Stefan Th. Gries

Methodological Skills in Corpus Linguistics: A Polemic and Some Pointers Towards Quantitative Methods

1. Introduction

For a variety of reasons, the (corpus) linguist's life is a hard one. One of these reasons is the complexity of the subject under investigation, language. Linguistic behaviour is influenced by a multitude of factors which can be categorized into different categories:

- general aspects of cognition having to do with attention span, working memory, general intelligence, etc.;
- specific aspects of the linguistic system: form, meaning, communicative pressures, etc.;
- other performance factors (e.g., visual distractions, etc.)

What makes it even harder, is that all of these factors influence language only probabilistically rather than deterministically, which makes it very difficult to precisely predict most aspects of human linguistic behaviour. In addition, and this leads to a second reason, the data on the basis of which we try to describe, explain, and predict linguistic behaviour is very fragmented and very noisy. With regard to observational data from corpora, the problem of fragmentation and noise manifests itself in different ways. On the one hand, corpora unfortunately are

- never infinite although language is in principle an infinite system;
- never really representative in the sense that they really contain all parts or registers or genres or varieties of human language;

- never really balanced in the sense that they contain these parts or registers or genres or varieties in exactly the proportions such parts make up in the language as a whole;
- never complete in the sense that they never contain all the contextual information that humans utilize in, say, conversation.

On the other hand, even if all of these issues *could* be addressed, corpora would still be at least one level of abstraction away from what many linguists are probably most interested in: sense, meaning, concepts, communicative function, etc. Rather, the inconvenient truth is that corpora exclusively contain (i) information on (relative) frequencies of occurrence: elements that occur x many times (with $x \geq 0$), (ii) information on dispersion: elements that occur $x \geq 0$ time in particular parts of corpora or at particular distances $d \geq 0$ from each other, (iii) information on (relative) frequencies of co-occurrence (collocation, colligation, etc.), and finally (iv) derivatives of the above (e.g., key words).

Thus, whatever the corpus linguist seeks to study, the first task is always to operationalize the phenomenon of interest in terms of frequencies of occurrence, co-occurrence, or dispersion. For example, a frequent way of operationalizing the semantic similarity of two words x and y (or the similarity of one word's senses x and y) using corpus data involves somehow quantifying the distributional similarity of x and y, say in terms of x's and y's collocates (cf., e.g., Church/Hanks 1989; Church et al. 1994; Gries 2003a; or Manning/Schütze 2000: ch. 5, 7, 14). For example, a frequent determinant for the semantic import of patterns or constructions is to identify the words that occur in particular syntactic slots of the patterns or constructions (cf., e.g., Hunston/Francis 2000 or Stefanowitsch/Gries 2003). It goes without saying that this operationalization is both crucial to the success of the study and tricky in that such operationalization can often provide only rather indirect access to what is being studied.

The second task always involves the recovery of the patterns that reflect the phenomenon of interest from the corpus/corpora. Unfortunately, this step can be just as tricky as the previous one because the distributions in corpora that linguists are interested in are not always easily obtainable because

- the relevant pattern is hard to define and/or can take on many different realizations;
- linguistic patterns can be messy;
- the corpus/corpora to be searched come in user-unfriendly formats and/or contain errors of annotation, transcription, etc.

The third task then involves counting the instances recovered from the corpus and/or analyzing the instances recovered from the corpus with regard to relevant characteristics. In both cases, the result will by definition result in the kind of information mentioned above: frequencies of (co-)occurrence, dispersion, and derivatives of these. To evaluate such information, practitioners of most sciences use the techniques or tools from the one scientific discipline that is concerned with distributions, probabilistic patterns, and (relative) frequencies, etc.: statistics (as well as the related disciplines of computer science and maybe data mining).

All of the above is probably relatively uncontroversial even among the various different kinds of corpus linguists. What may be more controversial, however, are the following two questions that I have been asking myself ever since I began to familiarize myself with corpus-linguistic approaches:

- Why is it that we corpus linguists look at something (language) that is completely based on distributional/frequency-based probabilistic data and just as complex as what psychologists, psycholinguists, cognitive scientists sociologists, etc. look at, but most of our curricula do not contain even a single course on statistical methods (while psychologists etc. regularly have two to three basic and one or two advanced courses on such methods)?
- And why is it that we corpus linguists are often called upon to retrieve complex patterns from gigabytes of messy data in various encodings and forms of organization, but most of our curricula do not contain even a single course on basic programming skills or relational databases (while psychologists, computational linguists, cognitive scientists etc. devote years to acquiring the required methodological skills)?

The answers to both questions are undoubtedly complex and interrelated. However, in this polemic I allow myself to simplify and polarize. My (overly simplistic and overly polemic) suggestion is that there are mainly two reasons for this.

Firstly, corpus linguistics is a divided discipline. One extreme consists of corpus linguists who are humanistically-oriented, do not go beyond what the text or the discourse has to offer, and may even oppose annotation because it renders a corpus impure. The other extreme consists of corpus linguists who view linguistics as part of the social sciences, feel free to speculate about what corpus data reveal about mechanisms and phenomena beyond the discourse itself (not to use the words *mind* or *cognition*), and adopt a more utilitarian approach to corpora so that the idea of annotation as such is no problem and, readers are advised to sit down and even argue in favour of using other methods in combination with corpora.

Secondly, many corpus linguists do not have the necessary methodological skills themselves (both in terms of programming/databases and statistical tools) and, thus, do not require their students to acquire a wide(r) range of methodological skills. I know corpus linguists whose corpus-linguistic skills are defined by what WordSmith Tools (or AntConc, or KwicFinder, or Concgram, etc.) or their web interface can do – if you take whatever program they use away from them, they are unable to pursue their corpus studies to any great extent. At the risk of redundantly mentioning the obvious, let me make even clearer what this means: If these corpus linguists' program(s) cannot lemmatize, neither can they. If their program(s) cannot do regular expressions, neither can they. If their program(s) cannot do keywords for bigrams, neither can they. And if their program(s) cannot do collocational or other statistics, neither can they. And so on. Note, by the way, that I am neither saying nor believe that these corpus linguists have not made or cannot make important contributions to linguistics in general and/or corpus linguistics in particular – they can and they have! But I *am* blaming part of the lack of methodological skills on the fact that some corpus linguists simply employ the easiest option open to them and as long as, say, Mike Scott's very comprehensive program suite does the job, they think, why do more and why require students to do more?

