pr(ExThere)
(Intercept)	-1.62	0.42	-3.89	1.0 e-4	0.17
GenreDrama_comedy	-0.76	0.37	-2.07	0.04	0.09
GenreFiction	-0.80	0.36	-2.21	0.03	0.08
GenreGeography	1.57	0.59	2.66	0.01	0.49
GenreLaw	-1.28	0.39	-3.28	1.0 e-3	0.05
GenrePhilosophy	1.61	0.37	4.30	1.0 e-3	0.50
GenreProcTrial	-0.81	0.36	-2.23	0.03	0.08
GenreScience	1.36	0.51	2.64	0.01	0.44

existential *there* anywhere in the range of 2% to 72% for the time periods.

Looking at the coefficients' estimated probabilities of existential *there*, it does indeed appear that some genres are more associated with existential *there* than others, namely `Geography`, `Science` and `Philosophy`. `Geography` and `Science` are found in Old English only,[3] whereas `Philosophy` is found in all three periods. If the upper and lower bounds are calculated for these estimates, however, they turn out more or less like expected; `Geography`: [0.23, 0.76], `Philosophy`: [0.32, 0.67], `Science`: [0.22, 0.68]. It is also evident that some genres which are uniquely associated with Early Modern English, such as `Drama_comedy` and `Fiction`, have low estimates for existential *there*. Thus, even among the few genres that have estimated mean values which are significant at the 5% level there is no clear correlation between existential *there* and time period (`Year25`).

The impression is further supported by a multiple correspondence analysis of time periods (coded as `Era`) and existential *there* vs. genre. The interaction between genre, existential *there* and corpus constitutes a three-way interaction. This could in principle be modeled using classical CA with an interactive coding. In such an approach every occurrence of locative or existential *there* would be tagged as belonging to an era (OE, ME, EME), resulting in a table with six rows and as many columns as there are genres, with frequencies for each cell. An alternative to this is using *Multiple Cor-*

[3]There are of course scientific texts in the later material, however, the genre coding varies somewhat and `Science` appears only in Old English.

respondence Analysis (MCA), which captures the variation within variables, not just between them. Since MCA is an extension of basic CA, no further detailed introduction will be given here, see Greenacre (2007, 137–144) for a further presentation and discussion. Some basic points, however, deserve mentioning: rather than working with frequencies directly (as in CA), MCA relies on a fully disjunctive binary matrix, where all the features to be analyzed are coded as present or missing (1 and 0) for each case in the dataset. This binary coding leads to much lower percentages of total explained inertia than with classical CA, something which must be kept in mind when interpreting the resulting plots (Greenacre, 2007, 144). Furthermore, there is no real distinction between rows and columns, rather, there are only columns. For this reason, the inertia does not have the same association-interpretation as in classical CA, although the total percentage of inertia accounted for in the map still gives an indication of the quality of representation. Figure 11.10 on page 267 shows an MCA map of the analysis.

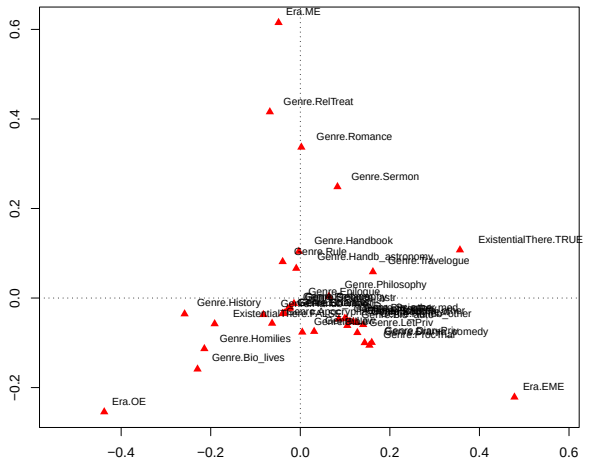


FIGURE 11.10: *Asymmetric MCA map of the interactions between existential* there*, time periods, and genres (scaling option:* `map="rowgreen"`*). Percentage inertia in map: 83.4%.*

The congestion of points near the origin of the plot is hard to read, but this is inconsequential. What is important is that the points near the origin are all genres, meaning that they all behave as expected. Note how the x-axis is defined by a continuum (from

left to right) of Old English and non-existential *there* (`ExistentialThere.FALSE`) to Early Modern English and existential *there* (`ExistentialThere.TRUE`). Middle English draws up the second axis, and is nicely situated more or less halfway between the other two periods. A look at the contributions to inertia reveals that the three time periods contribute most of the inertia, followed by `ExistentialThere.TRUE`, while `ExistentialThere.FALSE` and the genres contribute very little (see appendix A.4). Thus, the map in the MCA plot supports and illustrates the results from the logistic GLMM: there is a diachronic development taking place, and genre has little or nothing to do with it.

11.4.2 Translation

Translation is another factor which has not been included in the models so far. As with genre, I attempted to include translation in the early modeling stages (see also initial comments on translated material in Old English in section 8.2.1). However, this variable was always removed from the model during the model fitting process, since it never contributed to a significantly better fit. This lack of association between existential *there* and translated/original material can be illustrated using a Pearson chi-square test.

TABLE 11.5: *Occurrences of existential* there *for translated and non-translated texts in early English. The overall association between rows and columns in the table is negligible.*

	¬ Translated	Translated
¬ ExistentialThere	10986	4870
ExistentialThere	4017	1175

As table 11.5 shows, there are more examples of existential *there* in non translated than in translated text, which is hardly surprising since for the whole early English period the majority of corpus tokens are *not* translated. A Pearson chi-square test of independence shows that rows and columns are dependent ($\chi^2_{df(1)} = 124.43$, $p < 0.01$, $\phi = 0.08$). This should not come as a surprise, given that the test result of the Pearson chi-square is known to be highly sensitive to sample size and that there are almost 24 000 observations in the table. However, the association between rows and columns as measured with the ϕ coefficient shows that when we take sample size into account, there is virtually no association between the use and non-use of existential *there* for translated and non translated texts.

This result supports the impression from the model fitting phase for the regression models, namely that the there is no real difference in the use of existential *there* for translated and non-translated texts in early English.

11.5 A full diachronic model

The use of existential *there* can be described using the logistic model discussed previously. As before, the question is one of finding the best model which balances structural correctness and predictive capability against overfitting.

(6) `ExThere ~ LogComplexity + BeContext + (1|Year25), family = binomial`

(7) `ExThere ~ LogComplexity + BeContext + NomNP + (1|Year25), family = binomial`

(8) `ExThere ~ LogComplexity + BeContext + NomNP + SemClass + (1|Year25), family = binomial`

Consider the four logistic GLMMs above, each of which models the probability of existential *there* as a function of a number of predictors, with a random effect for the diachronic variable. A residual vs. fitted plot for model (6) is seen in the upper left corner of figure 11.11. Clearly, this model runs into problems. This is perhaps surprising since all the evidence for the individual corpora points to this being the simplest and best model. However, if the *input variables* are very different for the three corpora, then it seems natural that the model would run into problems, even if the structural form worked well for each of the three corpora in isolation. This suspicion is confirmed if we turn to the upper right panel of figure 11.11. Here, the same model has been fitted, but only to the data from PPME2 and PPEME, i.e. only the Middle and Early Modern English data. There are still problems, but the improvement in fit is considerable. This is not entirely unexpected, since these two corpora are more similar to each other than to YCOE with respect to existential *there*. Expanding the model improves the fit when modeling all three corpora, as shown in the two lower panels of figure 11.11. The right plot shows model (7), where information about the presence of a nominative/subject NP has been included. The outliers are now much closer to the confidence intervals. The lower left panel shows a similar plot for model (8), where the semantic class of the verb has been included, and very few points now fall outside the confidence intervals.

Clearly, to model *there* in the entire Early English period, we need more information than just co-occurrence with *be* and the log complexity ratio to obtain an acceptable fit. Despite the apparent superiority of the larger model in (8), it is in fact very close to

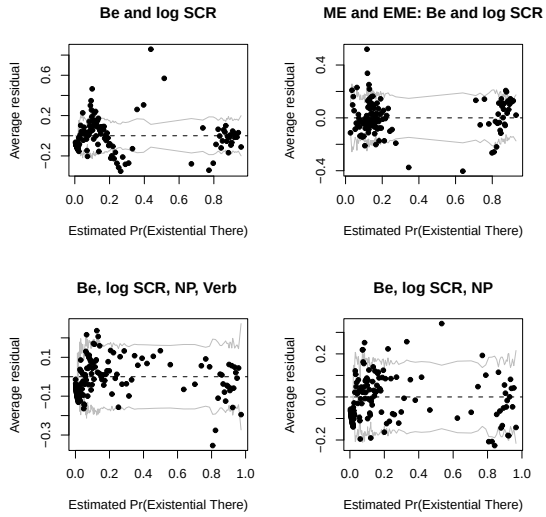


FIGURE 11.11: *Residuals vs. fitted plots for (clockwise from top left) the models in (6) run with all data; the model in (6) run with data from* PPME2 *and* PPEME *only; the model in (7) with all data; and the model in (8) with all data.*

the smaller model in (7). The larger model has a Nagelkerke R^2 of 0.54, whereas the corresponding value for the smaller model is 0.51. If we turn to calibration, that is, the models' predictive accuracy, the C index is 67.3 for the larger model and 67.5 for the smaller one. In other words, the smaller model does a slightly better job at predicting the correct response, as confirmed by the D_{xy} index: 34.6 for the larger model, and 35.1 for the smaller model without information about the semantic class of the verb. Thus, what the larger model in (8) gains in coverage over the smaller model in (7), it looses in predictive accuracy. The parsimonious choice is then to go for the smaller model.

However, as the C and D_{xy} indices showed, the predictive accuracy of the model is still not good. I will nevertheless use it to inspect the diachronic change in the mean probability of existential *there* in the entire Early English period, since it is still the best model discussed thus far.

Turning to the specific effects, shown in table 11.6, it is not surprising to see that co-occurrence with *be* is the strongest (and only substantial) effect in the model.[4] That is, co-occurrence with *be* is by far the best indicator of existential *there* throughout the Early English period.

TABLE 11.6: *Fixed effects for the model in (7). Note the negligible effects for all predictors save co-occurrence with* be *(`BeContext`).*

Fixed effects	Est. coef.	Std. Error	Pr(ExThere)
(Intercept)	-3.88	0.18	0.02
LogComplexity	0.39	0.02	0.03
BeContextTRUE	3.49	0.06	0.40
NomNPTRUE	0.88	0.05	0.05

Having seen the effects, it is interesting to consider the error margins. This can be achieved by looking at confidence intervals for the coefficients. Table 11.7 gives confidence interval for the fixed effects, following a procedure from Baayen (2008a, 283), based on 100 simulation runs. As the numbers show, the model gives reasonably small confidence intervals. `LogComplexity` and `NomNP` are consistently small effects, whereas `BeContext` gives the largest effect.

TABLE 11.7: *Bootstrap confidence interval for the model in (7). The values are log odds ratios. Note the small confidence interval for `BeContext`, for which the greatest effect is observed.*

Fixed effects	2.5%	50%	97.5%
Intercept	-4.32	-3.87	-3.38
LogComplexity	0.28	0.38	0.56
BeContextTRUE	3.19	3.46	3.88
NomNPTRUE	0.71	0.90	1.02

[4]The full output from the model can be found in appendix B.3.1.

11.5.1 Interpretation

Figure 11.13 on p. 273 shows the estimated mean probability of existential *there* by 25-year interval for all three periods (keep in mind that the data for Old English are based on back-estimates from Middle English). The estimates are obtained through the random intercepts for `Year25` from the model in (7). These effects are the so-called *Best Linear Unbiased Predictors*, or BLUPs, cf. Baayen (2008a, 247). So far the BLUPs have not been discussed in detail, and the assumptions they work under have not been explicitly mentioned. In chapter 4 it was explained how the random effects are not parameters of the model, but considered random samples from a normally distributed population of effects. Figure 11.12 shows a normal quantile-quantile plot of the random effects to check whether the assumption of normality holds, cf. the discussion in chapter 4. There is evidence of some problems in the tails, and the distribution seems closer to a Cauchy than a normal distribution. However, this is a mild form of non-normality, and with a large number of observations it becomes less important, cf. Faraway (2005, 59–60). Given the normality assumption, it is possible to predict the values of the BLUPs for the data, cf. Pinheiro and Bates (2000, 95). This is done through Bayesian statistics (cf. chapter 4), and allows for estimates that combine the information from the responses and the information about the whole model, cf. Everitt and Hothorn (2006, 168–169), increasing the precision of the model (Baayen, 2008a, 273–274).

As figure 11.13 shows, we get an S-shaped logistic curve (cf. the example in chapter 4), going from a mean probability of existential *there* below 0.5 in Old English, through a sharp increase during the Middle English period, to a higher stable probability closer to 1 for the Early Modern Period in PPEME. Recall that the plot in figure 11.13 shows the mean estimated predictions for the diachronic variable *above what is explained by the fixed effects*. The Nagelkerke values reported above refer to the improvement of the full model (i.e. a model with random effects for `Year25` and the fixed effects of the linguistic variables) over a model using only the BLUPs in figure 11.13.

