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
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.


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). 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

---

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.

| Table 1. Observational differences in collocation and feature analysis of corpora |
|---------------------------------------------|------------------|------------------|
| Stage 1: Analysis of data | objective | subjective |
| Stage 2: Interpretation of analysis | subjective | objective |
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

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

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 collocaional ‘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. 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/](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.

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.


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 comprehensive 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), Szmercsanyi (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


---

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).


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), Dalgard (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 users, 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 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>
<thead>
<tr>
<th>Technique</th>
<th>Type</th>
<th>Object / collocation</th>
<th>Example of application</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correspondence analysis</td>
<td>multivariate</td>
<td>associations btw. objects of multiple variables</td>
<td>synonymy, concepts – Szelld &amp; Geeraerts (2008)</td>
<td>Le Roux &amp; Rouanet (2010), Haugen et al. (2011), Glynyn (this volume)</td>
</tr>
<tr>
<td>Classification tree analysis</td>
<td>multivariate</td>
<td>identify factors that lead to an outcome / prediction</td>
<td>synonymy, lexical – Robinson (2012a), synonymy, lexical – Levshina et al. (this vol.)</td>
<td>Venables &amp; Ripley (2002), Everitt &amp; Hothorn (2010), Maindonald &amp; Braun (2010)</td>
</tr>
<tr>
<td>Technique</td>
<td>Function</td>
<td>Package</td>
<td>R code tutorial</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>Hierarchical cluster analysis</td>
<td>hclust</td>
<td>stats*</td>
<td>Crawley (2007: 738ff.); Zhao (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>agnes</td>
<td>cluster</td>
<td>Kaufman &amp; Rousseeuw (2005); Maechler (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pvclust</td>
<td>pvclust</td>
<td>Suzuki &amp; Hidetoshi (2006); Suzuki (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>clara</td>
<td>cluster</td>
<td>Kaufman &amp; Rousseeuw (2005); Maechler (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pamk</td>
<td>fpc</td>
<td>Hennig (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ca</td>
<td>ca</td>
<td>Greenacre (2007); Neandić &amp; Greenacre (2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>anacor</td>
<td>anacor</td>
<td>de Leeuw &amp; Mair (2009a, 2013a)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mjca</td>
<td>ca</td>
<td>Greenacre (2007); Neandić &amp; Greenacre (2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MCA</td>
<td>FactoMineR</td>
<td>Lé et al. (2008); Husson et al. (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sammon</td>
<td>MASS*</td>
<td>Maindonald &amp; Baun (2010: 284ff.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>smacofSym</td>
<td>smacof</td>
<td>de Leeuw &amp; Mair (2009b, 2013b)</td>
<td></td>
</tr>
<tr>
<td>Configural frequency analysis</td>
<td>cfa</td>
<td>cfa</td>
<td>Funke et al. (2007); von Eye &amp; Mair (2008)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cfa2</td>
<td>cfa2⁴</td>
<td>No tutorials available, cf. Schönbrodt (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>discrim</td>
<td>ade4</td>
<td>Chessel et al. (2004); Chessel &amp; Dufour (2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rda</td>
<td>klaR</td>
<td>Roever et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>Classification tree analysis / Random</td>
<td>rpart</td>
<td>rpart</td>
<td>Zhao (2012: 32ff.); Therneau et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>forest classification</td>
<td>tree</td>
<td>tree</td>
<td>Venables &amp; Ripley (2002: 266)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ctree</td>
<td>party</td>
<td>Zhao (2013: 29ff.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cforest</td>
<td>party</td>
<td>Strobl et al. (2009a, 2009b)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>quasipois</td>
<td>aod</td>
<td>Lesnoff &amp; Lancelot (2013)</td>
<td></td>
</tr>
</tbody>
</table>

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.
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


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.


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.


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


Dool: 10.1007/978-1-4612-0493-0


DOI: 10.1075/pbns.115
DOI: 10.1515/psicl-2012-0021
DOI: 10.1002/9780470238004
DOI: 10.1002/9781118596289


DOI: 10.1075/pbns.33


