Techniques and tools

Corpus methods and statistics for semantics

Dylan Glynn
University of Paris VIII

The use of corpora in semantic research is a rapidly developing method. However, the range of quantitative techniques employed in the field can make it difficult for the non-specialist to keep abreast with the methodological development. This chapter serves as an introduction to the use of corpus methods in Cognitive Semantic research and as an overview of the relevant statistical techniques and software needed for performing them. The discussion and description are intended for researches in semantics that are interested in adopting quantitative corpus-driven methods. The discussion argues that there are fundamentally two corpus-driven approaches to meaning, one based on observable formal patterns (collocation analysis) and another based on patterns of annotated usage-features of use (feature analysis). The discussion then introduces and explains each of the statistical techniques currently used in the field. Examples of the use of each technique are listed and a summary of the software packages available in R for performing the techniques is included.

Keywords: collocation analysis, corpus linguistics, semantics, statistics, usage-feature analysis (behavioural profile)

1. Introduction

This chapter offers an explanation of the corpus methods represented in the book and a brief overview of the various statistical techniques employed. It is designed as a resource for those less familiar with the field, but also as a reference for those already working with corpus-driven methods in Cognitive Semantics. Specifically, corpus-driven Cognitive Semantics is understood as the work beginning with Dirven *et al.* (1982), Schmid (1993, 2000), Geeraerts *et al.* (1994, 1999) and Gries (1999, 2003), and currently represented in the edited volumes of Gries and Stefanowitsch (2006), Stefanowitsch and Gries (2006), Lewandowska-Tomaszczyk and Dziwirek (2009), Glynn and Fischer (2010), Geeraerts *et al.* (2010), Divjak and Gries (2012), Gries and

Divjak (2012), Pütz *et al.* (2012), Reif *et al.* (2013), Glynn and Sjölin (2014), and in the monographs Hilpert (2008, 2012), Divjak (2010a), Gilquin (2010), Dziwirek and Lewandowska-Tomaszczyk (2011), Hoffmann (2011), and Glynn (forthc.).

In this chapter, corpus-driven Cognitive Semantics is argued to divide into two methodologies, or analytical approaches, based either on the formal analysis of collocations or the semantic analysis of features. This proposed distinction is described in Section 2. Following this, Section 3 describes the quantitative techniques used in such research. It lists and explains the techniques and offers examples of how they are used, giving detailed references on where the application of each technique is explained in the literature.

2. Collocations and features: Two approaches to corpora

A common misconception amongst cognitive linguists is that corpus-driven research, and indeed, the quantitative analysis of corpus data, does not involve any close analysis of actual examples. This is not necessarily true at all. Within Cognitive and Functional Linguistics, broadly speaking, there is a wide range of approaches to corpus data, from simply counting the number of occurrences of a given form in a given context to the development of complex computational models trained on enormous text banks. For corpus-driven research in semantics, where the 'meaning' of a given linguistic form is in question, it is possible to broadly identify at least two approaches. All the studies in the first section of this book fall into one of these two categories. The first of these is based on formal, and therefore, observable, patterns. We can term this approach 'collocation analysis'. Secondly, the corpus analysis can be based on patterns of annotated features, which we term 'feature analysis'. In the former, the analysis seeks to identify formal patterns so as to interpret them as indices of meaning structure and in the latter, the analysis seeks to directly identify semantico-pragmatic patterns through close manual annotation. Although the approaches can be combined (cf. Stefanowitsch and Gries 2008), they tend to be used separately and possess distinct strengths and weaknesses.

The first 'type' of corpus-driven research, collocation analysis, is more established and is typical of mainstream Corpus Linguistics. Collocation studies identify the co-occurrence of linguistic forms in a given sample of naturally occurring language. Firth's (1957: 179) now famous phrase, "you shall know a word by the company it keeps", is a succinct way of capturing the aim of this approach. When extended to other parts of language, such as syntactic patterns or indeed text types and genres, the large-scale study of collocation is a powerful tool for making generalisations about language use. Cognitive and Functional Linguistics are particularly concerned with why a given form is used and so it follows that in order to answer research questions of this nature, inferences as to the semantic, functional, or conceptual motivation for the collocation must be made in *post hoc* interpretation.

Despite this subjective step in the use of collocation analysis in Cognitive-Functional Linguistic research, the analytical approach has important advantages. To the extent that one can retrieve forms automatically, one can consider extremely large samples, making studies (relatively) representative of a given language or part of language. Secondly, forms are objectively identifiable, making this step largely independent of subjective analysis. However, this statement warrants qualification. Even if a form is objectively identifiable, linguists are typically interested in only certain uses of a given form and, often, these specific uses cannot be retrieved automatically. In such situations, the decision as to which occurrences are representative of the category is typically a question for debate (cf. Perek, this volume, 61–86).

Moreover, collocation studies rely on some measurement of association. Raw frequency of co-occurrence can be misleading because if one of the forms is extremely frequent, then relatively high co-occurrence may just be a result of the overall high frequency of that form. The problem of how to determine the degree of association, or 'attraction', is fundamental. Common ways of measuring the degree of association for lexical co-occurrence are the mutual information (MI) score, the *z*-score (standard score), the *t*-score and the log-likelihood. Many Corpus Linguistics programs, both on-line and stand-alone, automatically generate some of these scores. Collostructional analysis is one alternative to such measures. Developed by Stefanowitsch and Gries (2003, 2005) and Gries and Stefanowitsch (2004a, 2004b) and described in Hilpert (this volume, 391–404), it is a suite of methods that use the Chi-squared or Fisher exact test to compute degree of association. These techniques allow the researcher to consider the co-occurrence, not just of lexemes, but also of syntactic patterns. Collostructional analysis has proven popular in Cognitive Linguistics.

One of the newest advances in the use of collocation is the application of Word Space modelling to semantic research questions within computational linguistics. The principle is to extend the analysis of collocation beyond one or two words or even syntactic patterns, to whole lines, paragraphs and even entire texts. Such approaches give rich collocation-based behavioural profiles of a given linguistic form. The implications for such analytical techniques in semantics are only now being realised. This methodology is not represented in the volume. Peirsman *et al.* (2010) and Sagi *et al.* (2001) are examples of the application of these methods to research in semantic relations.

The number of studies employing a collocation approach, even restricted to Cognitive Linguistics, is enormous. A small sample of recent studies includes Newman and Rice (2004, 2006), Deignan (2005), Delorge (2009), Pęzik (2009), Van Bogaert (2010), Colleman (2010) and Zeschel (2010). Applications of collostructional analysis include Wulff (2006), Wulff *et al.* (2007), Hilpert (2008, 2009) and Gilquin (2010).

In general terms, it is possible to identify a second quantitative approach in corpus-driven Cognitive Semantics, one that focuses on the manual analysis of usage-features. Although less traditional in the mainstream of Corpus Linguistics, the general principle has a long tradition in Cognitive Linguistics (Dirven *et al.* 1982; Rudzka-Ostyn 1989, 1995; Fillmore and Atkins 1992; Geeraerts *et al.* 1994) and, more recently, is gaining currency in Functional Linguistics (Fischer 2000; Scheibman 2002; Kärkkäinen 2003; Pichler 2013). The principle of combining the results of this usage-feature analysis with multivariate statistics begins with Geeraerts *et al.* (1999) and Gries (2003). It is termed the behavioural-profile approach by Gries and Divjak (2009) and Divjak and Gries (2009) and multifactorial usage-feature analysis by Glynn (2009, 2010b).[1] The principle is simple: for a large sample of a given linguistic phenomenon, various formal, semantic, and/or social 'linguistic features' (or 'ID tags' in the terminology of Gries and Divjak 2009) are identified and ascribed to each occurrence. It is worth noting that the method *per se* has also been independently developed in social psychology and computational linguistics. In the former, it is termed the analysis of components (cf. Scherer 2005; Fontaine *et al.* 2013) and in the latter, sentiment analysis (Wiebe *et al.* 2005; Verdonik *et al.* 2007; Daille *et al.* 2011; Balahur and Montoyo 2012; Read and Carroll 2012; Taboada and Carretero 2012).

The approach consists of the repeated application of what is essentially a 'traditional' linguistic analysis to hundreds, or even thousands, of naturally occurring examples. This procedure results in a quantified usage-profile of the linguistic phenomenon in question. Usage-feature analysis is employed, with varying degrees of statistical sophistication, to examine phenomena of all kinds, from syntactic variation and semantics (Heylen 2005a; Bresnan *et al.* 2007; Speelman *et al.* 2009), to discourse studies and conversation analysis (Scheibman 2002; Kärkkäinen 2003; Flores Salgado 2011; De Cock 2014a, 2014b), and even gesture research (Zlatev and Andrén 2009; Morgenstern *et al.* 2011).

The limitations of the approach are twofold. Firstly, the detailed manual analysis is as subjective as any traditional linguistic analysis and is open to the same vagaries, theoretical biases and human error. Secondly, the manual analysis, or annotation, of examples is meticulous and laborious. This, combined with the simple practical reality of limited resources, means that samples are relatively small. The resulting sample

1. Since Multifactorial Usage-Feature Analysis (Behavioural Profile Approach) is less known in corpus circles, a selection of current examples of its use include: Geeraerts *et al.* (1999), Gries (1999, 2003, 2006, 2010), Szmrecsanyi (2003, 2010), Wulff (2003, 2009), Heylen (2005a), Divjak (2006, 2010a, 2010b), Divjak and Gries (2006, 2009), Bresnan *et al.* (2007), Grondelaers *et al.* (2007, 2008), Glynn (2009, 2010a, 2010b, 2014a, 2014b, forthc.), Janda and Solovyev (2009), Speelman *et al.* (2009), Speelman and Geeraerts (2010), Krawczak and Glynn (2011, in press), Krawczak and Kokorniak (2012), Levshina (2012), Levshina *et al.* (2013a, 2013b), Krawczak (2014a, 2014b), and Deshors (2014). Doctoral dissertations focusing on developing the method include Gries (2000), Grondelaers (2000), Heylen (2005a), Glynn (2007), Arppe (2008), Robinson (2010b), Deshors (2011), Levshina (2011), Barnabé (2012), Klavan (2012), and Diehl (2014).

Table 1. Observational differences in collocation and feature analysis of corpora

	Collocation	Feature
Stage 1: Analysis of data	objective	subjective
Stage 2: Interpretation of analysis	subjective	objective

size makes it more difficult to be sure of representativity and harder to obtain statistically significant results.

The advantages of the approach are also twofold. Firstly, the method allows the operationalisation and quantification of traditional linguistic analyses. This is no trivial matter because it permits hypothesis testing and produces falsifiable results for research questions not easily approached using traditional corpus methods (c.f. Geeraerts 2010; Glynn 2010b; Stefanowitsch 2010). Secondly, an important strength lies in the possibility of treating the results obtained through the usage-feature analysis with multivariate statistics. This is especially important for non-modular theories of linguistics, such as Cognitive Linguistics, because multivariate statistics permits an analysis to handle the complexity of the interaction of the different dimensions of language structure simultaneously (such as lexis, syntax, phonology, society, etc.), creating a multidimensional and socio-conceptually realistic profile of the use of a linguistic form or the role of a linguistic function.

Geeraerts (2011) compared the two corpus approaches, underlining that both are subjective, but are at different stages in their application. Table 1 summarises Geeraerts' point about subjectivity.

Juxtaposing the two analytical approaches like this is, of course, a simplification. At the first stage of analysis, collocation studies are often not entirely objective because of questions such as what constitutes a 'form'. Firstly, forms are polysemous and only certain uses may be relevant for a given study. In such a situation, manual selection is often the only solution. Secondly, the forms themselves are typically composite and so formal variation itself can cause category issues. In other words, is a given formal variant an example of the form in question or is it a 'different' form? Again, in such situations, subjective categorisation enters the analysis. Turning to feature analysis, the subjective first step is not always particularly subjective. Often, feature analysis is largely based on observable phenomena. For example, grammatical features can be crucial to usage-feature analysis and are annotated automatically, or if done manually, are done so objectively.