Why is this important? Figure 11.13 shows a clear shift in mean probabilities over the Early English period, following a classic S-curve. This pattern is widely attested in language change, cf. Pintzuk (2003, 512–513), and has been "established as a kind of template for change" (Chambers, 2002, 361). The S-curve is not found exclusively in linguistics. An early overview is found in **?**, who mentions both sociological and biological applications. **?** stresses the importance of the theoretical motivations for the curve (which seems to correspond to Goldthorpe's generative processes) over the goodness of fit itself. **?**, 53–55 discuss the S-curve in the context of historical sociolinguistics and point to research which suggests that such a curve might well consist of small, overlapping S-curves. The curve itself receives a sociolinguistics interpreta-

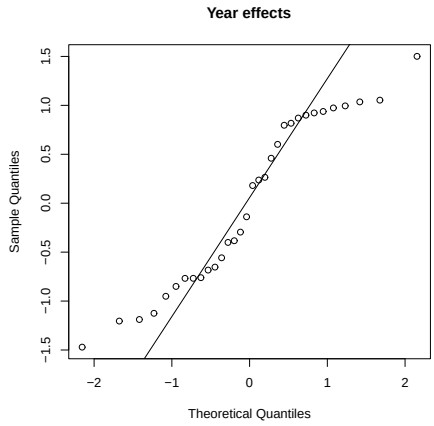


FIGURE 11.12: *Normal Q-Q plot of the random effects from model (7). Closeness to the solid line indicates a good fit to the normal distribution. The random effects follow a short-tailed distribution, but given the large number of observations this mild deviation from normality can safely be ignored.*

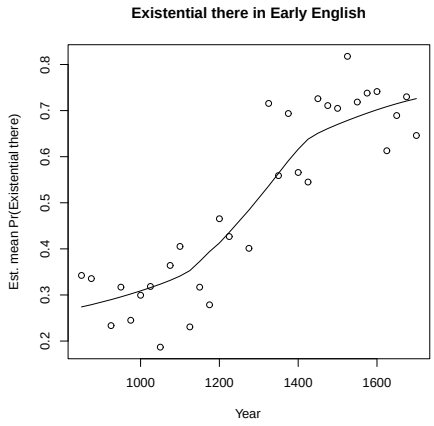


FIGURE 11.13: *Estimated mean probabilities by 25-year intervals for Early English from model (7), with a fitted lowess nonparametric regression line. Note the sharp rise during the Middle English period. The maximum estimated difference between 25-year intervals around the middle of the curve is* $\pm 22.75\%$.

tion in the form of social exposition and accommodation (**?**, 53–55). However, even breaking the curve down into an infinite recursion of smaller S-curves leaves open the question of whether social factors is the only process which can generate such a curve in linguistics. Given its prevalence in other, unrelated fields, this seems unlikely. Thus, it is difficult to see how the presence of such a curve in itself is an indication of the sociolinguistic or functional nature of an ongoing change.

It is perfectly reasonable to see the S-curve as the product of other, underlying generative processes which might be social, cognitive, or linguistic nature. One possible explanation might be the general increase in initial target words discussed in section 9.3 above, where it was shown that an increase in the mean probability of initial target words took place towards the end of PPME2. Another explanation might be the availability of more material. The latter explanation is made less plausible by the low tail in the Old English period: YCOE contains more material than PPME2, and if frequency were the most important factor we would expect to see a higher probability in

the estimates for Old English followed by a dip in the probabilities for the early Middle English period. The value of the S-curve is consequently its usefulness in describing rather than explaining a change, by quantifying the diachronic shifts in mean probabilities. The curve has the same shape as the graph for existential *there* shown in figure 1 of Breivik (1990, 226). However, while it confirms the conclusion in Breivik (1990, 227) that not much development takes place after 1550, the curve is nevertheless transposed to a later point in time. While Breivik's curve has a sharp rise between 1070 and 1225, the largest effect for my data are observed between 1200 and 1400.

Of course, frequency of use might influence the mean probability of using a construction, but just like we cannot assume full homogeneity in the data, we cannot expect such a pattern to account for the full variation in a case of syntactic change. Thus, it is of interest to establish to what degree this diachronic pattern can account for the change compared with the variation in the linguistic input variables.

In chapters 8, 9, and 10 it was shown that the three corpora are relatively homogenous. For YCOE, this was demonstrated both through a LNRE model and the DP measure. DP was also used to measure this for PPME2 and PPEME. Additionally, it was shown that, with regard to existential *there*, the mean difference between the 25-year intervals was small within all three corpora. That is, the input variables regarding *there* are comparable for each corpus, but there are some differences across corpora.

(9)
```
ExThere ~ LogComplexity + BeContext
+ NomNP + (1|Year25) + (1|Era), family = binomial
```

The model in (9) uses an extra intercept to add "Era" (i.e. the three corpora) as random effects for the model. The model is a reasonable fit and not a worse fit than the one in (7). Table 11.8 gives the fixed effects of the model, showing that co-occurrence with *be* is the single best predictor of existential *there*, the other predictors only make small contributions. Switching to the random effects, the standard deviation of the `Year25` term is 0.58 whereas it is 0.65 for the `Era` term. Following the by now familiar divide-by-four rule, we find the predicted maximum difference between categories on the probability scale near the center of the curve: $0.58/4 = 0.145$ or $\pm$14.5%, and $0.65/4 = 0.1625$ or around $\pm$16.3%.

The maximum predicted differences between existential and locative uses of *there* are in other words around $\pm$14.5% for the 25-year intervals and $\pm$16.3% for the three corpora. The corresponding value for the fixed effect of co-occurring with *be* is $3.50/4 = 0.875$ or $\pm$87.5%. In other words, the effect of switching from co-occurrence with *be* to non-co-occurrence with *be* corresponds to a maximum change in the probability of existential *there* of $\pm$87.5%, between 5 and 6 times larger than the maximum effect for switching between 25-year intervals or corpora. In other words, while there is

TABLE 11.8: *Fixed effects for the logistic GLMM in (9). Note the large effect of co-occurrence with* be.

Fixed effects	Est. coef.	Std. Error	Pr(ExThere)
(Intercept)	-3.82	0.40	0.02
LogComplexity	0.40	0.02	0.03
BeContextTRUE	3.50	0.06	0.42
NomNPTRUE	0.87	0.05	0.05

undoubtedly a diachronic change taking place, the most dramatic impact on the probability of existential *there* under the current model is not caused by the corpora, but by co-occurrence (or not) with *be*.

11.6 A cognitive explanation for the grammaticalization of *there*

The present section will take as its departure the attested association patterns in the corpora and attempt to account for them in terms of a generative process, to use the term from Goldthorpe (2001). By necessity, this section will be somewhat speculative in nature since it attempts to pull together corpus evidence with conclusions drawn from studies on language processing and cognition. However, I have attempted to make the line of reasoning coherent and specific, so that the proposals made here can at least in principle be testable.

It might seem superfluous to posit a cognitive explanation in the first place. Above, the hypothesis that a general increase in the probability of initial target words took place during the late Middle English period, corresponding to the increase in existential *there*. It is possible, then, that *there* was simply reanalyzed through some kind of analogy-based process. This would be very close to the argumentation in Breivik (1990), and the general scenario certainly seems plausible. Similar processes have been suggested for other areas of diachronic syntax by Fischer (2007a) and Barðdal (2008). However, just like with a grammaticalization-based explanation, the analogical reanalysis of *there* is something which itself is need of an explanation. It does not in itself answer the crucial questions *why* should *there* be reanalyzed, and *who reanalyzed what, how*? In brief, a generative process is required to connect the individual

instances with the broad (population) patterns described by terms such as grammaticalization and analogy. Thus, a cognitive, explanation involving language processing in individuals is still warranted, and it is offered here more as an addition to, rather than a replacement of, the more general descriptions.

There is arguably a high degree of semantic coherence between a locative deictic adverb and the verb *be*: for something to exist, it must exist somewhere in space and time, even if that existence is purely conjectural and restricted to a mental space of some sort.

To explain how this could lead to the situation described previously, two crucial assumptions will be made, based on Deacon (1997):

(i) Breakdown of referential competence leads to an orderly descent from symbolic to indexical to iconic [signs] (Deacon, 1997, 74).

(ii) Symbolic reference derives from *combinatorial* possibilities and impossibilies (Deacon, 1997, 83).

Thus, two opposing interpretations would be hypothesized to compete in early English: locative *there* and existential *there*. In Old English it was found that conditional probability provided a point of departure for this process, where *there* in clause-initial position was highly correlated with *be*. This should not in itself be taken as an immediate breakdown of symbolic reference, and many of those instances are no doubt very likely instances of locative *there*.

The crucial point is rather that at first an association with a specific context is established. For the symbol (locative *there*) to be re-interpreted as an indexical sign it is also necessary for the combinatorial relationship between *there* and *here* to break down. Or to put it differently, when *there* started to assume the role of subject in ECs, this was a first step towards a possible indexical interpretation. But as chapter 2 showed, dialects in the Scandinavian languages may use either *here* or *there* as existential subject. The role of existential subject is thus not sufficient for an indexical re-interpretation to take place. The existential subject role must be considered inappropriate for *here* before *there* could be interpreted as an indexical sign in the sense of Deacon (1997).

As it turns out, we need to keep the possibility open for an existential *here* in Old English. However, for Middle and Early Modern English this seems increasingly unlikely. Evidence from the regression model in (7) suggests that in the corpus material, a crucial shift took place in the transition to the Early Modern English period. As shown above, this cannot be attributed to differences in genre composition of the three corpora.

To properly evaluate the situation that might have led to the evolution of an existential *there*, it is necessary to keep in mind that evolutionary processes deal with

proximate, historical causes, not grand principles. To quote Lieberman (2000, 166): "Evolution does not give a damn about formal elegance". Taking Croft's notion of profile equivalence (2001, 257), discussed in section 2.6 above, we can predict that the symbolic interpretation of *there* should start to break down when *there* and *here* could no longer be interpreted as alternatives in the same syntactic position. Based on the suggestions from Deacon, this property is strictly distributional. That is, the symbolic interpretation of *there* rests on its distributional relationship with *here*. As shown above in sections 8.5 and 9.2.2, some indication was found supporting the view that a pseudo-existential *here* might have existed in Old and Middle English, with some association with *be*. However, it seems clear that already in Old and Middle English such an existential use of *here* must be considered marginal compared with *there*. Thus, the "reanalysis" of *there* as an indexical sign should properly be considered a description of the repeated interpretation of signs in context, cf. Deacon (1997, 75–76). The symbolic associations between deictic opposites like *here* and *there* would contribute to the mutual support the symbolic interpretation. Similarly, an increasing association with *be* and the subject position of the EC would work in the other direction, namely splitting up *there* and *here* and leading to an indexical interpretation of the former, but not of the latter as long as *there* continued to serve as a locative adverb in other syntactic positions. The next section will discuss some possible cognitive mechanisms to account for such a development.

11.6.1 Linguistics in Smallville

In the present section a crucial argument will be advanced regarding the cognitive mechanisms that could plausibly be behind the status change of *there* in Early English.

Chapters 8, 9, and 10 have presented strong evidence of two crucial facts for the three corpora in question. First, *there* is the most frequent locative adverb; second, *there* is highly correlated with *be*. However, in accordance with basic assumptions in cognitive linguistics, language is taken to be a cognitive symbolic system. This implies that somehow the *individual*'s cognition must be taken into account. Although the frequencies and correlations from the corpora are important as evidence, they do not readily translate into individual cognitive states and processes. Put differently, the facts assembled here on the basis of, say, YCOE were not directly available to individual speakers of Old English in their current form. How then, could speakers of various forms of Early English plausibly have made quantitative and probabilistic inferences about frequencies they could not have known? To answer this question a shift in focus is required, from large quantities to small ones.

It has been shown that with specific evidence about probabilities, people tend to put too much weight on small probabilities, see e.g. Tversky and Kahneman (1971).

However, when we are asked to make decisions based on experience, the effect is opposite; small probabilities tend to be ignored. The net effect of this is that small probabilities are felt to be observed a lot less than expected. As Hertwig and Pleskac (2008, 215–216) explain, with small samples effect sizes for the most likely outcome are large, because of missing data:

> An organism remembering only a small number of recent events—tantamount to drawing small samples from memory rather than from the environment—is better equipped to detect a change in its environment than it would be if it remembered all of its history.
>
> Hertwig and Pleskac (2008, 229)

There is some research suggesting that working memory may have developed in a way which increases the chances of early detection of covariation (Hertwig and Pleskac, 2008, 229). The attractiveness of such adaptation-based arguments notwithstanding there is a long tradition for criticizing appeals to adaptation as a motivation in research. One of the most well known examples is perhaps Gould and Lewontin (1979). No assumptions will be made about optimality in the present study. In other words: whether such an attention strategy is optimal for some purpose or not is not considered relevant here. The relevant issue is that the effect can be documented, not whether it is optimal.

Hertwig and Pleskac (2008, 229–230) point out that small samples are not better at representing realities than large samples, they are not. Their point is that small samples lead to *experienced* differences in sample means that are larger than the objective differences.

