Corpus Linguistics

An International Handbook

Edited by

Anke Lüdeling and Merja Kytö

Volume 2

Offprint

Walter de Gruyter · Berlin · New York

1. Introduction

As documented in the articles of this handbook, the central goal of corpus-based analysis is to describe and interpret generalizable patterns of language use. The subdisciplines of discourse analysis and functional linguistics are similar to corpus linguistics in having a primary interest in research questions relating to language use, but these disciplines form a cline in the extent to which they rely on quantitative methods and the value attached to the generalizability of findings. Discourse analytic studies often focus on detailed discussion of a few texts (see, e. g., the survey of studies contained in Schiffrin/Tannen/Hamilton 2001). Functional linguistic studies often focus on linguistic generalizations, with relatively little attention paid to the representativeness of the texts analyzed. In contrast, corpus linguistic investigations of language use are usually designed as quantitative studies with the goal of generalizable findings representing some domain of use.

The design of a corpus is a fundamentally important consideration for achieving this goal: the corpus must be representative of the target domain of use in order for subsequent analyses to be generalized to that domain (see also article 9). Quantitative analyses are also important for generalizable results, because they provide a measure of the extent to which a pattern holds in different domains of use. Computational techniques figure prominently in corpus analyses simply because they enable studies of a scope not feasible otherwise. Computers make it possible to identify and analyze complex patterns of language use, tracking the use of multiple linguistic features across large text collections. Further, computers provide consistent, reliable analyses; they do not change their minds or become tired during an analysis.

The present article provides a survey of the major research designs used in quantitative corpus-based analysis. The article begins with a brief overview of issues of corpus design, recognizing the central place of corpus representativeness for this research approach. We then organize the remainder of the article around the three major kinds of research questions investigated with corpus-based analysis: studies with the word as the primary unit of analysis; studies with a linguistic feature as the primary unit of analysis; and studies with texts as the primary unit of analysis. Finally, we briefly discuss the role of inferential statistics in corpus-based studies of language use.

2. Corpus design issues

To serve as the basis for linguistic investigations, corpora must be designed to represent particular registers, dialects, or other domains of use (see article 9). Part II of this handbook documents several of the most important domains of use that corpora have been designed to represent: written registers (article 10), spoken registers (article 11), historical genres (article 14), learner language (especially for learners of English; article 15), electronic texts (including computer-mediated communication and web registers; see articles 17–18), and translations of 'parallel texts' in two or more languages (article 16).

The representativeness of the corpus limits the kinds of research questions that can be addressed and the generalizability of the results of the research. For example, a corpus composed of newspaper texts would not provide the basis for a general investigation of

variation in English; it would not represent the patterns of use found in spoken registers or even in most written registers.

Representativeness is determined by two considerations: composition and size. The composition of a corpus refers to the text categories included in the design of the corpus. The size of a corpus refers to the number of words or number of texts in the corpus (see below).

Sampling methods are used to select texts in a way that represents the target text categories and size. There are two general approaches to sampling: proportional and stratified. A **proportional sample** represents groups to the extent that they occur in the larger population. For example, proportional samples are useful for political surveys, where predicting the final vote depends on the proportion of each demographic subgroup. However, proportional samples are usually not useful for studies of linguistic variation and use. For example, if we collected all the language produced and received for a week by residents of a city in the U.S., we could identify the actual proportions of language varieties that these people experienced – probably something like 80% conversation, 10% television shows, 1% newspapers, 1% novels, 2% meetings, 2% radio broadcasts, 2% texts that they wrote (memos, email messages, letters), and 2% other texts (signs, instructions, specialist written texts, etc.). A proportional corpus of this type would tell us how often a person is likely to encounter a certain word in the course of a typical week. However, a proportional corpus would be of limited use for studies of variation, because most of the corpus would be conversation. We could not use such a corpus to study language use patterns in other registers, because they would not be adequately represented. Even a popular register such as newspaper language would be minimally represented in such a corpus; specialist registers such as legal documents or medical/scientific research articles would be virtually nonexistent in the corpus. However, a general study of linguistic variation in English would normally want to consider the full range of registers considered to be important, regardless of their proportional use by typical speakers over the course of a normal week.

A **stratified sample** is constructed for these latter research purposes. A stratified corpus is designed to represent the full range of linguistic variation that exists in a language, not the proportions of variation. In a stratified corpus, the researcher begins by identifying the text categories that are the focus of research; texts are then selected from each of those categories so that each category is adequately represented. The text categories could be register differences, dialect differences, or other discourse domains, depending on the primary research questions.

Register variation is central to descriptions of language use: all speakers of a language control many registers, and every time people speak or write, they must choose a register. Therefore, a well-designed corpus will usually focus on a single specialized register, or it will be designed to represent a range of registers. Regional and/or social dialects are also important to consider if the research questions relate to dialect differences. In addition, more specific parameters, such as subject matter, can be important for some research questions.

