Article

Language & Linguistics 語言暨語言學

Language and Linguistics
16(1) 93–117

DOI: 10.1177/1606822X14556606
lin.sagepub.com

SAGE

Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions

Stefan Th. Gries

University of California, *Santa Barbara*

This paper surveys a variety of methodological problems in current quantitative corpus linguistics. Some problems discussed are from corpus linguistics in general, such as the impact that dispersion, type frequencies/entropies, and directionality (should) have on the computation of association measures as well as the impact that neglecting the sampling structure of a corpus can have on a statistical analysis. Others involve more specialized areas in which corpus-linguistic work is currently booming, such as historical linguistics and learner corpus research. For each of the problems, first ideas/pointers as to how these problems can be resolved are provided and exemplified in some detail.

Key words: association measures, mixed-effects/multi-level modeling, MuPDAR, token/type frequencies, variability-based neighbor clustering

1. Introduction

For several decades now, corpus linguistics has been among the fastest-growing methodological disciplines in linguistics. For instance, in his outgoing column as the editor of *Language*, Joseph (2004:382) comments explicitly on the increase of corpus and internet data; another example is Janda (2013), who discusses in detail the ways in which cognitive-linguistic theory in particular has made a 'quantitative turn'. Given this development and the somewhat obvious observation that corpora contain nothing but frequencies/probabilities—of occurrence or of co-occurrence—it is not surprising that linguistics in general has become much more quantitative/statistical in nature, a trend we also witness in corpus linguistics: For example, 10 or 15 years ago it would have been quite difficult to find papers with multifactorial statistical techniques in corpus-linguistic papers—now, monofactorial statistical tests at least are much more frequent, and multifactorial statistical methods are on the rise.

In spite of this welcome development, change in the field of linguistics is slow, and corpus linguistics in particular is limited in two ways: First, in computational ways in the sense that probably the majority of corpus linguists are still relying on a small set of often commercial and proprietary point-and-click kind of corpus search tools (such as WordSmith Tools, MonoConc Pro, or AntConc); given the severe constraints that this results in (see Clark-Sánchez 2013; Gries 2010a, 2011), it is gratifying to see how more and more practitioners now avoid all these constraints by switching to programming languages such as R or Python.

The second kind of limitation involves statistical methods: While the overall amount of statistical expertise in the field is growing, corpus linguists should both widen and deepen their expertise to go beyond the handful of widely used methods. By that I do not only mean that corpus linguists

need to use more different statistical tests (while that is generally true, the choice of a particular test is of course mostly dictated by the particular research question), but also that there needs to be a growing awareness that some choices that corpus linguists traditionally make may be problematic and would benefit from a different perspective. In the next section of this paper, I want to exemplify several such problems and survey some solutions to them. Specifically, I shall discuss potentially problematic choices or omissions in the area of general corpus statistics, in particular the choice of association measures for co-occurrence data, that is, measures with which corpus linguists quantify the degree of association between two linguistic expressions (e.g. two words or a word and a syntactic pattern/construction). In addition, I shall briefly comment on the underutilized notion of dispersion, that is, a measure that quantifies how evenly distributed elements are in a corpus, and thus also relates to the notion of corpus homogeneity. Finally, I shall demonstrate how the current typical neglect of the hierarchical structure of corpora poses severe problems. More specialized areas are currently booming, it seems: diachronic corpus linguistics, which needs to deal with the problem of how temporally-ordered corpus data are grouped into temporal stages for subsequent analysis; and learner corpus research, which needs to move on from decontextualized studies of over- and underuse to more comprehensive models of learner language and its differences to native language.

2. General corpus statistics

2.1 Co-occurrence information

One of the most fundamental notions in corpus linguistics is the *distributional hypothesis*, that is, the working assumption that linguistic elements that are similar in terms of their distributional patterning in corpora also exhibit some semantic or functional similarity. Firth (1957:11) captured this notion in his famous dictum '[y]ou shall know a word by the company it keeps', but Harris's (1970:785f.) following statement actually makes the same case much more explicitly:

> [i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

That is, a linguistic expression *E*—morphemes, words, constructions/patterns, . . .—can be studied by exploring what is co-occurring with *E* and how often. The simplest possible way to do this would be by raw co-occurrence frequency or, more likely, conditional probabilities such as *p*(function|*E*) or *p*(contextual element(s)|*E*). Since raw frequencies will be distorted by words that are highly frequent everywhere, a more frequent way is to use association measures (AMs), that is, statistics that quantify the strength of mutual association between two elements such as a function or a contextual element on the one hand and *E* on the other. Most AMs are based on co-occurrence tables of the kind exemplified in Table 1, which contain observed frequencies of (co-)occurrence of a linguistic expression *E* (e.g. a particular word) and functions/contexts *X* (e.g. a particular construction). In such a table, *a*, or *obs a* (for 'observed frequency *a*'), refers to the frequency with which

Table 1: Schematic co-occurrence frequency table

	E	Elements other than *E*	Totals
Function/context *X*	*a*	*b*	*a* + *b*
Functions/contexts other than *X*	*c*	*d*	*c* + *d*
Totals	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

E is observed with/in function/context *X*, etc.; examples for widely used AMs include Mutual Information (*MI*), the *t*-score, the *z*-score, log-likelihood G^2, $p_{\text{Fisher-Yates exact}}$, and many more (see Evert 2009 for how these measures are computed and discussion of their characteristics).

2.2 Problems with the quantification of co-occurrence

2.2.1 Problem: multi-word AMs are not conservative enough

Despite their frequency of use, AMs of the above kind are not unproblematic. One smaller problem is the fact that they do not easily generalize to *n*-grams (uninterrupted strings of *n* words), or multi-word units (such as *according to*, *in spite of*, etc.). At this point, *MI* for *n*-grams—$\log_2(^{\text{obs}\ a}/_{\text{exp}\ a})$—is often simply computed on the basis of complete conditional independence, which will tend to underestimate expected frequencies of *a* and, thus, overestimate the strength of association. If one computes the *MI* of *in spite of* in the untagged Brown corpus by comparing the observed frequency of *in spite of* of 54 against an expected frequency based on complete independence, *MI* becomes an extremely high value of 12.25. However, if one computes *MI* by comparing the same observed frequency of *in spite of* to the one expected from the occurrences of *in spite* and *of*, then that *MI*-value decreases to 4.76. Thus, corpus linguistics needs to explore more adequate and conservative ways to extend AMs to *n*-grams.

2.2.2 Problem: nearly all AMs are symmetric/bidirectional

An even more important problem is that nearly all AMs are symmetric: the association of expression *E* to context *C* is presumed to be symmetric/bidirectional. However, associations in general and associative learning are certainly not (always) symmetric, which is why, ideally, corpus linguistics would explore the use of directional AMs. Some work on this area exists, in particular Michelbacher et al. (2007, 2011), who explore two different conceptual options.

First, they explore the correlation of conditional probabilities from adjective—noun collocations with the University of South Florida Association Norms, but find the measure lacking in identifying symmetric associations; in addition, conditional probabilities do not normalize the observed percentage against any baseline.

Second, they explore a measure based on the differences of ranks of AMs (such as chi-squared values). For such rank measures, a collocation *x y* is explored by

- computing all AMs for collocations with *x*, ranking them, and noting the rank for *x y*;
- computing all AMs for collocations with *y*, ranking them, and noting the rank for *x y*;
- comparing the difference in ranks.

In tests analogous to those of conditional probabilities, this rank measure does not perform well with asymmetric associations but a little better with symmetric ones; in the additional classification task, the rank measure came with an even higher error rate than conditional probabilities. In Michelbacher et al.'s (2011) study, additional rank measures are also based on raw co-occurrence frequencies, G^2, and t, and the corpus-based data are compared to the results of a free association task undertaken specifically for that study. The results of the rank measures in that study are much more compatible qualitatively and quantitatively with the subjects' reactions in the experiment; of the rank measures, G^2 performs best.

While this sounds promising, the computational effort that goes into these calculations is immense, since the computation of one AM for the collocation *x y* requires the computation of all AMs for all collocations with *x* and then separately for all collocations with *y*. In addition, in spite of the huge computational effort involved in the thousands of ranked G^2-values, they do not perform better than conditional probability (Michelbacher et al. 2011:270). Finally, the rank-measure based approach is a very promising one, but probably not cognitively realistic in any sense. Against this background, the measure of ΔP from the associative learning literature seems a particularly interesting alternative (see Ellis 2006 for its introduction into linguistics). It, too, is based on tables such as Table 1, but can distinguish the association from *X* to *E* (see (1a)) from the association from *E* to *X* (see (1b)).