From my own experience, I know at least two reasons why one should do more: First, I believe that a scientist's analytical skills should not be dictated by limited and commercial software, and as useful as each of the above applications is, each is limited in terms of *availability*:

- several of the programs are only available for one operating system;
- many programs are commercial and, thus, not available to researchers from poorer countries;

limited in terms of *functionality*:

- several of the programs compute collocational statistics, but provide only one or two of the available measures of collocational strength;
- several programs compute collocational statistics, but only for words, not for bigrams, trigrams, etc.;
- several of the programs cannot handle Unicode, corpora with standoff annotation, or annotation on different tiers; etc;
- web interfaces do not make the whole corpus available and, thus, do not permit the analyst to perform larger-scale analyses that require the complete corpus;

limited in terms of *user control*:

- users are at the mercy of the developers. If, for instance, the creator of one of the above programs changed the way key words are computed by means of an update function but did not inform the users, then users would have no explanation for why the same data set suddenly yielded different results, and with non-open source software, no user would be able to find out exactly what had happened and how the software is changed. More trivially still, if one developer decided to discontinue a program, then what?

And, as if all the above were not enough, it is difficult to accept that scientists' possibilities of analysis are limited by a piece of commercial software as opposed to the limits of our knowledge, understanding and sources of data.

Second and maybe more obviously, more methodological knowledge is not only desirable in general, but can sometimes suddenly suggest completely new ways of analysis. Once one's thinking about a phenomenon is not determined anymore by the confines of one's out-of-the-box software, hitherto undiscovered ways of analysis may become possible.

In the remainder of this chapter, I will leave aside the issue of corpus-linguistic retrieval and data processing methods and focus on how better statistical methods provide us with more accurate, more reliable, and more comprehensive results. I will discuss several small case studies which, although often do not even involve complex statistical issues, should exemplify the benefits of more advanced quantitative approaches in corpus linguistics to such a degree that readers can at least not say anymore they were not warned! Hopefully, this chapter will therefore inspire many corpus linguists to delve more deeply into the distributional aspects of our trade and its related tools.

2. Case studies

In Section 2.1, I will be concerned with a few very brief and rather elementary examples of cases where (corpus) linguists have either not been as careful or not as comprehensive as they could have been; the examples all involve monovariate or bivariate distributions, i.e., they involve only one or two variables. In Section 2.2, I will then discuss a more complex example of where a multifactorial dataset has not been investigated multifactorially, why that is in general not a good idea, and what we can gain from pursuing the more appropriate multi-factorial methodology.

Before I begin, one final important comment: I would like to emphasize here that the point of this section is *not* to criticize the work of colleagues who I appreciate and whose work has provided good insights into the phenomena they study – the point is to show how one can avoid problems or go beyond the work that has already been done.

However, while it is of course possible to simply invent data to make a methodological point, I think that would invite readers to take my comments less seriously – if, on the other hand, I can show that there is actually published work out there that suffers from particular shortcomings, then this might be more of an incentive to consider acquiring the statistical kind of knowledge I have discussed in Section 1 above.[1] Also, I would like to point out that the reanalysis of data in Sections 2.1.1 and 2.1.2 is only possible because the cited scholars were conscientious enough to make the relevant data available in the study so that other researchers can make use of them.

2.1. Monovariate or bivariate data: pitfalls and how to avoid them

2.1.1. Monovariate data: comparing two hedges in English

An early corpus-based study of the English hedges *kind of* and *sort of* is Aijmer (1984). Rather than just reporting the frequencies of these hedges in different registers, she also studies the hedges' syntactic distributions in the London-Lund Corpus and reports the frequencies of *sort of* before major constituents shown in Table 1.

Table 1. The distribution of *sort of* before major constituents, based on Aijmer (1984: 121).

NP	PP	VP	AdjP	AdvP	Totals
302	8	145	19	8	482

In addition, she provides analogous data for *kind of*, as shown in Table 2.

---

[1] In addition, those readers who know my work are aware of the fact that I myself have been going through a painful methodology-learning process myself, which included having to admit in Gries (2003b) that part of the statistical discussion in Gries (1999) is rather problematic, and from what I know today, even Gries (2003b) could be improved upon. We're all in the same boat when it comes to acquiring methodological skills, and the alerting to, and talking about, the scope, implications, and potential shortcomings of our own studies comes with the job.

NP	PP	VP	AdjP	AdvP	Totals
73	0	5	3	0	81

Table 2. The distribution of *kind of* before major constituents, based on Aijmer (1984: 121).

While Aijmer (1984) makes many more observations, some of which are critically discussed in Gries and David (2007), here I am interested in the following statement made with regard to the data in Table 1 and Table 2. It is that "*sort of* is more common before noun-phrases than before other constituents" (Aijmer 1984: 121). I do not wish to challenge this claim because it is obviously correct since 302 is the largest number in Table 1. However, I think it is also correct that a paper whose overall purpose is to compare *sort of* and *kind of* should point out that *sort of* actually disprefers NPs, namely when it is compared to *kind of*. Consider Figure 1, which combines Table 1 and Table 2 into a crosstabulation plot (Gries 2009).

	NP	PP	VP	AdjP	AdvP
sort of	302	8	145	19	8
kind of	73	0	5	3	0

Figure 1. Crosstabulation plot of Aijmer's (1984) data on *sort of* and *kind of*.

In Figure 1, dark grey and light grey figures indicate observed frequencies that are larger or smaller than expected by chance, and the size of the figures is directly proportional to the degree to which observed and expected frequencies differ from each other. Not only is

the resulting distribution highly significant ($\chi^2=25.43$; $df=4$; $p<0.001$),[2],[3] but as is immediately obvious, two strong effects are in fact that, when *sort of* is compared to *kind of*, (i) the number of NPs following *sort of* is smaller than expected by chance and (ii) the number of NPs following *kind of* is much larger than expected by chance.