The second major consideration that determines corpus representativeness is size. Although corpus size is often measured by the total number of words in the corpus, other considerations are at least as important: the number of texts from different categories, and the number of words in each text sample. If too few texts are included, a single text can have an undue influence on the results of an analysis. Additionally, the number

of texts is the most important consideration for studies with the text as the unit of analysis (see section 5). Enough texts must be included in each category to capture variation across speakers or authors (see Biber 1990, 1993).

The number of samples from a text also deserves attention, because the characteristics of a text can vary dramatically internally. A clear example of this is experimental research articles, where the introduction, methods, results, and discussion sections all have different patterns of language use. Thus, sampling that did not include all of these sections would misrepresent the language patterns found in research articles.

Finally, the number of words in each sample is important for providing a reliable count of features in a text. Most earlier corpora used relatively small text samples (1,000–2,000 words), while recent corpora often include complete texts (sometimes over 50,000 words long).

Lexicographic studies (see section 4.2.) require particularly large corpora. Many words and collocations occur with low frequencies, and a corpus must contain many millions of words, taken from many different texts, to enable investigations of word use. However, for all kinds of research, both size and composition are important considerations.

Biber (1993) describes how the representativeness of a corpus can be investigated empirically, depending on the patterns of variation for target linguistic features. For example, rare linguistic features (e. g., clefts) and features that show greater variability (e. g., relative clauses) will require larger samples – both longer texts and a greater number of texts. In contrast, common features with stable distributions, such as pronouns, can be represented accurately in a relatively small sample. However, all linguistic features vary across registers (see the survey of grammatical features in the *Longman Grammar of Spoken and Written English*; Biber et al. 1999). Thus, although the distribution of common features like pronouns and nouns can be represented accurately with relatively short text samples, a single-register corpus (like newspapers) would tell us nothing about the use of nouns and pronouns in other registers (like conversation or fiction).

3. The unit of analysis in corpus-based studies

One of the first decisions required when carrying out a quantitative corpus-based analysis is to determine what the unit of analysis is. This is a crucial decision because it determines the object of the research and the way data should be collected and organized, which in turn limits the research questions that can be asked and the statistical techniques that can be applied (see Biber/Conrad/Reppen 1998, 269–274).

Corpus-based studies generally have one of two primary research goals: 1) to describe the variants and use of a linguistic structure, or 2) to describe differences among texts and text varieties, such as registers or dialects.

Three major types of research design have been employed in corpus research. The primary difference among these research design types is the unit of analysis, which in turn makes each design type appropriate for one of the above two research goals. In Type A studies, the unit of analysis is each occurrence of a linguistic feature. Type A studies are thus designed for Research Goal 1 (describing the variants of a linguistic structure). In Type B studies, the unit of analysis is each individual text. Type B studies

are thus designed for Research Goal 2 (describing the differences among texts and text varieties). Finally, in Type C studies, the unit of analysis is the entire corpus (or different subcorpora). Type C studies can be used for either Research Goal 1 or 2, but they do not permit the use of inferential statistics (see below).

These units of analysis are the 'observations' of the study. For the purposes of quantitative analysis, the main difference between the three research design types is the nature of these observations. That is, the observations in Type A studies do not have quantitative characteristics, while the observations in Type B studies are analyzed in terms of quantitative characteristics. Type C studies differ from both of the others in that there are actually very few observations – usually only 2 or 3 observations – because each subcorpus is treated as an observation.

For example, a Type A study of relative clauses might have the goal of predicting the choice of relative pronoun (*who*, *which*, *that*). For this purpose, we could analyze each relative clause to record characteristics such as whether the clause is restrictive or non-restrictive and whether the head noun is human, animate but not human, or inanimate. In this research design, there are three **variables**: relative pronoun type, head noun type, and clause type. All three variables are **nominal**, meaning that the values are simply categories. Most importantly for our purposes here, these variables are *not* numeric: they do not describe greater-than or lesser-than characteristics. For example, there is no sense in which the relative pronoun *who* is quantitatively greater or lesser than *which*.

In contrast, the observations in Type B studies (each text) have quantitative characteristics, so they can be analyzed with respect to true numeric variables. For example, in a Type B comparison of newspapers and academic prose, we would treat each text as an observation. These texts could be analyzed to determine the rate of occurrence of linguistic features, such as nouns, verbs, and relative clauses. In this case, these variables have an **interval** scale, meaning that they represent greater-than and lesser-than relationships, with a unit of 1 being fixed. For example, a text with 17.5 relative clauses per 1,000 words has a greater rate of occurrence than a text with 14.5 relative clauses per 1,000 words. Because there are many observations in a Type B study (i.e., the total number of texts in the corpus), it is possible to compute mean scores and standard deviations, and to use inferential statistics to compare the rates of occurrence across text categories (see 6 below).

Type C studies are similar to Type B studies in that the observations have quantitative characteristics. That is, in a Type C study, each subcorpus is treated as an observation, and it is possible to compute the rate of occurrence for linguistic features in each subcorpus. However, in this case there are only 2–3 observations included in the entire study; and each text variety is represented by only a single observation (i.e. the subcorpus for that variety). As a result, it is not possible to compute a mean score or standard deviation in a Type C study, and no inferential statistics are possible.