(1) a. $\Delta P_{E|X} = \frac{a}{a+b} - \frac{c}{c+d}$

b. $\Delta P_{X|E} = \frac{a}{a+c} - \frac{b}{b+d}$

For example, all traditional AMs would return a high value for *of course* (see Gries 2013:144), but it is ΔP that recognizes that the association between *of* and *course* is not symmetric: *of* is not a good predictor that *course* would follow whereas *course* is a strong predictor that *of* will precede. In fact, Gries (2013) finds that similarly strong asymmetric collocations are quite frequent—26% of his sample of 2-grams are exhibiting high G^2-values reflecting strong association, but are missing the fact that these are very asymmetric associations. (2a) lists some 2-grams in which the first word is much more predictive of the second than vice versa; (2b) lists some 2-grams in which the first word is much less predictive of the second than vice versa (as with *of course*).

(2) a. *apart from*, *according to*, *upside down*, *contrary to*, *ipso facto*, *irrespective of*
b. *at least*, *per annum*, *status quo*, *for instance*, *de facto*, *vice versa*

In sum, ΔP is by design more sensitive than traditional AMs since it can tease apart directionality effects; it is very easy to understand and compute; its computation/interpretation does not require assumptions (such as normality, which is very rare in corpus data); it avoids problems of the Null Hypothesis Significance Testing paradigm because it does not test the observed distributional data against an illusionary null hypothesis distribution; finally, it has received experimental support both in psychology and in linguistic work by Ellis and colleagues, and Gries (2013) at least mentions a way in which it could be used to explore *n*-grams. It would therefore behoove corpus linguists to

explore this measure in more detail; ultimately, maybe it can even help explore mismatches between corpus and experimental data of the type reported in Mollin (2009), for example, who finds a lack of correlation between association data from the Edinburgh Associative Thesaurus and co-occurrence data from the British National Corpus (BNC) explored bi- rather than unidirectionally.

2.2.3 Problem: nearly all AMs involve only token frequencies

The next AM problem to be discussed here is perhaps just as fundamental as the symmetry problem, but even less recognized and explored: namely that the computation of nearly all AMs involves only the four token frequencies represented in Table 1. That is, a crucial piece of information that none of the usual measures includes is

- minimally, the type frequencies that make up the frequencies *b* and *c*, that is, how many different elements not-*E* are there with the same function/context *X* (for *b*) and how many different functions/contexts not-*X* are there that *E* is used with? The answer to these two questions would be two numbers, the two type frequencies underlying *b* and *c*., for example 10 and 20.
- And even more useful would be the token frequencies of all the types that make up the token frequencies *b* and *c*. For *b*, that would mean how many different elements not-*E* there are with the same function/context *X* and how frequent each of them is with *X*, and the corresponding question for *c*. The answer to this question for *b* would be 10 token frequencies and, maybe, their entropy or some other summary statistic.

Given the importance of type frequencies or entropies for many domains (productivity, language change, language acquisition, . . .), it is amazing how little alternatives to AMs that utilize type frequencies or entropies have been explored in corpus linguistics proper. Studies from neighboring disciplines (Baayen 2010b; McDonald & Shillcock 2001; Recchia et al. 2008) all show that contextual-diversity measures, such as contextual distinctiveness and/or entropy-related measures, are better predictors of psycholinguistic behavioral data than token-frequency statistics alone, so corpus linguistics has its work cut out for it.

Within corpus linguistics, Daudaravičius & Marcinkevičienė (2004) were the first to make this topic known to a wider audience. They proposed a measure called lexical gravity *G* as defined in (3). As can be inferred from this equation, all other things being equal *G* increases as n_{w1w2}, $n_{\text{types after w1}}$, or $n_{\text{types before w2}}$ increases, and *G* decreases as n_{w1} or n_{w2} increases.

$$(3) \quad \text{Gravity } G\,(w_1, w_2) = \log\left(\frac{n_{w1w2} \cdot n_{types\ after\ w1}}{n_{w1}}\right) + \left(\frac{n_{w1w2} \cdot n_{types\ before\ w2}}{n_{w2}}\right)$$

Unfortunately, there has been very little follow-up on this notion. Two exceptions are Gries (2010b) and Gries & Mukherjee (2010). The former study uses a cluster analysis of sub-registers (of the BNC Baby) based on *G*-values for all 2-grams in the corpus and compares it to one based on *t*-values, and finds that the former is able to just about perfectly recreate the sampling decisions

of the corpus compilers (whereas the latter performs worse). Specifically, the *G*-based cluster analysis

- perfectly distinguishes speaking from writing;
- perfectly distinguishes fiction, news, and academic registers within writing;
- identifies even similar sub-registers within news and academic sub-registers.

The latter study explores an extension of *G* to the identification of *n*-grams in different varieties of English. More precisely, it shows how one can use *G* to identify *n*-grams, and how a *G*-based cluster analysis of spoken and written data from four different varieties (British, Hong Kong, Indian, and Singaporean English) perfectly distinguishes speaking from writing.

In sum, there are compelling arguments to include type frequencies from theoretical considerations as well as from neighboring disciplines such as psycholinguistics or computational linguistics, and there are promising first results within corpus linguistics proper, but more exploration is definitely required. In particular, all of the above approaches only deal with the minimal amount of information one should include—the more comprehensive information regarding token and type frequency distributions and entropies still awaits first exploration.

2.3 Problems with ignoring distribution in the structure of the corpus

2.3.1 Problem: (co-)occurrence may be underdispersed

The next AM problem to be discussed here concerns another important dimension of corpus data that the traditional kind of AM approach based on Table 1 does not reveal. Specifically, in the previous section it was shown how almost all AMs do not fully utilize the information that is summarized in *b* and *c* in Table 1 because *b* and *c* do not provide the type frequencies (let alone the entropies) making up the *b* and *c* tokens. However, another problem is that the co-occurrence frequency *a* in Table 1 does not provide the information of how (un)evenly across the corpus the *a* co-occurrences of element *E* and function/context *X* are found. Consider Figure 1 for an example in which *a* was arbitrarily set to 180: the upper panel shows that these 180 co-occurrences may be clustered with high frequencies in a very small section of a 500-part corpus (such as the British Component of the International Corpus of English, ICE-GB) or, as in the lower panel, much more widely distributed with smaller frequencies. This distributional notion is known as dispersion (see Gries 2008 for a recent overview of many dispersion measures) and not only can it be quantified (see the *DP*-value in Figure 1, which reflects clumpiness), but it also has important consequences for corpus-linguistic analysis as well as for psycholinguistic or more general applications.

As for the implications for corpus-linguistic analysis, consider the question of which verbs are likely to be used in imperatives. A perfectly normal traditional corpus-linguistic account could approach that question by computing for each verb lemma in a corpus that occurs in the imperative at least once an AM that quantifies the association between that lemma and the imperative based on tables such as Table 1 and then rank the verbs according to their association strength. Stefanowitsch & Gries (2003) did just that using the ICE-GB and obtained the ranking in (4):

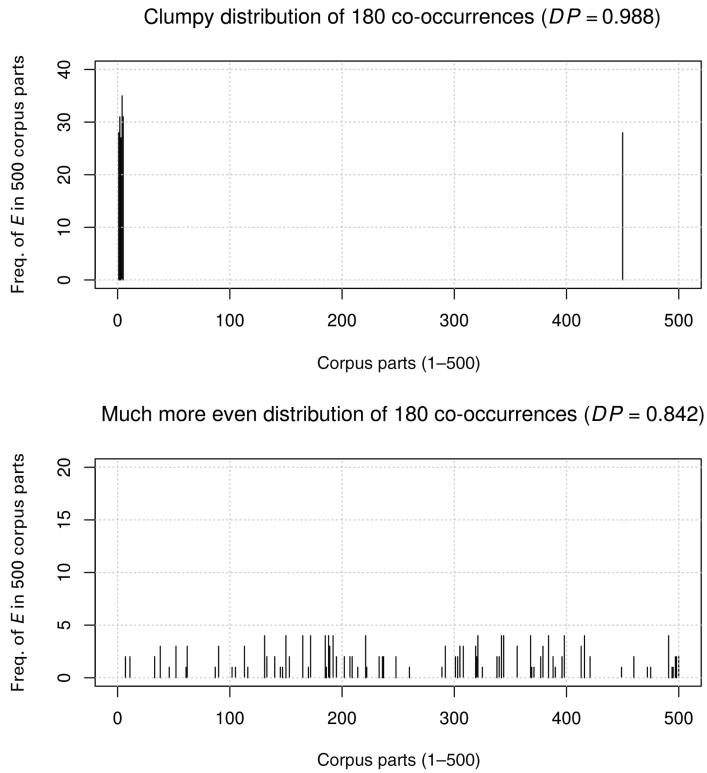

Figure 1: Two (extreme) ways in which 180 co-occurrences may be distributed across a corpus consisting of 500 parts (files): an extremely uneven/clumpy distribution (upper panel) and a much more even distribution (lower panel)

(4) *let*, *see*, *look*, *listen*, *worry*, *fold*, *remember*, *check*, *process*, *try*, *hang on*, *tell*, *note*, *add*, *keep*, ...