At the stage of interpretation the same objective-subjective blurring occurs. For collocation analysis, as Desagulier (this volume, 145–178) shows, statistical analysis can help add a degree of objectivity to interpreting the collocation patterns observed. A similar caveat is needed for the usage-feature method. Although multivariate statistics may help us to objectively distinguish semantico-pragmatic patterns from non-patterns, we still must decide if those patterns answer the research question at hand, which is an inherently subjective step.

3. Statistical techniques and tools

Often one of the most confusing issues in the application of quantitative techniques to linguistic research is the myriad of different techniques available. This section is primarily intended for the reader who has some experience with quantitative methods, presenting an overview of the techniques relevant to corpus linguistic research. For the reader who has little experience in quantitative techniques, the overview will be technical, but it is hoped, still informative.

It is important to understand that statistics is a rapidly growing science with constant new advances as well as many uncertainties and conflicts. Perhaps more importantly, we must also remember that statistical techniques are only analytical tools. No statistical technique will identify a linguistic fact or explain any linguistic structure. Nevertheless, statistical tools can be used by linguists to help look for language structure – assuming one knows where to look. They can also be used to confirm the probability that the results of an analysis are not a chance occurrence. Statistics can help linguists struggle with what they have been doing for centuries, describe and explain language, but they are only tools in that endeavour.

Just as there is sometimes a misconception that statistics can answer linguistic questions, there exists a misconception that quantitative corpus-driven research is devoid of 'real' linguistic analysis. Nothing is further from the truth. Corpus-driven linguists deal with real language and in large quantities. The 'numbers' presented in corpus-driven research are not the analysis; they are a quantitative summary of the analysis, which must, in turn, be interpreted. Corpus-driven linguists, for the most part, deal with language in a relatively close and fine-grained way; they just deal with large quantities of it.

One of the aims of this book is to showcase and explain the use of a small set of statistical techniques that can be helpful for traditionally trained linguists in their research. The aim is not to teach statistics or the computer programs for performing statistical analyses, but simply to introduce some of the possibilities. In this section, we begin with a short description of the computer applications available for performing statistics, and then briefly consider a fundamental theoretical question for the statistical sciences – type of data. This question is essential to understand before one can decide which statistical techniques are appropriate in a given situation. This is followed by a systematic summary of the techniques currently used in the field, examples of their use, as well as examples of texts that explain how they are used. The description ends with a detailed list of the different commands and packages for performing these statistical techniques in the programming suite R.

Statistical software

There are many computer applications, commercial and otherwise, that enable the researcher to perform statistical analysis. In this volume, the statistical program that is used by most authors is R. This program is, in fact, a powerful programming suite with enormous potential. The explanatory chapters all use R and the reader is taken step-by-step through the necessary "code", or command lines, needed to perform the analyses. No attempt is made to demonstrate the full functionality of the program, merely to offer a working knowledge of how to perform specific analyses.

This volume focuses upon R for three reasons. Firstly, it is a free and cross-platform program. Secondly, since it is open source, as soon as new statistical techniques develop, new software modules are written and uploaded for the public. Thirdly, the programme is one of the two most commonly used programs for statistics in the social sciences (there are, of course, many more, especially devoted to specific techniques). The other most frequently used program in the social sciences is SPSS. Like R, it is also an extremely powerful tool, as widely used, but also includes a graphic user interface (unlike R). Since R is equally powerful, arguably more up to date, entirely free and used by the majority of authors in the book, the only negative is its command-line interface. However, in the following chapters, the command-line is given simple step-by-step instructions and, it is hoped, will not pose too many problems for the beginner. It is true that the command-line may seem daunting at first, but if the steps are followed line-by-line, the only difference with 'button-for-button' (as in a graphic interface application) is one of familiarity.

Other important application suites include SAS, Statistica, and Stata, which are all powerful and versatile. SAS is command-line, like R and in some ways, R can be seen as the open-source version of SAS. It is arguably the most complete statistical programming suite, but is rarely used in the social sciences. Statistica and Stata are comparable to SPSS. They too have graphic user interfaces, are relatively user friendly and, just like SPSS, are costly. Statistica is restricted to the Windows operating systems, but has a relatively large and helpful online community. Stata is cross-platform, but is probably less common than Statistica. It is not really possible to say which suite is the best, since certain techniques are extremely well covered in one suite and not the other. Due to its being open source, R is surely the suite with the most options and also the quickest to respond to developments within the domain of statistics, but, of course, that does not mean its implementation of those techniques is the best.

If the reader is familiar with any of these other programs, the descriptions of the statistical techniques in the book, as well as their interpretation and application, will still be useful. Lastly, it should be noted that a graphic user interface is under development for R. This is not drawn upon because its development is not yet complete and the commands/R sessions described in this book are sufficiently straightforward that readers who are not familiar with statistics or command-line will not have problems following.

Types of data

Before choosing a statistical technique, one must first know what 'type' of data one is dealing with. This is because different types of data require different statistical techniques. The most basic distinction is between what is called *continuous* data and *categorical* data. The former typically come from measurements and therefore make a continuum, for example 1.0, 1.1, 1.2 … 1.8, 1.9, 2.0. This kind of data is probably the most common and comes from diverse sources such as age, time, height, dosage, temperature, response times, and, arguably, grammatical judgements. Continuous data are typical in psychology and psycholinguistics. The second kind of data is categorical, also called 'discrete data', 'tabular data' or 'count data'. It is this kind of data, as corpus linguists, with which we are most often concerned. Such data include, for example, the frequency of occurrence of a linguistic form, the number of times it occurs in a given tense, or in a given register. In these examples, the data are said to be *nominal* because each of the occurrences is independent from the other. However, categorical data can also be *ordered*. This is the case when, for example, the categories follow a natural sequence or ranking, such as young, middle-aged, and old or when a sentence is short, medium or long in length. Ordered categorical data share properties of both nominal categorical and continuous data. Grammatical judgements, on a scale of 1 to 7, for example, could be argued to be continuous or ordered categorical. Technically, it is ordered because a respondent cannot enter 3.5, for example, but is forced to make a discrete choice upon what is, in reality, a continuous scale of acceptability. However, if we assume that no respondent would perceive differences to the degree of 3.5, then we can treat the scale as a true measurement, and therefore, continuous.

Table 2. Types of data in statistics

Data type	Example of data	Description	Example of use
Continuous	1, 1.1, 1.2 … 1.8, 1.9, 2 1, 2, 3, 4, 5, 6, 7	Sequential (ordered) but non-discrete / continuous	Response times in Psycholinguistics
Ordered	short, medium, long cold, warm, hot	Sequential (ordered) but discrete / non-continuous	Different periods in diachronic linguistics
Nominal	apples, peaches, pears *y'all*, *you lot*, *youse*	Independent and discrete categories	Different lexemes in Corpus Linguistics

Although there is occasionally debate on the issue, most statistical techniques are designed for one of the kinds of data. For example, least squares estimation and linear regression are used for continuous data and maximum likelihood and logistic regression for categorical, just as principle components analysis is used for continuous data and correspondence analysis for categorical. Table 2 summarises the differences.

Statistical techniques for corpus linguistics

Statistics is an immense science – there are countless tests and corrections for those tests. There are even more exploratory techniques with various algorithms that each technique can employ and different ways for representing results of those exploratory techniques. Confirmatory analysis has again as many different techniques, but this time, seemingly endless sets of diagnostics to check the validity of the results. It must be stressed that the techniques presented here only scratch the surface of what is possible, but also of what problems exist.

We begin with significance tests and association measures. Although not statistical techniques *per se*, they are tools that are important to the field. We then cover exploratory methods, the results of which cannot be used to make claims about structure beyond the sample. In other words, what is found with these techniques may be restricted to the corpus or the extract of the corpus being examined. These exploratory techniques do not test hypotheses or make predictions about the population (real language). The description then turns to confirmatory techniques, which are more complex in their application but which make predictive claims and can test hypotheses in terms of statistical significance or the probability that observed structures exist in real language beyond the sample.

Sample, significance and independence

Establishing that the occurrence of something in a given sample is more or less common than would be expected by chance or that two sets of data are more different than would be expected by chance are basic steps in inductive research. Pearson's Chi-squared test and Fisher's Exact test are omnipresent in research based on samples of categorical data. Gries (this volume) explains these tests and shows how to apply them in R. Other tests useful for corpus data include the exact binomial test, McNemar's paired Chi-squared test, and the proportions test. These are used for investigating relations in frequency tables. An excellent explanation of these tests and their commands in R can be found in Dalgaard (2008: Ch. 8) and Baayen (2008: Section 4.1.1). See also Gries (2009b: 125–127, 158–176; 2013: 165–172), Everitt and Hothorn (2010: Ch. 3), and Adler (2010: 360–367).

Collocation and association measures

Within Cognitive Linguistics, collostructional analysis has proven to be one of the most important methods for investigating collocations. Developed by Stefanowitsch and Gries (2003, 2005) and Gries and Stefanowitsch (2004a, 2004b), the principle can be combined with a range of association measures for determining the degree of collocational 'strength' (the measure is typically calculated with a p-value obtained from a Fisher exact test, log-transformed). These calculations are not yet implemented in most corpus annotation or concordance software. However, Stefan Gries has developed R scripts (semi-automated sets of commands) for performing the tests.[2] Hilpert (this volume, 391–404) explains three varieties of collostructional analysis: collexeme analysis, distinctive collexeme analysis, and covarying collexeme analysis.

---

2. For more information, contact Stefan Th. Gries. His contact details can be found on his website: http://www.linguistics.ucsb.edu/faculty/stgries/.

Examples of use include Wulff (2003), Hilpert (2008), Stefanowitsch and Gries (2008), Colleman (2009), and Gilquin (2010).

The aim of quantifying degree of association between two forms in terms of frequency is not unique to the collostructional suite. Corpus Linguistics has developed an array of calculations to determine relative degree of association, especially between individual words. The most common are the mutual information (MI), the *z*-score, the *t*-score, and the log-likelihood. There is important variation in the results obtained from using any one test over another. Evert (2009) offers a detailed discussion on the matter; see also Wiechmann (2008), Wulff (2010) and Desagulier (this volume, 145–178). The *z*-score and the *t*-score are both explained with the R-code in Johnson (2008: Ch. 3) and Dalgaard (2008: Ch. 5). The freely available Ngram Statistics Package extracts sequences from a corpus and calculates a range of association measures. All these scores are used extensively in collocation-based corpus linguistics.

Cluster analysis

Cluster analysis is a diverse family of techniques, which, as the name suggests, cluster data. *K*-means clustering is used when one knows how many clusters there should be in advance; the technique 'sorts' the data accordingly. More common in semantic research is hierarchical clustering, which is used as an exploratory technique for the identification of clusters in the data. Importantly, by identifying clusters, it also sorts the data into the clusters it has 'discovered'. The technique begins with a set of features and then uses them to group the features of a given variable (for instance, a list of senses, concepts, words, or constructions). It represents the results in a dendrogram, a kind of plot that depicts groups in an intuitively transparent way as dependencies clustered in branches. Cluster analysis is an excellent technique for determining which forms are similar to each other and which are different.

It is explained by Divjak and Fieller (this volume, 405–442). Other explanations using R code include Crawley (2007: 742–744), Baayen (2008: 138–148), Johnson (2008: Ch. 6), and Ledolter (2013: Ch. 15). Härdle and Simar (2007: Ch. 11), Izenman (2008: Ch. 12), Drenan (2009: Ch. 25), Everitt and Hothorn (2010: 18), Afifi *et al.* (2011: Ch. 16) and Marden (2011: Ch. 12) represent detailed, yet approachable, explanations without R code. Everitt *et al.* (2011) is surely the most comprehensive work devoted to the technique, and although quite technical, is a systematic and excellent reference for using cluster analysis. The book provides no explanations for performing the analysis, but does give information on which software packages are available for many of the analyses it describes.