Notice that although the three research design types address fundamentally different kinds of research questions, they could all be based on the same corpus. In Type A studies, however, the focus is on accounting for the variants of a single linguistic feature. In contrast, Type B studies describe the differences among texts.

In general, most previous corpus-based studies have used either a Type A design or a Type C design. This handbook provides several examples of both types of research design. For example, the articles on historical corpora (14), learner corpora (15), discourse analysis (49), dialectology (53), contrastive studies (54), collocations (58), and

grammatical colligations (43) all illustrate research with Type A designs. Article 36, on "Statistical methods for corpus exploitation", provides detailed information on the statistical analysis of Type A designs. Several other articles in the handbook illustrate Type C designs, including the articles on language teaching (7), distributions in text (37) and recent language change (52). In contrast, corpus-based studies with Type B designs are less common; the article on multidimensional approaches (38) discusses several of those studies.

In the following sections, we provide more detailed descriptions and case studies of each research design type. In section 4, we describe Type A studies, which focus on the variants of a linguistic feature. In this section we also introduce studies that have individual words as the unit of analysis – a special case of Type A (see also article 37). Such lexicographic studies have been one of the most important applications of corpora, and a number of special analytical techniques have been developed specifically for this type of study; thus we treat these research designs in a separate subsection (4.2.). In section 5, we describe Type B studies, which compare the typical linguistic characteristics of different texts. Then, in section 5.3. we discuss Type C studies, which compare the typical linguistic characteristics of different subcorpora.

4. Type A designs: Corpus-based studies of a linguistic feature

Corpus-based studies that use Type A designs are focused on particular linguistic features. There are two main kinds of Type A design: those that focus on the contextual factors that lead language users to choose one variant of a linguistic feature over another, and those that focus on the co-occurrence of words (or *collocation*). In section 4.1. we describe studies of linguistic variants, and in section 4.2. we describe corpus-based studies of collocation.

4.1. Corpus-based studies of linguistic variants

4.1.1. The goal

Many functional studies of linguistic variation use a corpus-based approach. The subfield of functional linguistics is based on the premise that linguistic variability is not arbitrary; rather, there are systematic contextual factors that influence the choice of one linguistic variant over another. Corpus linguistics provides an ideal approach to the investigation of linguistic variation, because it allows the researcher to observe numerous tokens of the linguistic feature in natural contexts.

The goal of such studies is to analyze the distribution of linguistic variants across a range of contexts, in order to predict the choice of one variant over another (see also article 43). Several corpus studies have investigated the contextual and functional differences among seemingly equivalent linguistic variants, such as:

dative movement (*give John a ball* versus *give the ball to John*);
particle movement with phrasal verbs (*look up the answer* versus *look the answer up*);
raising and extraposed constructions (*John is difficult to please* versus *It is difficult to please John* versus *To please John is difficult*);
that-deletion (*I think that*/0 *I should be able to go*);

Similar studies have investigated the use of related constructions, even when these are not strictly equivalent. These include studies of:

active vs. passive constructions (*researchers consider many factors* versus *many factors were considered*);
that-clauses vs. *to*-clauses (*I hope that I can go* vs. *I hope to go*);
WH-clefts vs. *it*-clefts (*What happened three years ago was that I decided to go back to school* vs. *It was three years ago that I decided to go back to school*)

Corpus investigations are used to study natural occurrences of these constructions, to isolate the influence of factors such as phrase length, pronominal versus full noun objects, topic continuity, informational prominence, and purely lexical factors. Studies of this type include Thompson/Mulac (1991), Prince (1978), Collins (1991), and Oh (2000). The *Longman Grammar of Spoken and Written English* (LGSWE) also includes many corpus-based analyses of this type (Biber et al. 1999).

4.1.2. A case study: Complementizer *that* versus 0

An example of a Type A study is an analysis of the contextual factors that influence the choice between the complementizer *that* and 0 in *that*-clauses. These variants seem equivalent in many contexts, as in:

I do not think that the situation is slipping out of control.

versus

I don't think [ ] any of us would be willing to do that.

In a study of this linguistic choice, each occurrence of a *that*-complement clause is a separate observation. For each *that*-clause, we would record whether *that* or 0 was used; this is the variable that we are trying to predict. Other variables would record the characteristics of the context, so that we can determine which contextual factors are most strongly associated with each variant.

Several contextual factors might be influential in making the choice between these two variants (see Thompson/Mulac 1991; Biber et al. 1999, 681–683). For example, the complementizer *that* might be omitted more often with common matrix verbs, such as *think*, than with less common verbs, such as *show*. It also might be the case that the complementizer *that* is omitted more often with first person pronouns as subject than with other subjects. Finally, register can also be a contextual factor for individual linguistic features; thus, it might be that the complementizer *that* is omitted more often in conversation than in other registers. (Several other contextual factors turn out to be influential in this case. These include co-referential subjects in the matrix clause and *that*-clause; presence of an intervening noun phrase between the controlling verb and *that*-clause; and whether the controlling verb is in active or passive voice. For the sake of simplicity, the case study here is restricted to three variables.)