Again, I am not contesting the obvious truth of Aijmer's statement. But I am contesting the conversational implication that Aijmer's statement is the only important observation with regard to hedges and NPs. In a paper whose main objective appears to be a comparison of the two hedges, it should be mentioned what the preferences of the hedges look like in comparison, especially if it only takes a simple χ^2-test to make that point.

2.1.2. Bivariate data: *remember* and *forget* in English

The focus of Tao (2003) is the corpus-based observation that the mental-process verbs *remember* and *forget* take complements much less often than many previous intuition-based studies have assumed. His analysis of the two verbs' complementation patterns is based on partial data from three different corpora: a part of the Cambridge University Press/Cornell University Corpus, a sample from the Santa Barbara Corpus of Spoken American English, and a part of the Corpus of Spoken Professional American English. He discusses many different and interesting findings, but for present purposes I want to single out one part of his results and one statement made with regard to that part. Table 3 below summarizes the main results of Tao's (2003: 80) Tables 1 and 2, the distribution of postverbal elements in *remember* and *forget* clauses.

2   The frequencies in Figure 1 actually rule out the use of a chi-square test, which is why the reported *p*-value was also checked with an exact test, which also yielded a *p*-value smaller than 0.001.

3   All computations and graphs were performed and created with R for Windows 2.8.0; cf. R Development Core Team (2008).

	Non-complements	Complements	Totals
Verb: remember	295 (row perc.: 74%)	104	399
Verb: forget	131 (row perc.: 79%)	35	166
Totals	426	139	565

Table 3. Postverbal elements in *remember/forget* clauses (after Tao 2003: 80).

The sentence immediately following these data is "[c]omparing the postverbal elements in the two verbs, we can see that the proportion of non-complements for *forget* is higher than *remember*: 79% vs. 74%" (Tao 2003: 81). Just as with Aijmer's study, I do not wish to challenge this statement: of course, 79% is more than 74%. However, the question arises whether this relation is robust enough to be statistically significant and may thus be mentioned without any further qualification. But this is not the case. I see three obvious ways in which the above statement could be tested on the basis of Table 3. Either one performs a χ^2 test on Table 3 as a whole, or one performs a χ^2 test on whether the distribution obtained for *forget* is different from the one obtained for *remember* (since in Tao's statement *remember* is the standard of comparison), or one performs an exact binomial test testing the likelihood of getting between 131 and 166 non-complements after *forget* when the expected probability of a non-complement is approximately 73.93%.

The χ^2 test for the complete table shows that the distribution in Table 3 is not significantly different from chance and a rather weak effect ($\chi^2=1.57$; $df=1$; $p\approx0.21$; $\varphi=0.05$); the χ^2 test that tests whether the distribution of *forget* {131, 35} is significantly different from the distribution of *remember* {295, 104} – i.e., {73.93%, 26.07%} – also returns a non-significant result ($\chi^2=2.137$; $df=1$; $p\approx0.144$); the binomial test returns a $p_{two\text{-}tailed}$-value of approximately 0.083. Irrespective of how these data are inspected with regard to the summary statement, the observed distribution of non-complements after *forget* does not differ significantly from that of *remember*, and

any summary and/or interpretation of these data should make that very clear to avoid overstating the case.

2.1.3. *Bivariate data: the coupling of tense and grammatical aspect in Russian*

Stoll and Gries (2009) study the correlation between two verbal inflectional categories in Russian, namely the degree to which verbs with past tense or non-past tense marking are used in the imperfective or perfective aspect (cf. Shirai et al. 1998: 246). In general, one finds a strong correlation so that past tense preferably co-occurs with perfective aspect while non-past or present tense preferably co-occurs with imperfective aspect. Stoll and Gries are particularly interested in

- characterizing how Russian children often start out with an extremely strong correlation of past/perfective and non-past/imperfective, but then relax this correlation over time as they learn that it is possible to talk about past evens perfectively;
- comparing the strength of the tense-aspect correlations of children to those of their caretakers;
- measuring the tense-aspect correlation as a true correlation rather than as uncorrected frequencies of tenses and aspects.

To that end, they retrieved from the Stoll corpus of Russian first language acquisition all verb forms produced by the children studied and their caretakers and extracted the tense and grammatical aspect coding for each verb form as well as who produced the utterance. In this chapter, I will only focus on the data for the youngest child, child 3 in the study; for results regarding four other children, cf. Stoll and Gries (2009). For this child, 80 recordings from the age range of 1;11.28 to 4;3.12 with altogether 6,796 child utterances and 31,687 caretaker utterances were available. For each of the 80 recordings, a 2×2 table of the kind exemplified in Table 4 was created.

Table 4. Tense-aspect patterning of child 3 (at age 2;7.28) (from Stoll/Gries, forthcoming).

	Tense: non-past	Tense: past	Totals
Grm. asp.: imperf.	25	5	30
Grm. asp.: perf.	5	10	15
Totals	30	15	45

This distribution shows that, in this recording, the child exhibits a strong correlation of past/perfective and non-past/imperfective, which is statistically highly significant ($\chi^2=11.25$; $df=1$; $p<0.001$). Since, however, χ^2 values are dependent on sample sizes, the development of the child cannot be tracked by comparing the χ^2 values to each other, which is why Stoll and Gries computed an effect size from this χ^2 value ($\varphi/V=0.5$): φ/V ranges from 0 ('no association') to 1 ('perfect association'). Analogous computations for all 80 recordings separately for child 3 and her caretakers make it possible to plot how the V values change over time; cf. Figure 2 for the development of the V values for child 3 (left panel) and her caretakers (right panel), I added a linear regression line with a 95% confidence interval.
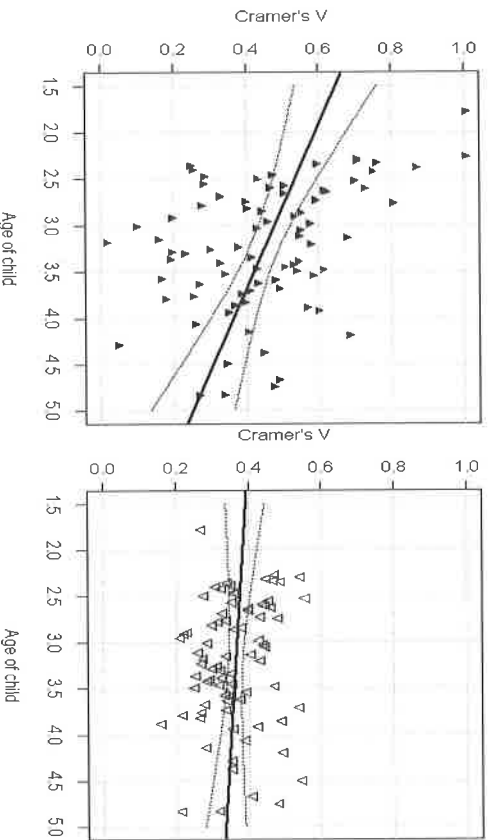

Figure 2. Cramer's V values for tense-aspect correlation (child 3 and her caretakers).