Examples of use in Cognitive Linguistics include Schulze (1991), Chaffin (1992), Myers (1994), Sandra and Rice (1995), Ravid and Hanauer (1998), Rice *et al.* (1999), Gries (2006), Divjak (2006, 2010a), Divjak and Gries (2006), Gries and Hilpert (2008), Valenzuela Manzanares and Rojo López (2008), Janda and Solovyev (2009), Louwerse and Van Peer (2009), Robinson (this volume, 87–116), Glynn (2010a, 2014a, 2014b, this volume, 117–144), Levshina (2012), Szmrecsanyi (2013), and Krawczak and Glynn (in press).

Correspondence analysis

Correspondence analysis is an exploratory technique that helps identify associations in the data, such as patterns in the combinations of linguistic features. The technique is designed for dealing with complex interactions where it is not known *a priori* which dimension, be that syntax, semantics, pragmatic, or social context, that structures the behaviour of the data. For instance, it can help find which semantic features typically occur with a set of grammatical forms or constructions, but also how these two dimensions interact relative to social variation. It visualises these associations in biplots, which, although arguably difficult to interpret, represent rich depictions of complex structures.

Glynn (this volume, 443–486) explains the application and interpretation of two varieties of correspondence analysis: binary correspondence and multiple correspondence analysis. There exist several comprehensive books devoted to the technique: Benzécri (1980, 1992), Murtagh (2005), Greenacre (2007 [1993], 2010), and Le Roux and Rouanet (2010). Amongst these, Greenacre (2007) is probably the standard book of reference. Useful introductions include Le Roux and Rouanet (2004: Chs. 2 and 5), Everitt (2005: Ch. 5), Härdle and Simar (2007: Ch. 13), Baayen (2008: Ch. 5), Izenman (2008: Ch. 17), and Husson *et al.* (2011: Chs. 2 and 3). The last of these, Husson *et al.* (2011), is particularly clear and includes some of the most recent developments.

Examples of use include Arppe (2006), Glynn (2007, 2009, 2010a, 2010b, 2014a, 2014b, this volume, 117–144), Szelid and Geeraerts (2008), Plevoets *et al.* (2008), Glynn and Sjölin (2011), Krawczak and Glynn (2011), Barnabé (2012), Krawczak and Kokorniak (2012), Nordmark and Glynn (2013), Levshina *et al.* (2013b), Desagulier (this volume, 145–178; in press), Delorge *et al.* (this volume, 39–60), Fabiszak *et al.* (this volume, 223–252), and Krawczak (2014a, 2014b; in press).

Multidimensional scaling

This technique is similar to correspondence analysis in its functionality and output. It identifies correlations between levels (features) in frequency tables. Explanation in R can be found in Rencher (2002: Ch. 15, Section 1), Everitt (2005: Ch. 5), Baayen (2008: 136–138), Drenan (2010: Ch. 23), Maindonald and Braun (2010 [2003]: 383–384), and Everitt and Hothorn (2009: Ch. 17; 2011: 121–127). A new volume, which is one of the most comprehesive applied works on the technique to date and one that includes explanation in R, is Borg *et al.* (2013). Adler (2010: 525, 541ff., 564) lists the wide range of functions in R for applying multidimensional scaling, but without examples of use. Härdle and Simar (2007: Ch. 15) and Izenman (2008: 13) offer more detailed explanations of how the technique functions. See Le Roux and Rouanet (2004: 12–14) and Cadoret *et al.* (2011) for comparison between multidimensional scaling and correspondence analysis. Borg and Groenen (2005) is a complete description, containing both mathematical theory and details of application and interpretation. Cox and Cox (2001) is equally detailed, though more concerned with mathematical theory. Nevertheless, the work includes helpful chapters on biplots and correspondence analysis. Examples of its use within the field include Bybee and Eddington (2006), Clancy (2006), Croft and Poole (2008), Szmrecsanyi (2010), Hilpert (2012), Heylen and Ruette (2013), and Ruette *et al.* (in press, forthc.). Although not a corpus study, Berthele (2010) is another recent example.

Configural frequency analysis

This is a simple and powerful technique, yet surprisingly uncommon outside the German linguistic tradition. It can be seen as a simplified log-linear analysis (see below) or as multiple Chi-squared tests; indeed, it functions by creating log-linear combinations of factors to predict cell frequencies typically based on Chi-squared tests. The technique offers possibilities for significance testing in multivariate models where no clear response variable exists, by identifying which correlations in a multiway frequency table are significant. The main limitation for the application of this technique is sample size. For a given analysis, all cells must have at least one occurrence and a minimum of 20% should have more than 5 occurrences. An excellent explanation, though with no R code, can be found in Tabachnick and Fidell (2007: Ch. 16). Gries (2009b: 240–252) offers a clear explanation of how to implement it, but note that this is omitted from the newest version of his book (Gries 2003). Von Eye (2002) is a textbook devoted to the subject and von Eye *et al.* (2010) represents the state-of-the-art. Hierarchical configuration frequency analysis has been used by Stefanowitsch and Gries (2005, 2008), Wulff *et al.* (2007), Hilpert (2009, 2012), Jing-Schmidt and Gries (2009), Schmidtke-Bode (2009), Berez and Gries (2010), Hoffmann (2011), and Kööts *et al.* (2012).

Linear discriminant analysis

Discriminant analysis is a classification technique that functions in a similar way to logistic regression and classification tree analysis (see below). However, linear discriminant analysis requires normally distributed data and continuous predictor variables, two conditions that are rarely met in Corpus Linguistics.[3]

Venables and Ripley (2002: 331–338), Crawley (2007: 744–747), Baayen (2008: 154–160), Adler (2010: 440–444) and Maindonald and Braun (2010: 385–391) offer explanations appropriate for the intermediate user. Everitt (2005: Ch 7), Härdle and Simar (2007: Ch. 12), Tabachnick and Fidell (2007: Ch. 9), Izenman (2008: Ch. 8) and Afifi *et al.* (2011: Ch. 11) offer more substantial descriptions of discriminant analysis,

3. Cf. Stevens (2001), Arppe (2008: 164), Baayen (2008: 154), Heylen *et al.* (2008), and Hoffman (2011: 95) for discussion on the problems associated with the implementation of discriminant analysis. See also Divjak (2010a: 138) who defends its use.

but offer no explanation for performing the analysis in R. Given the criteria are met, the method is a powerful classification technique and has been used by Gries (2003), Wulff (2003), and Divjak (2010a) in the field.

Classification tree analysis

An alternative to linear discriminant analysis is a data mining technique designed for categorical data called classification tree analysis. It is closely related to another technique termed regression tree analysis, which is used for continuous data. Together they are referred to as CART (or classification and regression tree analysis). The classification tree analysis technique employs an algorithm called recursive partitioning. For a given binary response variable (*a* vs. *b*), the algorithm begins with this alternation and asks which of the predictors (the other variables in the model) is best at predicting the choice between the two alternatives in the response variable. The algorithm continues this process for each of the two branches until all the predictor variables are 'used up'. This re-occurring branching gives us a 'tree' that shows how the different variables predict the outcome, *a* vs. *b*.

Classification tree analysis is explained and presented with R code in Crawley (2007: Ch. 21), Baayen (2008: 148–154), and Adler (2010: 406–117, 446–452). Other substantial descriptions include Venables and Ripley (2002: Ch. 9), Everitt and Hothorn (2010: Ch. 9), Maindonald and Braun (2010: Ch. 11), and Marden (2011: Ch. 11). The method has enjoyed some popularity in Cognitive Linguistic research, being both straightforward to apply and to interpret. Within the field, examples of its use include Klavan *et al.* (2011), Robinson (2012; this volume, 87–116), and Levshina *et al.* (this volume, 205–222).

Bootstrapping regression trees and, what is termed, the random forests technique, represent an important avenue for the development of these techniques. Bootstrapping is a widely used technique that randomises the data in order to test explanatory strength and, thus, to ascertain confidence scores for the observed data through comparison with the randomised version of the data. The application of such techniques to classification tree analysis is opening up a new set of statistical alternatives to logistic regression analysis (see below). See Everitt and Hothorn (2010: 170–173), Strobl *et al.* (2009a), Adler (2010: 414–417), and Maindonald and Braun (2010: 369–372) for a description. Such techniques have yet to be applied in the field.

Regression analysis

In its various forms, regression analysis is one of the most widely used and powerful techniques in statistics. The importance of regression techniques lies in their ability to 'predict outcomes'. The outcome is the term used to refer to a linguistic choice or a linguistic variant. This can be any kind of linguistic phenomenon, from lexemes, gestures, grammatical constructions and phonological patterns to the meanings of words, pragmatic functions, even gender, period, sociolect or dialect. The principle of how a regression analysis works is simple. The regression analysis takes our linguistic analysis of the data and builds a model that attempts to predict the behaviour of whatever phenomenon we are interested in explaining. If the model can predict which linguistic phenomenon (choice or variant, for example) is used, based on the linguistic analysis, then we can say that the analysis is accurate and, at least adequate, in distinguishing the phenomena under consideration.

The linguistic choice or variety is understood as the response variable, which is 'predicted' by the independent variables, or the factors and features of the linguistic analysis. The model provides a great deal of information about how the linguistic analysis predicts the behaviour of the response variable but three pieces of information are crucial. Firstly, it tells us which of the linguistic factors and features are statistically significant in predicting the outcome. Secondly, it tells us the effect size of those features and factors; in other words, the relative importance of that factor or feature in predicting the outcome. Lastly, it tells us how accurately a combination of all the significant factors and features distinguish between the linguistic phenomena (the forms, uses or varieties being investigated). The following sections summarise several types of regression that are designed for categorical outcomes. This family of regression techniques are typically referred to as logistic regression.

The standard references for logistic regression modelling include Agresti (2013 [1990, 2002], also 2007) and Hosmer and Lemeshow (2013 [1989, 2000]). Harrel (2001, also 2012) and Faraway (2006, also 2002) are also widely used reference books for the technique. Two other useful references include Hilbe (2009) and Menard (2010, also 2002). Once the basics have been mastered, and perhaps even before then, these books should be consulted. Especially useful is Thompson (2009), an unpublished and freely downloadable book that accompanies, step-by-step, Agresti's work, with the R code needed to perform most of what his books cover.

A note of caution is needed for the reader with little experience in statistics. None of the aforementioned books are designed for novice users, but they need to be consulted before regression analysis is used in research. Actually performing regression analysis is not particularly difficult. The complexity of confirmatory modelling lies not in applying the techniques (fitting the models), but in knowing which of the many algorithms and options one should use for the data and also applying and understanding the diagnostics of the model. Since confirmatory modelling tests hypotheses, one runs the risk of what is termed a Type I Error. This is statistics parlance, more or less, for demonstrating something to be true, when it is not. Before one reports findings obtained with regression modelling, one should always have the results thoroughly checked by a statistician.

Binary logistic regression

Currently, the most common regression analysis for categorical data is binary logistic regression. This technique takes one or more 'predictor' or 'explanatory' variables and attempts to predict the outcome of a binary response variable, such as the use of one sense or near-synonym over another (*start* vs. *begin*, for instance). The regression analysis 'models' the data, permitting it to indicate which features, or 'levels', are most important in distinguishing the binary outcome. It also indicates the statistical significance of each of these predictions. Finally, scores for the overall success of the model in predicting the outcome can be obtained.

As one of the most widely employed techniques in categorical statistics, there exists a diverse range of tutorials and textbooks devoted to it. Specifically designed for linguists, Speelman (this volume, 487–533) offers a concise introduction to applying the technique, so too does Baayen (2008: Ch. 6), Dalgaard (2008: Ch. 2008), Johnson (2008: Ch. 5), and Gries (2009b: 291–306; 2013: Ch. 5). Speelman and the latter two explanations include R code. Crawley (2005: Ch. 16) also includes lucid explanations of much of the R code needed.