To investigate the relative influence of these contextual factors, we would code a large sample of *that*-clause constructions, where each occurrence of a clause constitutes a separate observation. The output of such an analysis might look like Table 61.1.

Tab. 61.1: Coded observations for the analysis of *that*-omission

Complementizer	Matrix verb	Subject	Register
that	indicate	noun	academic
that	suggest	noun	academic
that	imply	noun	academic
0	say	pro-he	newspaper
that	argue	noun	newspaper
that	say	pro-I	newspaper
0	think	pro-I	newspaper
that	report	pro-they	newspaper
0	think	pro-I	conversation
0	say	pro-I	conversation
that	feel	pro-I	conversation
0	think	pro-I	conversation
0	think	pro-he	conversation
0	know	pro-I	conversation

Each line represents information about a single observation, i.e., a single occurrence of a *that*-clause. Each column represents the values for a different variable. The first column shows whether the complementizer *that* is present or omitted; this is the linguistic choice that we are trying to predict. The other columns represent contextual factors: the second column records the matrix verb; the third column records the type of matrix-clause subject; and the fourth column records the register that the example was taken from. For example, the first line in the output above gives the codes for the following sentence:

This analysis indicates that plant growth and nutrient uptake are directly linked. (Academic)

In this case, the complementizer *that* is present; the matrix verb is *indicates*; the matrix clause subject is a full noun phrase (*this analysis*); and the sentence is taken from an academic text. Note that the values for matrix verb have been consolidated, to represent the different verb lemmas (e.g., INDICATE), rather than the individual inflected forms (e.g., *indicated, indicates*).

Data such as these allow quantitative analyses to determine the association of different contextual factors with each structural variant. The simplest kind of analysis is to compute simple frequency counts. In research designs of this type, where each occurrence of a linguistic feature represents an observation, the variables have nominal rather than quantitative values; that is, the values represent different categories. For example, the variable 'matrix verb' has nominal values, such as 'indicate', 'suggest', and 'imply'.

The variable 'complementizer' has two values: 'that' and '0'. If we compute frequencies for these values from the data given in Table 61.1, we would obtain the results given in Table 61.2.

Tab. 61.2: Frequencies of complementizer variants from Table 61.1.

Complementizer	Frequency
0	7
that	7
Total:	14

We could similarly compute frequencies of the values for the variable 'Matrix verb', giving the results listed in Table 61.3.

Tab. 61.3: Frequencies of matrix verb variants from Table 61.1.

Matrix verb	Frequency
think	4
say	3
indicate	1
suggest	1
imply	1
argue	1
report	1
feel	1
know	1
Total:	14

In cases like this, we can simplify the results by grouping all values that occur only one time; see Table 61.4.

Tab. 61.4: Frequencies of matrix verb groups from Table 61.1.

Matrix verb	Frequency
think	4
say	3
other verbs	7
Total:	14

Simple frequency tables can be combined to produce 'cross-tabulation' tables (or 'cross-tabs'), which display the frequency counts for each combination of values across variables. Cross-tabs show the extent to which contextual factors are associated with each linguistic variant. Table 61.5 is a cross-tabulation table for the observations coded above.

Tab. 61.5: Cross-tabulation frequencies of complementizer choice by matrix verb

Complementizer	Matrix verb			
	think	say	other	Total
0	4	2	1	7
that	0	1	6	7
total	4	3	7	14

This distribution indicates that matrix verb is an important factor influencing the choice of complementizer. Six of the seven clauses with an omitted complementizer have the matrix verbs *think* or *say*; in contrast, only one of the seven clauses with *that* have the matrix verb *think* or *say*. (Large-scale corpus analysis of these constructions shows that the verbs *think* and *say* are by far the most common verbs controlling *that*-clauses. Thus, the zero-complementizer is favored by the presence of the most common controlling verbs, while *that*-complementizer is favored by the presence of other controlling verbs; see Biber et al. 1999, 680–683.)

Register can also be used as a variable in a cross-tabulation table. Table 61.6 shows the findings compiled from Table 61.1 above.

Tab. 61.6: Cross-tabulation frequencies of complementizer choice by register.

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
that	3	3	1	7
total	3	5	6	14

From this distribution we can see that retention is favored in academic prose (all 3 clauses), but omission is favored in conversation (5 of 6 clauses). Of course, many more observations are needed as the basis for an actual study of this type.

Given a large enough data set, we could also consider the influence of multiple factors at the same time. For example, we could contrast the influence of the matrix verb for *that*-clauses in news reportage with the influence of this variable in conversation.

Inferential statistical techniques can be used to analyze the significance and strength of these associations. The chi-squared test is the simplest of these techniques, while VARBRUL analysis allows for much more sophisticated investigations of this type. We briefly discuss the use of statistical tests in section 7 below.

4.2. Corpus-based studies of the co-occurrence of words

One important application of corpus-based research has been to study the extended meanings and patterns of use associated with individual words (see article 58). For example, corpus-based research has shown that the most obvious meaning of a word often turns out to *not* be the most common meaning.

