7 Variationist Analysis

Variability Due to Random Effects and Autocorrelation

Stefan Th. Gries

Introduction

The Overall Frequencies of (Co-)occurrence Approach

In contemporary linguistics, corpora are arguably one of the central methodological tools and one of the central sources of data. More and more linguists look into corpora for information on frequencies of occurrence of a particular expression or frequencies of co-occurrence of a particular linguistic expression with other expressions or contextual characteristics. While much earlier work in corpus linguistics involved questions of lexical semantics (often from a lexicographic perspective), for quite some time now, corpus-linguistic applications have become much wider in scope, covering questions from the domains of morphology, (morpho)syntax, and pragmatics from both synchronic and diachronic angles, in both native language and foreign/second languages. One area of research that has seen a particularly strong boost is the study of what I will broadly refer to here as lexicosyntactic alternations. With this term, I am referring to instances where speakers (have to) choose one out of a typically small set of several (nearly) equivalent lexical or syntactic options and typically do so without much or any awareness of the factors driving their choices. (1) Shows a few purely lexical choices (of near synonyms), whereas (2) exemplifies cases of either purely syntactic choices (e.g., (2a)) or of choices that involve both lexical and syntactic decisions (e.g., (2b–d)).

(1) a. La Forge couldn’t make sense of the symmetric/symmetrical pattern(s) b. Dr. Crusher was annoyed by the shouting/yelling children
c. Picard attempted/try to kill the Borg

(2) a. Picard picked up the tricorder versus Picard picked the tricorder up b. Worf will kill the Romulan versus Worf is going to kill the Romulan
c. Picard gave Riker his orders versus Picard gave his order to Riker
d. the admiral’s orders versus the orders of the admiral

Given the ease with which frequencies of choices involving lexical material can be extracted from corpora, it comes as no surprise that there are many reference works and studies that report and utilize frequencies of occurrence of the alternants that make up alternations. Maybe the most famous example of the former is Biber et al.’s (1999) comprehensive corpus-based reference grammar of English, which provides normalized frequencies of a large number of grammatical phenomena for different registers (conversation vs. fiction vs. news vs. academic prose) and modes (spoken vs. written). As for the latter, the following are examples from the domain of learner corpus research involving comparisons of native speaker (NS) and (different kinds of) non-native speaker (NNS) data, a very common application in that field:

- Hyland and Milton (1997) compare frequencies of epistemic modality expressions;
- Laufer and Waldman (2011) compare frequencies of V-N collocations across NS and differently proficient levels of NNS;
- Hasselgård and Johansson (2012) compare frequencies of quite (in isolation and in colligations).

That is, such studies usually provide (i) normalized frequencies of occurrence of particular expressions (per register, per mode, per L1, . . .) and/or (ii) normalized frequencies of how often particular expressions co-occur with some other (kind of) expression (sometimes explored statistically using many Χ²-tests or the related log-likelihood ratios). Given the regularity with which such frequencies of occurrence are reported, it is probably no exaggeration to assume that this is one of the corpus-based statistics most commonly used in the last three or so decades. However, as I will argue presently, they are also potentially very misleading.

The Variationist Case-by-Variable Approach

While the aforementioned kinds of frequencies of (co-)occurrence are very useful in reference works, their utility in research articles (in particular for learner corpus research but also more generally) is often much more doubtful given how raw/normalized frequencies of occurrence typically divorce the use of an expression from the rich context in which they are used. Consider the use of may and can by NS and NNS in the data of Gries and Deshors (2014). They show how a simple regression model trying to determine how frequently NS and NNS use may and can indicates that NS use may a bit more often than NNS. However, they proceed to show that this overall difference/effect is misleading because NS and NNS use the two modal verbs very differently depending on the aspect of, and the presence/absence of negation in, the verb phrase. Thus and more generally, an observed difference of frequencies of (co-)occurrence can have many reasons: if (i) the presence of negation leads to a preference of can over may in NS data and (ii) NNS use can more than NS, then either the NNS overuse
can (for reasons having to do with their non-native proficiency) or the NNS overuse negation and at the same time use can just like NS would if they also used negation more, namely more often. It is therefore necessary to recognize that overall frequencies of occurrence of some linguistic expression e that do not involve a detailed analysis of e's contexts are potentially useless and risky because they do not allow the analyst to determine which of the two mentioned explanations (or many other competing ones) is correct or at least more likely.