Several things are immediately obvious. First, there is a correlation of past/perfective and non-past/imperfective since nearly all V values are considerably larger than 0. Second, the regression for the caretakers shows an absence of development, which is to be expected since the caretakers are already fluent native speakers of Russian; the slope of the regression line in the right panel does not differ significantly from 0 ($p\approx0.276$). Third, the regression for the child shows a clear developmental trend: the older she becomes, the more she relaxes the overly rigid tense-aspect correlation from her earliest utterances, and the slope of the regression line in the left panel (-0.1129) differs significantly from 0 ($p<0.001$).

This is the point where one might actually complete the data analysis: the correlation of tense and aspect has been confirmed, the child exhibits the expected development (significantly so), the caretakers do not (also expected). However, Stoll and Gries show there is more to be seen, and this recognition arises from the fact that just because we can force a linear regression line through a messy cloud of points, this does not mean that that regression line is the best way to summarize the trend or even one that the data allow for. If one applies a non-parametric smoother to the data, then a very different developmental picture emerges, as is indicated in the left panel of Figure 3.

The finer-grained perspective shows that there is a much steeper developmental trend, but that that trend also ends much earlier, namely around age 3, when the child is already close to the adults' mean value. The right panel results from a regression with breakpoints (cf. Crawley 2007: Ch. 22 or Baayen 2008: Section 6.4) and shows that there is a steep and statistically significant learning process until age 3, and from that age, the child's second regression line levels off horizontally to approximate that of the adults.

In sum, while this section has discussed a case where statistical testing was used – contrary to the two case studies from Sections 2.1.1 and 2.1.2 – it has also showed that just because one particular statistical method yields significant results, one must be careful not to fall into the trap of prematurely accepting these results at face value. Often, only a more sophisticated study of the data can yield the more
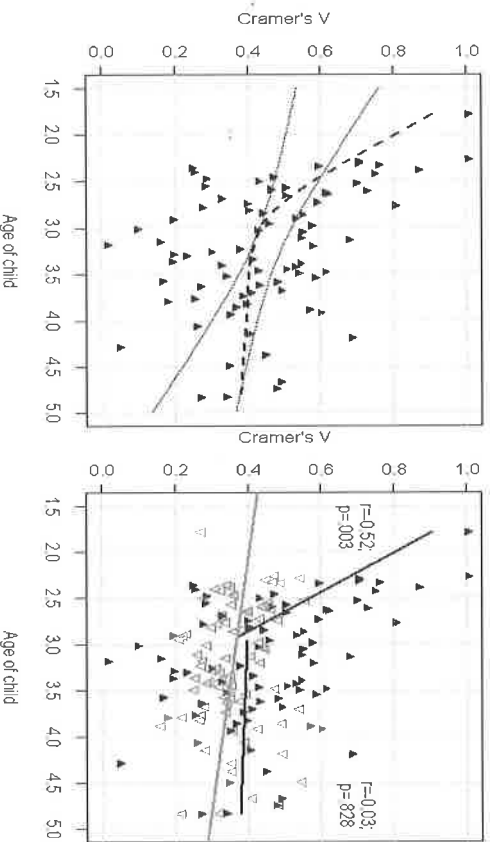
Figure 3. Non-parametric smoothing and regression with breakpoints for the data of child 3.

2.2. Multidimensional frequency tables

2.2.1. Problems of monofactorial approaches to multidimensional data

The above logic of applying more careful analysis of uni-/two-dimensional datasets applies even more strongly to multidimensional datasets. This is because the data are more complex so that (i) interesting patterns do not necessarily reveal themselves to the naked eye and (ii) higher-order interactions impossible to detect without statistical means qualify simple interactions or main effects to such a degree that monofactorial eyeballing may actually lead to interpretations that are the exact opposite of what is really happening. In this section, I will discuss an example from a recent corpus study published in the *ICAME Journal*, Hommerberg and Tottie (2007). Their study explores two complementation patterns of the verb *try* in British and American English: *try to* vs. *try and*. Their goal is "to

show how native speakers of present-day British and American English actually use the two constructions", and they use a data set from the Cobuild Direct Corpus, whose size and composition is summarized in Table 5.

Table 5. The data studied by Hommerberg and Tottie (2007).

VARIETY	American		British		Totals
MODE	spoken	written	spoken	written	
TRY: *and*	284	44	1663	217	2208
TRY: *to*	893	773	694	679	3039
Totals	1177	817	2357	896	5247

Each instance of *try to* and *try and* was also coded for the additional variable MORPH, which indicates the morphological form of *try* and identifies *imperative, infinitive, present,* and *past* (for details of the coding, cf. Hommerberg/Tottie 2007: 49f.). In this section, I will, however, only be concerned with the infinitive data they use to study the Rohdenburg's (2003) *horror aequi* principle. In this particular case, the *horror aequi* principle predicts that when the verb *try* is used in a *to*-clause, then an additional *to*-complement of *try* will be avoided. Thus, a is expected to be more likely than b (Hommerberg and Tottie's example):

    a. We understand the risks, and we're going to try and beat this thing.
    b. We understand the risks, and we're going to try to beat this thing.