Corpus-based analyses of individual words often rely on the construct of 'collocation': how a word tends to occur together with other specific words. Often, words that are supposedly synonymous are shown to be strikingly different in terms of their collocations. For example, the verbs *turn*, *go*, and *come* can all be used as resulting copulas, with a meaning of 'to become' or 'to change to a different state'. However, a consideration of their collocates immediately shows that these copular verbs are dramatically different in their extended meanings (see the *Longman Grammar of Spoken and Written English*, Biber et al. 1999, 444–445). The copular verb *turn* usually collocates with color adjectives (e.g., *turn white*) or adjectives that describe physical appearance (e.g., *turn pale*).

The copular verb *go* usually collocates with negative adjectives, like *bad, crazy,* and *nuts.* In contrast, the copular verb *come* collocates with *true* and with adjectives like *alive* and *clean.* By considering the set of common collocations, it is often easy to distinguish among related words that had previously been regarded as synonymous.

Corpus-based studies of collocation make use of a Type A design, where each occurrence of the target word is treated as an observation. The word immediately preceding the target word might be one variable, and the following word could be a second variable. For example, coded observations for the target word *blue* might look like those given in Table 61.7.

Tab. 61.7: Coded observations for an analysis of the collocates of *blue.*

Preceding word	Target word	Following word
your	blue	eyes
the	blue	sky
had	blue	and
her	blue	eyes
a	blue	napkin

Once a large number of observations have been coded, it is possible to identify important collocates of the target word by computing frequencies for each of the values of a variable. It is important to note such frequencies are not variables. Rather, frequencies are simply counts of how often each value occurs for a variable. The above example included two variables: the word positions immediately preceding and immediately following the target word. The actual words that precede or follow the target word are the values of those variables; *eyes, sky, and,* and *napkin* are values of the variable 'following word' in the example above. Frequency counts are a kind of descriptive statistic that tells us how often each of these values occurs.

To illustrate, consider the three most frequent 'following words' (or right collocates) of *blue* in a 1.67-million-word corpus of fiction; see Table 61.8.

Tab. 61.8: Frequencies of the three most frequent right collocates of *blue.* (Note the corpus is a sub-sample from the *Longman Spoken and Written English Corpus.*)

Target word	Following word	Frequency
blue	eyes	39
blue	and	25
blue	sky	11

Such frequency information identifies the most common combinations, which can be interpreted as collocations. Other combinations, such as *blue napkin,* rarely occur and so should not be regarded as collocations.

Simple frequency information can present a biased measure of the strength of a collocation, because very frequent words are likely to occur together simply by random chance. For example, the combination *and the* occurs 3,248 times in this fiction corpus, even though we would not normally consider this to be a strong collocation.

An alternative method to assess the strength of collocations is to compare the frequency of a combination with the likelihood that the words will occur together simply by chance. The most commonly used statistical measure for this purpose is the 'mutual information score': a ratio of the observed frequency (f_o) of the combination divided by the expected frequency (f_e) of the combination:

Mutual Information Score = f_o / f_e

(The original formula converts the result to a base-2 logarithm; see Church/Hanks 1990; Church et al. 1991. Many practitioners, however, use the simpler version given above, because the results are more easily interpretable; see Stubbs 1995; Barnbrook 1996. Article 58 provides a detailed discussion of collocation statistics.)

The expected frequency is the frequency if the combination were to occur merely by chance; it is computed by multiplying the total frequencies of the two words, divided by the corpus size (1,672,055 words in the present case):

Expected frequency (fe) = (Target word frequency * Collocate word frequency) / Total corpus size

The observed frequencies of the individual words in the above example are:

Target word:
blue 371

Possible collocates:
eyes 1,649
and 49,598
sky 379

Using the formula above, we can compute the expected frequency (f_e) for *blue eyes*:

f_e (*blue eyes*) = (371 * 1649) / 1,672,055 = .37

The expected frequencies for the other two combinations are:

f_e (*blue and*) = (371 * 49598) / 1,672,055 = 11.0

f_e (*blue sky*) = (371 * 379) / 1,672,055 = .08

Notice that the expected frequency reflects the absolute frequencies of the individual words. For example, *and* is extremely common by itself, and therefore we can expect to find the combination *blue and* occurring together by random chance (shown by the expected frequency of 11.0). In contrast, both *blue* and *sky* rarely occur by themselves, and it is therefore extremely unlikely that we would find this combination by random chance (shown by the expected frequency of only .08).