The solution to this problem is to adopt an approach that is variationist in nature (i.e., is compatible with the work done for a long time in variationist sociolinguistics) and requires what is often referred to in statistics as the case-by-variable format: typically, every occurrence in the corpus to be studied—each case—is annotated (in a spreadsheet) for a variety of variables or predictors that are likely to affect the linguistic choice under investigation (in the 'Match' column), as represented in Table 7.1.

Note that it is often useful to also add a column, which might be called "Alternate", that indicates for each case whether each of the alternants would have been possible or not because, depending on one's goals, subsequent statistical analyses may be run either on all instances of the competing linguistic choices or only on those that could alternate. Either way, the next step is often a statistical analysis to determine which of the many annotated predictors (and, ideally, their interactions) are correlated with the linguistic choice and how so. Interestingly, for many of the syntactic alternations whose studies dominate the literature such as those listed earlier in (2), the linguistic factors that govern them in English at least are similar (and often related):

- **information-structural factors** having to do with the givenness, or degree of discourse activation, of the referents of noun phrases such that, usually, given/inferable elements precede new referents;
- **weight-related factors** having to do with the length/weight/complexity of the phrases whose ordering is studied such that, usually, short elements precede long elements;
- **animacy-related factors** having to do with what degree of animacy the referents of various noun phrases (NPs) in the relevant verb phrases have;

<table>
<thead>
<tr>
<th>Case</th>
<th>Preceding</th>
<th>Match</th>
<th>Subsequent</th>
<th>Predictor 1</th>
<th>Predictor 2</th>
<th>Predictor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Worf</td>
<td>will</td>
<td>kill the</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Romulan</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

On the basis of fine-grained annotation of the aforementioned kind, multifactorial statistical analyses—currently these are frequently regression models—can be applied to see which of these factors are correlated with, and thus likely causes of, the relevant alternation. This kind of analysis is hugely superior to overall frequencies of (co-)occurrence because it allows one to distinguish many different potential causes for what may seem like over-/underuse of a particular expression in some groups of speakers (e.g., learners of different L1s, speakers of different dialects, speakers using language in different registers).

In much recent work, the aforementioned approach was already implemented and has yielded results that improve considerably upon the more traditional approach of the preceding section. In the remainder of this paper, I want to draw attention to a small set of additional factors whose inclusion would benefit corpus-based research on alternative linguistic choices.

**Case Studies**

In this section, I will discuss how the study of the to-be-explained variability in the data can benefit from taking more into consideration than the usual linguistic determinants discussed earlier, namely, by exploring effects that, in the language of statistics, could be characterized as

- **random effects**, i.e., the role of factors whose levels in the current corpus sample do not exhaust the range of possible levels in the population ('out there in the language'); these include speaker-specific variation (because typically a corpus does not contain all speakers of the language) and lexically specific variation (because typically a corpus does not contain all, say, verbs, that can occur with, say, a particular tense);
- **autocorrelation**, i.e., the fact that earlier linguistic behavior co-determines later linguistic behavior (by the same speaker or others) as when, by virtue of a process often referred to as structural priming, the use of a passive structure by a speaker makes it more likely that that speaker will use a passive again in the near future (see Schenken 1980; Weimer & Labov 1983; Estival 1985 for the earliest observational studies).
The linguistic choice I will use to exemplify the large amount of variability covered by such factors is future choice as shown in (2b). Such an alternation is an interesting question for the aforementioned kinds of effects because not only are there a variety of linguistic factors governing future choices, but there are also a range of studies that have revealed sometimes marked differences in alternation behaviors/preferences of specific lexical items but also between native and indigenized varieties of English (see Mukherjee & Hoffmann 2006; Mukherjee & Gries 2009). That, in turn, makes a corpus that includes different (kinds of) varieties and topics, which may give rise to different kinds of verbs, a prime test case. Thus I used R to retrieve candidates of future choices from the Q+A corpus using the regular expression shown in (3):

\[(\text{wilshia|illwo}_*)(\text{going|vkgk})-\text{to}_{-}\text{to};([_]+[(_v)[(_]+)[(_]+[(_]+[_[_]+v[^]+]