The frequency of *there* revisited

An important prerequisite for a distinction to arise is thus that *there* is very frequent in a specific environment (as has been shown), and that this environment is a low frequency phenomenon.

This immediately begs the question of how many words we really are exposed to on a daily basis. Rayson, Leech, and Hodges (1997) report frequencies for men and women from the demographic part of the spoken component of the British National Corpus. These frequencies are the total number of words produced, both by the speakers themselves and the people they interact with, so they should give a reasonable starting point. According to Rayson et al. (1997, 136), there were a total of 2 593 452 words recorded for the female participants and 1 714 443 for the male participants. This

gives a total of 4 307 895 words uttered and perceived by 148 speakers, over a period of two or three days. As Mark Liberman comments in his *Language Log*, the estimate seems rather on the low side,[5] so it seems safe to make the simplifying assumption that the recording period was two days for all the participants. If we divide the total of 4 307 895 words by the 148 speakers and again divide by two, we get an estimate of the number of words the participants were exposed to per day:

$$\frac{4307895}{148} \times \frac{1}{2}\ days = 14553.7 \tag{11.2}$$

Taking this as an acceptable estimate, the next step is to compute the co-occurrence rates for *þær* and *beon* in YCOE. As reported in section 8.6, there are 846 tokens with *þær* followed by *beon* in the corpus, or 0.06% of the corpus. Another way of calculating this would be to take the co-occurrence frequencies for the 25-year intervals and dividing by the number of words per interval, and take the median proportion as our departure. The latter approach should be less influenced by large differences in the number of words per 25-year interval in the corpus. Normalizing this median gives an estimate of 3 co-occurrences of *þær* and *beon* per 1 000 words in YCOE, i.e. half as many as with the total frequency divided by the total number of words (which equals 6 per 1 000 words).

If we instead use the notional number of daily words as the normalizing factor, we get respectively 8.5 for the total and 4.5 for the median as the estimated daily co-occurrence rates of *þær* and *beon*. Thus, assuming that the numbers from YCOE and the BNC are representative of individual exposure to language (which might be a big assumption to make), and furthermore that the number of words we are exposed to through *personal* interaction is more or less on the same scale today as in the Middle Ages, we can conservatively estimate the notional co-occurrence rate of *þær* and *beon* to about five to ten per day. This raises the question of whether this is much or not.

For comparison, I searched YCOE (using Perl) for the number of co-occurrences of determiners immediately followed by common nouns, which seems like a good candidate for a high-frequency phenomenon. This resulted in 80 529 hits, or 5.6% of all words in the corpus. Rescaling this using the notional daily word count of 14 554 gives 808.4. In other words, making the same assumptions as above, we would expect a notional daily co-occurrence rate for DETERMINER + NOUN of 808 per day. Compared with this, the co-occurrence rates for *þær* and *beon* are clearly on the low side.

Of course, these estimates must be taken with more than a grain of salt. They use text frequencies over large time intervals to approximate the individual exposure to word combinations. However, the calculations serve to illustrate the main point,

[5] http://itre.cis.upenn.edu/ myl/languagelog/archives/003420.html.

namely that the combination *þær* and *beon* in YCOE, though frequent in the corpus compared with many other possible combinations, is not really a high frequency phenomenon. If YCOE is taken as an acceptable approximation to the co-occurrence rates for *þær* and *beon* in Old English at large, and if we assume that the number of words we are exposed to through personal communication has not changed too dramatically, then it would seem that the co-occurrence of *þær* and *beon* is not really a high frequency phenomenon from the speaker's point of view.

There and the EC

However, as it stands this is still only half a cause. Why should the co-occurrence of *there* and *be* lead to *there* being integrated into the EC? Two steps are necessary to adequately account for this process. In addition to a plausible attention mechanism which notices the pattern (outlined above), we need a cognitive mechanism which perpetuates and solidifies the new pattern, once the connection between *there* and the EC has been made. Stewart and Cohen (1997, 167) argue that

> brains do not represent the outside world *as it is*, but in terms of a 'chunked' model in which certain types of stimulus are lumped together and interpreted as being 'the same'. These chunks stand out from the rest because the brain perceiving them has evolved detectors for such stimuli. [emphasis in original]

For the present study it is of less consequence whether such stimulus-detection is based on an evolutionary advantage, represents exaptation (a consequence of another development, cf. Gould and Vrba (1982)) or a special analogy-mechanism. It is sufficient for our purposes to note that there is reason to believe that human brains detect patterns by processing input in chunks. This chunking capacity, Stewart and Cohen argue, is primarily directed at detecting *differences*: the default option, so to say, is to "make everywhere the same as everywhere else" (Stewart and Cohen, 1997, 171). Although looking for differences rather than what is actually "there" might have the evolutionary advantage of freeing up capacity for other tasks (Stewart and Cohen, 1997, 171) this claim is also secondary to the present discussion.

Thus, we have two plausible cognitive mechanisms to account for the evolution of *there*: first, in a low frequency phenomenon, high frequency outcomes have an advantage over low frequency outcomes and differences tend to be amplified. Second, once the high-frequency outcome is established as the default in a particular context, it could easily be perceived as different from the low-frequency outcome(s). Following this step, it would be natural to merge the high-frequency outcome with the general

context it appears in (such as the EC in the case of *there*). This could possibly take place by analogy or some general "make everywhere the same as everywhere else" tendency as noted by Stewart and Cohen (1997).

The outline above fits well with the corpus evidence. However, it does not answer the question of why the development took place at the time it did. That is, the *proximate* cause can plausibly be said to be the mechanisms outlined above, but the *underlying* cause might be better accounted for by other means. Of course, the possibility that there are no "underlying" causes at all must be kept in mind: the process might have taken place when it did by pure chance. However, a range of non-linguistic factors can influence linguistic change, and the next paragraph will discuss one such factor.

11.6.2 Population and language

The section above dealt with the possible cognitive underpinnings of the evolution of existential *there*. However, the cognitive mechanisms obviously do not operate in a vacuum. Although a plausible cognitive mechanism for the described change was offered, this is not necessarily the cause (or the only cause) of the evolution of existential *there* at some specific time. While small variations could easily lead to a skewness which amplifies small differences, it is also worthwhile to ask if there are any external (or social) factors which might affect any "naturally occurring" variations and feed an amplification process.

It can be instructive to change focus once again and look at the development from the point of view of society, or "population" in the sense of "people living in an area". What, if any, demographic changes took place in England during the period under investigation?

Johansson (1997) discusses language as a population process, where language is learnt through reproduction of perceived distinctions. Based on simulation studies, he argues that the Black Death and subsequent plagues could have had an impact on the transmission of language as reproduction process and thus contributed to the loss of case marking in several European languages after the plague.

Hatcher (1977), in his study of the interaction of epidemics with English economy and demographics, argues that in the mid-15th century, the English population reached a nadir, cf. Hatcher (1977, 69), after a protracted decline which had already begun when the great plague struck England in 1348. The end of that century, as well as the first half of the 16th, saw a gradual recovery to the point that before 1600 "the real wages had plunged to the lowest level ever and commentators were increasingly arguing that England was overpopulated" (Hatcher, 1977, 67).

Population estimates for England before the mid 16th century are very uncertain, as discussed in Hatcher (1977). However, while the actual figures (especially for specific

years, decades, or areas) are tentative, there appears to be some consensus about the main trends. The population in England appears to have peaked around 1300, and then to have fallen, before recovery started in the Early Modern period.

Although the observations on *there* fit the population trend, this does not, of course, indicate that the plague played a role. However, if the plague did play a role, we would expect more or less the observed pattern. After initial variation where the (proto) existential *there* is a minor variant in Old and Middle English, a change takes place during the 15th century i.e. about 100 to 150 years after the plague, when the effects to language transmission would start to be fully felt.

A possible topic for future studies could be to use the approach presented in Johansson (1997) to model the potential effects of demographic changes on the use of *there* using simulations, incorporating the information from the models established in the present study.

11.6.3 How *there* became existential

Some uncontroversial facts have been established through corpus data so far. First, in all three corpora *there* tends to co-occur with *be*. Second, there is a gradual increase in proportion of existential vs. locative *there* to the point where the existential use becomes the predominant one in Early Modern English, a situation which persists in Present-day English. Although it is perfectly natural to *describe* this as a grammaticalization of *there*, this does not offer any immediate explanatory advantage over other descriptions, as pointed out in chapter 3. One or more mechanisms, which can be plausibly assumed to cause (or "generate" to use the terminology in Goldthorpe (2001)) the observed effect with a high probability is needed. To attempt this, it is necessary to situate the corpus findings within a linguistic theory. As discussed in chapter 3, the choice of theory is not neutral with respect to the possible explanations, nor can the choice of theory be settled on purely empirical grounds alone.

Croft (2000, 176) notes that "there appears to be a natural human tendency for a community to select one alternative as the conventional signal for a recurrent coordination problem", which he refers to as the *First law of propagation*. Croft (2000, 176) points out that it is not immediately obvious why this law should hold, since humans have the memory capacity to handle multiple linguistic variants. He suggests that as forms are gradually associated with specific social groups, the combined pressure to at the same time communicate, establish, and confirm a social or communal identity leads to selection of one form over the other.

However, it is not clear how such a selective mechanism was to work in the case of *there*. Did *there* function differently with respect to different social and geographical groups? There is no real evidence for this in the corpora. It was shown above that a po-

tential marker of different social ties, genre, plays no role in predicting whether a given target word will be existential *there* or not. A mechanism such as grammaticalization – or the associated term "metanalysis" used in Croft (2000, 130–134) – cannot offer any suggestions for generative processes beyond general descriptions such as "the speaker reanalyses x as y for pragmatic reasons in context z". However, such a description merely restates the probabilistic facts: there is a likely correlation between x and y in context z. As such, the grammaticalization-description statement is merely an unquantified probabilistic statement with an unspecified empirical foundation.

Following the criticism set out in Lass (1997, 341), I would add that it seems unlikely that language – and, by implication, the social and contextual *use* of language – constantly poses problems and challenges for speakers and hearers. This leads to a view of language change where these problems in turn are overcome through changes, which offers an implied (optimum based) teleology in linguistic change, as rightly pointed out by Lass. Furthermore, as emphasized in Fischer (2007a), it is important not to underestimate the effects of syntactic structure.

The following explanation attempts to accurately account for the corpus data, integrate the cognition and language processing of the language user while at the same time maintaining the social implications of language change.

Explaining the process

The initial situation in Old English is one of variation, with either an existential *there* or a proto-existential *there* occurring as a low frequency phenomena. Locative adverbs frequently occur initially in YCOE, and *there* is the most frequent locative adverb, especially when co-occurring with *be*. Following **?**, there was a *convention* of using *there*, in the sense that *there* could be used or not. There are some indications that *there* might have had a status as a non-central subject in *that*-clauses in Old English, on a par with accusative and dative NPs. This situation is found well into the Middle English period, as illustrated in figure 11.7 on p. 263 above. However, during the 15th century, there is an increase in the proportions of existential *there*. This increase could be explained by a gradual *perceived* amplification of the mean probability of finding *there* in the canonical subject position, followed by *be*. This amplification would be expected if *there* was the most frequent adverb occurring with *be* and their co-occurrence was a low frequency phenomenon, as seems likely. The timing of the change itself might have come about due to some distortion of the probabilities of co-occurrence, which might be attributed to either population change (cf. the previous section) through normal variation coupled with chance, or their combination with one or more other factors. It is well established that in small populations, small changes can easily be amplified (Guttman, 2005, 73–74).

The effect of such a process would be to tie one use of *there* to a specific context. As Johansson (1997, 93) suggests, this could be seen as a simplifying move by ruling out a number of alternatives. That is, instead of the pattern ADVERB + *be*, we get *there* + *be*. However, Johansson (1997) points out that such a move cannot be seen as caused by syntax; rather, it is the result of specific selections. In this context, this could be seen as a side effect of the cognitive processing mechanisms mentioned above, rather than a goal in itself. This allows for a genuinely simplifying effect to arise through a cognitive mechanism, but avoids the notions of striving towards a goal associated with a scenario in which speakers actively attempt to make language simpler. Once *there* was associated with a specific context, the deictic distinction with *here* would start to break down for this particular use of the morpheme.

Given the existential-like uses of *here* found in Old and to some extent Middle English, it seems that this could have happened gradually. Once the distinction was gone (that is, when *here* was no longer used as an alternative to *there* in this position), *there* would no longer qualify as a fully symbolic sign according to the criteria of Deacon (1997). If Deacon (1997) is correct in assuming that there is a *structured descent* from one category of signs to another, *there* would be reclassified as an *indexical sign*. As such, it would be left without symbolic meaning, but instead point to the co-occurrence with *be*, most notably in the EC, corresponding to the signal function discussed in Breivik (1997). At no point would *there* be devoid of meaning, since the meaning in any case would be contextual and depend on the construction. However, existential *there* would come to have a less readily *compositional* meaning, rendering its meaning more opaque and difficult to distinguish from that of the EC itself.

