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/tried 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 χ^2-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 7.1 A partially schematic concordance display of future choice in the case-by-variable format

Case	*Preceding*	*Match*	*Subsequent*	*Predictor 1*	*Predictor 2*	*Predictor 3*
1	Worf	will	kill the Romulan	...	...	...
...	...	...	...	...	...	...

- various other *semantic factors* having to do with aspects, aktionsart, general semantic categories, case roles, and many other phenomenon-specific ones;
- *processing-related factors* having to do with the distribution of the information provided by upcoming linguistic material (often measured in information-theoretic terms);
- *phonological factors* having to do with how much competing constituent orders violate near-universal preferences, such as rhythmic alternation or preferred syllable structure.

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 Schenkein 1980; Weiner & 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):

(3) (((wilsha)lll|wo)_vm|going_vvgk·to_to)·([^_]+_[^v][^]+·){0,2}[^_]+_v[^]+

This retrieved 2,329 matches of

- *will* or *shall* or *wo* (for *won't*).[1] Followed by the tag *vm* OR *going* followed, by the tag *vvgk*, 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

Table 7.2 Frequencies of *will* and *going to* across varieties and variety types in the Q+A corpus

Type	*Variety*	going to	will	*Total per variety*	will *per type*
Indigenized	IN	38	561 (93.7%)	599	92.6%
	PH	43	529 (92.5%)	572	
Native	UK	72	354 (83.1%)	426	83.1%
	US	81	554 (87.2%)	635	

Table 7.3 Frequencies of future choices depending on negation in the Q+A corpus

Future	*Affirmative*	*Negative*	*Total*
going to	194	40	234
shall	75	22	97
will	1731	267	1998
Total	2000	329	2329

cross-tabulation). For instance, future choice is said to be affected by the presence of negation (see Szmrecsanyi 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 χ^2-test, as many would do (not that one should, see the following section), returns a significant p value (χ^2-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 χ^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'

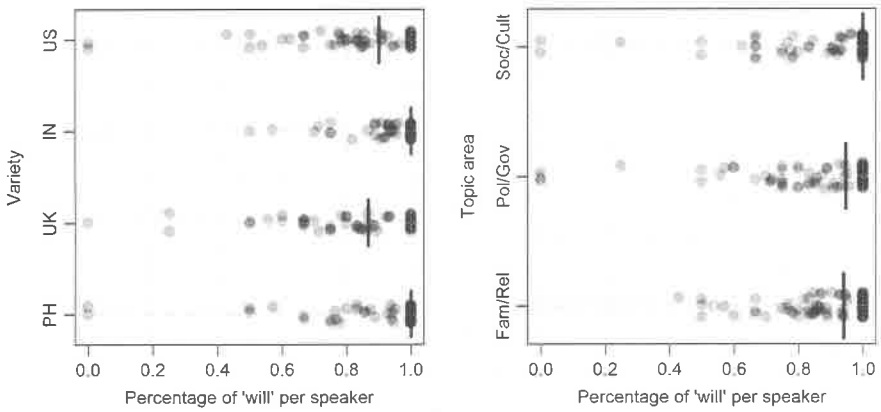

Figure 7.1 Percentages of use of *will* per file/speaker by variety (left) and by topic (right)

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 $^{2009}/_{2232} = 90\%$ 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.[2] 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 *colligation* 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' constructional 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

	Construction c_1	*Construction* c_2	*Total*
word lemma w	a	b	$a + b$
other word lemmas	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