This retrieved 2,329 matches of
- will or shall or wo (for won’t). Followed by the tag vnm OR going followed, by the tag vkgk, followed by to tagged as to, followed by;
- between zero and two tagged ‘things’ that are not tagged as verbs (each followed by a space);
- followed by something tagged as a verb;
- within one line.

This (then slightly cleaned and homogenized) concordance constitutes the data on which the following sections are based. One traditional kind of approach discussed earlier would consist of providing overall frequencies of, say, will and going to in the corpus as a whole or in variety-/register-topically restricted parts of the corpora. Table 7.2 is an example of the kind of overall frequency data that much work (especially in learner corpus research) has provided but that, given its neglect of context, cannot really reveal that much.

Another frequent instantiation of the traditional approach would be to annotate the concordance lines with regard to some features likely to affect future choice and then study each feature (often done in isolation using cross-tabulation). For instance, future choice is said to be affected by the presence of negation (see Szirecseyi 2005, 2006 for an excellent analysis). A quick classification of whether the future verb phrase (VP) is negated or not in the whole corpus yields Table 7.3, which, if tested with a \(\chi^2\)-test, as many would do (not that one should, see the following section), returns a significant \(p\) value \((\chi^2\text{-test } = 8.509, df = 2, p = 0.014)\) of a rather weak effect \((V=0.06)\) such that, with negated VPs, the proportion of going to and shall are higher for than for will.

One major shortcoming of such analyses is that they are not multifactorial because each such predictor is studied in isolation, which by definition already leads to an incomplete picture. However, the next few sections will show how such analyses also miss a lot of variability by neglecting sources of variability other than the ‘regular’ linguistic predictors discussed earlier.

### Speaker-Specific Variation

The first kind of random effect that distorts all overall frequencies but is usually readily available from the fully annotated, case-by-variable format is speaker-file-specific variation. This refers to the fact that speakers may differ considerably and systematically in terms of their future choices, which rules out the aforementioned \(\chi^2\)-test. The fact that the overall percentages mask a considerable amount of speaker-specific variation is represented in Figure 7.1. Both panels represent the percentage of uses of will (as opposed to going to, shall has been omitted here because of its low overall frequency) on the x-axis, the variety (left panel), and the topic (right panel) are on the y-axis, and every gray point indicates one speaker’s overall preference of will with darker grays reflecting overplotting and short vertical lines indicating group medians.

Several observations are immediately obvious: (i) there is a large amount of variability between speakers; (ii) this is true even if the notable differences of the variety-specific medians suggest that, on the whole, the native varieties use will less than the indigenized varieties (see the following section for more discussion); (iii) the topic-specific medians differ much less from each other than the variety-specific medians; and (iv) there are many speakers (124 in fact, nearly half of all speakers) who invariably use will and a few speakers (5) who invariably use going to, which means that these speakers’
behavior, if not controlled for, can potentially distort the analysis of any factor affecting future choice simply because these speakers might weaken any factor’s impact (since that factor would potentially not explain any variation in those speakers’ choices).

For instance, if a foreign language learner of English does not know the going-to future yet, then he is not going to use it even when negation is present, thereby seemingly weakening the statistical effect that negation has on going to when the real reason is that the speaker does not even know he has a choice in the first place. In fact, if one tries to predict every future choice in the corpus and does so just by choosing the construction that each speaker prefers in general and chooses will when a speaker uses both futures equally often (because will is generally so much more frequent), then one can predict \( \frac{2309}{2232} = 90\% \) of all instances of will and going to correctly on the basis of speaker-specific effects alone and will’s general predominance.