Most verbs in (4) make perfect sense as lemmas to be associated with the imperative, but *fold* and *process* are somewhat surprising. Closer inspection reveals that the high frequency of each of these two verbs in the imperative that is responsible for them ending up in the top 10 list is due to just a single one out of 500 files, namely a file with an excerpt from an origami book (for *fold*) and a file with an excerpt from a cook book (for *process*). Clearly, that shows that the AM-based ranking can be quite misleading in the sense that *fold* and *process* appear to be more strongly associated to the imperative than *remember* or *try*, whereas this is a register artifact that can be recognized once the dispersions of the *a* co-occurrences are studied. While high frequencies of co-occurrence will in general be correlated with a wider dispersion, this correlation is never perfect and cannot be taken for granted, as we have just seen. Thus, it stands to reason that the analysis of co-occurrence data using AMs can benefit considerably from taking dispersion into consideration. This could be done, for instance, as demonstrated above, by computing AMs for co-occurrence data, but also dispersion measures so that one can compare the elements' dispersion values to their

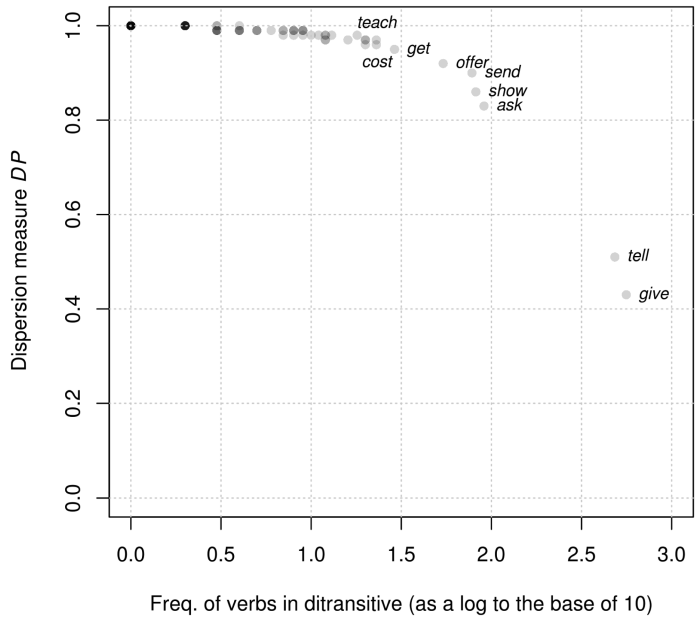

Figure 2: Verbs' attraction to the ditransitive in the ICE-GB: dispersion (on the *y*-axis) plotted against logged co-occurrence frequencies (on the *x*-axis)

AM-values and/or their frequencies of co-occurrence with/in the function/context *X* in question. If one does the latter for the verbs in the ditransitive in the ICE-GB, Figure 2 is a clear case showing a correlation between co-occurrence frequency (on the *x*-axis) and dispersion (on the *y*-axis).

Thus, in this case and unlike in the above imperative example, co-occurrence frequencies, frequency-based AMs, and dispersion measures yield quite similar verb rankings—however, only by exploring all these dimensions can we be certain that the different dimensions present in the corpus data do in fact converge.

As for the implications for psycholinguistic and more general (theoretical) applications, dispersion has by now been shown to be relevant in domains other than core corpus linguistics, too. For instance, Simpson-Vlach & Ellis (2005) and Ellis et al. (2007) have shown that even the simplest conceivable dispersion measure—range, the (normalized) number of corpus parts in which (co-)occurrences are attested—has significant predictive power above and beyond raw frequency in the study of academic formulas; Casenhiser & Goldberg (2005) have found that the evenness of the distribution of verb types in a novel construction (which, in fact, amounts to its entropy) is correlated with how well children and adults learn a novel syntactic construction; Gries (2010c) has shown how many dispersion measures or related adjusted frequencies are better predictors of psycholinguistic behavioral data than corpus frequencies, etc. To the extent that corpus linguists want their work to be interdisciplinary, to also impact neighboring fields, they should add the exploration of dispersion measures to any study of co-occurrence data, if only to protect themselves against invalid generalizations based on overly clumpy, and thus non-representative, data. In that sense, exploring dispersion offers necessary protection against biases due to corpus heterogeneity.

2.3.2 Problem: ignoring the hierarchical structure of the corpus

The final problem to be discussed in this section is concerned with the fact that the vast majority of statistical analyses in corpus linguistics—be they chi-squared tests, simple correlations, generalized linear models (GLM, e.g. binary logistic regressions), . . .—violate a fundamental assumption of these statistical methods: that the data points are independent of each other. Rather, there are three different ways in which many corpus data points can be seen as related to each other, the first two of which are well-known from psycholinguistic work:

- Speakers/writers in corpus data/files often provide more than one data point in a concordance so that all data points from a particular speaker/writer are related to each other (as they may reflect that speaker's idiosyncratic behavioral patterns). In psycholinguistics, this is often addressed with F_1-or related ANOVA statistics.
- For many grammatical patterns, concordance lines will involve the same lexical item so that all data points with that lexical item are related to each other (as they may reflect that lexical item's idiosyncratic patterning). In psycholinguistics, this is often addressed with F_2-or related ANOVA statistics.
- Corpora often come with a hierarchically-nested structure in which speakers are nested into files, which in turn are nested into sub-registers, which in turn are nested into registers, which in turn are nested into modes (e.g. spoken versus written). Thus, there are multiple levels of corpus organization at which effects may be located, but these levels are typically not all tested.

While it is usually freely admitted that corpus data are much more messy/noisy than (often carefully) controlled psycholinguistic experimental data, the massive interrelatedness of corpus data along the above three lines is typically ignored. In this section, I exemplify how this is problematic by comparing an analysis that, as usual, ignores this interrelatedness to one that takes it into consideration. As a small example, whose actual linguistic implications I shall not be concerned with, let us consider the question of who is more likely to use *I* or *you*—men or women—and where/when (early/late in a conversation and/or early/late in a sentence); maybe there is an assumption that women are generally less likely to use *I* . . . Using an R script (R Core Team 2014), I extracted all instances of *I* and *you* (when tagged as PNP) from all 21 files of the British National Corpus World Edition (XML) whose names begin with 'KR'. For each instance, I retrieved/annotated the following variables:

- MATCH: whether the speaker used *I* or *you*;
- FILE: the name of the file in which a speaker used *I* or *you*;
- SPEAKER: a unique identifier for the speaker who used *I* or *you*;
- SEX: the sex of the speaker, *female* versus *male*;
- SENTENCE: the square root of the ID number (from 1 to *n*) of the sentence in the files in which a speaker used *I* or *you* (the square root transformation was used to make the distribution of SENTENCE more normal);

Table 2: Results of the final model of the generalized linear model (rounded)

Predictor	*b*	*se*	*z*	*p*	$p_{deletion}$
Intercept	–0.742	0.044	–16.721	<0.0001	
$\text{SEX}_{\text{unknown versus female/male}}$	0.044	0.032	1.391	0.164	<0.0001
$\text{SEX}_{\text{female versus male}}$	0.197	0.034	5.761	<0.0001	
DISTANCE	0.193	0.008	24.925	<0.0001	<0.0001
SENTENCE	–0.004	0.001	–3.961	<0.0001	<0.0001
$\text{SEX}_{\text{unknown versus female/male}}$:SENTENCE	0.003	0.001	4.139	<0.0001	<0.0001
$\text{SEX}_{\text{female versus male}}$:SENTENCE	–0.008	0.001	–10.365	<0.0001	

– DISTANCE: the natural log of the number of characters in the sentence before the *I* or *you* in question (after tags etc. had been removed; the log transformation was used to make the distribution of DISTANCE more normal).