As for the presentation of the data, Hommerberg and Tottie again laudably provide their data in a way that makes their analysis replicable. As for the evaluation, however, they fall short of the mark. Note that their data set is multidimensional in nature since it involves the impact of three independent variables VARIETY (*American* vs. *British*), MODE (*spoken* vs. *written*), and CLAUSETYPE (*to* vs. *other*) on one dependent variable TRY (*and* vs. *to*). However, while Hommerberg and Tottie do perform significance tests and are thus

already ahead of some of the studies discussed in Section 2.1, they restrict their analysis to (i) χ^2-tests (all $dfs=1$; all $ps<0.003$) of the interactions listed below and (ii) summarize the findings from this rather complex data set in one paragraph, stating that *horror aequi* is stronger in written British English than in spoken British English and stronger in British English than in American English:

- CLAUSETYPE:TRY only for VARIETY: *British* and MODE: *spoken* ($\chi^2=12.11$);
- CLAUSETYPE:TRY only for VARIETY: *British* and MODE: *written* ($\chi^2=54.24$);
- CLAUSETYPE:TRY only for VARIETY: *American* and MODE: *spoken* ($\chi^2=8.92$);
- CLAUSETYPE:TRY only for VARIETY: *American* and MODE: *written* ($\chi^2=11.13$).

Now why is this a problem? This is a problem because this data set involves some complexities that cannot be addressed with such brief treatment. For example, Hommerberg and Tottie basically study only the effect of one independent variable (CLAUSETYPE) on the dependent variable (TRY) *separately* for the four different data sets defined by VARIETY and MODE. First, however, this approach does not allow for the comparison of the strength of the effects: χ^2 values are dependent on sample sizes so one cannot simply compare them with each other, and measures of effect size such as φ would have to be used instead (for the spoken British data, $\varphi=0.31$; for the written British data, $\varphi=0.31$; for the spoken American data, $\varphi=0.11$; for the written American data, $\varphi=0.16$; and it is plain to see that the χ^2 values and the φ-values are not directly related!).

Second and even more importantly, what Hommerberg and Tottie do *not* do is test

- whether CLAUSETYPE interacts with VARIETY (i.e., we do not know whether the *horror aequi* effect of the clause type on *try*'s complementation structure is different in each variety);
- whether CLAUSETYPE interacts with MODE (i.e., we do not know whether the effect of the clause type on *try*'s com-plementation structure in speaking is different from its effect in writing);

- whether CLAUSETYPE interacts with VARIETY *and* MODE (i.e., we do not know whether the effect of the clause type on *try*'s complementation structure is different in each mode in each variety).

This in turn has several undesirable consequences: we do not know whether all variables Hommerberg and Tottie included in their study need to be included – Occam's razor dictates that unnecessary variables or interactions must be discarded, but that is only possible when we know each variable's and each interaction's effect. For example, note that while Hommerberg and Tottie compare the two modes within British English and the two varieties, they do not mention that, while *horror aequi* appears to be stronger in British English, that is not true in general: *horror aequi* effects in American English are weaker than that in written British English ($0.11<0.31$ and $0.16<0.31$), but stronger than that in spoken British English ($0.11>0.09$ and $0.16>0.09$), and one needs to show whether this is a significant effect or not. Note also that not including all variables from the very start theoretically allows even for the worst case scenario, namely the one that, when all variables and their interactions are included, CLAUSETYPE suddenly plays no role anymore. Alternatively, one may find that the interaction of two variables practically reverses the monofactorial effect of one of the variables in isolation.

Second, the description of how *try*'s complementation pattern is chosen is incomplete, which means that we also do not know the degree to which the choice of *try to* vs *try and* can be 'predicted' given the data we have. Obviously, if the prediction accuracy was close to that expected by chance then we would know that the variables included in the analysis are not well enough correlated with *try*'s complementation structure to really make a difference.

In sum, if the data are multifactorial in nature, one *must* perform a multifactorial analysis involving *all* variables and *all* two-way and higher-order interactions – higher-order interactions can hardly ever be identified by eyeballing data, nor (usually) by the more compre-hensive procedure of inspecting several two-dimensional tables in isolation – for this, a truly multifactorial approach is required, which will be exemplified in the next section.

2.2.2. The analysis of multidimensional frequency data: an example

The data provided by Hommerberg and Tottie (2007) were analyzed with a binary logistic regression, in which

- TRY was the dependent variable (the predicted level was *try to*);
- VARIETY, CLAUSETYPE, and MODE were the independent variables included;
- VARIETY:CLAUSETYPE, VARIETY:MODE, CLAUSETYPE:MODE, VARIETY:CLAUSETYPE:MODE were the interactions included.

The maximal model including all these variables and interactions already provided a good and highly significant fit, but the three-way interaction turned out to be insignificant ($p \approx 0.63$) and was therefore discarded. This means that the effect of *horror aequi* is rather similar in the eight groups consisting of the combinations of two varieties, two modes, and two clause types.

The second model without the three-way interaction was still highly significant, but now the interaction VARIETY:CLAUSETYPE turned out insignificant ($p \approx 0.97$) and was again discarded. This indicates that CLAUSETYPE has the same kind of effect in both British and American English so that clause types in the varieties need not be distinguished.

The third model without these two interactions is also the minimally adequate model: each predictor is significantly correlated with TRY. The overall correlation is rather strong (Model L.R. $\chi^2 = 1{,}279.23$; $df=5$; $p<0.001$; Nagelkerke $R^2 = 0.436$) and the model's classificatory power is quite good ($C=0.831$; classification accuracy≈78.2%). Table 6 summarizes the coefficients and effect sizes, where positive/negative coefficients mean that the variables/interactions of the first column increase/decrease the likelihood of TRY: *to*, respectively; the statistics in the three right columns are just a different way to express this and are for those used to odds ratios.

Variable / interaction	Coeff.	Wald Z	p	Odds ratio or upper or CI95%	lower or CI95%	
CLAUSETYPE: to	-0.47	-4.56	<0.001	0.623	0.508	0.763
VARIETY: British	-2.35	-22.26	<0.001	0.095	0.077	0.117
MODE: written	2.05	8.64	<0.001	7.772	4.944	12.554
VARIETY: British MODE: written	0.67	2.95	<0.004	1.962	1.242	3.044
CLAUSETYPE: to MODE: written	-0.91	-4.42	<0.001	0.404	0.269	0.601

Table 6. Statistical results of the minimal adequate model.