More general explanations, which remain accessible to the relative beginner, include Chatterjee and Hadi (2006: Ch. 12), Faraway (2006: Chs. 2–4), Gelman and Hill (2007: Ch. 5), Sheather (2009: Ch. 8), Everitt and Hothorn (2010: Ch. 7), Maindonald and Braun (2010: Ch. 8), Azen and Walker (2011: Chs. 8, 9), and Field *et al.* (2012: Ch. 8). As mentioned above, the 'standard' references for the technique include Harrell (2001), Faraway (2006), Hilbe (2009), Menard (2010: Chs. 8, 9), Agresti (2013: Chs. 4–7; 2007: Chs. 4, 5), and Hoshmer and Lemshow (2013).

The technique is widely used in sociolinguistics and has a well-established tradition in Cognitive Linguistics. A few examples of use include Szmrecsanyi (2003, 2006), Heylen (2005b), Grondelaers *et al.* (2007, 2008), Speelman *et al.* (2009), Divjak (2010a), Glynn (2010b, this volume, 117–144), Robinson (2010a, 2010b, this volume, 87–116), Speelman and Geeraerts (2010), Deshors (2011, 2014), Levshina (2011), and Deshors and Gries (this volume, 179–204).

Loglinear analysis

Multiway frequency analysis or loglinear analysis is a technique not yet widely used in the field. Unlike binary logistic regression, loglinear analysis is not limited to determining the difference between a maximum of two possibilities. Therefore, it can be used to predict the behaviour of several senses, lexemes, or constructions. The technique is similar to configural frequency analysis, described above. Where configural frequency analysis examines configurations of sets of cells in a multiway frequency table, log-linear analysis looks at the interaction of variables that make up the multiway frequency table. Another way to think of loglinear analysis is to think of it as a logistic regression analysis without a response variable (*start* vs. *begin*, for instance).

Instead of this response variable, one attempts to predict the actual frequencies for each variable with the minimal number of factors.

Gries (this volume) offers a brief introduction to the technique, where it is termed "Poisson regression". Adler (2010: 394–395, 444) offers a very short explanation, but suggests a range of functions in R that can be used for fitting loglinear models (Adler 2010: 227, 425, 437–438, 543, 557–558, 569). Thompson's (2009: Chs. 8, 9) R manual for Agresti (2002) has two detailed chapters devoted to the technique. Short explanations include Oakes (1998: Ch. 5), Agresti (2007: Ch. 7; 2013: Chs. 9, 10), Faraway (2006: 61–67, 93–95), Dalgaard (2008: Ch. 15), Gries (2009b: 240–248; 2013: 324–327), Tarling (2009: Ch. 7), Braun (2010: 258–266), Afifi *et al.* (2011: Ch. 17), Azen and Walker (2011: Ch. 7), Smith (2011: Ch. 4), Field *et al.* (2012: Ch. 18), and Ledolter (2013: Ch. 7). Von Eye and Mun (2013) is a new volume devoted to the technique and includes practical explanations in R. However, the book is relatively theoretical and may prove challenging for learners. For users of SPSS, Tabachnick and Fidell (2007: Ch. 16) present a thorough explanation. Kroonenberg (2008) is an approachable, non-technical, volume devoted to the topic, and Christensen (1997) is older and more technical, but comprehensive. Finally, Hilbe (2011) offers a less orthodox discussion, contextualising loglinear modelling as a means for identifying multivariate dependencies. With an example-based discussion, the author reveals how the approach ties in with other techniques. Within the field of Cognitive Linguistics, Krawczak and Glynn (in press) and Glynn (forthc.) are examples of its use.

Multinomial logistic regression

This extension of binary logistic regression (explained above) is also called polychotomous logistic regression, or polytomous logistic regression. The principle is the same as for binary logistic regression, save that there are multiple nominal outcomes. The technique, however, still requires a base line for the model, that is, an outcome that serves as the point of reference for the 'other' outcomes (*start* vs. *begin, set off* and *commence*, for example).

Arguably the most approachable descriptions to date are Hilbe (2009: Ch. 10), Orme and Combs-Orme (2009: Ch. 3), and Ledolter (2013: Ch. 11), but see also Agresti (2007: Ch. 6). Arppe (2008) represents a detailed study on possible alternatives to this technique. For SPSS users, Tarling (2009: Ch. 6) and Azen and Walker (2011: Ch. 10) include a step-by-step example-based explanation. For Stata uses, Long and Freese (2006) is clear; its explanations are also useful independent of the statistical package used. The application of multinomial logistic regression is not straightforward and the technique has not yet enjoyed wide use in the field. However, as quantitative approaches to semantics continue, its application is likely to be an important contribution. Arppe (2008), Nordmark and Glynn (2013), Krawczak (2014a, 2014b, in press), and Glynn (forthc.) represent examples of its application in Cognitive Linguistics.

Ordinal logistic regression

Also referred to as ordered multinomial logit regression or proportional odds regression, the technique is a special case of logistic regression where the response is multiple and ordered, such as 'short', 'medium', 'long' or 'young', 'older', and 'oldest'. At least three ways of modelling ordinal regression exist; the most common is called the proportional method. The principle is straightforward. Rather than a binary response, one has a series of response variables. For example, for an ordered list of choices A, B, C or D, one attempts to predict the outcome of A versus B, C, or D, then in turn A or B versus C and D, and finally A or B or C versus D. If these response variables A, B, C, and D are ordered, this can be interpreted as determining what factors predict that ordering.

The most accessible explanations of such modelling can be found in Baayen (2008: Ch. 6), Hilbe (2009: Ch. 9), Orme and Combs-Orme (2009), and Tarling (2009: Ch. 8). O'Connell (2006) is a user-friendly textbook devoted to the technique, but intended for users of SPSS. Long and Freese (2006) is comparable for users of Stata. Agresti (2013: 86–98) offers a description of some of the basic issues and tests involved with ordered categories, and Agresti (2007: Ch. 6) offers a more detailed description, though somewhat theoretical. In terms of theory, Agresti (2010) represents a comprehensive work of reference. Johnson and Albert (1999) is a detailed and somewhat technical book devoted to the subject. This is a good reference, but has little explanation on application and only includes a software guide for program MATLAB.

Mixed-effects logistic regression

Sometimes also called multilevel modelling or hierarchical modelling, this technique is similar to 'normal' logistic regression, except that the model accounts for both 'fixed' effects (that is, the predictors in the model) and 'random' effects (or factors we know *a priori* are 'noise' in the model). For example, if one is looking at examples from a small set of sources, such as a set of authors in a diachronic corpus or speakers in discourse analysis, one does not want the individual traits of those authors or speakers influencing the outcome of the analysis. These unwanted effects are treated as 'random' in the model. Put simply, mixed-effects regression analysis accounts for those 'unwanted' factors, and 'neutralises' their effects, preventing them from skewing results. The principle can be applied to any form of regression, including the ordinal and multinomial regression explained above. Speelman (this volume, 487–533) offers a succinct explanation.

An older, but thorough, description can be found in Edwards (2000: Ch. 4). Gellman and Hill (2006) offer an extremely detailed, yet approachable, book on the matter. Crawley (2007: Ch. 19), Baayen (2008: Ch. 7), Maindonald (2008: Ch. 10), Sheather (2009: Ch. 10), and Tarling (2009: Ch. 9) give clear introductions to the method, as does Johnson (2008: 255–260). See also Frawley (2007: Ch. 19), who gives one of the clearest explanations on how to distinguish random variables from fixed variables, and Maindonald and Braun (2010: Ch. 10), who offer a thorough description of the interpretation of the output in R. Finally, Hox (2010) is a work devoted to the technique. It is broad in its coverage, with a theoretical orientation, but it remains approachable for the *faux-débutant*, serving as an excellent book of reference. Mixed models are beginning to become more common in the Cognitive Linguistic literature; examples include Bresnan *et al.* (2007), Divjak (2010b), Klavan (2012), Levshina *et al.* (2013a; this volume, 205–222); Krawczak and Glynn (in press), and Glynn (2014a).

Table 3 summarises the different techniques described here. Although the table systematically covers the techniques for categorical data, it does not include any techniques for continuous data. Moreover, it does not include many of the recent advances and variants, such as random forest classification or hierarchical configural frequency analysis. Tabachnick and Fidell (2007: 29–31) offer an excellent breakdown of many of the multivariate techniques available; so too does Baayen (2008: Appendix B). Tummers *et al.* (2005), Heylen *et al.* (2008) and Gilquin and Gries (2009) offer extensive discussions on the quantitative state-of-the-art in Cognitive Linguistics.

Just as the number of different statistical techniques can be overwhelming for someone first learning, so too can the number of packages and commands available for performing them in R. Packages are modules that expand R's functionality and the commands are the computer prompts to make them operate. One of R's most important strengths is the fact it is a vibrant community, with countless active internet fora and just as many people writing packages to refine and advance the application of every imaginable statistical technique. The downside to this, of course, is that a simple search request on the Internet can result in in an overload of information and options. In response to this problem, Table 4 represents a concise reference list for the functions and packages in R for performing the multivariate techniques described above. It is far from complete, being designed as a quick reference for the intermediate user who wishes to get started on a method with which he or she is not yet familiar. Also included are references for tutorials and textbooks on the functions and packages. A complete list would be impossible since many of the techniques have a number of packages devoted, or partially devoted, to them and other techniques have many variants. Moreover, it must be remembered that for the confirmatory techniques, there also exist large numbers of diagnostic and visualisation options, most of which are performed with the use of other more general or more specific packages and functions.

Certain books can be recommended for the reader who wishes to go back and investigate the basics that this volume skips, and also for the reader who wishes to delve deeper into the kinds of methods presented here. Baayen's (2008) *Analyzing Linguistic Data* is an excellent place to start. Another highly recommended guide for starting statistical analysis using R in Linguistics is Dalgaard's (2008)'s *Introducing Statistics with R*. If used in combination with Baayen (2008), one should be able to move on

Table 3. Quantitative techniques and their usage in corpus-based Cognitive Linguistics