It is for this reason that corpus-linguistic analyses should always explore speaker/file-specific effects of the aforementioned kind.\(^3\) In fact, an even better kind of analysis would also take into consideration the fact that speakers/files are nested into varieties (because each speaker is only attested in one variety), which are in turn nested into variety types (because each variety in this corpus is either native or indigenized), and variability in future choice can be manifested at each of these levels of resolution.

**Lexically Specific Variation**

The second kind of random effect that distorts overall frequencies but is readily available from concordance data is how grammatical constructions can exhibit preferences to particular lexical items; this may often be due
to the lexical items’ semantics (and, thus, their correlations with semantic factors discussed earlier). In corpus linguistics, this notion has been captured under the notion of collocation and also during the last ten-plus years under that of collostruction, a blend of collocation and construction (see Stefanowitsch & Gries 2003). The family of methods called collostructional analysis includes the method of distinctive collexeme analysis, a straightforward application of association measures from collocation research to co-occurrence of a word \( w \) and two constructions \( c_1 \) and \( c_2 \); the analyst creates tables of the kind of Table 7.4 for every word occurring at least once in either \( c_1 \) or \( c_2 \) and computes an association measure from that table such as Mutual Information \( MI \), \( t \), log-likelihood, or \( p_{\text{Fisher-Yates exact test}} \).

Gries and Stefanowitsch (2004) applied this method to contrast will- and going-to futures in the ICE-GB and found that many verbs attracted to the will- future are characterized by relative non-agentivity and low dynamicity, including perception/cognition events and states whereas the opposite is found for verbs attracted to the going-to future.

An extension of this method, multiple distinctive collexeme analysis, can compare how much a word \( w \) is attracted to, or repelled by, more than two constructions such as the three future choices will, going to, and shall.\(^3\) Given the strong predominance of will in the present corpus, the results will be less revealing semantically because so few verbs occur significantly more frequently with will than the overall high baseline already leads one to expect. However, the point is, as before, to show that much variability that can easily and prematurely be attributed to linguistic factors, learners’ lack of proficiency, etc., may in fact consist (in part) just of lexical preferences (and whatever these ‘operationalize’ semantically).

If such a multiple distinctive collexeme analysis is applied to all 422 verb lemmas occurring with at least one future choice once in the corpus as a whole, then only a few lemmas, 32, reach significant levels of attraction; however, these 32 lemmas account for nearly half the data, namely, 1,030 future choices. Consider Figure 7.2 for a visual representation of verbs’ construc
tional preferences.

As is obvious, many of these verb lemmas have quite distinct preferences. The three future choices are symbolized by the three differently colored segments, the sizes of which represent the percentage of times the relevant

---

### Table 7.4 Schematic co-occurrence table for measuring the association between a word lemma \( w \) and each of two constructions \( c_1 \) and \( c_2 \) in some corpus

<table>
<thead>
<tr>
<th></th>
<th>Construction ( c_1 )</th>
<th>Construction ( c_2 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>word lemma ( w )</td>
<td>( a )</td>
<td>( b )</td>
<td>( a + b )</td>
</tr>
<tr>
<td>other word lemmas</td>
<td>( c )</td>
<td>( d )</td>
<td>( c + d )</td>
</tr>
<tr>
<td>Total</td>
<td>( a + c )</td>
<td>( b + d )</td>
<td>( a + b + c + d )</td>
</tr>
</tbody>
</table>
one can predict that future choice. For instance, *happen* is most strongly attracted to the **will** form, whereas *come* and *do* are most strongly attracted to the **to** form. While the dataset is too small to make meaningful comparisons between all verbs, it is reassuring to see that several of the verbs that are most strongly attracted to **will** are also high-frequency verbs in the corpus, whereas **to** is associated with more specialized contexts. Any attempt to predict future choice in the corpus and participant variability would require additional data (in particular, verbs specific to specific, as before, the speakers' preferences, if we use a model trained on the corpus and participant variability, our model would need to be trained on a larger, more diverse dataset).
Figure 7.3 Switch-rate plot for *will* futures