At any one point in the process, there would be bound to be blurry cases. However, the exact meaning of *there* would matter less in the account presented here than patterns of use. As long as the pattern, or construction, was reasonably clear, there is no reason why a potentially unclear status pertaining to *there* should matter. The explanation presented here rests on a clear division of labor, where the separate components take care of different aspects of the process. The cognitive attention mechanism described above is responsible for *selecting* the morpheme *there* as a salient feature of a pattern of co-occurrence with *be*. Once the pattern *there* + *be* was noted, it could easily be propagated by several converging processes. First, *there* would occur as a non-prototypical subject, easing its integration into the EC through structural factors. Second, this move might lead to less perceived complexity by reducing syntactic variation. Whether this would be felt as an improvement is difficult to tell; however, it seems likely that it would not be a significantly *worse* solution, thus preventing its removal. Third, as *there* came to be increasingly associated with *be* and the EC, the distinction with *here* would disappear, thus reinforcing this particular use of there as something tied to the subject position of ECs.

A population change, such as e.g. the Black Death, might have been one of the contributing factors in initiating the process, that is, placing it at the time in history where it actually occurred. No explicit appeal to the semantic or pragmatic motivations of speakers need be made, since *there* would in any case inherit semantics from the EC. The signal function of the indexical/existential use of *there* would arise naturally once *there* was sufficiently associated with *be* and the EC, and the deictic contrast with *here* had broken down.

The solution offered here meets a number of the criteria set out by Lass (1997, 349)

- no appeal is made towards optimality in any qualitative sense
- no appeal is made towards teleological causes
- the claim is falsifiable

Note that the proposed explanation rests on the following premises, some of which are purely descriptive. Others are tested and found to be highly likely. Finally, some are more speculative, but have been tested in other fields and are in principle falsifiable:

(i) *there* in initial position co-occurs with *be* in YCOE far more than would be expected if they were unrelated (fact);

(ii) the presence of *be*, and to some lesser extent other factors, can predict the existential use of *there* in early English independent of position in the sentence (fact);

(iii) *there* and *be* appear together at moderate to low rates in YCOE, compared with high frequency phenomena (fact);

(iv) small samples maximize the subjective probability of the most frequent outcome (strong claim, tested, and potentially falsifiable);

(v) based on estimates from present day usage documented in Rayson, Leech, and Hodges (1997), estimates show that the co-occurrence of *there* and *be* in Old English would be a low frequency phenomenon (conjecture);

(vi) with *there* and *be* co-occurring as the most salient outcome in small samples, a contextual probabilistic processing of the sign *there* could be gradually enforced, whereby the perceived deictic opposition is lost leading to a referential breakdown with *here* and a descent from symbolic to indexical sign and perceived semantic "emptiness" (proposed explanation).

Itkonen (1981, 694–695) argues that "linguistic change ... is not action on a par with speaking ... The choice between two equally good alternatives (here it is between two innovations, between innovation and non-innovation, or between acceptance and non-acceptance) may be random". However, simply accepting randomness (or possibly arbitrariness) when functional or sociolinguistic explanations fail is not necessary. As I have attempted to show, there are clear tendencies to be found in the corpus material. However, as discussed above, small (possibly random) differences can be amplified in small populations. As **?**, 86 puts it: "A convention is produced when a big enough fluctuation meets strong enough amplifying forces". The focus of the present study has been the variation and the forces that could amplify and reproduce it, not on how the variation itself arose.

11.7 Summary

In this chapter, it was shown that the models discussed in previous chapters could be used to estimate the distribution of existential *there* in Old English. The estimates, which are likely to be on the low side, showed a small number (400–500) occurrences of existential *there* in the Old English material. A diachronic model of all the three corpora shows that the Middle and Early Modern data are more similar to each other with respect to *there* than either is to the Old English data. The evolution of *there* follows an S-shaped curve, where the probability of existential *there* increases during the late Middle English period before it stabilizes at a relatively high probability. A comparison with a Present-day English treebank showed that the proportion of existential *there* was comparable to that found in Early Modern English.

Possible confounding factors such as translation status and genre were investigated. It was found that these factors were not good predictors of existential *there*. Based on the study of these factors, it seems reasonable to conclude that the syntactic and diachronic factors take precedence when it comes to explaining the evolution of *there*.

Finally, some space was devoted to possible causes and mechanisms involved. A cognitive explanation would need to explicitly relate the corpus frequencies to individual cognition via some plausible mechanism. By making some simplifying assumptions, it was shown that the estimated co-occurrence of *there* and *be* is relatively infrequent. Thus, the corpus data show that *there* and *be* frequently co-occur, but that the daily estimated exposure to this co-occurrence should be much lower than for high frequency linguistic patterns such as DETERMINER + NOUN. Hertwig and Pleskac (2008) among others have argued that in a small sample, high frequencies tend to be amplified. This suggests a reason why *there* (rather than *here* or some other adverb) evolved into an existential marker: *there* occurred frequently with *be*, a co-occurrence which was

amplified through general cognitive attention-mechanisms within a construction which itself was infrequent. It was pointed out that the diachronic development coincides with large population changes in England. Based on Johansson (1997), it was suggested that *one* possible cause for shifts in attention and exposure could be found in the decline in the English population caused by the bouts of plague that hit England in the 14th century.

On the basis of the factors outlined in the preceding paragraphs, an attempt was made to outline a coherent story, where the evolution of *there* is seen as a result of the interplay of linguistic and cognitive factors with historical contingencies. The distribution of locative adverbs in YCOE favored *there* as the dominant (or only) existential subject. Locative adverbs were frequently found initially in Old English, with *there* being the most frequent. At some point the co-occurrence of *there* and *be* was amplified and became a fixed pattern. The reasons for this change needs further investigation, but population changes could well be a potential candidate. Once the amplification of the *there* + *be* pattern started, other adverbs would start to disappear from this construction. Consequently, *there* in this position would lose its contrastive function with *here*, causing a descent from fully symbolic to indexical sign, which accounts for the putatively empty semantics of *there*. Under a RCG analysis, it is not necessary to posit a single point at which *there* finally changes its meaning/function. Furthermore, it is not necessary to motivate a detailed process through semantics or pragmatics. Instead, the semantics of *there* is supplied by the existential construction which it occurs with/within. This approach allows a description and explanation in very simple terms, namely:

- syntax and co-occurrence rates;
- constructional semantics;
- general cognitive characteristics.

Notably, this means that the explanation proposed here circumvents the problem of classifying individual occurrences of *there* manually. In this way, one of problems discussed in Breivik (1990, 181–188) and section 1.1.1 above regarding the status of there is avoided. Instead of classifying single occurrences of *there* as *either* locative or existential, the status of *there* is estimated in each case through probabilities derived from the whole population (or more accurately: the corpus selection) of utterances with *there*. Furthermore, the explanation is falsifiable on a number of crucial points. Although the testing of all these points falls outside the scope of the present study, the number of words we are typically exposed to in face to face communication, the amplification effects in low-frequency constructions, as well as the effects of population changes on syntax can be investigated independently of the present study.

Chapter 12

Conclusions

> history . . . must always end with questions. Conclusions are much too convenient
>
> Antony Beevor

12.1 Introduction

In this chapter I will summarize the overall results and conclusions of my investigation of *there*. The main findings will be briefly outlined and commented upon. Subsequently, the goals set out in the introductory chapter will be discussed. Finally, some concluding remarks, including suggestions for further research, will be presented.

12.2 Summary of goals

The main goals set out in chapter 1 were to contribute to the study of the evolution of existential *there* in the fields of data, methodology, and theory. As far as theory is concerned, it was shown that a RCG approach to *there* has a number of benefits. Specifically,

the RCG approach is non-compositional and the semantics of the part is inherited from the construction itself. This takes care of the problem of assigning a compositional meaning to existential *there*: the meaning is inherited from the EC, but it is not

necessarily a compositional meaning. This would account for the *signal function* associated with *there* discussed in Breivik (1990) and Breivik (1997): *there* signals the introduction of new material precisely because of its association with the EC. As an extension of this, I argued, following Deacon (1997), that existential *there* should be regarded as a linguistic sign, but an indexical rather than a fully symbolic sign. This is compatible with a RCG approach and solves the problem of explaining the status of *there*. Once referential opposition with *here* has broken down, the fully symbolic locative adverb descends the hierarchy of signs proposed by Deacon (1997) in a systematic way to indexical status, and takes up a new function as an indicator (i.e. the signal function discussed by) of the EC. To explain this, it was necessary to describe and analyze data and processes which could plausibly and coherently account for the development where *there* lost its symbolic status.

The chapters on corpus data and analysis were devoted to analyzing the almost 24 000 sentence-tokens extracted from three corpora (the *York-Toronto-Helsinki Corpus of Old English*, the *Penn-Helsinki Parsed corpus of Middle English*, and the *Penn-Helsinki Parsed corpus of Early Modern English*). This analysis was carried out using state of the art statistical techniques illustrating three important methodological points:

(i) The choice of statistical test requires careful consideration – not all tests and methods are suitable for all questions;

(ii) using advanced statistical techniques can give rich and insightful information about corpus data beyond raw frequencies, percentages, or p-values from null hypothesis tests;

(iii) analyzing large amounts of linguistic material with high accuracy is quite feasible (once the data has been collected).

A major obstacle was, in fact, to extract and process data from the treebanks into a format suitable for statistical analysis. Once this task was completed, the actual analysis was comparatively swift and easy.

12.3 Main findings

The main findings from the corpus investigation can be summarized as follows:

- *There* was the most frequent locative adverb in all three corpora;
- *there* co-occurs with *be* to a much greater extent than would be expected in all three corpora;

- it is likely that the existential use of *there* existed as a marginal or low frequent phenomenon already in Old English;
- an increase in the mean probability of existential *there* takes place at the end of the Middle English period following an S-shaped pattern;
- the mean probability of existential *there* has changed very little since the 16th century;
- the changes seem to be largely connected with linguistic (in the narrow sense) variables, rather than with variables such as genre, author, or translation status;
- the change could have been generated, or greatly affected, by general cognitive attention and classification processes.

Needless to say, the final point is a conjecture based on the corpus analysis and research by others on processing mechanisms, since corpus data themselves do not provide direct evidence on language processing mechanisms. Further simulation-based and/or experimental research would be needed to verify this interpretation. However, a number of important conclusions were reached regarding the hypotheses presented in chapter 2:

- The hypothesis that *there* could function as a subject in Old English was strengthened;
- the hypothesis that there was a general increase of initial adverbs during the Old English period was rejected;
- the hypothesis that existential *there* became more prevalent during the Old English period was rejected: the largest change seems instead to have taken place in the late Middle English period;
- the hypothesis that there is a correlation between adverbs in initial position and existential *there* was strengthened, but this too took place in the late Middle English period.

Based on this, some of the general properties ascribed to the evolution of existential *there* in Breivik (1990) and Butler (1980) seem to hold; however, I have shown that although this must have been a gradual process which had already begun in Old English, it only gained momentum in late Middle English and was not completed until the beginning of the Early Modern English period.

12.4 Concluding remarks

In the present work, I have demonstrated the benefits of working with large amounts of data using sophisticated state of the art statistical models. Far from being reductive in nature, I have shown that such models can provide a rich source of insight to diachronic syntactic processes. By using large amounts of data, the distinction between what is central and what is less central becomes clearer, and the main trends can be readily distinguished from natural variation. The statistical techniques needed to handle such amounts of data efficiently present a detailed picture where interactions and correlations can be quantified.

Furthermore, I have shown that such an approach is fully compatible with a cognitive-oriented theoretical frame work, in this case RCG, which explicitly includes the language processing of the speaker in its models. The evolution of existential *there* can easily seem incomprehensible from a formalized perspective which sees syntax as first and foremost a rule-based affair: why should *there* be inserted (if it is inserted at all) over some other morpheme, what is its meaning, and why should it be necessary at all? Conversely, its evolution might easily seem inevitable from a functional perspective which sees language as primarily based on meaning and speaker intentions: *there* could be suitably exploited to express pragmatic functions. Instead, I have attempted to outline a third alternative.

The evolution of existential *there* seems perfectly natural given the fact that *there* was the most frequent locative adverb occurring with *be*. A reclassification of *there* as an indexical, rather than symbolic, sign accounts for its signal function, whereas the cognitive mechanisms involved in this was hypothesized to be general classification and attention traits that have been independently investigated by others. In short, I propose that the evolution of existential *there* was a process driven by a combination of factors. These factors were hypothesized to be formal and functional properties of the existential construction, frequencies of occurrence, and general language processing factors, and I have been able to estimate these factors and their effects using the available data and statistical techniques.