Let us inspect the variables' and interactions' effects. Beginning at the top, CLAUSETYPE: has a significant effect, and it is in the direction expected from *horror aequi*: the coefficient for CLAUSETYPE: *to* is negative, which means when CLAUSETYPE is *to*, then the probability of the predicted pattern, *try to*, becomes smaller. Then, there is a very strong effect of VARIETY such that British English strongly disprefers *try to*. The final main effect shows that the written mode prefers *try to*. Since odds ratios are not particularly intuitively interpretable, it is often useful to represent the observed relative frequencies in bar plots, and Figure 4 does just that for the three main effects just discussed. The dark grey and light grey bars represent the percentages of *try* and *try to* respectively, the plotted numbers are the observed frequencies.
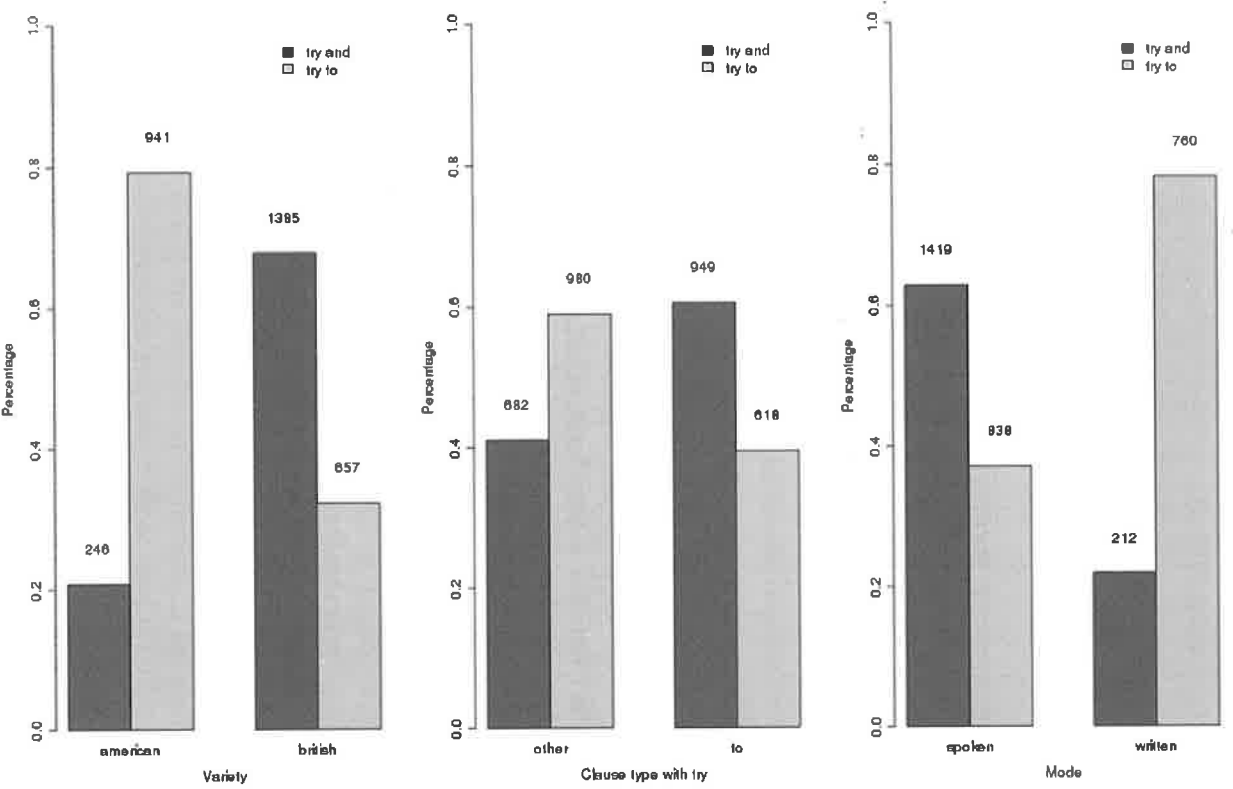
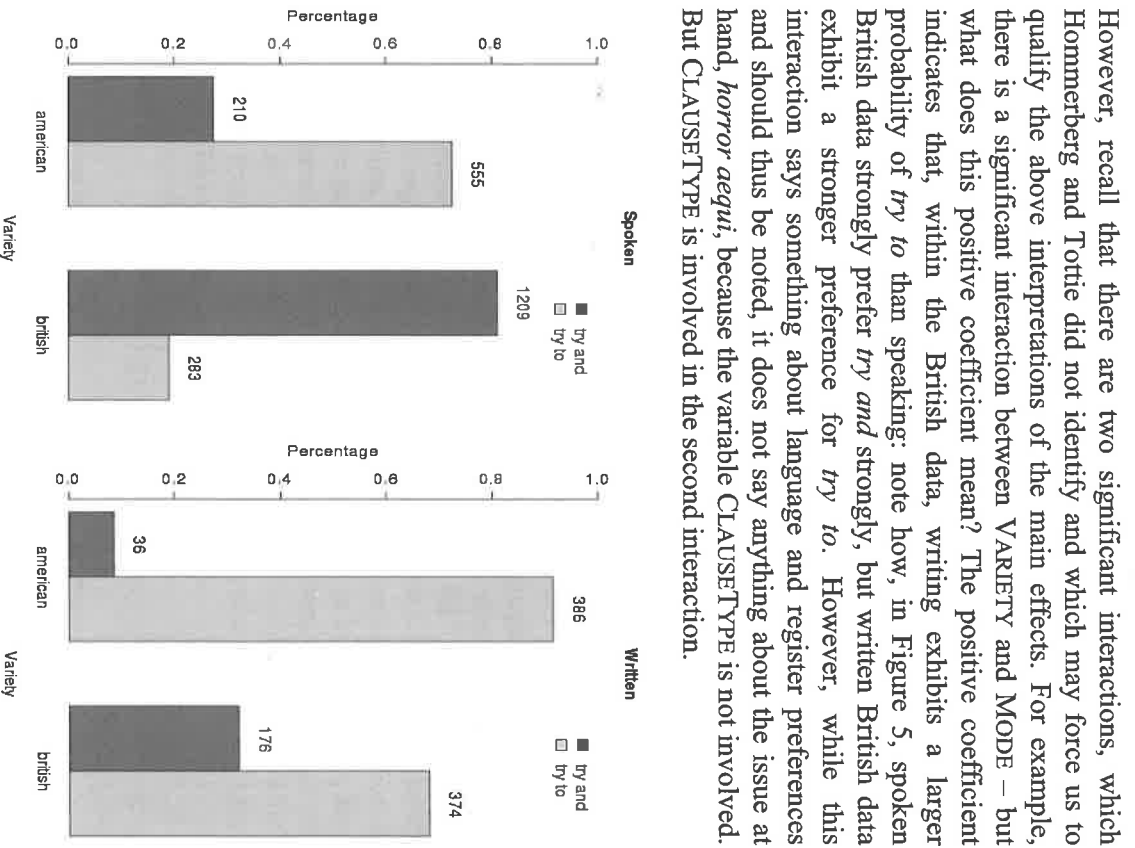Figure 4. Bar plots of observed absolute and relative frequencies (main effects).