The result is very straightforward: switch rates to *will* are overwhelmingly lower than the frequency of *will* would lead one to expect. Speakers switch less, i.e., repeat more, i.e., exhibit priming effects. However, the overplotting makes it very difficult to explore the results in more detail (e.g., by variety or by topic), which is what Figure 7.3 allows one to do, which represents for each speaker the subtraction from the x-axis value in Figure 7.4 from the corresponding y-axis value: the smaller a plotted value, the more different the switch rate to *will* is from the frequency of *will* for that speaker and the more the results are compatible with priming effects.

The results are again quite clear but now come with the finer resolution of varieties and topics. The left panel shows that priming exists (given that so many values are much smaller than zero) and that it affects the speakers of the native varieties (UK and US) less than the speakers of the indigenized varieties (IN and PH): the difference between observed and expected switch rate is closer to zero for the former than for the latter, a finding that researchers may try to integrate with regard to different degrees of evolution of different varieties (as in Schneider's 2007 model) or with regard to different susceptibilities toward priming of varieties differently entrenched in speakers' minds. However, and as might be expected, the right panel suggests strongly that the three topic areas exhibit priming effects, too, but do not differ from each other at all. Also, the findings show that the Q+A corpus seems to be more similar to spoken than to written registers, given the high overall degree of priming even in the native varieties since priming has been found to be weaker in writing.

As before, let us briefly consider how predictive priming is on its own: if one tries to predict every future choice in the corpus and does so just by choosing the construction that the speakers used last time and chooses *will* for a speaker's first future (because *will* is generally so much more frequent), then one can predict \( \frac{1884}{2329} = 80.9\% \) of all instances of *will*, *shall*, and *going to* correctly just on the basis of what the speaker did last time around, a finding that should again be a strong incentive to always explore priming effects.

**Concluding Remarks**

As mentioned initially, corpora and the frequency data that they offer to corpus linguists have become an ever-more important tool for theoretical and applied linguistics alike and various kinds of frequency information have provided immensely useful information. However, I hope to have shown (i) that overall frequencies of occurrence—absolute or relative—such as in Table 7.2, while useful in the context of surveys and overall reference works, are from my point of view most useful for exploratory purposes because such frequencies are typically both decontextualized and zero-monofactorial in nature, whereas linguistic choices are not. Ignoring—i.e., not annotating and statistically analyzing—contextual and other features of a phenomenon of
interest means the researcher cannot, by definition, distinguish between different explanations for whatever over- or underuse frequencies he found and reported, which in turn virtually guarantees that monofactorial results will over- or underestimate the actual trends in the data.

I also hope to have shown (ii) that even if linguistic features from the context of a linguistic choice are included—information-structural, weight-related, animacy, and other semantic factors, etc.—there are also other sources of variation that commonly remain underanalyzed: variation due to (a) speakers and (b) lexical items (and not discussed in great detail), variation due to (c) the hierarchical structure of most corpora, as well as (d) priming/autocorrelation effects, each of which has considerable predictive power on its own. That in turn means that studies ignoring such effects run the risk of (i) misidentifying the reasons for linguistic choices—the reason for a particular choice may not have been information-structural or weight-related but simply that speaker’s preferred choice—and/or (ii) failing to find an explanation for what appear to be inexplicable linguistic choices—maybe the explanation for a speaker’s inexplicable choice of a construction is nothing that can be seen in the current (concordance) context but is quite obvious from the previous one. Ideally, of course, all four effects discussed earlier would be included at the same time as the contextual features with, for instance, mixed-effects/multi-level modeling (see Gries 2015b for recent explanation in a corpus-linguistic context). If a multi-level model involving all four aforementioned effects is applied to the present data to determine whether the weak but significant correlation between negation and future choice apparent from Table 7.3, the risks associated with the cross-tabulation of frequencies becomes apparent: a model with all random effects and priming as a predictor is hugely more preferable (evidence ratio_{AICc} > 10^{15}) than a model that also involves negation—thus the simple cross-tabulation leads one to believe in an effect that better analysis shows to be non-existent.