Appendices

Appendix A

CA output

A.1 Old English: Semantic class and adverb position

```
Principal inertias (eigenvalues):

      dim    value       %    cum%   scree plot
[1,]  1      0.062391   92.0   92.0   ************************
[2,]  2      0.005439    8.0  100.0
[3,]         --------   -----
[4,]  Total: 0.067830  100.0
```

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FIN	35	1000	107	−250	303	35	−379	697	930
2	INI	256	1000	681	425	999	740	−10	1	5
3	MID	709	1000	212	−141	975	225	22	25	66

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	Abl	16	1000	1	−28	228	0	52	772	8
2	App	45	1000	19	165	941	20	41	59	14
3	Asp	19	1000	3	−30	101	0	−91	899	28
4	BdP	5	1000	6	302	999	7	10	1	0
5	COP	54	1000	61	275	991	66	−26	9	7
6	COS	23	1000	2	30	155	0	−70	845	21
7	Cmb	4	1000	6	−284	874	6	108	126	9
8	Cmm	91	1000	142	−324	992	153	−29	8	14
9	Cnc	2	1000	1	−140	475	1	−147	525	7
10	Cnt	7	1000	4	11	3	0	192	997	50
11	Crt	14	1000	8	−201	996	9	12	4	0
12	Ctt	2	1000	1	−25	28	0	−146	972	7
13	Dst	4	1000	15	−459	912	14	142	88	16
14	Dsp	3	1000	28	834	946	29	−198	54	19
15	Ems	2	1000	4	−66	41	0	−319	959	43

16	Exs	335	1000	78	125	990	84	13	10	10
17	Grm	0	1000	1	−563	774	1	304	226	2
18	HlK	19	1000	39	−371	987	42	42	13	6
19	Ing	4	1000	6	−230	481	3	239	519	39
20	Int	10	1000	32	−479	995	35	34	5	2
21	Jdg	5	1000	4	−187	732	3	113	268	12
22	Kll	1	1000	6	−563	774	5	304	226	17
23	Lrn	1	1000	3	−337	626	2	260	374	14
24	Mrk	6	1000	14	−403	981	15	56	19	3
25	Mtn	82	1000	52	206	994	56	−15	6	4
26	Occ	7	1000	30	−516	990	32	51	10	4
27	Prc	24	1000	67	−431	995	73	31	5	4
28	Pok	1	1000	17	−651	408	7	−785	592	124
29	PsS	52	1000	80	−314	955	83	−68	45	45
30	PsP	1	1000	4	−397	748	3	−231	252	12
31	Ptt	24	1000	87	−482	947	90	−114	53	57
32	Rmv	4	1000	28	344	272	8	−564	728	256
33	Src	3	1000	6	−344	844	6	−148	156	12
34	SnC	9	1000	2	18	19	0	126	981	27
35	ScI	72	1000	97	303	999	105	9	1	1
36	SpC	44	1000	35	−217	870	33	84	130	57
37	Thr	4	1000	13	−422	699	10	277	301	49

A.2 Middle English: Semantic class and adverb position

```
Principal inertias (eigenvalues):

     dim    value       %    cum%   scree plot
[1,] 1      0.044858   74.2   74.2  ************************
[2,] 2      0.015563   25.8  100.0
[3,]        --------  -----
[4,] Total: 0.060421  100.0
```

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FIN	100	1000	234	48	16	5	373	984	895
2	INI	275	1000	533	−341	994	713	−27	6	12
3	MID	626	1000	233	142	898	282	−48	102	93

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	Abl	29	1000	6	−113	997	8	−6	3	0
2	Act	11	1000	13	−8	1	0	270	999	52
3	App	41	1000	66	283	825	73	−130	175	45
4	Asp	26	1000	252	−769	994	337	61	6	6
5	Ass	2	1000	2	−241	944	3	−59	56	1
6	BdP	9	1000	31	−419	877	37	157	123	15
7	COP	37	1000	29	−216	978	38	−32	22	2
8	COS	14	1000	1	31	162	0	70	838	4
9	Cmb	1	1000	9	218	115	1	605	885	30
10	Cmm	107	1000	105	243	999	141	9	1	1
11	Cnc	3	1000	6	−321	936	7	84	64	1
12	Cntc	1	1000	2	−90	70	0	−327	930	8
13	Cntn	38	1000	43	247	893	52	86	107	18
14	Crt	16	1000	13	−209	920	16	62	80	4
15	Ctt	3	1000	6	0	0	0	−334	1000	22
16	Dcs	2	1000	0	−58	993	0	5	7	0
17	Dst	0	1000	12	448	106	2	1305	894	40
18	Dsp	0	1000	2	671	753	2	−385	247	2
19	Ems	2	1000	0	8	68	0	−31	932	0
20	Eng	1	1000	4	−307	494	3	−311	506	8
21	Exs	368	1000	59	−70	514	41	−68	486	111
22	HlK	6	1000	9	294	992	12	26	8	0
23	Ing	3	1000	7	368	999	10	10	1	0
24	Int	59	1000	4	−62	932	5	17	68	1
25	Jdg	7	1000	26	394	654	23	287	346	35
26	Kll	3	1000	12	−189	166	3	424	834	38
27	Lrn	2	1000	24	−787	799	25	394	201	18
28	Ldg	1	1000	7	671	753	7	−385	247	7
29	Mrk	4	1000	1	92	388	1	115	612	3
30	Mtn	40	1000	20	141	659	18	−101	341	27
31	Obl	3	1000	4	282	841	5	123	159	2
32	Occ	4	1000	17	360	498	12	−361	502	34
33	Prc	15	1000	6	110	519	4	106	481	11
34	Pok	1	1000	29	−693	199	8	1391	801	91
35	Prv	1	1000	2	126	130	0	326	870	6

36	PsS	52	1000	78	213	505	53	211	495	151
37	PsP	0	1000	27	226	6	0	2994	994	106
38	Ptt	9	1000	12	89	104	2	261	896	40
39	Rmv	1	1000	7	−544	809	7	265	191	5
40	Src	2	1000	13	−614	999	17	16	1	0
41	SnC	10	1000	8	198	824	8	−92	176	5
42	ScI	40	1000	0	0	0	0	−12	1000	0
43	SpC	20	1000	23	−198	557	17	176	443	40
44	Thr	5	1000	2	37	64	0	141	936	7
45	Wth	0	1000	2	−470	712	2	−299	288	2

A.3 Early Modern English: Semantic class and adverb position

```
Principal inertias (eigenvalues):

     dim    value       %    cum%   scree plot
[1,] 1      0.068766   91.2  91.2   ************************
[2,] 2      0.006619    8.8 100.0
[3,]        --------- -----
[4,] Total: 0.075386  100.0
```

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FIN	118	1000	135	−202	473	70	−213	527	812
2	INI	210	1000	714	506	999	783	−15	1	7
3	MID	672	1000	150	−123	894	147	42	106	181

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	Abl	25	1000	19	−195	692	14	130	308	65
2	Act	16	1000	37	−406	946	38	97	54	23
3	App	45	1000	7	−43	163	1	97	837	63
4	Asp	10	1000	0	−48	893	0	−17	107	0
5	Ass	2	1000	9	−499	855	8	205	145	14
6	BdP	3	1000	3	223	604	2	−181	396	14
7	COP	27	1000	4	−85	640	3	−64	360	17
8	COS	11	1000	5	190	999	6	6	1	0
9	Clr	0	1000	1	−468	449	0	519	551	5
10	Cmb	3	1000	2	−147	365	1	194	635	18
11	Cmm	74	1000	105	−322	974	112	52	26	30
12	Cnc	0	1000	6	1130	998	6	49	2	0
13	Cntc	2	1000	11	−483	630	8	370	370	48
14	Cntn	57	1000	96	−349	968	102	−63	32	35
15	Crt	12	1000	16	22	5	0	−313	995	177
16	Ctt	1	1000	0	−49	37	0	−250	963	5
17	Dcs	0	1000	1	−468	449	0	519	551	5
18	Dst	0	1000	1	−468	449	1	519	551	9
19	Ems	1	1000	3	−202	173	1	441	827	29
20	Eng	0	1000	3	−569	538	2	−527	462	14
21	Exs	411	1000	384	264	993	418	21	7	29
22	HlK	4	1000	10	−433	918	10	129	82	9
23	ImC	0	1000	1	−468	449	0	519	551	5
24	Ing	5	1000	1	105	745	1	−62	255	3
25	Int	39	1000	16	−169	916	16	−51	84	15
26	Jdg	2	1000	1	−206	966	1	39	34	0
27	Kll	0	1000	6	1130	998	6	49	2	0
28	Lrn	1	1000	4	−169	65	0	−645	935	42
29	Ldg	2	1000	4	270	383	2	−344	617	30
30	Mrk	4	1000	6	−323	964	7	62	36	3
31	Mtn	65	1000	64	−272	990	70	27	10	7
32	Obl	8	1000	6	−28	13	0	−238	987	68
33	Occ	8	1000	14	−378	997	16	−22	3	1
34	Prc	27	1000	6	−106	697	4	−70	303	20
35	Pok	1	1000	5	730	950	5	167	50	3

36	PsS	58	1000	86	−330	989	93	−35	11	11
37	PsP	2	1000	5	−512	981	6	71	19	1
38	Ptt	15	1000	17	240	692	13	−160	308	59
39	Rmv	1	1000	4	−354	650	3	−260	350	15
40	Src	1	1000	2	−319	963	2	−63	37	1
41	SnC	7	1000	2	−127	788	2	66	212	5
42	ScI	30	1000	11	−98	348	4	−134	652	82
43	SpC	16	1000	17	266	888	17	−95	112	22
44	Thr	1	1000	0	93	781	0	−49	219	0

A.4 Early English: MCA of genre, *there* and period

```
Principal inertias (eigenvalues):

      dim     value       %    cum%   scree plot
 [1,] 1       0.266791   61.6   61.6   ************************
 [2,] 2       0.094108   21.7   83.4   *********
 [3,] 3       0.001587    0.4   83.7
 [4,] 4       00000000    0.0   83.7
 [5,] 5       00000000    0.0   83.7
 [6,] 6       00000000    0.0   83.7
 [7,] 7       00000000    0.0   83.7
 [8,] 8       00000000    0.0   83.7
 [9,] 9       00000000    0.0   83.7
[10,] 10      00000000    0.0   83.7
[11,] 11      00000000    0.0   83.7
[12,] 12      00000000    0.0   83.7
[13,] 13      00000000    0.0   83.7
[14,] 14      00000000    0.0   83.7
[15,] 15      00000000    0.0   83.7
[16,] 16      00000000    0.0   83.7
[17,] 17      00000000    0.0   83.7
[18,] 18      00000000    0.0   83.7
[19,] 19      00000000    0.0   83.7
[20,] 20      00000000    0.0   83.7
[21,] 21      00000000    0.0   83.7
[22,] 22      00000000    0.0   83.7
[23,] 23      00000000    0.0   83.7
[24,]         --------- -------
[25,] Total:  0.432935  <NA>
```

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	ExT.FALSE	259	2247	20	−194	2132	19	−35	115	1
2	ExT.TRUE	74	2247	71	675	2132	66	121	115	4
3	Gnr.Apocrypha	2	1811	4	−737	1226	2	−392	585	1
4	Gnr.Bible	19	2074	2	17	12	0	−171	2062	2
5	Gnr.Bio_auto	5	1675	9	789	1392	6	−274	283	1
6	Gnr.Bio_lives	28	1789	45	−713	1393	27	−293	396	8
7	Gnr.Bio_other	3	1718	8	893	1528	5	−243	190	1
8	Gnr.Chart_wills	1	1860	1	−764	1272	1	−401	589	0
9	Gnr.DiaryPriv	12	1644	23	763	1336	13	−282	308	3
10	Gnr.Drama_comedy	11	1569	20	717	1220	11	−296	349	3
11	Gnr.EccLaws	1	1859	1	−763	1271	1	−400	589	0
12	Gnr.EducTreat	5	1713	12	913	1539	8	−237	174	1
13	Gnr.Epilogue	0	1904	0	−797	1316	0	−411	588	0
14	Gnr.Fiction	11	1511	12	586	1351	7	−155	159	1
15	Gnr.Geography	0	1454	1	−617	931	0	−356	523	0
16	Gnr.Handb_astronomy	0	2147	1	−243	65	0	1063	2082	1
17	Gnr.Handb_med	4	1753	5	−683	1564	4	−183	189	0
18	Gnr.Handb_other	6	1708	15	926	1544	10	−233	164	1
19	Gnr.Handbook	1	2441	3	−58	4	0	1119	2437	3
20	Gnr.History	73	1689	48	−494	1670	34	−40	19	0
21	Gnr.Homilies	27	1722	37	−670	1477	24	−210	245	4
22	Gnr.Law	4	1645	3	240	368	0	−344	1277	2
23	Gnr.LetNonPriv	3	1700	6	818	1444	4	−265	256	1

24	Gnr.LetPriv	7	1658	14	773	1360	8	−279	298	2
25	Gnr.Philosophy	7	928	5	390	926	2	14	2	0
26	Gnr.Preface	0	1904	0	−797	1316	0	−411	588	0
27	Gnr.ProcTrial	12	1580	23	723	1237	12	−294	343	3
28	Gnr.RelTreat	27	2370	54	−214	102	2	778	2268	53
29	Gnr.Romance	8	2476	32	13	0	0	1141	2476	35
30	Gnr.Rule	3	2095	3	−371	595	1	454	1500	2
31	Gnr.Science	1	1572	1	−650	1024	1	−366	548	0
32	Gnr.Science_astr	0	1859	1	−763	1271	0	−400	589	0
33	Gnr.Science_med	3	1716	8	900	1532	5	−240	184	1
34	Gnr.Science_other	4	1719	10	880	1518	6	−247	201	1
35	Gnr.Sermon	15	2396	22	354	379	4	630	2017	19
36	Gnr.Travelogue	30	1593	21	486	1478	14	104	115	1
37	Era.EME	127	1650	186	692	1464	118	−190	186	15
38	Era.ME	77	2435	111	−90	25	1	681	2410	116
39	Era.OE	129	1750	157	−629	1458	99	−217	292	20