Technique	**Type**	**Object** collocation	**Example of application**	**Explanation**
T-score, Z-score, MI score	measure	strength	identifying multi-word patterns – Biber (2009) identifying constructional variants – Wong (2009)	Evert (2009), Biber & Jones (2010)
Chi-squared test, Fisher's exact Test	univariate	probability / independence	synonymy, constructional – Wulff (2006) polysemy, lexical – Robinson (2010)	Dalgaard (2008), Everitt & Hothorn (2009), Gries (this volume)
Collostructional analysis	univariate	collocation strength	synonymy, constructional – Hilpert (2008) synonymy, constructional – Gilquin (2010)	Stefanowitsch & Gries (2003), Gries & Stefanowitsch (2004a), Hilpert (this volume)
Hierarchical cluster analysis	multivariate	associations btw. objects of single variable	polysemy, lexical – Gries (2006) synonymy, lexical – Divjak (2010a)	Baayen (2008), Everitt *et al.* (2011), Divjak & Fieller (this volume)
Multidimensional scaling	multivariate	associations btw. objects of multiple variables	synonymy, morphological – Croft & Poole (2008) relations between variants Berthele (2010)	Baayen (2008), Izenman (2008), Everitt & Hothorn (2010)
Correspondence analysis	multivariate	associations btw. objects of multiple variables	synonymy, concepts – Szelid & Geeraerts (2008) polysemy, constructional – Glynn (2009)	Le Roux & Rouanet (2010), Husson *et al.* (2011), Glynn (this volume)
Configural frequency analysis	multivariate	associations btw. objects of multiple variables	polysemy, constructional – Hilpert (2009) synonymy, constructional – Hoffmann (2011)	von Eye (2002), Tabachnick & Fidell (2007), Gries (2009b)
Discriminant analysis	multivariate	identify factors that lead to an outcome / prediction	synonymy, constructional – Gries (2003) synonymy, lexical – Divjak (2010a)	Tabachnick & Fidell (2007), Baayen (2008), Maindonald & Braun (2010)
Classification tree analysis	multivariate	identify factors that lead to an outcome / prediction	polysemy, lexical – Robinson (2012a) synonymy, lexical – Levshina *et al.* (this vol.)	Venables & Ripley (2002), Everitt & Hothorn (2010), Maindonald & Braun (2010)
Loglinear analysis	multivariate	predict correlation multiple response variables	synonymy, constructional – Krawczak & Glynn (in press) polysemy, lexical – Glynn (forthc.)	Kroonenberg (2008), Hilbe (2011), Smith (2011)
Binary logistic regression	multivariate	predict outcome binary response variable	synonymy, constructional – Szmrecsanyi (2003) synonymy, lexical – Speelman & Geeraerts (2010)	Orme & Combs-Orme (2009), Everitt & Hothorn (2010), Speelman (this volume)
Ordinal logistic regression	multivariate	predict outcome ranked response variable	synonymy. lexico–constructional – Klavan (2012) synonymy, constructional – Glynn & Krawczak (forthc.)	Baayen (2008), Tarling (2009), Orme & Combs-Orme (2009)
Multinomial logistic regression	multivariate	predict outcome multiple response variables	synonymy, lexical Arppe (2008) synonymy, lexical – Krawczak (2014a)	Long & Freese (2006), Tarling (2009), Orme & Combs-Orme (2009)
Mixed-effects logistic regression	multivariate	predict outcome include random variables	synonymy, constructional – Divjak (2010b) synonymy, constructional – Levshina *et al.* (this vol.)	Baayen (2008), Maindonald & Braun (2010), Smith (2011)

Table 4. Functions and packages for categorical multivariate statistics in R

Technique	Function	Package	R code tutorial
Hierarchical cluster analysis	`hclust`	stats*	Crawley (2007: 738ff.); Zhao (2013)
	`agnes`	cluster	Kaufman & Rousseeuw (2005); Maechler (2013)
	`pvclust`	pvclust	Suzuki & Hidetoshi (2006); Suzuki (2013)
K-means cluster analysis	`kmeans`	stats*	Crawley (2007: 742ff.); Zhao (2013)
	`clara`	cluster	Kaufman & Rousseeuw (2005); Maechler (2013)
	`pamk`	fpc	Hennig (2013)
Binary correspondence analysis	`corresp`	MASS*	Venables & Ripley (2002: 326ff.); Ripley (2013)
	`ca`	ca	Greenacre (2007); Neandić & Greenacre (2007)
	`anacor`	anacor	de Leeuw & Mair (2009a, 2013a)
Multiple correspondence analysis	`mca`	MASS*	Venables & Ripley (2002: 329f.); Ripley (2013)
	`mjca`	ca	Greenacre (2007); Neandić & Greenacre (2007)
	`MCA`	FactoMineR	Lê *et al.* (2008); Husson *et al.* (2013)
Multidimensional scaling	`cmdscale`	stats*	Baayen (2008: 136ff.); Johnson (2008: 208ff.)
	`sammon`	MASS*	Maindonald & Baun (2010: 284f.)
	`smacofSym`	smacof	de Leeuw & Mair (2009b, 2013b)
Configural frequency analysis	`cfa`	cfa	Funke *et al.* (2007); von Eye & Mair (2008)
	`hcfa`	cfa	Gries (2010: 248ff.); von Eye *et al.* (2010: 265ff.)
	`cfa2`	cfa2[4]	No tutorials available, cf. Schönbrodt (2013)
Linear discriminant Analysis	`lda`	MASS*	Baayen (2008: 167ff.); Maindonald & Braun (2010: 385ff.)
	`discrim`	ade4	Chessel *et al.* (2004); Chessel & Dufour (2013)
	`rda`	klaR	Roever *et al.* (2013)
Classification tree analysis / Random forest classification	`rpart`	rpart	Zhao (2012: 32ff.); Therneau *et al.* (2013)
	`tree`	tree	Venables & Ripley (2002: 266)
	`ctree`	party	Zhao (2013: 29ff.)
	`cforest`	party	Strobl *et al.* (2009a, 2009b)
	`randomForest`	randomForest	Maindonald & Braun (2010: 351ff.); Liaw & Wiener (2002)
Loglinear analysis	`glm`	MASS*	Maindonald & Braun (2010: 258ff.); Baguley (2012)
	`loglm`	MASS*	Thompson (2009: 142ff.); Baguley (2012)
	`quasipois`	aod	Lesnoff & Lancelot (2013)

4. The package `cfa2` is not currently in the CRAN repository for R but can be found in the RForge repository. This repository is typically used for packages still under development. A simple command listed on the Rforge site for the package will install a package as effortlessly as installation using the 'normal' method in R.

Table 4. (*continued*)

Technique	Function	Package	R code tutorial
Binary logistic regression	`glm`	MASS*	Baayen (2008: 195ff.); Everitt & Hothorn (2010: 122ff.)
	`lrm`	rms†	Harrell (2001: 257ff.; 2012: 221ff.); Baayen (2008: 195ff.)
	`MCMClogit`	MCMCpack	Martin *et al.* (2010)
Ordinal logistic regression	`polr`	MASS*	Faraway (2006: 117ff.); Maindonald & Braun (2010: 270ff.)
	`lrm`	rms†	Baayen (2008); Johnson (2008)
	`clm`	ordinal	Christensen (2012)
Multinomial logistic regression	`multinom`	nnet	Faraway (2006: 106); Thompson (2009: 118)
	`polytomous`	polytomous	Arppe (2014)
	`mlogit`	mlogit	Field *et al.* (2012: 325); Croissant (2013)
Multilevel logistic regression (mixed effects)	`lmer`	lme4	Baayen (2008: 278ff.); Bates (forthc.: 1ff.)
	`glmmPQL`[5]	MASS*	Johnson (2008: 255ff.); Thompson (2009: 179ff.)
	`MCMCglmm`	MCMCglmm	Hadfield (2010)

* Recommended package for the base installation. This means it comes 'pre-installed'.

† In older textbooks and tutorials, this package is called `Design`. The package `rms` is simply a new version. The command line to use the package is unchanged and so older descriptions remain helpful.

from what is covered in this volume on all three fronts – developing knowledge of R, the basic statistical principles and tests, as well as advanced statistical analysis.

Gries' (2009a) *Quantitative Corpus Linguistics with R* is another book to consider. Although an excellent book, it is designed more for corpus linguistics *per se* than multivariate analysis. More in line with the focus of this volume is Gries' (2009b) *Statistics for Linguistics Using R*. It covers the basics thoroughly and introduces some multivariate statistical techniques. A new edition, Gries (2013), expands the chapter on precisely the techniques covered in this volume

Johnson's (2008) *Quantitative Methods in Linguistics* is good for a debutant level statistics textbook using R – it explains both the command line and statistics lucidly and concisely. However, it 'orders' the different statistical techniques relative to different subfields of linguistics. This could be misleading for the novice and not particularly logical for the reader with some knowledge in the field, since most of the techniques are not at all restricted to the subfield Johnson ascribes to them. However, the expla-

5. The function `glmmPQL` uses a so-called penalised quasi-likelihood, which has lost favour in the research community (Crawley 2007: 655). Although the functions `lmer` and `MCMCglmm` are more up to date in this regard, `glmmPQL` in the MASS package still works perfectly well, especially when learning since some of the command line is closer to other regression functions a learner may have already mastered.

nations of the techniques are clear, especially concerning the issues that lie between the very basic and more advanced study, such as understanding data distribution and samples. Slightly more advanced books, though approachable, are Everitt and Hothorn (2010; 2011). These volumes are excellent textbooks for researchers with an introductory knowledge in statistics and/or with R, but who wish to adopt multivariate techniques – veritable handbooks. Although the examples are not linguistic, they are clear and well chosen. The statistical techniques covered are all explained through the use of examples. The demonstration of the R code is systematic and complete. Finally, Keen (2010) offers a thorough coverage of the graphic possibilities in R. Appropriate for novice and expert alike, the book is practically orientated with detailed examples of the R code.

References

Adler, J. (2010). *R in a nutshell: A desktop quick reference*. Sebastopol: O'Reilly Media.

Afifi, A., May S., & Clark, V. A. (2011). *Practical multivariate analysis* (5th ed.). London: Chapman & Hall.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken: John Wiley. DOI: 10.1002/0470114754

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken: John Wiley. DOI: 10.1002/9780470594001

Agresti, A. (2013) [1990, 2002]. *Categorical data analysis* (3rd ed.). New York: John Wiley.

Arppe, A. (2006). Frequency considerations in morphology: Finnish verbs differ, too. *SKY Journal of Linguistics*, 19, 175–189.

Arppe, A. (2008). Univariate, bivariate and multivariate methods in corpus-based lexicography – A study of synonymy. Unpublished PhD dissertation, University of Helsinki.

Azen, R., & Walker, C. (2011). *Categorical data analysis for the behavioral and social sciences*. New York & Hove: Routledge.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511801686

Baguley, T. (2012). *Loglinear models. Online Supplement 5 to Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave. Available at: http://www.palgrave.com/psychology/baguley/students/supplements.html.

Balahur, A., & Montoyo, A. (2012). Semantic approaches to fine and coarse-grained feature-based opinion mining. In H. Horacek, E. Métais, R. Muñoz, & M. Wolska (Eds.), *Natural language processing and information systems* (pp. 142–153). Berlin: Springer.

Barnabé, A. (2012). Le schème du chemin en grammaire et sémantique anglaises. Unpublished PhD dissertation, Université Bordeaux 3.

Bates, D. (Forthcoming). *lme4: Mixed-effects modeling with R*. Heidelberg & New York: Springer. Preprints available at: http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf.

Benzécri, J.-P. (1980). *Pratique de l'analyse des donnees*. Paris: Dunod.

Benzécri, J.-P. (1992). *Correspondence analysis handbook*. New York: Dekker.

Berthele, R. (2010). Investigations into the folk's mental models of linguistic varieties. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in cognitive sociolinguistics* (pp. 265–290). Berlin & New York: Mouton de Gruyter.
Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14, 275–311. DOI: 10.1075/ijcl.14.3.08bib
Biber, D., & Jones, J. (2009). Quantitative methods in Corpus Linguistics. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An international handbook*. Vol. 2. (pp. 1287–1304). Berlin & New York: Mouton de Gruyter.
Borg, I., Groenen, & Mair, P. (2013). *Applied multidimensional scaling*. Heidleberg & New York: Springer. DOI: 10.1007/978-3-642-31848-1
Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling* (2nd ed.). Heidelberg & New York: Springer.
Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation alternation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Arts and Sciences.
Bybee, J., & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82, 323–355. DOI: 10.1353/lan.2006.0081
Cadoret, M., Lê, S., & Pagès, J. (2011). Multidimensional scaling versus multiple correspondence analysis when analyzing categorization data. In B. Fichet, D. Piccolo, R. Verde, & M. Vichi (Eds.), *Classification and multivariate analysis for complex data structures* (pp. 301–308). Heidleberg & New York: Springer. DOI: 10.1007/978-3-642-13312-1_31
Chaffin, R. (1992). The concept of a semantic relation. In A. Lehrer, & E. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organisation* (pp. 253–288). London: Lawrence Erlbaum.
Chatterjee, S., & Hadi, A. (2006). *Regression analysis by example*. London: John Wiley. DOI: 10.1002/0470055464
Chessel, D., & Dufour, A.-B. (2013). Analysis of ecological data: Exploratory and Euclidean methods in environmental sciences. Available at: http://cran.r-project.org/web/packages/ade4/ade4.pdf.
Chessel, D., Dufour A.-B, & Thioulouse, Y. (2004) The ade4 package – I: One-table methods. *R News*, 4, 5–10.
Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). Heidleberg & New York: Springer.
Christensen, R. (2012). A tutorial on fitting cumulative link models with the ordinal package. Available at: http://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf.
Clancy, S. (2006). The topology of Slavic case: Semantic maps and multidimensional scaling. *Glossos*, 7, 1–28.
Colleman, T. (2009). The semantic range of the Dutch double object construction. A collostructional perspective. *Constructions and Frames*, 1, 190–221. DOI: 10.1075/cf.1.2.02col
Colleman, T. (2010). Beyond the dative alternation: The semantics of the Dutch *aan*-Dative. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 271–304). Berlin & New York: Mouton de Gruyter.
Cox, T., & Cox, M. (2001). *Multidimensional scaling* (2nd ed.). Boca Raton: Chapman & Hall.
Crawley, M. (2005). *Statistics: An introduction using R*. Southern Gate & Hoboken: John Wiley. DOI: 10.1002/9781119941750

Crawley, M. (2007). *The R book*. Chichester: John Wiley. DOI: 10.1002/9780470515075

Croft, W., & Poole, K. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics*, 34, 1–37. DOI: 10.1515/THLI.2008.001

Croissant, Y. (2013). Estimation of multinomial logit models in R: The mlogit packages. Available at: cran.r-project.org/web/packages/mlogit/mlogit.pdf.