However, recall that there are two significant interactions, which Hommerberg and Tottie did not identify and which may force us to qualify the above interpretations of the main effects. For example, there is a significant interaction between VARIETY and MODE – but what does this positive coefficient mean? The positive coefficient indicates that, within the British data, writing exhibits a larger probability of *try to* than speaking: note how, in Figure 5, spoken British data strongly prefer *try and* strongly, but written British data exhibit a stronger preference for *try to*. However, while this interaction says something about language and register preferences and should thus be noted, it does not say anything about the issue at hand, *horror aequi*, because the variable CLAUSETYPE is not involved. But CLAUSETYPE is involved in the second interaction.

Figure 5. Bar plots of observed absolute and relative frequencies (VARIETY:MODE).

This second interaction is a little bit stronger, has a negative coefficient, and is represented in Figure 6. Hommerberg and Tottie did stipulate that *horror aequi* was stronger in written British English than in spoken British English, but this interaction shows that this is true of both varieties, not just British English. In writing, *horror aequi* – the avoidance of *try to* in *to*-clauses – is stronger (e.g., the proportion of *try* and increases more strongly in writing [12% → 33%] than in speaking [54% → 72%, i.e., an increase by 33%]), which may well be a result of the fact that writing is more premediated and that the kind of conscious processes involved in writing are more likely to pick up the undesirable structural repetition.
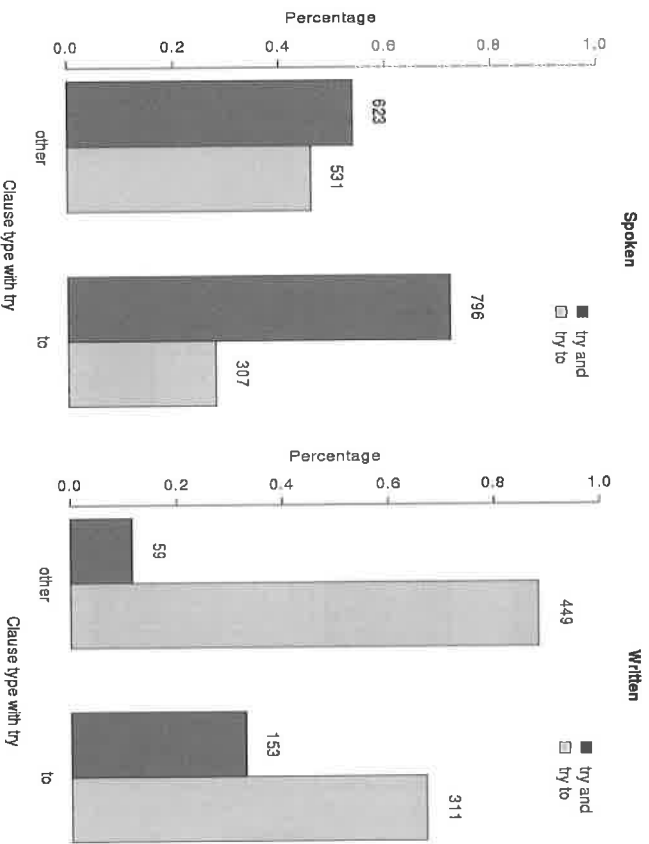

Figure 6. Bar plots of observed absolute and relative frequencies (CLAUSETYPE:MODE).

Finally, let us briefly explore which kinds of situations were difficult for the logistic regression to predict – where did most misclassifications arise? It turns out that the misclassifications are distributed rather randomly across British and American English and across speaking

and writing, but that CLAUSETYPE: *to* yielded a significantly larger number of misclassifications, as did TRY: *try and*. These variable levels would therefore be the natural point for follow-up analyses.

This small example does not of course constitute a full-fledged analysis of *try*'s two complementation patterns – obviously, a lot more needs to be done. For instance, additional future work on this pattern could involve a more serious study of the collostructional preferences than both Gries and Stefanowitsch (2004: section 4.2) and Hommerberg and Tottie (2007: section 4) undertook, maybe along the lines of Wulff's (2006) study of *go* V and *go and* V as well as Wulff (2008) on *try* V and *try and* V, which also use semantic classes à la Levin (1993) or Aktionsarten à la Vendler (1967).

Nevertheless, I hope to have shown that multifactorial data must be analyzed multifactorially: as mentioned in Section 1, the complexities of linguistic data do not reveal themselves easily either to the naked or to the monofactorial eye. Recall that Hommerberg and Tottie discussed their findings in their single summary paragraph, but the truly multifactorial study of the same data provided a much more descriptively accurate picture: we now know which variables and which interactions influence the choice of the complementation structure (cf. significance values), how strong each of these variables and interactions is (cf. the coefficients and odds ratios) and that *horror aequi* is more pronounced in writing but at work in both varieties, how well all of these variables explain/predict the constructional choice (cf. the C and R^2 values and the classification accuracy), and which variables or constructions are still hardest to predict and merit more attention (CLAUSETYPE: *to* and *try-and* patterns).