This is a data set that requires a multifactorial method of analysis such as a binary logistic regression. Let us assume that one decided to begin with a first maximal model that tries to predict MATCH, that is, the choice of *I* and *you* on the basis of all fixed-effects predictors—SEX, SENTENCE, and DISTANCE—as well as their pairwise interactions, and that one used a backwards model selection process in which the least significant predictor is deleted till only significant predictors are left. It turns out that this model selection process involves the elimination of the interactions SENTENCE:DISTANCE ($p = 0.058$) and SEX:DISTANCE ($p = 0.05$) and results in a highly significant model (L.R chi-squared 881.9; $df = 6$, $p < 0.0001$); the coefficients of this model are listed in Table 2.

Note that, while the regression model is highly significant, its predictive power is extremely weak: $R^2 = 0.055$, $C = 0.613$, and the classification accuracy is a mere 58.3%, which is not significantly better than chance. The nature of the effects is somewhat clear from Table 2, but for ease of interpretation is also visually represented in Figure 3: speakers are more likely to use *you* later in the utterance (left panel) and women are more likely to use *you* later in the conversation whereas men and speakers of unknown sexes are more likely to use *I* later in the conversation (right panel).

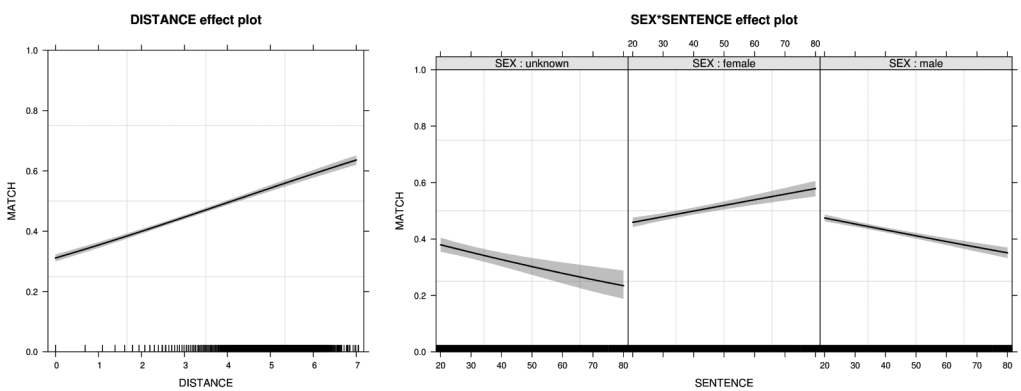

Figure 3: The significant effects of the final model of a GLM, which does not take the relatedness of data points into consideration: the effect of predictors on the predicted probability of using *you* (rather than *I*) (on the *y*-axis)

While this procedure is what most corpus linguists would do—those that have moved beyond chi-squared tests, that is—it is, strictly speaking, flawed because it does not take into consideration that the data points are not independent of each other. A much better (though still improvable) approach would be a generalized linear mixed-effects model (GLMEM) in which these interdependencies were taken into consideration. Since the speakers are nested into the files—each speaker occurs in one and only one file—one might choose the same maximal fixed-effects structure as above—SEX, SENTENCE, and DISTANCE as well as their pairwise interactions—but also include what are called *random effects* into the analysis. Random effects can be defined as effects whose levels in the sample do not cover all possible levels in the population, as opposed to *fixed effects*, whose levels in the sample cover all possible levels in the population. Typical examples of the former include SPEAKER (because not all speakers of a language are part of the sample), LEXICALITEM (because not all lexical items that can be used in a pattern will occur in a sample), TEXTSOURCE (because not, say, all newspapers from which one could have sampled are in the sample), etc.; examples of the latter include SEX (*female* versus *male*; there are no other levels), PREVIOUSLYMENTIONED (*no* versus *yes*, there are no other options), etc. While a traditional GLM returns only a regression equation that includes one intercept as well as one coefficient for each predictor, a GLMEM allows the researcher to be more flexible and, essentially, also obtain for every level of every random effect included adjustments to the overall intercept as well as adjustments to differences of means and slopes. This way, the relatedness of the data points, speaker-specific, lexical-item-specific, . . . effects, are taken into consideration.

When one then does an analogous model selection process by eliminating non-significant fixed effects, once the same interactions are deleted as before—with very different *p*-values, though: SENTENCE:DISTANCE ($p = 0.216$) and SEX:DISTANCE ($p = 0.224$)—and one arrives at a final model with the coefficients represented in Table 3.[1]

Table 3: Results of the final model of the GLMEM (rounded)

Fixed-effects predictors	*b*	*se*	*z*	*p*	$p_{deletion}$
Intercept	–0.982	0.106	–9.245	<0.0001	
$\text{SEX}_{\text{unknown versus female/male}}$	0.0026	0.083	0.031	0.975	0.523
$\text{SEX}_{\text{female versus male}}$	0.099	0.085	1.163	0.245	
DISTANCE	0.23	0.009	26.874	<0.0001	<0.0001
SENTENCE	–0.001	0.002	–0.444	0.657	0.525
$\text{SEX}_{\text{unknown versus female/male}}$:SENTENCE	0.002	0.002	0.866	0.386	0.002
$\text{SEX}_{\text{female versus male}}$:SENTENCE	–0.004	0.001	–3.469	0.0005	
Random effects (varying intercepts)					
FILE	*sd* = 0.026				
FILE/SPEAKER	*sd* = 0.821				

[1] For the sake of simplicity, I did not also trim down the random-effects structure. For all intents and purposes, the results are identical; see Gries (forthcoming) for discussion of such modeling and the corresponding R code.

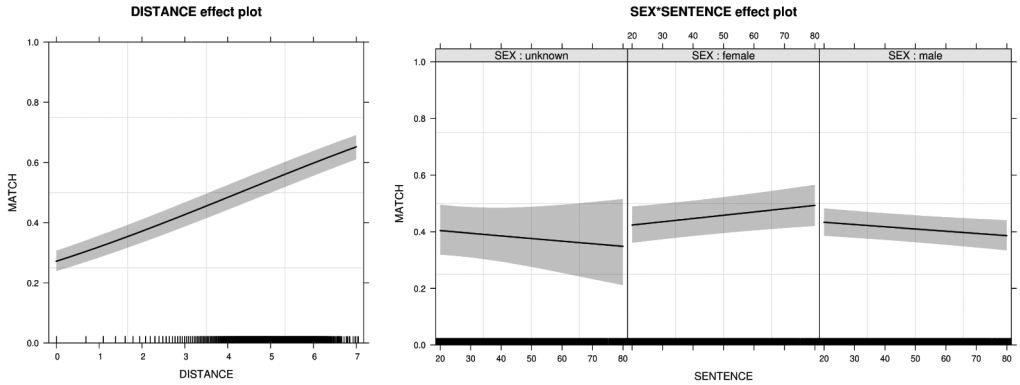

Figure 4: The significant main effects of the final model of a GLMEM, which takes the relatedness of data points into consideration: the effect of predictors on the predicted probability of using *you* (rather than *I*) (on the *y*-axis)

What about the classificatory power of this model? While it is still not as good as one would theoretically want it to be, it is much higher than the previous one: marginal R^2 = 0.044 and conditional R^2 = 0.24, C = 0.717, and the classification accuracy is now at 65.7%, which is now highly significantly better than chance.[2] Before we compare the two models, let us again first look at the visualization of the significant highest-order effects, which are shown in Figure 4.