Daille, B., Dubreil, E. Monceaux, L., & Vernier, M. (2011). Annotating opinion–evaluation of blogs: The Blogoscopy corpus. *Language Resources and Evaluation*, 45, 409–437. DOI: 10.1007/s10579-011-9154-z

Dalgaard, P. (2008). *Introductory statistics with R* (2nd ed.). Dordrecht: Springer. DOI: 10.1007/978-0-387-79054-1

De Cock, B. (2014a). *A discourse-functional analysis of speech participant profiling in spoken Spanish*. Amsterdam & Philadelphia: John Benjamins.

De Cock, B. (2014b). The discursive effects of Spanish impersonals *uno* and *se*. In D. Glynn, & M. Sjölin (Eds.), *Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance (pp. 103–120)*. Lund: Lund University Press.

De Leeuw, J., & Mair, P. (2009a). Simple and canonical correspondence analysis using the R package anacor. *Journal of Statistical Software*, 31, 1–18.

De Leeuw, J., & Mair, P. (2009b). Multidimensional scaling using majorization: The R package smacof. *Journal of Statistical Software*, 31, 1–30.

De Leeuw, J., & Mair, P. (2013a). anacor: Simple and canonical correspondence analysis. Available at: cran.r-project.org/web/packages/anacor/anacor.pdf.

De Leeuw, J., & Mair, M. (2013b). SMACOF for multidimensional scaling. Available at: http://cran.r-project.org/web/packages/smacof/smacof.pdf.

Deignan, A. (2005). *Metaphor and Corpus Linguistics*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/celcr.6

Delorge, M. (2009). A diachronic corpus study of the constructional behaviours of reception verbs in Dutch. In B. Lewandowska-Tomaszczyk, & K. Dziwirek (Eds.), *Studies in Cognitive Corpus Linguistics* (pp. 249–272). Frankfurt/Main: Peter Lang.

Desagulier, G. (In press). Le statut de la fréquence dans les Grammaires de Constructions: 'simple comme bonjour'? *Langages*.

Desagulier, G. (Submitted). Quite new methods for a rather old issue: Exploring and visualizing collocation data from the BNC with correspondence analysis.

Deshors, S. (2011). A multifactorial study of the uses of *may* and *can* in French-English interlanguage. Unpublished PhD dissertation, University of Sussex.

Deshors, S. (2014). Identifying different types of non-native co-occurrence patterns: A corpus-based approach. In D. Glynn, & M. Sjölin (Eds.), *Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance (pp. 387–412)*. Lund: Lund University Press.

Diehl, H. (2014). On modal meaning in the uses of quite, rather, pretty and fairly as degree modifiers in British English. Unpublished PhD dissertation, Lund University.

Dirven, R., Goossens, L., Putseys, Y., & Vorlat, E. (1982). *The scene of linguistic action and its perspectivization by speak, talk, say, and tell*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/pb.iii.6

Divjak, D. (2006). Ways of intending: A corpus-based Cognitive Linguistic approach to near-synonyms in Russian. In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis* (pp. 19–56). Berlin & New York: Mouton de Gruyter.

Divjak, D. (2010a). *Structuring the lexicon: A clustered model for near-synonymy*. Berlin & New York: Mouton de Gruyter.

Divjak, D. (2010b). Corpus-based evidence for an idiosyncratic aspect-modality relation in Russian. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 305–331). Berlin & New York: Mouton de Gruyter.

Divjak, D., & Gries, St. Th. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2, 23–60. DOI: 10.1515/CLLT.2006.002

Divjak, D., & Gries, St. Th. (2009). Corpus-based Cognitive Semantics: A contrastive study of phrasal verbs in English and Russian. In B. Lewandowska-Tomaszczyk, & K. Dziwirek (Eds.), *Studies in Cognitive Corpus Linguistics* (pp. 273–296). Frankfurt/Main: Peter Lang.

Divjak, D., & Gries, St. Th. (Eds.). (2012). *Frequency effects in language learning and processing*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110274059

Drenan, R. (2009). *Statistics for archaeologists: A common sense approach* (2nd ed.). Heidelberg & New York: Springer.

Dziwirek, K., & Lewandowska-Tomaszczyk, B. (2011). *Complex emotions and grammatical mismatches: A contrastive corpus-based study*. Berlin & New York: Mouton de Gruyter.

Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). Heidelberg: Springer. DOI: 10.1007/978-1-4612-0493-0

Everitt, B. S. (2005). *An R and S-PLUS companion to multivariate analysis*. London: Springer.

Everitt, B. S., & Hothorn, I. (2010). *A handbook of statistical analyses using R* (2nd ed.). Boca Raton: Taylor & Francis. DOI: 10.1201/9781420079340

Everitt, B. S., & Hothorn, I. (2011). *An introduction to applied multivariate analysis with R*. Munich: Springer. DOI: 10.1007/978-1-4419-9650-3

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester: John Wiley. DOI: 10.1002/9780470977811

Evert, S. (2009). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An international handbook* (pp. 1212–1249). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110213881.2.1212

Faraway, J. (2002). Practical regression and anova using R. Available at: cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

Faraway, J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. London: Taylor & Francis.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London & Thousand Oaks: Sage.

Fillmore, C., & Atkins, B. (1992). Toward a frame-based lexicon: The semantics of risk and its neighbours. In A. Lehrer, & E. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organisation* (pp. 75–102). London: Lawrence Erlbaum.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Oxford: Basil Blackwell.

Fischer, K. (2000). *From Cognitive Semantics to Lexical Pragmatics: The functional polysemy of discourse particles*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110828641

Flores Salgado, E. (2011). *The pragmatics of requests and apologies: Developmental patterns in Mexican students*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/pbns.212

Fontaine, J., Scherer, K., & Soriano, C. (Eds.). (2013). *Components of emotional meaning: A sourcebook*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199592746.001.0001

Funke, S., Mair, P., & von Eye, A. (2007). cfa: R package for the analysis of configuration frequencies. Available at: http://cran. R-project.org.

Geeraerts, D. (2010). The doctor and the semantician. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 63–78). Berlin & New York: Mouton de Gruyter.

Geeraerts, D. (2011). Entrenchment, conventionalization, and empirical method. *Presented at the 44th Meeting of the Societas Linguistica Europaea*, Logroño.

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The structure of lexical variation: Meaning, naming, and context*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110873061

Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Geeraerts, D., Kristiansen, G., & Peirsman, Y. (Eds.). (2010). *Advances in cognitive sociolinguistics*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226461

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gilquin, G. (2010). *Corpus, cognition and causative constructions*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/scl.39

Glynn, D. (2007). Mapping meaning: Toward a usage-based methodology in Cognitive Semantics. Unpublished PhD dissertation, University of Leuven.

Glynn, D. (2009). Polysemy, syntax, and variation: A usage-based method for Cognitive Semantics. In V. Evans, & S. Pourcel (Eds.), *New directions in Cognitive Linguistics* (pp. 77–106). Amsterdam & Philadelphia: John Benjamins.

Glynn, D. (2010a). Synonymy, lexical fields, and grammatical constructions: A study in usage-based Cognitive Semantics. In H.-J. Schmid, & S. Handl (Eds.), *Cognitive foundations of linguistic usage-patterns: Empirical studies* (pp. 89–118). Berlin & New York: Mouton de Gruyter.

Glynn, D. (2010b). Testing the hypothesis: Objectivity and verification in usage-based Cognitive Semantics. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 239–270). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226423

Glynn, D. (2014a). The conceptual profile of the lexeme *home*: A multifactorial diachronic analysis. In J. E. Díaz-Vera (Ed.), *Metaphor and metonymy across time and cultures* (pp. 265–293). Berlin & New York: Mouton de Gruyter.

Glynn, D. (2014b). The social nature of ANGER: Multivariate corpus evidence for context effects upon conceptual structure. In I. Novakova, P. Blumenthal, & D. Siepmann (Eds.), *Emotions in discourse* (pp. 69–82). Frankfurt/Main: Peter Lang.

Glynn, D. (Forthcoming). *Mapping meaning: Corpus methods for Cognitive Semantics*. Cambridge: Cambridge University Press.

Glynn, D., & Sjölin, M. (2011). Cognitive Linguistic methods for literature: A usage-based approach to metanarrative and metalepsis. In A. Kwiatkowska (Ed.), *Texts and minds: Papers in cognitive poetics and rhetoric* (pp. 85–102). Frankfurt/Main: Peter Lang.

Glynn, D., & Krawczak, K. (Forthcoming). Social cognition, Cognitive Grammar and corpora: A multifactorial approach to epistemic modality. *Cognitive Linguistics*.

Glynn, D., & Fischer, D. (Eds.). (2010). *Quantitative Cognitive Semantics: Corpus-driven approaches*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226423

Glynn, D., & Sjölin, M. (Eds.). (2014). *Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance*. Lund: Lund University Press.

Greenacre, M. (2007) [1993]. *Correspondence analysis in practice* (2nd ed.). London: Chapman & Hall.