The mutual information score compares the size of the actual observed frequency of a word combination to its expected frequency. For example, the mutual information score for *blue eyes* is:

Mutual info (*blue eyes*) = f_o (*blue eyes*) / f_e (*blue eyes*) = 39 / .37 = 105.4

This is a relatively large mutual information score, showing that this combination represents a strong collocation. In contrast, the mutual information score for *blue and* shows that this combination represents a much weaker collocation:

Mutual info (*blue and*) = 25 / 11.0 = 2.3

At the other extreme, the mutual information score for *blue sky* indicates that this combination represents an even stronger collocation than *blue eyes*:

Mutual info (*blue sky*) = 11 / .08 = 137.5

Although the mutual information index gives us a measure of the strength of association between two words, it can be misleading with particularly infrequent words. For example, imagine that the pair of words *blue marlin* appears once in our corpus, and that the word *marlin* appears only twice. We can then calculate the expected frequency of *blue marlin*:

f_e (*blue marlin*) = (371 * 2) / 1,672,055 = .0004

Because the expected frequency of *blue marlin* is so low (due to the low frequency of the word *marlin*), the mutual information index for *blue marlin* is misleadingly high:

Mutual info (*blue marlin*) = 1 / .0004 = 2500

The t-score is a second statistic used with collocations. While the mutual information index gives a measure of the strength of association between two words, the t-score is used to contrast the collocates of two supposedly synonymous words. Church et al. (1991) use t-scores to contrast the collocates of *strong* and *powerful*. For example, the words *showing*, *support*, *defense*, and *economy* are much more likely to occur as collocates of *strong* than with *powerful*. In contrast, the words *figure*, *minority*, *military*, and *presidency* are much more likely to occur as collocates of *powerful*.

Mutual information scores are the most commonly used measure of collocation, because they are easy to interpret: they simply measure the strength of association between a target word and a potential collocate. T-scores are more difficult to interpret because they are measures of dissimilarity, contrasting the possible collocates of two target words.

5. Type B designs: Corpus-based studies of texts and text categories

The second major type of research design in corpus linguistics examines differences between texts and text categories. Texts can have many nominal characteristics, such as the register category of the text, whether the text is spoken or written, whether the author is female or male, etc. However, texts also have quantitative characteristics: rates of occurrence for linguistic features. It is these quantitative characteristics that distin-

guish Type A and Type B research designs. In the following subsection, we discuss the methods for computing normed variables, which express the rates of occurrence for linguistic features in texts. Then, in section 5.2. we present a case study with a Type B research design. In section 5.3., we discuss a special kind of Type B design that treats subcorpora as units of analysis.

5.1. Normed rates of occurrence

When corpus-based studies examine counts of features across texts, it is important to make sure that the scores are comparable (see also articles 36, 37, and 41). In particular, if the texts in a corpus are not all the same length, counts from those texts are not directly comparable. For example, imagine that you analyzed two texts and found that each one has 20 modal verbs. It might be tempting to conclude that modals are equally common in the texts. However, further imagine that the first text has a total length of 750 words, and the second text is 1,200 words long. Because the second text is longer, there are more opportunities for modals to occur; therefore simply comparing the raw counts does not accurately represent the relative rate of occurrence of modals in the two texts.

'Normalization' is a way to convert raw counts into rates of occurrence, so that the scores from texts of different lengths can be compared. Normalization takes into account the total number of words in each text. Specifically, the raw counts are divided by the number of words in the text and then multiplied by whatever basis is chosen for norming. To continue with the example above, the counts in the two texts could be normed to a basis per 1,000 words of text as follows:

Text A:
(20 modals / 750 words) × 1000 = 27.5 modals per 1,000 words

Text B:
(20 modals / 1200 words) × 1000 = 16.7 modals per 1,000 words

You can see from these normed rates of occurrence that the raw counts are very misleading in this case; that is, modal verbs are actually considerably more common in Text A than in Text B.

In the above example, counts were normed to a basis of 1,000 words since both texts were approximately this long. In a corpus with shorter texts, counts might be normed to rates per 500 words of text. When working with very short texts, such as the writing of children, it might even be necessary to norm counts to rates per 100 words of text. If a higher basis is adopted, the counts for rare features can be artificially inflated – sometimes dramatically so. For example, if a student text of 80 words happened to have one passive construction (a generally rare feature in elementary student writing), and the texts were normed to a basis of 100 words, the text would have a normed score of 1.25 passives per 100 words. However, if that same count was normed to a basis of 1,000 words, the normed value would be 12.5 passives per 1,000 words, which represents a rate of occurrence unlikely to be achieved in any extended elementary student writing. Thus, counts should be normed to the typical text length in a corpus.

5.2. A case study: Newspaper and conversation texts

Normalized rates of occurrence for any linguistic feature can be analyzed as 'interval' variables (see section 3 above). Coded observations in this type of study would look like the display in Table 61.9. Each line in the display represents information about one text, and each column gives the values for a variable. In this case, two of these variables are nominal: text identification and register. The other four variables are interval, giving the total word count and quantitative rates of occurrence for past tense verbs, attributive adjectives, and first person pronouns. The scores for the three linguistic features have been normalized to a rate per 1,000 words of text.