Appendix B

Regression output

B.1 Middle English

B.1.1 A model of existential *there*

```
Generalized linear mixed model fit by the Laplace approximation
Formula: ExTag ~ I(LogComplexity - mean(LogComplexity)) + BeContext + SemClass + NomNP + (1 | Year25)
   Data: me
  AIC  BIC logLik deviance
 4029 4353  -1966     3931
Random effects:
 Groups Name        Variance Std.Dev.
 Year25 (Intercept) 0.40497  0.63637
Number of obs: 5448, groups: Year25, 12

Fixed effects:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -1.99717    0.29097  -6.864 6.71e-12 ***
I(LogComplexity - mean(LogComplexity))  0.26838    0.05070   5.293 1.20e-07 ***
BeContextTRUE                         2.80195    0.10406  26.926  < 2e-16 ***
SemClassActivity                     -0.63465    0.45522  -1.394 0.163271
SemClassAppearance                    1.30094    0.25537   5.094 3.50e-07 ***
SemClassAspect                       -1.59465    0.43187  -3.692 0.000222 ***
SemClassAssessment                   -0.75462    0.96903  -0.779 0.436131
SemClassBodyProcess                  -1.29019    0.58819  -2.194 0.028271 *
SemClassChangeOfPossession           -2.63750    0.56986  -4.628 3.69e-06 ***
SemClassChangeOfState                -0.58066    0.45996  -1.262 0.206800
SemClassCombining                    -1.33719    1.51675  -0.882 0.377983
SemClassCommunication                -0.94871    0.24915  -3.808 0.000140 ***
SemClassConcealment                  -0.75637    0.86153  -0.878 0.379980
SemClassContact                      -0.14212    1.13029  -0.126 0.899940
SemClassContain                      -0.73341    0.30897  -2.374 0.017609 *
SemClassCreation                     -1.73965    0.51882  -3.353 0.000799 ***
SemClassCutting                       0.70422    0.59304   1.187 0.235040
```

SemClassDecision	−0.75111	1.11097	−0.676	0.498986	
SemClassDestruction	−13.61139	1636.53451	−0.008	0.993364	
SemClassDisappearance	−13.31631	3419.50669	−0.004	0.996893	
SemClassEmission	−1.51641	1.43233	−1.059	0.289736	
SemClassEngender	−13.98160	1076.05013	−0.013	0.989633	
SemClassExistence	−0.50123	0.22495	−2.228	0.025870	*
SemClassHoldKeep	−2.87901	1.17505	−2.450	0.014281	*
SemClassIngesting	−1.63507	1.03337	−1.582	0.113589	
SemClassIntention	−0.18570	0.25814	−0.719	0.471908	
SemClassJudgment	0.19812	0.51454	0.385	0.700200	
SemClassKilling	−1.40573	1.09377	−1.285	0.198715	
SemClassLearning	−13.75247	706.52810	−0.019	0.984470	
SemClassLodge	−14.12423	1066.22901	−0.013	0.989431	
SemClassMarking	−0.74684	0.74261	−1.006	0.314564	
SemClassMotion	−0.63094	0.30655	−2.058	0.039568	*
SemClassObligation	−1.04075	1.08755	−0.957	0.338584	
SemClassOccurrence	0.56397	0.56278	1.002	0.316294	
SemClassPerception	−0.89182	0.42691	−2.089	0.036707	*
SemClassPoke	−13.40408	1078.58378	−0.012	0.990085	
SemClassPrevent	−13.81045	1214.25690	−0.011	0.990925	
SemClassPsychologicalState	−0.69144	0.28170	−2.455	0.014105	*
SemClassPushPull	−13.85515	2072.60861	−0.007	0.994666	
SemClassPutting	−2.02296	0.81983	−2.468	0.013604	*
SemClassRemoving	−13.59033	1015.26686	−0.013	0.989320	
SemClassSearching	−1.02140	1.01438	−1.007	0.313971	
SemClassSendCarry	−2.78788	0.81919	−3.403	0.000666	***
SemClassSocialInteraction	−0.64465	0.30221	−2.133	0.032916	*
SemClassSpatialConfiguration	−1.20777	0.43206	−2.795	0.005184	**
SemClassThrowing	−0.51063	0.62638	−0.815	0.414948	
SemClassWeather	−13.58594	2149.03291	−0.006	0.994956	
NomNPTRUE	0.91763	0.08762	10.473	< 2e−16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

B.2 Early Modern English

B.2.1 A model with Nodes as predictor

```
Generalized linear mixed model fit by the Laplace approximation
Formula: ExTag ~ I(Nodes/3) + BeContext + (1 | Year25)
   Data: eme
  AIC  BIC logLik deviance
 7318 7347  -3655     7310
Random effects:
 Groups Name        Variance Std.Dev.
 Year25 (Intercept) 0.089676 0.29946
Number of obs: 9087, groups: Year25, 9

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.897336   0.111468  -17.02  < 2e-16 ***
I(Nodes/3)    0.003770   0.001037    3.63 0.000279 ***
BeContextTRUE 3.774259   0.069664   54.18  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

B.2.2 A model with NP as predictor

```
Generalized linear mixed model fit by the Laplace approximation
Formula: ExTag ~ NP + BeContext + (1 | Year25)
   Data: eme
  AIC  BIC logLik deviance
 7308 7337  -3650     7300
Random effects:
 Groups Name        Variance Std.Dev.
 Year25 (Intercept) 0.091474 0.30245
Number of obs: 9087, groups: Year25, 9

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.943129   0.113328  -17.15  < 2e-16 ***
NP            0.010714   0.002221    4.82 1.41e-06 ***
BeContextTRUE 3.777704   0.069714   54.19  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

B.2.3 A model with IP as predictor

```
Generalized linear mixed model fit by the Laplace approximation
Formula: ExTag ~ IP + BeContext + (1 | Year25)
   Data: eme
  AIC  BIC logLik deviance
 7321 7349  -3656     7313
Random effects:
 Groups Name        Variance Std.Dev.
 Year25 (Intercept) 0.088104 0.29682
Number of obs: 9087, groups: Year25, 9
```

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) −1.909584   0.112355 −17.00  < 2e−16 ***
IP           0.007380   0.002303   3.20  0.00135 **
BeContextTRUE 3.773025  0.069672  54.15  < 2e−16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

B.3 Early English

B.3.1 A diachronic model of *there* with one random effect

```
Generalized linear mixed model fit by the Laplace approximation
Formula: ExTag ~ LogComplexity + BeContext + NomNP + (1 | Year)
   Data: ee
   AIC   BIC logLik deviance
 12995 13036 −6493    12985
Random effects:
 Groups Name        Variance Std.Dev.
 Year   (Intercept) 0.82259  0.90697
Number of obs: 23341, groups: Year, 32

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) −3.88401    0.17742 −21.89   <2e−16 ***
LogComplexity 0.39395   0.02233  17.64   <2e−16 ***
BeContextTRUE 3.49281   0.05581  62.59   <2e−16 ***
NomNPTRUE     0.87769   0.05124  17.13   <2e−16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix C

Perl scripts

This appendix gives examples of Perl scripts used to extract and process data from the corpus files. The Perl scripts were run under Windows XP with *ActiveState Perl* version 5.10.

C.1 A script to extract data from YCOE

```
#!/usr/bin/perl

use warnings;

# variable declarations follow here in original script

# ...

# ---- Begin program code ---- :

while(<>){
    # input is corpus data without line breaks
        # print out the header line:
        print "Adverb\tContext\tConTag\tConFiller\tContext2\tConTag2\tConFiller2\tTHTclause\t";
print "AdvLength\tText\tID\tInitialAdv\tConjClause\tVerbTag\tVerbRaw\tVerbType\tBeContext\tObject\t";
print "Locative\tEmptyExpl\tProSubj\tThere\tNP\tNodes\tIP\tBeCount\tInitialThere\tInitialBe\n";

        # find each tree, beginning with "(IP" and ending with "*/":
        while(m/(\(((IP-(MAT|SUB))|FRAG).*?)(\/\*)/igs){

                $raw = $1;


                        #while inside a tree, look for locative adv's and what follows next:
                        if($raw =~ m/(.*?)?\(ADV\^L\s+\$?(.+?)\)*?\s+(\(.+?\))\s+(\(.+?\))/igs){
```

```
                        $lcont = $1;
                        $adv = $2;
                        $next = $3;
                        $next2 = $4;
                # print the adverb of the IP in lower case:
                        print lc "$adv\t";

                        print "$next\t";

                        &con_tag();
                # print the POS tag of the left(1) context of the adverb:
                        print "$contag\t";
                # print the realization of the left(1) context of the adverb
                        print "$confiller\t";

                        print "$next2\t";

                        &con2_tag();
                        print "$con2\t";
                        print "$confiller2\t";

                        &left_con();
                        print "$that\t";


                }
                # invoke subroutines for gathering info about each tree:
                  &count_loc();
                  &count_be();
                  &adv_init();
                  &count_cp();
                  &count_ip();
                  &count_np();
                  &verb();
                  &conj_cl();
                  &be_init();
                  &count_nodes();
                  &dir_obj();
                  &expletive();
                  &pro();
                  &id();
                  &text();
                  &there();
                  &be();
                  &adv_Length();
                  &there_init();
                  &verb_tag();


                 print "$advLength\t$text\t$id\t$init\t$conjcl\t$vtag\t$vb\t$verbtype\t";
        print    "$nextbe\t$obj\t$countloc\t$expl\t$pro\t$there\t$countnp\t$countnodes\t";
        print "$countip\t$countbe\t$thereinit\t$beinit\n";


}
# —— End of main script ——
```

```
# —— Subroutines —— :
# (only three examples included)

# check if there is an initial locative ADVP. If yes, set $init to TRUE, if not FALSE.
# ignore initial conjunctions and count the following locative adverb as initial.
sub adv_init{
        if($raw =~ m/(\(IP-MAT(-SPE)?(-\d)?\s\(ADVP-LOC\s)/gs){
                $init = "T"; # get initial locatives
        }

        elsif($raw =~ m/(\(IP-MAT(-SPE)?(-\d)?\s\(CONJ\s((a|A|o|O)(nd|c)|\&)\)\s+\(ADVP-LOC\s)/gs){
                $init = "T"; # ignore initial conjunctions
        }
        else{
                $init = "F";
        }

}

# count the number of IPs in each tree.
sub count_ip{

        my $countmat = () = ($raw =~ m/\(IP-MAT(-\w\w\w)?\s/g);
        my $countsub = () = ($raw =~ m/\(IP-SUB(-SPE)?\s/g);
        my $countinf = () = ($raw =~ m/\(IP-INF(-\w\w\w)?(-.)?\s/g);
        my $countx = () = ($raw =~ m/\(IPX-(SUB|MAT).*?\s/g);
        $countip = $countmat + $countsub + $countinf + $countx;

}


# extract the content of the ID tag, and put the content in $id.
sub id{
        $raw =~ m/\(ID\s(.*?)\)/igs;
                $id = $1;
}
```

C.2 A simple KWIC concordance script

```
#!/usr/bin/perl

  use warnings;

  open (OUTPUT, ">be_con.txt");
  open (INPUT, "oe_be_1line.txt");
  # input is corpus data without line breaks

$raw = "";
$l_con = "";
$r_con = "";
$rightcon = "";
$leftcon = "";
$be = "";

  while(<INPUT>){

          print OUTPUT "LeftCon\tBe\tRightCon\n";

          while(m/(\(((IP-(MAT|SUB))|FRAG).*?)(\/\*)/igs){

                $raw = $1;

          if($raw =~ m/((.+?)\s+\(?(BE\w?\w?\s+.+?)\)\)?\s+(\(.*))/gs){
                $l_con = $2;
                $be = $3;
                $r_con = $4;

        # Invoke subroutines:
                &lcut();
                &rcut();

        # Print $be with left and right context:
                print OUTPUT "$leftcon\t$be\t$rightcon\n";
                }
          }

  }
# ---- End of main script ----

# ---- Subroutines ---- :

sub rcut{
        my $right = $r_con;

        $right =~ m/\((\w+)\s/g;

        $rightcon = $1;

}

sub lcut{
        my $left = $l_con;

        if($left =~ m/(\s+)?\(IPX?-(MAT|SUB|FRAG)(-SPE)?(\=\d)?(-\d?)\s?\s?$/g){
```

```
                $leftcon = "IP";
        }
        else{

        $left =~ m/.*\((.+?\s+.+?)\)?\)?$/gs;