As for the commonalities: both models contain the same fixed effects and in both models the effect of DISTANCE is probably the same. However, there are also many (more) marked differences. The most obvious was already mentioned: the GLMEM achieves a much higher and highly significant classification accuracy. Then, the GLMEM can see that, once file and speaker information is included, SENTENCE is not significant, whereas it is significant in the GLM. Most important, however, are the differences for the crucial interaction most of interest, SEX:SENTENCE. First, the GLM assigns to this interaction a *p*-value that is 24 orders of magnitude smaller (i.e. more significant) than the GLMEM. Second and more interestingly, the above two models were fitted with user-defined orthogonal contrasts—something else that happens way too rarely in corpus linguistics—to see easily (i) whether the speakers of an unknown sex are different from those where the sex is known, and (ii) whether female and male speakers behave differently. Since the GLM does not take the relatedness of the data points of each speaker into account, it returns results that are quite different from the more precise GLMEM:

– With regard to the contrast of *female* versus *male*, the GLM returns a highly significant coefficient that is ≈2 times as high as the non-significant coefficient for *female* versus *male* from the GLMEM. In other words, the GLM strongly overestimates this contrast, much of which is in fact due to speaker-specific behaviors.

[2] Marginal and conditional R^2 were computed following the logic of Nakagawa & Schielzeth (2013); marginal R^2 quantifies the fit based on only the fixed effects, conditional R^2 quantifies the fit based on all effects.

– With regard to the contrast of *female* versus *male*, the GLM returns a highly significant coefficient for *female* versus *male* that is >2 times as high as the highly significant coefficient for *female* versus *male* from the GLMER. Again, while the contrast is significant in both models, the GLM strongly overestimates its strength.

Space does not permit a more detailed discussion of these data or of the specifics of mixed-effects and multi-level modeling here (see Gries forthcoming for some more details in a corpus-linguistic context). It should have become clear, however, that much of what happens in corpus data is a result of word-/speaker-/file-/register-specific random effects rather than of the fixed effects we as corpus linguists are usually interested in. GLMs or any other statistical tool that does not take the relatedness of data points into consideration run the risk of severely overestimating the size and significance of effects. But to make matters worse, it is just as possible that GLMs *under*estimate the size and significance of effects—the problem is there is no way of knowing the direction of error of GLMs ahead of time. It is therefore imperative that corpus linguists follow the lead of recent developments in psycholinguistics and make mixed-effects/multi-level modeling a central analytical tool: without it, we will never know how much of an effect is interesting, and how much is just due to particular speakers sampled in a corpus.

2.4 Interim summary

Given the distributional hypothesis discussed above, the quantitative exploration of co-occurrence data is the most fundamental methodological tool in corpus linguistics and the last few decades have produced a plethora of papers and findings that are based on co-occurrence frequencies, co-occurrence probabilities, association measures, and other statistical approaches (most often regression-analytic methods). While much of that work has, of course, been successful because, for example, high token frequencies in *b* and *c* are positively correlated with high type frequencies, and high token frequencies in *a* are negatively correlated with clumpy distributions, it is unclear how potentially skewed the results are for cases where those correlations do not hold. A study that tries to identify multi-word units while at the same time trying to address all these AM issues mentioned above is Wahl (in progress).

In addition, ignoring the repeated-measurements nature as well as the hierarchical structure of the corpus data not only violates the fundamental assumptions of most statistical methods—the independence of data points—but also distorts our results in unpredictable ways. Thus, most of the approaches above are relatively easy ways in which we can try to make our co-occurrence-based studies more robust; there is no reason not to pursue those strategies if corpus linguistics as a whole wants to evolve in tandem with what happens in other disciplines.

3. More specialized applications

The three problems discussed above have implications for most corpus-linguistic studies: the issue of underdispersion, or clumpiness in distribution, is a threat to any statistic based on frequency data—because they all involve frequencies of occurrence and of co-occurrence. Likewise, the lack of bidirectionality and of type frequencies and their distributions in the computation of AMs

is a threat to virtually all studies based on co-occurrence data. However, at this point in time, quantitative corpus linguistics is becoming more and more established also in specific linguistic subdisciplines, which raise their own, more specialized problems. In this section, I shall discuss one example each from two areas in which corpus research is currently booming. In §3.1, I shall discuss the issue of studying temporally-ordered corpus data in a way that is both bottom-up/exploratory and principled/objective; in §3.2, I shall turn to the field of learner corpus research and the question of how to make the best use of what native and non-native learner corpora have to offer.

3.1 Temporally-ordered data and the problem of identifying stages

Temporally-ordered corpus data play an important role in two different areas in linguistics. On the one hand, there is the area of first language acquisition. In that area, corpus data are both longitudinal and cross-sectional and in order: (i) to discern longitudinal trends in the data for one or more children, (ii) to identify children at comparable levels of development for cross-sectional analysis, or (iii) to increase sample sizes and/or filter out outliers, it is often useful to be able to group the temporal data for children into different stages.

On the other hand, there is the area of diachronic historical corpus linguistics, in which corpus data are—given the relevant time spans—usually cross-sectional, covering, for instance, several centuries of the history of a language. Given that historical data are not collected in the carefully controlled ways in which psycholinguists (try to) collect language acquisition corpus data, such historical data are often quite heterogeneous so that here, too, it is useful to be able to group temporal data and at the same time clean the data of outliers in a principled fashion. Figure 5 exemplifies these challenges. The left panel shows the change of the mean length of utterances (MLU) in words of one Russian child from age 2 to age 4.5 years from Sabine Stoll's Russian first language acquisition corpus (see Stoll & Gries 2009 for details), and while it is clear that there is the expected overall increase over time, it comes with many ups and downs and no clear separation into stages. The right panel shows the change of the proportion of third person singular (*e*)*s* out of third person singular (*e*)*s* and (*e*)*th* over more than two centuries in the Parsed Corpus of Early English Correspondence and, again, there is the expected increase to the contemporary form, but again with many ups and downs and different possibilities to divide the time points into stages (see Gries & Hilpert 2010 for details).[3]

[3] One may wonder whether, following the logic of Baayen (2010a), discretization of numeric data (such as TIME or AGE) into a factor with ordinal levels is ever useful. As usual, the answer depends on what one wants to do with the data. While I agree with Baayen that, in most cases, discretization is probably not necessary and may even be detrimental, in cases where a regression is to be fit that includes some version of TIME or AGE as a predictor, it seems that the messiness of the raw TIME or AGE values (see again Figure 5) poses problems for regression-analytic approaches. Gries & Hilpert (2010) compared a model fit with raw values of TIME to a model fit with the five stages of TIME they arrived at using variability-based neighbor clustering (VNC) and the latter model fit was better. Still, this is not to be understood as a blanket one-size-fits-all recommendation—such decisions need to be made on a case-by-case basis. For instance, in the case of the much more monotonous trend represented in Figure 6 below, for regression-analytic purposes at least, using the raw values of TIME may prove just as useful as using VNC-derived mean frequencies.

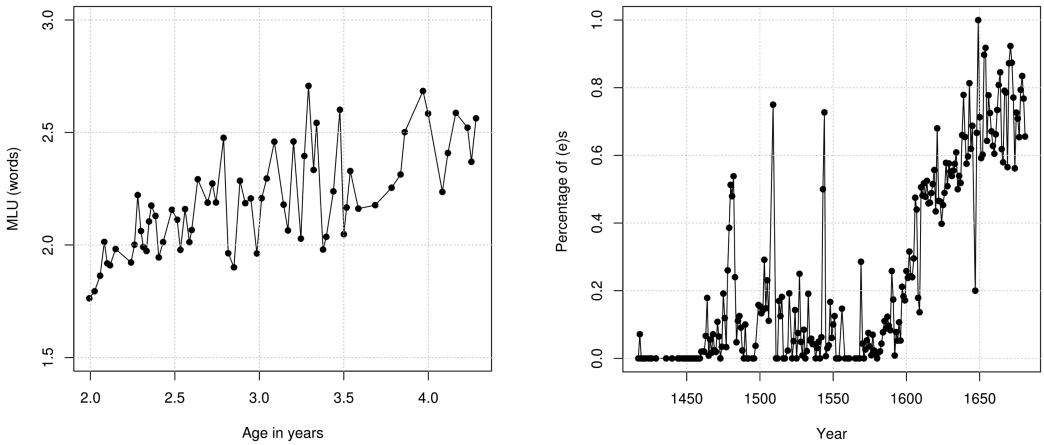

Figure 5: Examples of heterogeneous temporal corpus data: MLU data in first language acquisition (left panel) and proportions of third person singular (*e*)*s* (right panel)

The fact that there are overall increasing trends can be easily tested with correlation coefficients such as Kendall's τ or others. However, not only can such data violate the assumptions of frequently used statistical tests such as linear regression, but many frequently used statistics also provide too little information about the data. In particular, such statistics do not necessarily answer questions such as: (i) Are there different stages in the data, and if so, how many?; (ii) Do these different stages exhibit kinds of trends?