While the exposition here could only scratch the surface, I hope that the empirical issues and methodological strategies discussed in this chapter to tackle these kinds of problems will stimulate researchers to pay closer attention to these important factors: studies of different varieties need to look beyond the immediate context to more widespread preferences of people and words, as well as previous contexts to avoid potentially misinterpreting results.

Postscript

To me, this experiment was a very interesting experience for mainly two reasons. On the one hand, I was (positively) surprised by the whole range of areas that were explored, many of which are outside my areas of expertise and thus exposed me to research that I had not known (well) before; in that connection, I have to admit I was struck by a feeling that my chapter didn’t fit the rest of the volume as well as I had hoped to be able to achieve because (i) most other chapters focused on lexical items/bundles as well as (e.g., semantic) characteristics of theirs and their distributions across varieties and topics and (ii) how the papers were located on a (simplistic) continuum from mostly/exclusively qualitative to mostly/exclusively quantitative work. My own submission was narrower in scope than many others in how it focused on one small and lexicogrammatical alternation—future choice of will versus going to (vs. shall)—as opposed to a larger range of (lexical) expressions, and my submission was more on the (less populated) quantitative side of the spectrum (together with, say, Friginal & Biber’s or Egberts’ chapters).

On the other hand, and this is not to criticize any other submission(s) their relevance for valuable exploratory purposes, many other submissions also reaffirmed my aforementioned views on (i) the importance, if not (often) indispensability, of context annotation of current or previous instances for the study of any frequency data (or statistics derived from them such as keywords or co-occurrence strengths) and (ii) the subsequent statistical analysis of the degree to which such annotated characteristics affect, or at least correlate with, the phenomenon of interest, and I am not implying I myself have always done this to the extent that I now consider essential! It is hard to see which, if any, of the case studies in this volume would not be affected by at least one of the three factors discussed here: any frequency can be affected by dispersion (e.g., speaker- or, here, thread-specific variation), and many frequencies of occurrences of lexicogrammatical choices will also be affected by autocorrelation/priming, which makes it ever-more important to control for such factors (using good sampling, controlling for contexts, and/or appropriate statistics). To mention but one example, do keyword statistics change if particular parts of the reference corpora are omitted, where ‘parts’ can be defined on any level of granularity, thread, variety, topic, etc.?

Thus, while my chapter’s contribution to the identification and understanding of differences between varieties and topics in the Q+A corpus is perhaps more limited than that of many other chapters, I hope that it is still worthwhile as a perhaps cautionary but certainly complementary follow-up to the many discoveries my co-contributors have made.

Notes

1 Given the inconsistent use of apostrophized forms, for the sake of simplicity, no forms such as I’ll, he’ll, etc., were explored; this has no effect on the overall argument.

2 There are already some studies that adopt an approach similar to the aforementioned by computing, for instance, normalized frequencies per file (as in Figure 7.1) and then compute means, standard deviations, or more complex statistics based on all by-file normalized frequencies. This indeed addresses the role of speaker-specific variation but still usually faces problems. First, the role of context is still unclear, which means that essentially no, even only potentially causal, claims can
be made; second, the usual kinds of parametric statistics (such as means, standard deviations, etc.) must not actually be applied to such data because they are typically not normally distributed. In the present data, the seven by-speaker percentages of will-futures across varieties and topics are all non-normal (all seven Shapiro-Wilk test \( p < 0.05 \)). Third, this approach cannot easily accommodate multiple kinds of random effects at the same time.

4 A Kolmogorov–Smirnov test comparing the plotted differences for the native speakers to those of the indigenized speakers returns a significant result \( (D = 0.249, \ p < 0.001) \).

5 Kolmogorov–Smirnov tests comparing the three topics to each other return only non-significant results (all \( D < 0.1 \), all \( p \) adjusted for three tests > 0.9).

References