        $leftcon = $1;
        }
}
```

Bibliography

Aarts, B. (2000). Corpus linguistics, Chomsky and fuzzy tree fragments. In C. Mair and M. Hundt (Eds.), *Corpus linguistics and linguistic theory: Papers from the twentieth International Conference on English Langage Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, pp. 5–13. Amsterdam: Rodopi.

Aitchison, J. (1991). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Aitchison, J. (2003). Psycholinguistic perspectives on language change. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*, pp. 736–743. Malden, MA.: Blackwell.

Andersen, H. (1973). Abductive and deductive change. *Language 49*(4), 765–793.

Apollon, D. (1990). Dataanalytiske metoder i filologien. In O. E. Haugen and E. Thomassen (Eds.), *Den filologiske vitenskap*, pp. 181–208. Oslo: Solum forlag.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika 68*(1), 13–20.

Baayen, H. (1992). Quantitative aspects of morphological productivity. In G. Booij and J. van Marle (Eds.), *Yearbook of morphology 1991*, pp. 109–149. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, and S. Jannedy (Eds.), *Probabilistic linguistics*, Chapter 7, pp. 229–287. Cambridge, MA.: The MIT Press.

Baayen, R. H. (2008a). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge Univeristy Press.

Baayen, R. H. (2008b). *languageR: Data sets and functions with* Analyzing Linguistic Data: A practical introduction to statistics using R. R package version 0.95.

Baker, P. S. (2007). *Introduction to Old English* (2nd ed.). Malden, MA: Blackwell.

Barðdal, J. (2008). *Productivty: Evidence from case and argument structure in Icelandic*. Amsterdam: John Benjamins Publishing Company.

Baroni, M. and S. Evert (2006). *The zipfR package for lexical statistics: A tutorial introduction*. http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf.

Bates, D., M. Maechler, and B. Dai (2008). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-28.

Bayley, R. (2002). The quantitative paradigm. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), *The hanbook of language variation and change*, pp. 117–141. Malden, MA.: Blackwell.

Bech, K. (2008). Verb types and word order in Old and Middle English non-coordinate and coordinate clauses. In M. Gotti, M. Dossena, and R. Dury (Eds.), *English Historical Linguistics 2006*, Volume I: Syntax and morphology, pp. 49–67. Amsterdam: John Benjamins.

Berger, J. O. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association 82*(397), 112–122.

Biber, D. (1992). Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In J. Svartvik (Ed.), *Directions in corpus lingusitics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*, pp. 213–252. Berlin: Mouton de Gruyter.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1997). *Longman grammar of spoken and written English*. London: Longman.

Bilisoly, R. (2008). *Practical text mining with perl*. Hoboken, NJ.: Wiley.

Blasius, J. (1994). Correspondence analysis in social science research. In M. Greenacre and J. Blasius (Eds.), *Correspondence analysis in the social sciences*, pp. 23–52. London: Academic Press.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.

Bod, R. (2003). Introduction to elementary probability theory and formal stochastic language theory. In R. Bod, J. Hay, and S. Jannedy (Eds.), *Probabilistic linguistics*, Chapter 2, pp. 11–38. Cambridge, MA.: The MIT Press.

Bolinger, D. (1977). *Meaning and form*. London: Longman.

Breivik, L. E. (1981). On the interpretation of existential *there*. *Language 57*(1), 1–25.

Breivik, L. E. (1990). *Existential* there*: A synchronic and diachronic study* (2nd ed.). Oslo: Novus Press.

Breivik, L. E. (1997). *There* in space and time. In H. Ramisch and K. Wynne (Eds.), *Language in time and space: studies in honour of Wolfgang Viereck on the occation of his 60th birthday*, pp. 32–45. Stuttgart: Franz Steiner Verlag.

Burrow, J. A. and T. Turville-Petre (2005). *A book of Middle English* (3rd ed.). Oxford: Blackwell.

Butler, M. C. (1980). *Grammatically motivated subjects in Early English*. Ph. D. thesis, University of Texas at Austin. Printed 1995 by UMI Dissertation Services, Ann Arbor, MI.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*, pp. 602–623. Malden, MA.: Blackwell.

Cameron, A. C. and P. K. Travin (1998). *Regression analysis of count data*. New York: Cambridge University Press.

Campbell, L. (2004). *Historical linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Cedergren, H. J. and D. Sankoff (1974). Variable rules: Performance as a statistical reflection of competence. *Language 50*(2), 333–355.

Chambers, J. (2002). Patterns of variation including change. In J. Chambers, P. Trudgill, and N. Schilling-Estes (Eds.), *The handbook of language variation and change*, pp. 349–372. Oxford: Blackwell.

Chomsky, N. (1959). Untitled review of *Verbal behavior*. *Language 35*(1), 26–58.

Chomsky, N. (1961). Some methodological remarks on generative grammar. *Word 17*, 219–239.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA.: The MIT Press.

Chomsky, N. (1975). *The logical structure of linguistic theory*. New York: Plenum Press.

Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.

Chomsky, N. (2002). *Syntactic structures* (2nd ed.). Berlin: Mouton de Gruyter.

Church, K. W. (2000). Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of the 18th conference on Computational linguistics*, Morristown, NJ, pp. 180–186. Association for Computational Linguistics.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics – Theory and Methods A9*(10), 1025–1041.

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior 15*(3), 261–262.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American psychologist 49*(12), 997–1003.

Comrie, B. and T. Kuteva (2005). The evolution of grammatical structures and 'functional need' explanations. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution*, pp. 185–207. Oxford: Oxford University Press.

Coopmans, P. (1989). Where stylistic and syntactic processes meet: Locative inversion in english. *Language 65*(4), 728–751.

Coseriu, E. (1987). Linguistic change does not exist. In J. Albrecht (Ed.), *Schriften von Eugenio Coseriu (1965–1987)*, Volume I, pp. 147–157. Tübingen: Gunter Narr Verlag.

Cowie, F. (1999). *What's within? Nativism reconsidered*. Oxford: Oxford University Press.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

Crawley, M. J. (2005). *Statistics: An introduction using R*. Chichester: Wiley.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. London: Longman.

Croft, W. (2001). *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, W. (2007). Form, meaning and speakers in the evolution of language: Commentary on kirby, smith and brighton. In M. Penke and A. Rosenbach (Eds.), *What counts as evidence in linguistics*, pp. 139–142. Amsterdam: John Benjamins Publishing Company.

Croft, W. and D. A. Cruse (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.

Dahl, D. B. (2009). *xtable: Export tables to LaTeX or HTML*. R package version 1.5-5.

Damasio, A. R. and D. Tranel (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences, USA 90*, 4975–4960.

Danielsson, P. (2004). Simple Perl programming for corpus work. In J. M. Sinclair (Ed.), *How to use corpora in second language teaching*, pp. 225–246. Amsterdam: John Benjamins Publishing Company.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. New York: W.W. Norton & Company.

Ebeling, J. (1999). *Presentative constructions in English and Norwegian: A corpus-based contrastive study*. Ph. D. thesis, University of Oslo, Oslo.

Enkvist, N. E. (1972). Old English adverbial *þā* – an action marker? *Neuphilologische Mitteilungen 73*, 90–96.

Everitt, B. S. and T. Hothorn (2006). *A handbook of statistical analyses using R*. Boca Raton, Fl.: Chapman & Hall/CRC.

Evert, S. (2007). How random is a corpus? The library metaphor. *Zeitshcrift für Anglistik und Amerikanistik 54*(2), 177–190.

Evert, S. and M. Baroni (2008). *zipfR: Statistical models for word frequency distributions*. R package version 0.6-5.

Falk, C. (1993). *Non-referential subjects in the history of Swedish*. Lund: Department of Scandinavian Languages, University of Lund. Ph.D. thesis.

Faraway, J. J. (2005). *Linear models with R*. Boca Raton, FL: Chapman & Hall/CRC.

Faraway, J. J. (2006). *Extending the linear model with R*. Boca Raton, FL: Chapman & Hall/CRC.

Faraway, J. J. (2009). *faraway: Functions and datasets for books by Julian Faraway*. R package version 1.0.4.

Faverty, F. E. (1928). Legends of Joseph in Old and Middle English. *PMLA 43*(1), 79–104.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*, pp. 35–60. Berlin: Mouton de Gruyter.

Fischer, O. (2007a). *Morphosyntactic change: Functional and formal perspectives*. Oxford: Oxford University Press.

Fischer, O. (2007b). What counts as evidence in historical linguistics? In M. Penke and A. Rosenbach (Eds.), *What counts as evidence in linguistics*, pp. 249–282. Amsterdam: John Benjamins.

Fitch, W. T. (2007). Linguistics: An invisible hand. *Nature 449*(7163), 665–667.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis*, pp. 245–260. New York: Russel Sage Foundation.

Fleiss, J. L., B. Levin, and M. C. Paik (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives*, pp. 129–189. Cambridge: Cambridge University Press.

Freeze, R. (1992). Existentials and other locatives. *Language 68*(3), 553–595.

Freeze, R. (2001). Existential constructions. In M. Haspelmath (Ed.), *Language typology and language universals: An international handbook*, Volume 2, Chapter 70, pp. 941–953. Berlin: Walter de Gruyter.

Friendly, M. (1995). Conceptual and visual models for categorical data. *The American statistician 49*(2), 153–160.

Gale, W. (1994). Good-Turing smoothing without tears. Statistics Research Reports from AT&T Laboratories 94.5, AT&T Bell Laboratories.

Gale, W. A. and G. Sampson (1995). Good-Turing frequency estimation without tears. *Journal of quantitative linguistics 2*(3), 217–237.

Geeraerts, D. (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Clarendon Press.

Geeraerts, D. (2006a). Methodology in cognitive linguistics. In G. Kristiansen, M. Achard, R. Dirven, and F. J. Ruiz de Mendoza Ibáñez (Eds.), *Cognitive linguistics: Current applications and future perspectives*, pp. 21–50. Berlin: Mouton de Gruyter.

Geeraerts, D. (2006b). *Words and other wonders: Papers on lexical and semantic topics*. Berlin: Mouton de Gruyter.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL.: Chapman & Hall/CRC.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel / hierarchical models*. Cambridge: Cambridge University Press.

Gibbs, R. W. (2007). Why cognitive linguistics should care more about empirical methods. In M. Gonzales-Marquez, I. Mittelberg, S. Coulson, and M. J. Spivey (Eds.), *Methods in cognitive linguistics*, pp. 2–18. Amsterdam: John Benjamins Publishing Company.

Gibson, E. (2000). The Dependency Locality Theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium*, pp. 95–126. Cambridge, Mass.: The MIT Press.

Gill, J. (2006). *Essential mathematics for political and social research*. Cambridge: Cambridge University Press.

Glymour, C. (1986). Comment: Statistics and metaphysics. *Journal of the American Statistical Association 81*(396), 964–966.

Gold, E. M. (1967). Language identification in the limit. *Information and Control 10*(5), 447–474.

Goldthorpe, J. H. (2001). Causation, statistics, and sociology. *European Sociological Review 17*(1), 1–20.

Gorard, S. (2003). *Quantitative methods in social science*. New York: Continuum.

Gould, S. J. (1978). Sociobiology: The art of storytelling. *New Scientist 80*, 530–533.

Gould, S. J. (1993). *Eight little piggies: Reflections in natural history*. London: Vintage.

Gould, S. J. and R. Lewontin (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptionist programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences 205*(1161), 581–598.

Gould, S. J. and E. S. Vrba (1982). Exaptation–A missing term in the science of form. *Paleobiology 8*(1), 4–15.

Greenacre, M. (1994). Correspondence analysis and its interpretation. In M. Greenacre and J. Blasius (Eds.), *Correspondence analysis in the social sciences*, pp. 3–22. London: Academic Press.

Greenacre, M. (2006). Tying up the loose ends in simple, multiple, joint correspondence analysis. In A. Rizzi and M. Vichi (Eds.), *Compstat – Proceedings in computational statistics: 17th symposium held in Rome, Italy, 2006*, Heidelberg, pp. 163–185. Physica-Verlag.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). Boca Raton, FL.: Chapman & Hall/CRC.

Greenacre, M. and O. Nenadic (2007). *ca: Simple, Multiple and Joint Correspondence Analysis*. R package version 0.21.

Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus linguistics and linguistic theory 1*(2), 277–294.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13*(4), 403–437.

Gries, S. T. (2009). What is corpus linguistics? *Language and linguistics compass 3*, 1–17.

Gries, S. T. and A. Stefanowitsch (2004). Extending collostructional analysis. *International journal of corpus linguistics 9*(1), 97–129.

Grondelaers, S., D. Geeraerts, and D. Speelman (2007). A case for a cognitive corpus linguistics. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, and M. J. Spivey (Eds.), *Methods in Cognitive Linguistics*, pp. 149–169. Amsterdam: John Benjamins Publishing Company.

Grondelaers, S., D. Speelman, and D. Geeraerts (2002). Regressing on *er*: Statistical analysis of texts and language variation. In A. Morin and P. Sébillot (Eds.), *6th international conference on the statistical analysis of textual data*, Rennes, pp. 335–346. Institut Nationale de Recherche en Informatique et en Automatique.

Guiraud, P. (1959). *Problèmes et Méthodes de la statistique linguistique*. Dordrecht: Reidel.

Guttman, B. S. (2005). *Evolution: A beginner's guide*. Oxford: Oneworld.

Guy, G. R. (2003). Variationist approaches to phonological change. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*, pp. 369–400. Malden, MA.: Blackwell Publishing.