A frequent exploratory method to answer the first question, namely to discern sub-structure(s) in corpus data, is hierarchical cluster analysis, a statistical tool that groups data points into clusters on the basis of the points' pairwise similarity (such as the differences between MLU values or differences between percentages of (*e*)*s*). However, such cluster analyses cannot straightforwardly be applied to such temporally-ordered data: The computation of the similarity matrix of, say, the percentages of (*e*)*s* will return extremely high similarities for data points 150 years or more apart. However, a cluster analysis should not group such distant data points together given that, in historical data, grouping data points that might be 150 or more years apart makes little sense linguistically just as, in language acquisition data, grouping data points that might be 2 or more years apart makes little sense cognitively. Thus, what is required is a modification of the cluster-analytic approach that makes it operate locally, rather than allow it to merge data points that are too far apart.

One such approach is variability-based neighbor clustering (VNC; see Gries & Hilpert 2008). VNC differs from traditional clustering approaches in that it only permits temporally adjacent data points to be clustered together. Specifically, it is an iterative approach which, during each iteration, tests all adjacent (clusters of) points of time for their similarity, determines which two (clusters of) points of time are most similar to each other, merges those into one new cluster of (cluster of) points of time, and iterates. This way, widely disparate time periods cannot be merged into a (diachronically or acquisitionally) unrealistic cluster, but stages and outliers can be identified in a principled and replicable way.

Consider Figure 6 as a simple example. The left panel shows the development of the frequencies/10K words of *just because* in the *Time* magazine corpus. Obviously, there is a trend

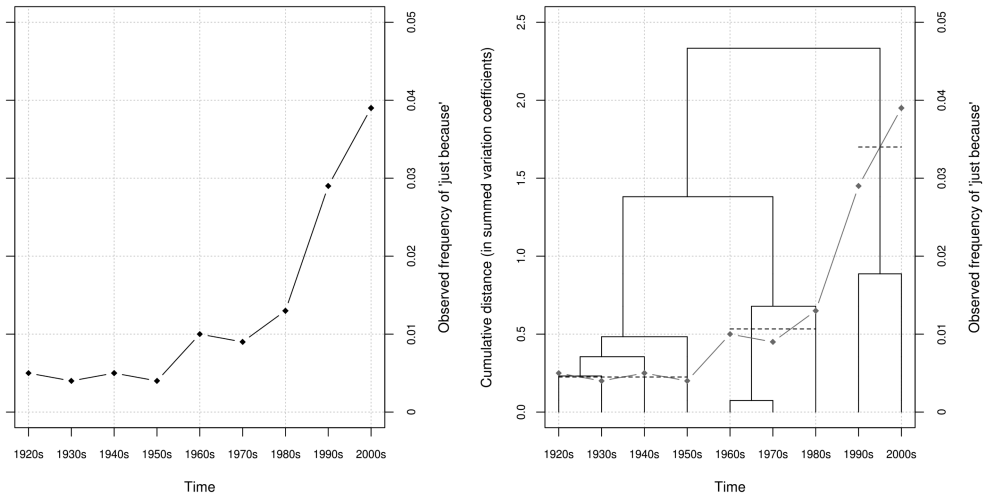

Figure 6: The development of the frequency of *just because* in several decades of *Time* magazine: frequency/10K words (left panel) and a three-cluster VNC dendrogram with cluster mean frequencies (right panel)

such that *just because* is becoming more frequent (and a rank correlation would reveal this trend to be significant ($\tau = 0.743$, $p = 0.005$). The right panel still shows the observed frequencies of *just because* (greyed out) but overlays the result of a VNC analysis. As is typical in hierarchical cluster analyses, the analyst has to choose a similarity metric and an amalgamation rule, and this analysis used variation coefficients for the former and concatenation for the latter. The VNC algorithm then returns three clusters (1920s–1950s, 1960s–1980s, and 1990s–2000s) and allows the analyst to compute (and represent with dashed horizontal lines) the mean observed frequency of *just because* in each time period.

This kind of approach has interesting potential. It can be used just to identify stages in historical data, which can be interesting in its own right. Then, as alluded to earlier in fn. 3, such stages can also in turn be utilized for subsequent analysis such as in regression-analytic approaches. Obviously, the method can also be applied to language acquisition data to identify developmental stages of children or to identify recordings that behave out of the ordinary given all the other recordings before and after them.

For example, Figure 7 shows the results of an application of VNC to frequencies of grammatical patterns in 13 recordings of a Korean child (in chronological order) from Patricia Clancy's Korean first language acquisition corpus (see Clancy 2003). The left panel shows a VNC dendrogram that not only identifies three distinct multi-recording clusters, but also shows that the first recording, where the child is youngest, is somewhat of an outlier. Once three clusters are assumed, then one can compute for each cluster mean (normalized) frequencies of occurrence. In this case, one can see the following tendencies:

- Zero becomes less frequent over time;
- P becomes more frequent over time;
- N and DP do not change much/markedly.

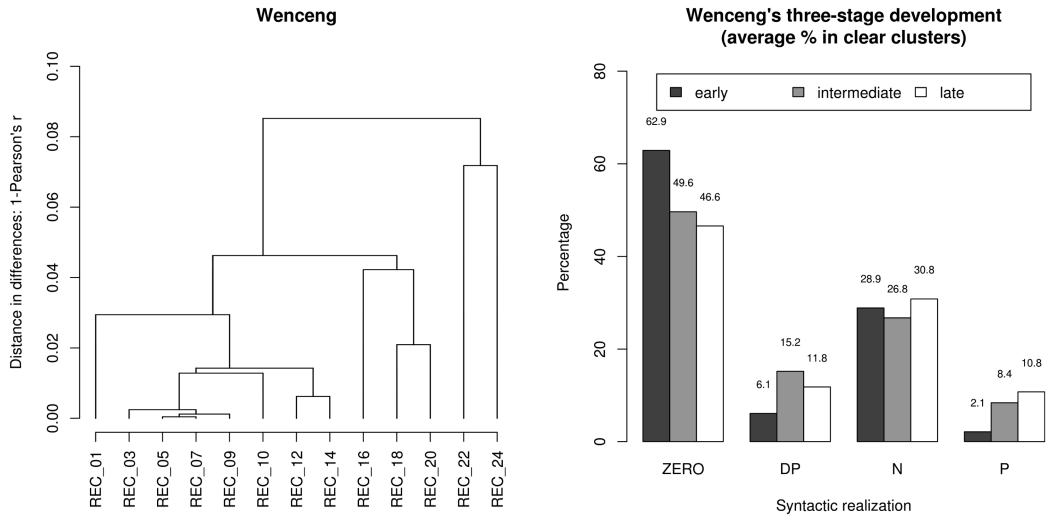

Figure 7: The application of VNC to language acquisition data: VNC analysis of frequencies of grammatical patterns of 13 recordings of a child called Wenceng (left panel); bar plots of mean percentages of patterns in the three age clusters identified in the VNC analysis (right panel) (data courtesy of P.M. Clancy)

In all of the above, VNC was used on data in which the measured data could be univariate (just one frequency as in the case of *just because*) or multivariate (several frequencies (of grammatical patterns) as in the language acquisition data), but where the dimension along which the clustering happened and along which VNC restricted it to neighboring elements was one-dimensional: time. Another interesting extension is using VNC for the analysis of data where there is more than one dimension, as when one studies geographical data in a quantitative dialectology setting and wants to prevent a regular hierarchical cluster algorithm from merging geographically very distant regions. The VNC algorithm can be adjusted correspondingly. Figure 8 shows an application of VNC to a matrix that provides normalized frequencies for 62 lexico-grammatical features for more than 30 regions in the U.K. If one wants to determine which regions emerge from the frequency data, however, one would probably not want to cluster Banffshire (BAN) together with South Devon (DEV). Thus the VNC algorithm is tweaked so that it only allows the clustering of counties that are next to other counties, whereas other counties can occur either in isolation or as part of an already merged cluster of counties.