Greenacre, M. (2010). *Biplots in practice*. Bilbao: Fundación BBVA.
Gries, St. Th. (1999). Particle movement: A cognitive and functional approach. *Cognitive Linguistics*, 10, 105–145. DOI: 10.1515/cogl.1999.005
Gries, St. Th. (2000). Towards multifactorial analyses of syntactic variation: The case of particle placement. Doctoral dissertation, University of Hamburg.
Gries, St. Th. (2003). *Multifactorial analysis in Corpus Linguistics: A study of particle placement*. London: Continuum Press.
Gries, St. Th. (2006). Corpus-based methods and Cognitive Semantics: The many senses of *to run*. In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis* (pp. 57–99). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197709
Gries, St. Th. (2009a). *Quantitative Corpus Linguistics with R: A practical introduction*. London: Routledge. DOI: 10.1515/9783110216042
Gries, St. Th. (2009b). *Statistics for Linguistics with R: A practical introduction* (1st ed.). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110216042
Gries, St. Th. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus based lexical semantics. *The Mental Lexicon*, 5, 323–346. DOI: 10.1075/ml.5.3.04gri
Gries, St. Th. (2013). *Statistics for linguistics with R: A practical introduction* (2nd ed.). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110307474
Gries, St. Th., & Divjak, D. (2009). Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In V. Evans, & S. Pourcel (Eds.), *New directions in Cognitive Linguistics* (pp. 57–75). Amsterdam & Philadelphia: John Benjamins.
Gries, St. Th., & Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora*, 3, 59–81. DOI: 10.3366/E1749503208000075
Gries, St. Th., & Stefanowitsch, A. (2004a). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9, 97–129. DOI: 10.1075/ijcl.9.1.06gri
Gries, St. Th., & Stefanowitsch, A. (2004b). Co-varying collexemes in the *into*-causative. In M. Achard, & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–36). Stanford: CSLI.
Gries, St. Th., & Divjak, D. (Eds.). (2012). *Frequency effects in language representation*. Berlin & New York: Mouton de Gruyter.
Gries, St. Th., & Stefanowitsch, A. (Eds.). (2006). *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197709
Grondelaers, S. (2000). De distributie van niet-anaforisch er buiten de eerste zinsplaats: Sociolexicologische, functionele en psycholinguïstische aspecten van *er*'s status als presentatief signaal. Doctoral dissertation, University of Leuven.
Grondelaers S., Geeraerts, D., & Speelman, D. (2007). A case for a cognitive Corpus Linguistics. In M. Gonzalez-Marquez, I. Mittleberg, S. Coulson, & M. Spivey (Eds.), *Methods in Cognitive Linguistics* (pp. 149–169). Amsterdam & Philadelphia: John Benjamins.
Grondelaers S., Speelman, D., & Geeraerts, D. (2008). National variation in the use of *er* "there": Regional and diachronic constraints on cognitive explanations. In G. Kristiansen, & R. Dirven (Eds.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems* (pp. 153–204). Berlin & New York: Mouton de Gruyter.
Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33, 1–22.

Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis*. Heidelberg & New York: Springer.

Harrell, F. (2001). *Regression modeling strategies: With Applications to linear models, logistic regression, and survival analysis*. Heidelberg & New York: Springer.

Harrell, F. (2012). Regression modeling strategies. Unpublished manuscript, available at: www.biostat.mc.vanderbilt.edu/wiki/pub/Main/RmS/rms.pdf.

Hennig, C. (2013). Flexible procedures for clustering. Available at: http://cran.r-project.org/web/packages/fpc/fpc.pdf.

Heylen, K. (2005a). A quantitative corpus study of German word order variation. In St. Kepser, & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 241–264). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197549.241

Heylen, K. (2005b). Zur Abfolge (pro)nominaler Satzglieder im Deutschen: Eine korpusbasierte Analyse der relativen Abfolge von nominalem Subjekt und pronominalem Objekt im Mittelfeld, 264. Doctoral dissertation, University of Leuven.

Heylen, K., & Ruette, T. (2013). Degrees of semantic control in measuring aggregated lexical distances. In L. Borin, A. Saxena, A., & T. Rama (Eds.), *Approaches to measuring linguistic differences* (pp. 353–374). Berlin & New York: Mouton de Gruyter.

Heylen, K., Tummers, J., & Geeraerts, D. (2008). Methodological issues in corpus-based Cognitive Linguistics. In G. Kristiansen, & R. Dirven (Eds.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems* (pp. 91–128). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110199154.2.91

Hilbe, J. (2009). *Logistic regression models*. London: Chapman & Hall.

Hilbe, J. (2011) [2007]. *Negative binomial regression* (2nd ed.). Cambridge: Cambridge University Press.

Hilpert, M. (2008). *Germanic future constructions: A usage-based approach to language change*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/cal.7

Hilpert, M. (2009). The German *mit*-predicative construction. *Constructions and Frames*, 1, 29–55. DOI: 10.1075/cf.1.1.03hil

Hilpert, M. (2012). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University.

Hoffmann, Th. (2011). *Preposition placement in English: A usage-based approach*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511933868

Hosmer, D., & Lemeshow, S. (2013) [1989, 2000]. *Applied logistic regression*. Hoboken: John Wiley. DOI: 10.1002/9781118548387

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Hove & New York: Routledge.

Husson, F. Josse, J., Lê, S., & Mazet, J. (2013). Multivariate exploratory data analysis and data mining with R. Available at: http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf.

Husson, F., Lê, S., & Pagès, J. (2011). *Exploratory multivariate analysis by example using R*. London: Chapman & Hall.

Izenman, A. (2008). *Modern multivariate statistical techniques: Regression, classification and manifold learning*. Heidelberg & New York: Springer. DOI: 10.1007/978-0-387-78189-1

Janda, L., & Solovyev, V. (2009). What constructional profiles reveal about synonymy: A case study of the Russian words for sadness and happiness. *Cognitive Linguistics*, 20, 367–393. DOI: 10.1515/COGL.2009.018

Johnson, K. (2008). *Quantitative methods in linguistics*. Oxford: Blackwell.

Johnson, V., & Albert, J. (1999). *Ordinal data modeling*. Heidelberg & New York: Springer.

Kärkkäinen, E. (2003). *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/pbns.115

Kaufman, L., & Rousseeuw, P. (2005) [1990]. *Finding groups in data: An introduction to cluster analysis*. Hoboken: John Wiley.

Keen, K. (2010). *Graphics for statistics and data analysis with R*. Boca Raton: CRC Press.

Klavan, J. (2012). Evidence in linguistics: Corpus-linguistic and experimental methods for studying grammatical synonymy. Doctoral Dissertation, University of Tartu.

Klavan, J, Kesküla K., & Ojava, L. (2011). Synonymy in grammar: The Estonian adessive case and the adposition *peal* 'on'. In S. Kittilä, K. Västi, & J. Ylikoski (Eds.), *Studies on case, animacy and semantic roles* (pp. 1–19). Amsterdam & Philadelphia: John Benjamins.

Krawczak, K. (2014a). Shame and its near-synonyms in English: A multivariate corpus-driven approach to social emotions. In I. Novakova, P. Blumenthal, & D. Siepmann (Eds.), *Emotions in discourse* (pp. 84–94). Frankfurt/Main: Peter Lang.

Krawczak, K. (2014b). Epistemic stance predicates in English: A quantitative corpus-driven study of subjectivity. In D. Glynn, & M. Sjölin (Eds.), *Subjectivity and epistemicity: Corpus, discourse, and literary approaches to stance* (pp. 355–386). Lund: Lund University Press.

Krawczak, K. (In press). Corpus evidence for the cross-cultural structure of social emotions: Shame, embarrassment, and guilt in English and Polish. *Poznań Studies in Contemporary Linguistics*.

Krawczak, K., & Glynn, D. (2011). Context and cognition: A corpus-driven approach to parenthetical uses of mental predicates. In K. Kosecki, & J. Badio (Eds.), *Cognitive processes in language* (pp. 87–99). Frankfurt/Main: Peter Lang.

Krawczak, K., & Kokorniak, I. (2012). Corpus-driven quantitative approach to the construal of Polish 'think'. *Poznań Studies in Contemporary Linguistics*, 48, 439–472. DOI: 10.1515/psicl-2012-0021

Krawczak, K., & Glynn, D. (In press). Operationalising construal: *Of/about* prepositional profiling for cognitive and communicative predicates. In C. M. Bretones Callejas (Ed.), *Construals in language and thought: What shapes what?* Amsterdam: John Benjamins.

Kroonenberg, P. (2008). *Applied multiway data analysis*. New York: John Wiley. DOI: 10.1002/9780470238004

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25, 1–18.

Le Roux, B., & Rouanet, H. (2004). *Geometric data analysis: From correspondence analysis to structured data analysis*. Dordrecht: Kluwer.

Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. London & Thousand Oaks: Sage.

Ledolter, J. (2013). *Data mining and business analytics with R*. Hoboken: John Wiley. DOI: 10.1002/9781118596289

Lesnoff, M., & Lancelot, R. (2013). Analysis of overdispersed data. Available at: http://cran.r-project.org/web/packages/aod/aod.pdf.

Levshina, N. (2011). A usage-based study of Dutch causative constructions. Doctoral dissertation, University of Leuven.

Levshina, N. (2012). Comparing constructicons: A usage-based analysis of the causative construction with *doen* in Netherlandic and Belgian Dutch. *Constructions and Frames*, 4, 76–101. DOI: 10.1075/cf.4.1.04lev

Levshina, N., Geeraerts, D., & Speelman, D. (2013a). Towards a 3D-grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics*, 52, 34–48. DOI: 10.1016/j.pragma.2012.12.013

Levshina, N., Geeraerts, D., & Speelman, D. (2013b). Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics*, 51, 825–854. DOI: 10.1515/ling-2013-0028

Lewandowska-Tomaszczyk, B., & Dziwirek, K. (Eds.). (2009). *Studies in Cognitive Corpus Linguistics*. Frankfurt/Main: Peter Lang.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2, 18–22.

Long, J. S., & Freese, J. (2006) [2001]. *Regression models for categorical dependent variables using Stata*. College Station: Stata Press.

Louwerse, M., & Van Peer, W. (2009). How cognitive is cognitive poetics? The interaction between symbolic and embodied cognition. In G. Brône, & J. Vandaele (Eds.), *Cognitive poetics goals, gains and gaps* (pp. 423–444). Berlin & New York: Mouton de Gruyter.

Maechler, M. (2013). Cluster analysis extended. Available at: http://cran.r-project.org/web/packages/cluster/cluster.pdf.

Maindonald, J. (2008). Using R for data analysis and graphics: Introduction, code and commentary. Available at: http://www.maths.anu.edu.au/˜johnm/r/usingR.pdf.

Maindonald, J., & Braun, J. (2010) [2003]. *Data analysis and graphics using R* (3rd ed.). Cambridge: Cambridge University Press.

Marden, J. (2011). *Multivariate statistical analysis: Old school.* Department of Statistics, University of Illinois at Urbana-Champaign. Available at: istics.net/pdfs/multivariate.pdf.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010). Markov chain Monte Carlo (MCMC) package. Available at: http://mcmcpack.wustl.edu/.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). London & Thousand Oaks: Sage.

Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. London & Los Angeles: Sage.

Morgenstern, A., Blondel, M., Caët, S., & Boutet, D. (2011). Hearing children's use of pointing gestures: From pre-linguistic buds to the blossoming of communication skills. *Presentation at SALC III, Copenhagen*.

Murtagh, F. (2005). *Correspondence analysis and data coding with R and Java*. London: Chapman & Hall. DOI: 10.1201/9781420034943

Myers, D. (1994). Testing for prototypicality: The Chinese morpheme *gong*. *Cognitive Linguistics*, 5, 261–280. DOI: 10.1515/cogl.1994.5.3.261

Neandić, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca Package. *Journal of Statistical Software*, 20, 1–13.

Newman, J., & Rice, S. (2004). Patterns of usage for English sit, stand, and lie: A cognitively-inspired exploration in corpus linguistics. *Cognitive Linguistics*, 15, 351–396. DOI: 10.1515/cogl.2004.013

Newman, J., & Rice, S. (2006). Transitivity schemas of English eat and drink in the BNC. In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis*. (pp. 225–260). Berlin & New York: Mouton de Gruyter.

Nordmark, H., & Glynn, D. (2013). anxiety between mind and society: A corpus-driven cross-cultural study of conceptual metaphors. *Explorations in English Language and Linguistics*, 1, 107–130.

O'Connell, A. (2006). *Logistic regression models for ordinal response variables*. London & Thousand Oaks: Sage.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Orme, J., & Combs-Orme, T. (2009). *Multiple regression with discrete dependent variables*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780195329452.001.0001

Peirsman, Y. Heylen, K., & Geeraerts, D. (2010). Applying word space models to sociolinguistics: Religion names before and after 9/11. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 111–139). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226461

Pęzik, P. (2009). Extraction of multiword expressions for corpus-based discourse analysis. In B. Lewandowska-Tomaszczyk, & K. Dziwirek (Eds.), *Studies in Cognitive Corpus Linguistics* (pp. 249–272). Frankfurt/Main: Peter Lang.