Halliday, M. (1992). Language as system and language as instance: The corpus as theoretical construct. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*, pp. 61–78. Berlin: Mouton de Gruyter.

Hammond, M. (2003). *Programming for linguists: Perl for language researchers*. Oxford: Blackwell.

Harrell, F. E. (2009). *Hmisc: Harrell Miscellaneous*. R package version 3.7-0.

Harrell, F. E., K. L. Lee, and D. B. Mark (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*(4), 361–387.

Hatcher, J. (1977). *Plague, population and the English economy 1348–1530*. London: Macmillan.

Hauser, M. D., N. Chomsky, and W. T. Fitch (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science 298*(5598), 1569–1579.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hertwig, R. and T. J. Pleskac (2008). The game of life: How small samples render choice simpler. In N. Chater and M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science*, pp. 209–235. Oxford: Oxford University Press.

Hill, A. A. (1961). Grammaticality. *Word 17*, 1–10.

Hinton, P. R. (2004). *Statistics explained* (2nd ed.). London: Routledge.

Hirt, H. (1934). *Handbuch des Urgermanischen*, Volume III: Abriss der Syntax. Heidelberg: Carl Winter.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Hopper, P. J. and E. C. Traugott (1993). *Grammaticalization*. Cambridge: Cambridge Univeristy Press.

Horning, J. J. (1969). *A study of grammatical inference*. Ph. D. thesis, Stanford, CA.

Hovorka, D. S., M. Germonprez, and K. R. T. Larsen (2002). Explanation in information systems. *Sprouts: Working papers on information environments, systems and organizations 3*(3), 169–187.

Itkonen, E. (1981). Untitled review of *On explaining language change*. *Language 57*(3), 688–697.

Janda, R. D. (2001). Beyond "pathways" and "unidirectionality": On the discontinuity of language transmission and the counterability of grammaticalization. *Language sciences 23*, 265–340.

Jenset, G. B. (2005). "That ther lakke no word…": A cognitive study of existential *there* in the works of Geoffrey Chaucer. Master's thesis, english language and linguistics, University of Bergen, Bergen.

Jenset, G. B. (2008). Existential *there* beyond grammaticalization. In G. B. Jenset, Ø. Heggelund, M. D. Cardona, S. Wold, and A. Didriksen (Eds.), *Linguistics in the making: Selected papers from the second Scandinavian PhD conference in linguistics and philology in Bergen 4–6 June, 2007*, pp. 57–75. Oslo: Novus Press.

Jespersen, O. (1924). *The philosophy of grammar*. London: Allen & Unwin.

Jespersen, O. (1969). *Analytic syntax* (Reprint ed.). New York: Holt, Rinehart and Winston.

Johansson, C. (1997). *A view from language*. Lund: Lund University Press.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The journal of wildlife management 63*(3), 763–772.

Johnson, K. (2008). *Quantitative methods in linguistics*. Oxford: Blackwell Publishing.

Joseph, B. D. (2001). Is there such a thing as "grammaticalization?". *Language sciences 23*, 163–186.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, and S. Jannedy (Eds.), *Probabilistic linguistics*, pp. 39–96. Cambridge, MA.: The MIT Press.

Jurafsky, D., A. Bell, M. Gregory, and W. D. Raymond (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*, pp. 229–254. Amsterdam: John Benjamins publishing company.

Just, M. A. and P. A. Carpenter (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review 99*(1), 122–149.

Keller, R. (1994). *On language change: The invisible hand in language*. London: Routledge. Translated by Brigitte Nerlich.

Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association 61*(313), 11–34.

Kempthorne, O. (1979). In dispraise of the exact test: Reactions. *Journal of statistical planning and inference 3*, 199–213.

Kempthorne, O. and T. Doerfler (1969). The behaviour of some significance tests under experimental randomization. *Biometrika 56*(2), 231–248.

Ker, N. R. (1957). *Catalogue of manuscripts containing Anglo-Saxon*. Oxford: Clarendon Press.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory 1*(2), 263–276.

Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language variation and change 1*, 199–244.

Kroch, A., B. Santorini, and L. Delfs (2004). *Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/.

Kroch, A. and A. Taylor (2000). *Penn-Helsinki Parsed Corpus of Middle English* (2nd ed.). http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/.

Kroeber, A. L. and C. D. Chrétien (1937). Quantitative classification of Indo-European languages. *Language 13*(2), 83–103.

Labov, W. (2008). Triggering events. In S. M. Fitzmaurice and D. Minkova (Eds.), *Studies in the history of the English language*, Volume IV: Empirical and analytical advances in the study of English language change, pp. 11–54. Berlin: Mouton de Gruyter.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Langacker, R. W. (2008). *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Lehmann, C. (1985). Grammaticalization: Synchronic variation and diachronic change. *Lingua e stile 20*(3), 303–318.

Lehmann, W. (1975). The challenge of history. In R. Austerlitz (Ed.), *The scope of American linguistics: The first golden anniversary symposium of the Linguistic Society of America*, pp. 41–58. Lisse: Peter de Ridder Press.

Lemay, L. and R. Colburn (2002). *Sams teach yourself Perl in 21 days* (2nd ed.). Indianapolis, Ind: Sams.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics 20*(1), 1–31.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Lieberman, E., J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak (2007). Quantifying the evolutionary dynamics of language. *Nature 449*(7163), 713–716.

Lieberman, P. (2000). *Human language and our reptilian brain: The subcortical bases of speech, syntax, and thought*. Cambridge, MA: Harvard Universit Press.

Lightfoot, D. (2006). *How new languages emerge*. Cambridge: Cambridge University Press.

Lightfoot, D. (2007). Abstraction and performance: Commentary on Fischer. In M. Penke and A. Rosenbach (Eds.), *What counts as evidence in linguistics*, pp. 283–286. Amsterdam: John Benjamins.

Limpert, E., W. A. Stahel, and M. Abbt (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience 51*(5), 341–352.

Long, J. S. and J. Freese (2001). *Regression models for categorical dependent variables using STATA*. College Station, Tex.: Stata Press.

Lüdeling, A. and S. Evert (2005). The emergence of productive non-medical *-itis*: Corpus evidence and qualitative analysis. In S. Kepser and M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, pp. 351–370. Berlin: Mouton de Gruyter.

Mair, C. (2004). Corpus linguistcs and grammaticalization theory: Statistics, frequencies, and beyond. In H. Lindquist and C. Mair (Eds.), *Corpus approaches to grammaticalization in English*, pp. 121–150. Amsterdam: Jonn Benjamins Publishing Company.

McCawley, J. D. (1982). *Thirty million theories of grammar*. Chicago: University of Chicago Press.

McEnery, T. and A. Wilson (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McMahon, A. (1994). *Understanding language change*. Cambridge: Cambridge University Press.

Mitchener, W. G. (2006). A mathematical model of the loss of verb-second in Middle English. In N. Ritt, H. Schendl, C. Dalton-Puffer, and D. Kastovsky (Eds.), *Medieval English and its heritage*, pp. 189–202. Frankfurt am Main: Peter Lang.

Mithun, M. (2003). Functional perspectives on syntactic change. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*, pp. 552–572. Malden, MA.: Blackwell.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association 321*(63), 1–28.

Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological review 67*(1), 16–31.

Nagashima, D. (1992). *A historical study of the introductory* there (Reprint ed.). Osaka: Kansai Gaidai University.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*(3), 691–692.

Neath, I. (1993a). Contextual and distinctive processes and the serial position function. *Journal of Memory and Language 32*(6), 820–840.

Neath, I. (1993b). Distinctiveness and serial position effects in recognition. *Memory & cognition 21*(5), 689–698.

Nenadic, O. and M. Greenacre (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software 20*(3), 1–13.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, somers' D and median differences. *Stata Journal 2*(1), 45–64.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Pagel, M., Q. D. Atkinson, and A. Meade (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature 449*(7163), 717–720.

Panther, K.-U. and L. Thornburg (1998). A cognitive approach to inferencing in conversation. *Journal of pragmatics 30*, 755–769.

Pearl, J. (2000, January). Simpson's paradox: An anatomy. Department of Statistics Papers 2000010109, Department of Statistics, UCLA.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South Central SAS User's Group (SCSUG-96) Conference*, pp. 188–200.

Pedersen, T. and R. Bruce (1996). What to infer from a description. Technical report, Southern Methodist University.

Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society 358*, 1239–1253.

Pérez-Guerra, J. (1999). *Historical English syntax: A statistical corpus-based study on the organisation of Early Modern English sentences*. Number 11 in LINCOM studies in Germanic linguistics. München: LINCOM.

Pfenninger, S. E. (2009). *Grammaticalization paths of English and High German existential constructions: A corpus-based study*. Bern: Peter Lang.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer Verlag.

Pintzuk, S. (2003). Variationist approaches to syntactic change. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*, pp. 509–528. Malden, MA.: Blackwell Publishing.

Pullum, G. (2009). Computational linguistics and generative linguistics: The triumph of hope over experience. In T. Baldwin and V. Kordoni (Eds.), *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, Athens, Greece, pp. 12–21. Association for Computational Linguistics.

Quirk, R. and C. L. Wrenn (1957). *An Old English grammar* (2nd ed.). London: Routledge.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Radford, A. (1997). *Syntactic theory and the structure of English: A minimalist approach*. Cambridge: Cambridge University Press.

Rayson, P., G. Leech, and M. Hodges (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics 2*(1), 133–152.

Rissanen, M. (1990). On the happy reunion of English philology and historical linguistics. In J. Fisiak (Ed.), *Historical linguistics and philology*, pp. 353–370. Berlin: Mouton de Gruyter.

Rissanen, M. (1992). The diachronic corpus as a window to the history of English. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*, pp. 185–205. Berlin: Mouton de Gruyter.

Rosch, E. (1975). Cognitive reference points. *Cognitive psychology 7*, 532–547.

Rosenthal, R., R. L. Rosnow, and D. B. Rubin (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge: Cambridge University Press.

Roy, J. R. (2006). The constant effect in the spread of syntactic change: Applications of alternating logistic regressions in historical linguistic repeated response data. Ms., The University of Texas at San Antonio, San Antonio, TX.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association 81*(396), 961–962.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of statistical planning and inference 25*, 279–292.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance 16*(2), 26–32.

Sampson, G. R. (2001). *Empirical linguistics*. London: Continuum.

Sampson, G. R. (2005). Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics 10*, 10–36.

Sampson, G. R. (2007). Grammar without grammaticality. *Corpus linguistics and linguistic theory 3*(1), 1–32.

Sankoff, D. and W. Labov (1979). On the uses of variable rules. *Language in society 8*(2), 189–222.

Saussure, F. d. (1983). *Course in general linguistics*. London: Duckworth. Translated & annotated by Roy Harris.

Stefanowitsch, A. (2005). New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus linguistics and linguistic theory 1*(2), 295–301.

Stefanowitsch, A. (2006). Distinctive collexeme analysis and diachrony: A comment. *Corpus linguistics and linguistic theory 2*(2), 257–262.

Stefanowitsch, A. (2007). Linguistics beyond grammaticality. *Corpus linguistics and linguistic theory 3*(1), 57–72.

Stefanowitsch, A. and S. T. Gries (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics 8*(2), 209–243.

Steup, M. (2005). Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Stewart, I. and J. Cohen (1997). *Figments of reality: The evolution of the curious mind*. Cambridge: Cambridge University Press.

Stratmann, F. H. (1940). *A Middle-English dictionary* (New ed.). London: Oxford University Press.

Szmrecsányi, B. M. (2004). On operationalizing syntactic complexity. In *7^{th} international Conference on Textual Data Statistical Analysis (*JADT*): March 10-12 2004*, Louvain-la-Neuve, pp. 1031–1038. Journées internationales d'Analyse statistique des Données Textuelles.

Talmy, L. (2007). Foreword. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, and M. J. Spivey (Eds.), *Methods in Cognitive Linguistics*, pp. xi–xxi. Amsterdam: John Benjamins Publishing Company.

Taylor, A., A. Warner, S. Pintzuk, and F. Beths (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. http://www-users.york.ac.uk/ lang22/YCOE/YcoeHome.htm.

Taylor, J. R. (2002). *Cognitive Grammar*. Oxford: Oxford University Press.

Therneau, T. M. and B. Atkinson (2009). *rpart: Recursive Partitioning*. R package version 3.1-45; R port by Brian Ripley.

Traugott, E. C. (1992). Syntax. In R. Hogg (Ed.), *The Cambridge History of the English Language*, Volume I: Old English, pp. 168–289. Cambridge: Cambridge University Press.

Tversky, A. and D. Kahneman (1971). Belief in the law of small numbers. *Psychological bulletin 76*(2), 105–110.

Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society. Series A(Statistics in society) 155*(3), 395–402.

Warren, T. and E. Gibson (2002). The influence of referential processing on sentence complexity. *Cognition 85*, 79–112.

Woodward, J. (2003). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Index