Figure 8 shows three different steps in the iteration schedule:

- In the left panel, some first smaller clusters have emerged mainly in the south (one in the Cornwall and Devon regions and one in the Kent, East Suffolk, and London regions) as well as one small one involving Dumfriesshire and a larger one around Manchester.
- In the center panel most of the south is now interconnected (although Cornwall/Devon remain separate from the rest); not much has changed in the middle area.
- In the right panel, most of the country is now inter-connected apart from the very north—around Banffshire, Sutherland, Ross, and the Hebrides.

Figure 8: The extension of VNC to two-dimensional geographical data: three iterations (#14 (left), #20 (center), #31 (right)) from clustering British counties on the basis of frequencies of lexico-grammatical features (data courtesy of B. Szmrecsanyi; see Szmrecsanyi & Wolk 2011 for different analyses, discussion, and more references)

Thus, VNC can contribute to the (methodologically already quite sophisticated) domain of quantitative dialectology by helping to identify structures in corpus-linguistically described regions of a country or other larger regions that can then be interpreted against the background of other empirical or theoretical work. Given the increasing availability of historical corpora and regionally-stratified corpora, this method may therefore be a useful addition to the corpus-linguistic toolkit.

3.2 Learner corpus research and the problem of missing/impoverished context

The final corpus-linguistic domain to be discussed here is learner corpus research, that is, the branch of corpus linguistics exploring corpora containing non-native speaker (NNS) speech and/or writing. This field has become increasingly vibrant over the last 15 years or so, given the increasing availability of learner corpora. Much of this work is contrastive in the sense that NNS language is compared to the target of the learner as well as his L1(s), and an increasing amount of work approaches learner corpus data from a cognitively-informed perspective. Unfortunately, many studies in this field are quantitatively quite simplistic and restricted to the description of over- and underuses of linguistic elements in NNS language, accompanied by univariate or bivariate chi-squared tests. Examples include:

- Aijmer (2002), who explores the frequencies of use of modal verbs in NS English (in the LOCNESS corpus) and NNS English (in the Swedish component of the ICLE corpus) with multiple chi-squared tests.
- Altenberg (2002), who discusses frequencies/percentages of uses of English *make* and Swedish *göra* in four different constructional patterns and an 'other' category.
- Hasselgård & Johansson's (2011) case study of the use of *quite* in the LOCNESS corpus and four components of the ICLE Corpus (Norway, Germany, France, and Spain) involving chi-squared tests comparing *quite*'s frequency (both on its own and with a colligation) from the ICLE components to its LOCNESS frequency.

Typically, such quantitative analyses are lacking not only because of all the issues raised above, but also because they are not 'comparing/contrasting what non-native and native speakers of a language do *in a comparable situation*' (Péry-Woodley 1990:143, quoted from Granger 1996:43, our emphasis). This is because many studies reduce the notion of *comparable situation* to a single co-occurring factor/predictor, such as when Altenberg (2002) explores the use of *make* based on one predictor—patterns that *make* co-occurs with—or when Hasselgård & Johansson (2011) explore the use of *quite* based on one predictor—its colligation. Given the many factors that co-determine, say, which word of a set of near synonymous words is chosen, or which of two or more grammatical constructions is chosen, such studies cannot be anything but severely impoverished.

Thus, if the goal of learner corpus research is to determine how native speaker (NS) language and NNS language differ, a more comprehensive definition of comparable situation is needed, which will typically require the annotation of multiple features of the instances of the word/pattern in question. This in turn means that all these multiple features have to be included in the statistical analysis so as to determine which of these features has what kind of effect in the company of all other characteristics. Two main possibilities to do all this are available: both require corpus data on the element *E* under consideration that come from both NS and NNS data and that have been annotated with regard, ideally, to all the features that one has reason to believe affect the choice of *E*. Then, first, one can fit a regression in which:

- The dependent variable is either a binary or polytomous choice (for a binary or multinomial logistic regression) or a frequency (for a Poisson regression); for the choice of *of*- versus *s*-genitives, this would be the binary variable Genitive: *of* versus *s*.
- The predictors are all the annotated features as well as their statistical interactions (usually only up to the second or third degree); for the choice of *of*- and *s*-genitives, these may include the animacy of the possessor and the possessed, the length of the possessor and the possessed, the givenness of the possessor and the possessed, and many more; ideally, this would be a mixed-effects/multi-level model with random effects as required by the data/question(s).
- All the predictors from the previous bullet point are also allowed to interact with a predictor called Corpus or L1.

What is the rationale for the latter two guidelines? The rationale for the second guideline is that if one does not include the interaction, say, AnimacyPossessor:AnimacyPossessed, then one has

no way of finding out whether the preference of animate possessors for *s*-genitives holds regardless of whether the possessed is concrete or not. The rationale for the third guideline is that if one does not include the interaction, say, AnimacyPossessor:L1, then one has no way of finding out whether the preference of animate possessors for *s*-genitives holds in both NS and one or more NNS groups to the same degree (given the presence of all other (significant) predictors), which is precisely the kind of question that much learner corpus research is interested in but can often not answer because too few relevant predictors have been included (see Gries & Wulff 2013 and Gries & Deshors 2014 for examples and discussion).

There is a second approach (called MuPDAR, for Multifactorial Prediction and Deviation Analysis with Regressions) that is even more promising. It involves the following steps:

(i) Fit a first regression R_1 that conforms to the first two bullet points above, but only to the NS data.
(ii) If and only if R_1 results in a good fit and classification accuracy, then apply the regression equation thus obtained from R_1 to the NNS data to obtain for every NNS data point a prediction of what a NS would have done in the very same situation, which will serve as the gold standard.
(iii) If and only if R_1's NS regression equation also results in a relatively good fit with the NNS data, fit a second regression R_2 in which the dependent variable now is either a binary variable specifying whether the NNS made the same choice as a NS (*yes* versus *no*) would have made, or a continuous variable quantifying how much of the NNS choice was compared to what an NS was expected to say/write (this variable is 0 if the NNS made the NS choice, and a number other than zero, but between –1 and +1 if not).

It is this regression approach that precisely answers the core question of learner corpus research—in this linguistically and maybe contextually complex situation where the NNS had to make a choice, did he make a nativelike choice, 'Yes or no?'. And it is this regression approach that requires and at the same time guarantees a comprehensive definition of comparable situation—a hopefully large number of annotated factors describing the situation in which the NNS had to make a choice.

Gries & Adelman (2014) is a study using this approach:

(i) Fit a first mixed-effects regression R_1 that models whether Japanese NS realize a subject in a sentence on the basis of whether the referent of the subject is contrastive (a variable called Contrast) and how given it is (a variable called Givenness).
(ii) Apply the regression equation thus obtained from R_1 to the non-native speakers of Japanese corpus data to obtain for every NNS data point a prediction of whether an NS would have realized the subject there, *yes* or *no*.
(iii) Fit a second mixed-effects regression R_2 in which the dependent variable is a binary variable specifying whether the NNS made the same choice as an NS (*yes* versus *no*).

Using a polynomial to the second degree to model the predictor Givenness, they find that the NNS are on the whole quite close to the NS behavior, but (i) different speakers exhibit quite

different degrees of proficiency, and (ii) all NNS struggle most with making nativelike choices with intermediate degrees of givenness and non-contrastive referents:

- When the referent is contrastive, they realize it in the subject position as NS would.
- When the referent is non-contrastive and highly given or completely new, they do not realize it in the subject position or realize it in the subject positions as NS would.
- When the referent is non-contrastive and somewhat given, then faced with this middle-ground degree of givenness, their degree of nativelikeness decreases.

This approach, too, needs to be refined and developed further, however. It goes without saying that it is cognitively and contextually much more realistic and statistically more appropriate than decontextualized frequencies and/or chi-squared tests. So, again, it remains to be hoped that analytical strategies like this one will gain more ground in learner corpus research, the research on varieties, and any other domain where one part of the corpus data can be considered a standard or target with which the others can be meaningfully compared.