Tab. 61.9: Linguistic data for twelve texts

Text ID	Register	Word count	Past tense	Attrib adjs	1st person pronouns
n1.txt	news	2743	47.4	68.1	3.1
n2.txt	news	1932	49.2	63.0	9.2
n3.txt	news	2218	42.2	74.8	7.1
n4.txt	news	2383	45.3	72.1	2.2
n5.txt	news	1731	47.1	67.3	5.4
n6.txt	news	2119	51.2	70.0	5.2
c1.txt	conv	2197	32.2	43.1	62.6
c2.txt	conv	2542	37.4	36.3	59.1
c3.txt	conv	2017	36.8	39.7	58.7
c4.txt	conv	1896	29.2	35.2	65.5
c5.txt	conv	1945	31.3	34.0	58.2
c6.txt	conv	2072	23.8	38.3	60.4

Because the observations in this type of study (i.e., the texts) are described with respect to quantitative variables, it is possible to compute descriptive statistics like a mean score (expressing the central tendency) and standard deviation (expressing dispersion). For example, the mean score for past tense verbs in the data above is:

$$(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2 + 32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 12 = 39.4$$

It is similarly possible to compute separate mean scores for each register, as in:

Mean score of past tense verbs for newspapers:
$(47.4 + 49.2 + 42.2 + 45.3 + 47.1 + 51.2) / 6 = 47.1$

Mean score of past tense verbs for conversations:
$(32.2 + 37.4 + 36.8 + 29.2 + 31.3 + 23.8) / 6 = 31.8$

A second descriptive statistic – the 'standard deviation' – indicates the extent to which texts are dispersed away from the mean score. Descriptive statistics like these can also be used for inferential statistical techniques like t-test or ANOVA, to test the statistical 'significance' of observed differences among registers (i.e., by comparing the size of the mean differences among registers relative to the size of the differences among the texts within a register). Inferential statistics are discussed further in section 7 below.

5.3. Type C designs: Corpus-based studies with subcorpora as the unit of analysis

It is also possible to use different subcorpora as observations, to compute rates of occurrence of linguistic features in each subcorpus (Research Design Type C). In this case, each subcorpus is treated as if it were a large text. This is the design used for the most part in the *Longman Grammar of Spoken and Written English*, where the rates of occurrence for grammatical features (per one million words) are compared across conversation, fiction, newspapers, and academic prose. For example, the following calculations compare the overall normed rates of occurrence for past tense verbs in the subcorpora of conversation and newspapers:

Conversation (corpus size = 3,929,500 words):
(113,170 past tense verbs / 3,929,500) * 1,000,000 = 28,800 per million words

Newspapers (corpus size = 5,432,800 words):
(204,273 past tense verbs / 5,432,800) * 1,000,000 = 37,600 per million words

Studies with this research design give similar kinds of findings to those where each text is treated as an observation. There is a major difference though: in this design, we compute a single, overall rate of occurrence for each register, based on the subcorpus for that register. In contrast, in studies with texts as the unit of analysis, we compute a rate of occurrence for each text, which provides the basis for computing mean scores and standard deviations. Thus, when the text is the unit of analysis, we can compute both the typical rate of occurrence (the mean score) and the extent to which individual texts vary away from that typical score (the standard deviation). This provides the basis for inferential statistical tests (such as ANOVA and correlational techniques); such tests are not possible with designs that use subcorpora as the units of analysis. Despite this fact, subcorpora are commonly used as the units of analysis in studies where the differences across registers are so large that inferential statistics are not essential (see section 7 below).

6. Comparing Type A and Type B designs for register analyses

Because the observations in Type B designs are texts, this kind of analysis is especially well suited to comparing the linguistic characteristics of registers (or other text categories), as in section 5.2. above. However, Type A designs can also be used to study register differences. For example, Table 61.6 above compares the use of *that* and 0 complementizers across three registers (academic prose, newspapers, and conversations).

Although both design types can be used to analyze register differences, there is a crucial distinction in the information that they provide: Type B designs are based on rates of occurrence, and so they tell us how common a linguistic feature is. In contrast, Type A designs tell us the relative preference for one variant over another, but we have no way of knowing how common the features actually are.

Tab. 61.10: Cross-tabulation frequencies of complementizer choice by register

Complementizer	Register			
	Academic	News	Conversation	Total
0	0	2	5	7
that	3	3	1	7
total	3	5	6	14

For example, Table 61.6 in section 4.1. above (repeated here as Table 61.10) shows that *that*-retention is favored in academic prose (3 out of 3 clauses), while *that*-omission is favored in conversation (5 of 6 clauses). Based on this table, we might be tempted to conclude that *that*-retention occurs more commonly in academic prose than in conversation. However, a Type A design does not provide the basis for such conclusions: this study provides no information about the corpora used for academic prose and conversation, and so we are unable to determine the actual rates of occurrence. That is, Table 61.6 only gives us proportional information: when a *that*-clause occurs in conversation, it is likely to omit the complementizer. When a *that*-clause occurs in academic prose, it is likely to retain the complementizer. However, Table 61.6 does not tell us the actual rates of occurrence for *that*-clauses in each register.

In fact, it is possible to imagine a scenario where *that*-retention occurs more commonly in conversation, even though it is the dispreferred variant in that register.

Imagine, for example, that the data in Table 61.6 are based on a 50,000-word corpus of academic prose and a 10,000-word corpus of conversation. With this background information, we can compute rates of occurrence for each corpus:

Normed rate of occurrence of *that*-retention in academic prose:
(3 clauses / 50,000 words) × 100,000 = 6 clauses per 100,000 words

Normed rate of occurrence of *that*-retention in conversation:
(1 clause / 10,000 words) × 100,000 = 10 clauses per 100,000 words

In this case, the actual distribution turns out to be the exact opposite of the apparent difference seen in Table 61.6: the rate of occurrence for clauses with *that*-retention is actually higher in conversation than in academic prose. This happens because *that*-clauses overall are many times more common in conversation than in academic prose. As a result, even the dispreferred variant in conversation (with *that*-retention) has a relatively high rate of occurrence.