Pichler, H. (2013). *The structure of discourse-pragmatic variation*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/silv.13

Plevoets, K., Speelman, D., & Geeraerts, D. (2008). The distribution of T/V pronouns in Netherlandic and Belgian Dutch. In K. Schneider, & A. Baron (Eds.), *Variational pragmatics: Regional varieties in pluricentric languages* (pp. 181–209). Amsterdam & Philadelphia: John Benjamins.

Pütz, M, Robinson, J. A., & Reif, M. (Eds.) (2012). *Cognitive Sociolinguistics: Social and cultural variation in cognition and language use*. (Special edition of *Annual Review of Cognitive Linguistics*, 10.)

Ravid, D., & Hanauer, D. (1998). A prototype theory of rhyme: Evidence from Hebrew. *Cognitive Linguistics*, 9, 79–106. DOI: 10.1515/cogl.1998.9.1.79

Read, J., & Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46, 421–447. DOI: 10.1007/s10579-010-9135-7

Reif, M., Robinson, J. A., & Pütz, M. (Eds.). (2013). *Variation in language and language use: Linguistic, socio-cultural and cognitive perspectives*. Frankfurt/Main: Peter Lang.

Rencher, A. (2002). *Methods of multivariate analysis* (2nd ed.). New York: John Wiley. DOI: 10.1002/0471271357

Rice, S., Sandra, D., & Vanrespaille, M. (1999). Prepositional semantics and the fragile link between space and yime. In M. Hiraga, C. Sinha, & S. Wilcox (Eds.), *Cultural, typology and psycholinguistic issues in Cognitive Linguistics* (pp. 107–127). Amsterdam & Philadelphia: John Benjamins.

Ripley, B. (2013). Support functions and datasets for Venables and Ripley's MASS. Available at: http://cran.r-project.org/web/packages/MASS/MASS.pdf.

Robinson, J. A. (2010a). *Awesome* insights into semantic variation. In D. Geeraerts, G. Kristiansen, & Y. Piersman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 85–109). Berlin & New York: Mouton de Gruyter.

Robinson, J. A. (2010b). Semantic variation and change in present-day English. Doctoral dissertation, University of Sheffield.

Robinson, J. A. (2012). A sociolinguistic perspective on semantic change. In K. Allan, & J. A. Robinson (Eds.), *Current methods in Historical Semantics* (pp. 191–231). Berlin & New York: Mouton de Gruyter.

Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G., & Zentgraf, M. (2013). Classification and visualization. Unpublished manuscript available at: http://cran.r-project.org/web/packages/klaR/klaR.pdf.

Rudzka-Ostyn, B. (1989). Prototypes, schemas, and cross-category correspondences: The case of *ask*. In D. Geeraerts (Ed.), *Prospects and problems of prototype theory* (pp. 613–661). Berlin & New York: Mouton de Gruyter.

Rudzka-Ostyn, B. (1995). Metaphor, schema, invariance: The case of verbs of answering. In L. Goossens, P. Pauwels, B. Rudzka-Ostyn, A.-M. Simon-Vandenbergen, & J. Vanparys (Eds.), *By word of mouth: Metaphor, metonymy, and linguistic action from a cognitive perspective* (pp. 205–244). Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/pbns.33

Ruette, T., Ehret, K., & Szmrecsanyi, B. (In press). *Frequency effects in lexical sociolectometry are insubstantial*. In H. Behrens, & S. Pfänder (Eds.), *Again on frequency effects in language*. Berlin & New York: Mouton de Gruyter.

Ruette, T., Geeraerts, D., Peirsman, Y., & Speelman, D. (Forthcoming). Semantic weighting mechanisms in scalable lexical sociolectometry. In B. Szmrecsanyi, & B. Waelchli (Eds.), *Aggregating dialectology and typology: Linguistic variation in text and speech, within and across languages*. Berlin & New York: Mouton de Gruyter.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. In K. Allan, & J. Robinson (Eds.), *Current methods in Historical Semantics* (pp. 161–183). Berlin & New York: Mouton de Gruyter.

Sandra, D., & Rice, S. (1995). Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics*, 6, 89–130. DOI: 10.1515/cogl.1995.6.1.89

Scheibman, J. (2002). *Point of view and grammar: Structural patterns of subjectivity in American English conversation*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/sidag.11

Scherer, K. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44, 693–727. DOI: 10.1177/0539018405058216

Schmid, H.-J. (1993). *Cottage and co., idea, start vs. begin: Die kategorisierung als grundprinzip einer differenzierten bedeutungsbeschreibung*. Tübingen: Max Niemeyer. DOI: 10.1515/9783111355771

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110808704

Schmidtke-Bode, K. (2009). *Going-to*-V and *gonna*-V in child language: A quantitative approach to constructional development. *Cognitive Linguistics*, 20, 509–553. DOI: 10.1515/COGL.2009.023

Schönbrodt, F., Collins, L., & Stemmler, M. (2013). cfa2: Configuration frequency analysis with a design matrix. Available at: http://www.rforge.net/cfa2/.

Schulze, R. (1991). Getting round to *(a)round:* Towards the description and analysis of a 'spatial' predicate. In G. Rauh (Ed.), *Approaches to prepositions* (pp. 253–74). Tubingen: Günter Narr.

Sheather, S. (2009). *A modern approach to regression with R*. New York: Springer. DOI: 10.1007/978-0-387-09608-7

Smith, R. (2011). *Multilevel modeling of social problems: A causal perspective*. Heidelberg: Springer. DOI: 10.1007/978-90-481-9855-9

Speelman, D., & Geeraerts, D. (2010). Causes for causatives: The case of Dutch 'doen' and 'laten'. In T. Sanders, & E. Sweetser (Eds.), *Causal categories in discourse and cognition* (pp. 173–204). Berlin & New York: Mouton de Gruyter.

Speelman, D., Tummers, J., & Geeraerts, D. (2009). Lexical patterning in a Construction Grammar: The effect of lexical co-occurrence patterns on the inflectional variation in Dutch attributive adjectives. *Constructions and Frames*, 1, 87–118. DOI: 10.1075/cf.1.1.05spe

Stefanowitsch, A. (2010). Empirical Cognitive Semantics: Some thoughts. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 355–380). Berlin & New York: Mouton de Gruyter.

Stefanowitsch, A., & St. Th. Gries. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8, 209–243. DOI: 10.1075/ijcl.8.2.03ste

Stefanowitsch, A., & St. Th. Gries. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1, 1–43. DOI: 10.1515/cllt.2005.1.1.1

Stefanowitsch, A., & St. Th. Gries. (2008). Register and constructional meaning: A collostructional case study. In G. Kristiansen, & R. Dirven (Eds.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems* (pp. 129–152). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110199154.2.129

Stefanowitsch, A., & Gries, St. Th. (Eds.). (2006). *Corpus-based approaches to metaphor and metonymy*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110199895

Stevens, J. 2001. *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah: Lawrence Erlbaum.

Strobl, C., Hothorn, T., & Zeileis, A. (2009a). Party on! A new, conditional variable importance measure for random forests available in the party package. *The R Journal*, 1, 14–17.

Strobl, C., Malley, J., & Gerhard T. (2009b). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323–348. DOI: 10.1037/a0016973

Suzuki, R. (2013). Hierarchical clustering with p-values via multiscale bootstrap resampling. Available at: http://cran.r-project.org/web/packages/pvclust/pvclust.pdf.

Suzuki, R., & Hidetoshi, S. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22, 1540–1542. DOI: 10.1093/bioinformatics/btl117

Szelid, V, & Geeraerts, D. (2008). Usage-based dialectology: Emotion concepts in the Southern Csango dialect. *Annual Review of Cognitive Linguistics*, 6, 23–49. DOI: 10.1075/arcl.6.03sze

Szmrecsanyi, B. (2003). *Be going to* versus *will/shall*: Does syntax matter? *Journal of English Linguistics*, 31, 295–323. DOI: 10.1177/0075424203257830

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English: A corpus study at the intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197808

Szmrecsanyi, B. (2010). The English genitive alternation in a cognitive sociolinguistic perspective. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 141–166). Berlin & New York: Mouton de Gruyter.

Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects*. Cambridge: Cambridge University Press.

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). London: Pearson.

Taboada, M., & Carretero, M. (2012). Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish. *Linguistics and the Human Sciences*, 6, 275–295.

Tarling, R. (2009). *Statistical modelling for social researchers: Principles and practice*. London & New York: Routledge.

Therneau, T., Atkinson, E., & Foundation, M. (2013). An introduction to recursive partitioning using the RPART routines. Available at: http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Thompson, L. (2009). *S-PLUS (and R) manual to accompany Agresti's categorical data analysis (2002)*. Available at: home.comcast.net/~lthompson221/Splusdiscrete2.pdf.

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1, 225–261. DOI: 10.1515/cllt.2005.1.2.225

Valenzuela Manzanares, J., & Rojo López, A. M. (2008). What can language learners tell us about constructions? In S. De Knop, & T. De Rycker (Eds.), *Cognitive approaches to pedagogical grammar? A volume in honour of René Dirven* (pp. 197–230). Berlin & New York: Mouton de Gruyter.

Van Bogaert, J. (2010). A constructional taxonomy of *I think* and related expressions: Accounting for the variability of complement-taking mental predicates. *English Language and Linguistics*, 14, 399–428. DOI: 10.1017/S1360674310000134

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S* (4th ed.). Heidelberg: Springer. DOI: 10.1007/978-0-387-21706-2

Verdonik, D., Rojc, M., & Stabej, M. (2007). Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language. *Language Resources and Evaluation*, 41, 147–180. DOI: 10.1007/s10579-007-9035-7

von Eye, A. (2002). *Configural frequency analysis: Methods, models, and applications*. Mahwah: Erlbaum.

von Eye, A., & Mair, P. (2008) A functional approach to configural frequency analysis. *Austrian Journal of Statistics*, 37, 161–173.

von Eye, A, Mair, P., & Mun, E.-Y. (2010). *Advances in configural frequency analysis*. London: Guilford Press.

von Eye, A, & Mun, E.-Y. (2013). *Log-linear modeling: Concepts, interpretation, and application*. Hoboken: John Wiley.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39, 165–210. DOI: 10.1007/s10579-005-7880-9

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, 4, 253–290. DOI: 10.1515/CLLT.2008.011

Wong, M. (2009). *Gei* constructions in Mandarin Chinese and *bei* constructions in Cantonese: A corpus-driven contrastive study. *International Journal of Corpus Linguistics*, 14, 60–80. DOI: 10.1075/ijcl.14.1.04won

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8, 245–82. DOI: 10.1075/ijcl.8.2.04wul

Wulff, S. (2006). *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis* (pp. 101–126). Berlin & New York: Mouton de Gruyter.

Wulff, S. (2009). *Rethinking idiomaticity: A usage-based approach*. London: Continuum.

Wulff, S. (2010). Marrying cognitive-linguistic theory and corpus-based methods: On the compositionality of English V NP-idioms. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 223–238). Berlin & New York: Mouton de Gruyter.

Wulff, S., Stefanowitsch, A., & Gries, St. Th. (2007). Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In G. Radden, Köpcke, K.-M., Berg, Th., & Siemund, P. (Eds.), *Aspects of meaning construction* (pp. 265–281). Amsterdam & Philadelphia: John Benjamins.

Zeschel, A. (2010). Exemplars and analogy: Semantic extension in constructional networks. In D. Glynn, & K. Fischer (Eds.), *Quantitative Cognitive Semantics: Corpus-driven approaches* (pp. 201–221). Berlin & New York: Mouton de Gruyter.

Zhao, Y. (2013). R and data mining: Examples and case studies. Unpublished manuscript. Available at: http://www.rdatamining.com.

Zlatev, J., & Andrén, M. (2009). Stages and transitions in children's semiotic development. In J. Zlatev, M. Andrén, C. Lundmark, & M. Johansson Falck (Eds.), *Studies in language and cognition* (pp. 380–401). Newcastle: Cambridge Scholars.