4. Concluding remarks

By way of a brief conclusion, corpus linguistics has made enormous headway in the recent past. To grow from a not particularly widely used method, geographically somewhat restricted to several Northern and Central European countries, to one of the most widely applied methods in linguistics of all sorts of theoretical persuasions worldwide in 15 to 20 years is no small feat. However, this is no time to rest on our laurels—now that corpus linguistics has become mainstream, and that's a good thing, we too must continue to refine our methods just as other fields have to. Many areas in psycholinguistics and computational linguistics have made interesting discoveries, have developed useful tools, have adopted great methods from neighboring fields, but corpus linguistics is unfortunately not leading the pack and must take care not to lose momentum either in terms of its own evolution or in terms of how it helps to shape linguistics as a whole. The present paper is an attempt to provide a snapshot of current problems, both in corpus linguistics in general and in selected hot topic areas, as well as to provide ideas and (first) suggestions about how to cope with these problems; I hope it will succeed as a call to (methodological) arms, and thus trigger developments that will help our field advance once more.

References

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. *Computer Learner Corpora*, *Second Language Acquisition and Foreign Language Teaching*, ed. by Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson, 55–76. Amsterdam & Philadelphia: John Benjamins.

Altenberg, Bengt. 2002. Using bilingual corpus evidence in learner corpus research. *Computer Learner Corpora*, *Second Language Acquisition and Foreign Language Teaching*, ed. by Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson, 37–54. Amsterdam & Philadelphia: John Benjamins.

Baayen, R. Harald. 2010a. A real experiment is a factorial experiment? *The Mental Lexicon* 5.1: 149–157.

Baayen, R. Harald. 2010b. Demythologizing the word frequency effect: a discriminative learning perspective. *The Mental Lexicon* 5.3:436–461.

Casenhiser, Devin, & Adele E. Goldberg. 2005. Fast mapping between a phrasal form and meaning. *Developmental Science* 8.6:500–508.

Clancy, Patricia M. 2003. The lexicon in interaction: developmental origins of Preferred Argument Structure in Korean. *Preferred Argument Structure: Grammar as Architecture for Function*, ed. by John W. Du Bois, Lorraine E. Kumpf & William J. Ashby, 81–108. Amsterdam & Philadelphia: John Benjamins.

Clark-Sánchez, Victoria. 2013. Review of *Quantitative Corpus Linguistics with R: A Practical Introduction*. *Corpora* 8.2:269–272.

Daudaravičius, Vidas, & Rūta Marcinkevičienė. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics* 9.2:321–348.

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27.1:1–24.

Ellis, Nick C., Rita Simpson-Vlach, & Carson Maynard. 2007. The processing of formulas in native and L2 speakers: psycholinguistic and corpus determinants. Paper presented at the UWM Linguistics Symposium on Formulaic Language, April 16–21, 2007. Milwaukee: University of Wisconsin-Milwaukee.

Evert, Stefan. 2009. Corpora and collocations. *Corpus Linguistics: An International Handbook*, Vol. 2, ed. by Anke Lüdeling & Merja Kytö, 1212–1248. Berlin & New York: Mouton de Gruyter.

Firth, John R. 1957. A synopsis of linguistic theory 1930–55. *Studies in Linguistic Analysis*, 1–32. Oxford: Basil Blackwell.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*, *Lund*, *4–5 March 1994*, ed. by Karin Aijmer, Bengt Altenberg & Mats Johansson, 37–51. Lund: Lund University Press.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13.4:403–437.

Gries, Stefan Th. 2010a. Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods. *Corpus Linguistics in Language Teaching*, ed. by Tony Harris & María Moreno Jaén, 121–146. Frankfurt am Main: Peter Lang.

Gries, Stefan Th. 2010b. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. Paper presented at the Corpus Linguistics 2009, July 20–23, 2009. Liverpool: University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009.

Gries, Stefan Th. 2010c. Dispersions and adjusted frequencies in corpora: further explorations. *Corpus Linguistic Applications: Current Studies*, *New Directions*, ed. by Stefan Th. Gries, Stefanie Wulff & Mark Davies, 197–212. Amsterdam & New York: Rodopi.

Gries, Stefan Th. 2011. Methodological and interdisciplinary stance in corpus linguistics. *Perspectives on Corpus Linguistics: Connections and Controversies*, ed. by Vander Viana, Sonia Zyngier & Geoffrey Barnbrook, 81–98. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan Th. 2013. 50-something years of work on collocations: what is or should be next … *International Journal of Corpus Linguistics* 18.1:137–165.

Gries, Stefan Th. (forthcoming). The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10.1.

Gries, Stefan Th., & Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, 35–54. Berlin & New York: Springer.

Gries, Stefan Th., & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9.1:109–136.

Gries, Stefan Th., & Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3.1:59–81.

Gries, Stefan Th., & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14.3:293–320.

Gries, Stefan Th., & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: an ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15.4:520–548.

Gries, Stefan Th., & Stefanie Wulff. 2013. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of *context* in learner corpus research. *International Journal of Corpus Linguistics* 18.3:327–356.

Harris, Zellig S. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel.

Hasselgård, Hilde, & Stig Johansson. 2011. Learner corpora and contrastive interlanguage analysis. *A Taste for Corpora: In Honour of Sylviane Granger*, ed. by Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot, 33–61. Amsterdam & Philadelphia: John Benjamins.

Janda, Laura A. (ed.) 2013. *Cognitive Linguistics: The Quantitative Turn*. Berlin & New York: De Gruyter Mouton.

Joseph, Brian. 2004. On change in *Language* and change in language. *Language* 80.3:381–383.

McDonald, Scott A., & Richard C. Shillcock. 2001. Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech* 44.3: 295–322.

Michelbacher, Lukas, Stefan Evert, & Hinrich Schütze. 2007. Asymmetric association measures. Paper presented at the International Conference on Recent Advances in Natural Language Processing (RANLP 2007), September 27–29, 2007. Borovets, Bulgaria.

Michelbacher, Lukas, Stefan Evert, & Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7.2:245–276.

Mollin, Sandra. 2009. Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5.2: 175–200.

Nakagawa, Shinichi, & Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4.2:133–142.

Péry-Woodley, Marie-Paule. 1990. Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching* 23.3:143–151.

R Core Team. 2014. R: a language and environment for statistical computing. R Foundation for statistical computing. Vienna, Austria. http://www.R-project.org/.

Recchia, Gabriel, Brendan T. Johns, & Michael N. Jones. 2008. Context repetition benefits are dependent on context redundancy. *Proceedings of the Annual Conference of the Cognitive Science Society* 30:267–272.

Simpson-Vlach, Rita, & Nick C. Ellis. 2005. An academic formulas list (AFL): extraction, validation, prioritization. Paper presented at Phraseology 2005, October 13–15, 2005. Louvain-la-Neuve: Catholic University of Louvain.

Stefanowitsch, Anatol, & Stefan Th. Gries. 2003. Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8.2:209–243.

Stoll, Sabine, & Stefan Th. Gries. 2009. How to measure development in corpora? An association strength approach. *Journal of Child Language* 36.5:1075–1090.

Szmrecsanyi, Benedikt, & Christoph Wolk. 2011. Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics* 11.2:561–592.

Wahl, Alexander R. (in progress). *New Approaches to Extracting Multi-word Expressions from Corpora: Unprespecified Ngram Lengths*, *Long-distance Dependencies*, *and Enhanced Association Measures*. Santa Barbara: University of California at Santa Barbara dissertation.

[Received 30 December 2013; revised 18 April 2014; accepted 27 June 2014]

Department of Linguistics
University of California at Santa Barbara
Santa Barbara, CA 93106-3100
USA
stgries@linguistics.ucsb.edu

語料庫語言學量化研究的問題及其解決方案

Stefan Th. Gries
加州大學聖塔芭芭拉分校

目前迅速發展的語料庫語言學面臨眾多研究方法的挑戰。與語料庫語言學本身相關的方法論問題——資料的散布離差、詞種頻率／亂度、與資料隱含的方向性問題等，直接影響語料相關性指標的計算；忽視語料庫的資料取樣結構也與之後的統計分析結果直接相關。歷史語言學與學習者語料庫研究等領域應用語料庫語言學時，也有方法論的問題。本文詳細討論以上所提到的問題，並具體提出實例演示相對應的解決方法。

關鍵詞：關連性度量，詞次／詞種頻率，混合效應／多層次模型，以變異性為本的連結群聚法，MuPDAR