The important point here is that Type A research designs do not provide the basis for determining rates of occurrence, so they cannot be used to determine if a feature or variant occurs more commonly in one register or another. This is potentially confusing, and even published research studies sometimes make this mistake. Type A studies do tell us what the preferred variant is in a register, and how registers differ in their reliance on a particular variant. For example, Table 61.6 above shows that when a *that*-clause is used in academic prose, it will usually retain the complementizer. When a *that*-clause is used in conversation, it will usually omit the complementizer. This is a genuine register difference. However, it would be incorrect to therefore conclude that *that*-retention is

more common in academic prose. Even though a much higher proportion of *that*-clauses retain the complementizer in academic prose, the actual rate of occurrence for this variant could be higher in conversation – Table 61.6 does not tell us one way or the other.

7. The role of inferential statistics in corpus linguistics

Inferential statistics can be used with all corpus-based research designs to assess whether observed differences might have occurred simply due to chance (see articles 36 and 37 for more detailed discussion). In the case of a Type A study, only non-parametric techniques can be applied, because all variables are nominal. The most commonly used non-parametric technique in this type of study is the chi-squared test. It is also possible to use multivariate statistical techniques for this type of design, such as loglinear regression and VARBRUL.

In contrast, it was pointed out in section 5 that Type B research designs have true numeric variables, which permit descriptive statistics such as mean scores and standard deviations. These designs allow the use of parametric statistical techniques, such as t-test and ANOVA to test for differences across categories, and Pearson correlations to test for relationships among linguistic variables. Multivariate statistical techniques used with these designs include multiple regression, factor analysis, and discriminant analysis.

Inferential statistical tests help to identify meaningful differences, as opposed to differences that occur just due to random chance. However, we would argue that inferential statistics should be used and interpreted with caution in corpus-based research. Tests of statistical significance depend on the sample size (N): as the sample size becomes larger, the difference among groups required to achieve significance becomes smaller. For very large samples – the normal case in corpus-based research studies – relatively small differences between groups are considered significant.

A complementary statistic is the measure of strength, which indicates the importance of a quantitative difference or relationship. With very large samples, it is easy to find small linguistic differences that are statistically significant but not strong; we would argue that these differences are often not interesting, because they do not reflect the important differences across text categories. By also considering measures of strength, researchers can identify the linguistic differences that are important and therefore more interesting for interpretation.

8. Conclusion

In this article, we have attempted to provide an overview of some methodological issues for doing quantitative corpus-based research. Arguably, the most important of these is also the one least often recognized: determining the 'unit of analysis' and the appropriate research design required for a particular research question. Thus, most of the article has focused on those considerations. By considering the design requirements of a research project before embarking on corpus construction, data collection, and/or linguistic coding and analysis, the researcher can ensure that the effort invested in a major research project will in fact result in the intended outcomes.

9. Literature

Barnbrook, G. (1996), *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Biber, D. (1990), Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. In: *Literary and Linguistic Computing* 5, 257–269.

Biber, D. (1993), Co-occurrence Patterns among Collocations: A Tool for Corpus-based Lexical Knowledge Acquisition. In: *Computational Linguistics* 19, 549–556.

Biber, D. (1993), Representativeness in Corpus Design. In: *Literary and Linguistic Computing* 8(4), 243–257.

Biber, D./Conrad, S./Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Church, K. W./Hanks, P. (1990), Word Association Norms, Mutual Information, and Lexicography. In: *Computational Linguistics* 16, 22–29.

Church, K. W./Gale, W. A./Hanks, P./Hindle, D. (1991), Using Statistics in Lexical Analysis. In: Zernick, U. (ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Erlbaum, 115–164.

Collins, P. (1991), *Cleft and Pseudo-cleft Constructions in English*. London: Routledge.

Nakamura, J./Sinclair, J. (1995), The World of *Woman* in the Bank of English: Internal Criteria for the Classification of Corpora. In: *Literary and Linguistic Computing* 10, 99–110.

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

Oh, S-Y. (2000), *Actually* and *in fact* in American English: A Data-based Analysis. In: *English Language and Linguistics* 4, 243–268.

Prince, E. F. (1978), A Comparison of *wh*-clefts and *it*-clefts in Discourse. In: *Language* 54, 883–906.

Schiffrin, D./Tannen, D./Hamilton, H. E. (eds.) (2001), *The Handbook of Discourse Analysis*. Oxford: Blackwell.

Stubbs, M. (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.

Thompson, S. A./Mulac, A. (1991), The Discourse Conditions for the Use of the Complementizer *that* in Conversational English. In: *Journal of Pragmatics* 15, 237–251.

Douglas Biber and James K. Jones, Flagstaff, AZ (USA)