3. Conclusion

While my assessment of many corpus linguists' methodological knowledge may have been harsh – but then, this *is* partly a polemic – I hope to have shown how easy it is to commit errors in data analysis

that result in researchers' (i) failing to report important effects in their data (Sections 2.1.1 and 2.2), (ii) overstating their case (Section 2.1.2), (iii) believing nice results too quickly (Section 2.1.3), and (iv) not treating multifactorial data in a multifactorial way (Section 2.2). Just like members of every other discipline – and maybe even more so – we as corpus linguists must learn to (more) fully utilize the potential of the methods that are already there and just waiting for us – relying on commercial and limited corpus processing software and underutilizing available quantitative methods cannot be the way to use the ever increasing amount of great corpora and integrate/address new and exciting findings from neighbouring disciplines that support the frequency-based perspective that is at the core of our field. If this chapter stimulates at least a few researchers to embrace more suitable tools and methods, then it has fulfilled its main goal.

Acknowledgements

References

Aijmer, Karin, 1984. *Sort of* and *kind of* in English Conversation. *Studia Linguistica.* 38/2, 118-128.

Baayen, R. Harald 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press.

Church, Kenneth W. / Hanks, Patrick 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics.* 16/1, 22-9.

Church, Kenneth W. / Gale, William / Hanks, Patrick / Hindle, Donald / Moon, Rosamund 1994. Lexical Substitutability. In Atkins, Beryl T. Sue / Zampolli, Antonio (eds) *Computational Approaches to the Lexicon.* Oxford, New York: Oxford University Press, 153-177.

Crawley, Michael J. 2007. *The R Book.* Chichester: John Wiley.

Gries, Stefan Th. 1999. Particle Movement: A Cognitive and Functional Approach. *Cognitive Linguistics.* 10/2, 105-145.

Gries, Stefan Th. 2003a. Testing the Sub-test: A Collocational-overlap Analysis of English *-ic* and *-ical* Adjectives. *International Journal of Corpus Linguistics.* 8/1, 31-61.

Gries, Stefan Th. 2003b. *Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement.* London, New York: Continuum Press.

Gries, Stefan Th. 2009. *Statistics for Linguists with R.* Berlin, New York: Mouton de Gruyter.

Gries, Stefan Th. / David, Caroline V. 2007. This is Kind of/Sort of Interesting: Variation in Hedging in English. In Pahta, Päivi / Taavitsainen, Irma / Nevalainen, Terttu / Tyrkkoö, Jukka (eds) *Towards Multimedia in Corpus Linguistics.* Studies in variation, contacts and change in English 2, University of Helsinki, <http://www.helsinki.fi/varieng/journal/volumes/02/gries_david/>.

Gries, Stefan Th. / Stefanowitsch, Anatol 2004. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. *International Journal of Corpus Linguistics.* 9/1, 97-129.

Hommerberg, Charlotte / Tottie, Gunnel 2007. *Try to* and *Try and?* Verb Complementation in British and American English. *ICAME Journal.* 31, 45-64.

Hunston, Susan / Francis, Gill 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam and Philadelphia: John Benjamins.

Levin, Beth 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago, IL: The University of Chicago Press.

Manning, Christopher D. / Schütze, Hinrich 2000. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The M.I.T. Press.

R Development Core Team 2008. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing; ISBN 900051-07-0; <http://www.R-project.org>.

Rohdenburg, Günter 2003. Cognitive Complexity and Horror Aequi as Factors Determining the Use of Interrogative Clause Linkers in English. In Rohdenburg, Günter / Mondorf, Britta (eds) *Determinants of Grammatical Variation.* Berlin, New York: Mouton de Gruyter, 205-249.

Shirai, Yasuhiro / Slobin, Dan I. / Weist, Richard E. 1998. Introduction: The Acquisition of Tense-aspect Morphology. *First Language.* 18/54, 245-253.

Stefanowitsch, Anatol / Gries, Stefan Th. 2003. Collostructions: Investigating the Interaction between Words and Constructions. *International Journal of Corpus Linguistics.* 8/2, 209-243.

Stoll, Sabine / Gries, Stefan Th. 2009. How to Characterize Development in Corpora: An Association Strength Approach. *Journal of Child Language.* 46/5, 1075-1090.

Tao, Hongyin 2003. A Usage-based Approach to Argument Structure: 'Remember' and 'Forget' in Spoken English. *International Journal of Corpus Linguistics.* 8/1, 75-95.

Vendler, Zeno 1967. *Linguistics in Philosophy.* Ithaca, NY: Cornell University Press.

Wulff, Stefanie 2006. *Go-V* vs. *go-and-V* in English: A Case of Constructional Synonymy? In Gries, Stefan Th. / Stefanowitsch, Anatol (eds) *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis.* Berlin, New York: Mouton de Gruyter, 101-125.

Wulff, Stefanie 2008. *V-and-V* und *V-V* im Englischen: eine konstruktionsgrammatische Analyse. [*V-and-V* and *V-V* in English: A Constructionist Approach.] In Fischer, Kerstin / Stefanowitsch, Anatol (eds) *Konstruktionsgrammatik II: Von der Konstruktion zur Grammatik.* Tübingen: Stauffenburg, 189-201.

MARÍA ELENA RODRÍGUEZ MARTÍN

Comparing Parts of Speech and Semantic Domains in the BNC and a Micro-corpus of Movies: Is Film Language the 'Real Thing'?

1. Introduction

The present investigation aims at expanding two previous exploratory studies which compared conversational processes and collocations in the British National Corpus (BNC) (face-to-face conversations) and a micro-corpus of movies (Rodríguez Martín 2010 in press; Rodríguez Martín/Moreno Jaén 2009). The purpose of the comparison was to explore the similarities and differences between real spoken language and screen dialogue in order to provide empirical evidence to support two research projects which focus on the use of films for the development of speaking competence (Moreno Jaén/Pérez Basanta 2009; Pérez Basanta/Rodríguez Martín 2007).

As the issue of film language authenticity has hitherto been a topic of considerable debate in ELT, there are reasons to believe that more research based on empirical data is required. To this end, we carried out the aforementioned corpus-based studies. The first one (Rodríguez Martín 2010 in press) compared conversational processes in the BNC (spoken component: face-to-face conversations) and a corpus of film transcripts using Rühlemann's situational framework for conversation (2007), a model which makes an attempt to define the main aspects of the construct of conversation. In order to carry out the comparison, two software tools were used: Wordsmith Tools (Scott