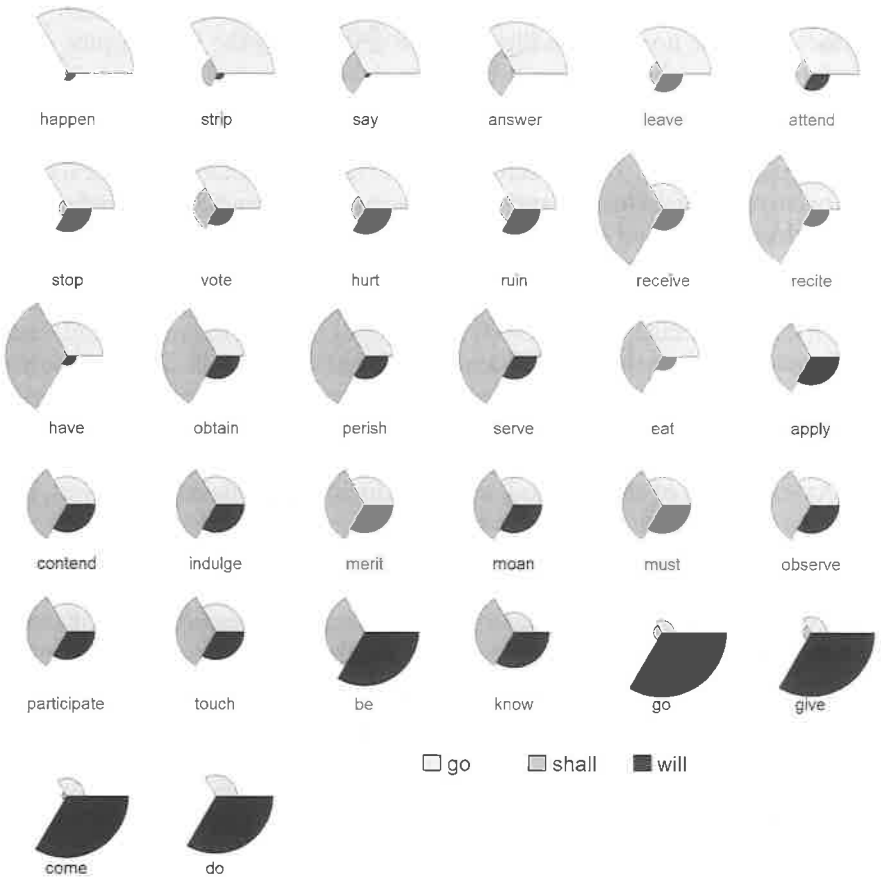

Figure 7.2 The degrees of attraction of significantly attracted verbs to futures

verb occurs with that future choice. For instance, *happen* is most strongly attracted to the *going-to* future, whereas *come* and *do* are most strongly attracted to *will*. While the dataset is too small to make meaningful comparisons between varieties or topics, it is reassuring to see that several of the earlier findings of Gries and Stefanowitsch are supported even in this more specialized corpus: *going to* is used more with rare but more specific verbs (in particular verbs of communication), whereas *will*'s default status emerges from the general high-frequency verbs it prefers; the verbs preferring *shall* are mostly rare verbs.

For the present purposes, it is most central to again point out the predictive power of these verb-specific preferences: as before with speaker-specific preferences, if one tries to predict every future choice in the corpus and does so just by choosing the construction that each verb prefers most, then

one can predict $^{2042}/_{2329}$ = 87.7% of all futures correctly just on the basis of verb-specific effects. That in turn means that, if a researcher finds differences in future use between varieties or topics, he can only be confident that these are in fact due to variety- or topic-specific effects if the more general confound of lexically specific effects is not responsible for the future choices.

Persistence/Priming

After two random-effect factors, the final important source of variability to be discussed here is different in nature. In the previous two sections, the idea was to discuss annotated factors that characterize a constructional choice in the data to see how, if at all, they were related to the constructional choice. In the language of spreadsheets, this means the column of some factor or independent variable/predictor was correlated with the column that represents the dependent variable/response, here the constructional choice. In the current section, we deal with the case where what might affect a constructional choice at time t_x is in the same column; namely, a previous choice at time $t_{y < x}$. In other words, the dependent variable is potentially correlated with (an earlier value of) itself, hence the term in statistics for this is *autocorrelation*.

As mentioned earlier, this phenomenon is referred to as structural priming and has been observed in a huge number of studies from production to production, from comprehension to production, in various experimental tasks (picture description, sentence completion, dialog tasks, etc.), in observational/corpus data, in many languages, and between languages. Overwhelmingly, a certain structural choice at some point of time increases the probability that the same speaker or another speaker who heard the previous structural choice will use the same construction the next time he makes a choice from the same set of alternants. That of course means that structural priming can often be orthogonal to other linguistic factors and, therefore, make it harder to determine how much of the variability in the data can be attributed to linguistic predictors describing the utterance currently under investigation and how much is just due to something that happened a minute ago and is, correspondingly, far away from the current concordance line.

Observational studies of structural priming have become quite sophisticated in the past few years (see Gries 2015a for an overview), but exploring priming can also be achieved more simply by, for instance, switch-rate plots proposed by Sankoff and Laberge (1978). Such plots plot the rates of switches from one of the alternants to the other against the relative frequency of the latter alternant per speaker; low switch rates are compatible with priming. Consider Figure 7.3, which represents the frequencies of *will*-futures on the *x*-axis, the switch rate toward *will* on the *y*-axis, and every letter is one speaker (with letters representing varieties: *N* for IN, *H* for PH, *K* for UK, and *S* for US). The dashed line is the null hypothesis that the

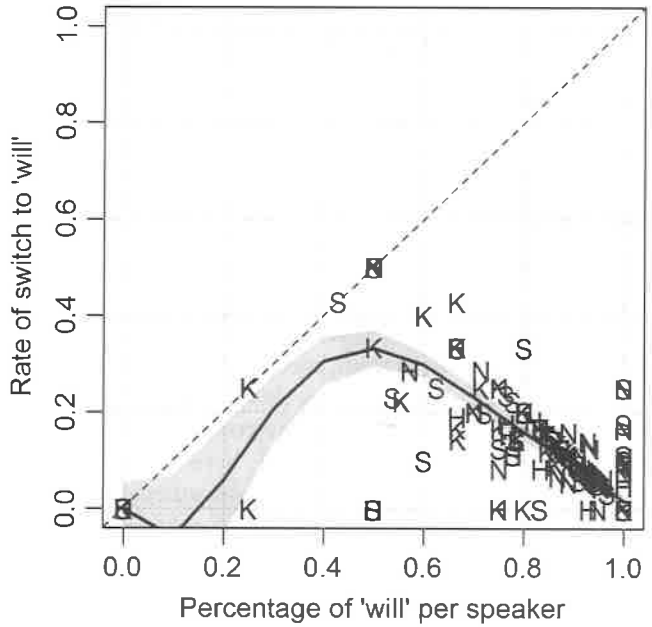

Figure 7.3 Switch-rate plot for *will-* futures

switch rate toward *will* is proportional to *will*'s frequency and the line with the confidence interval summarizes the points.

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

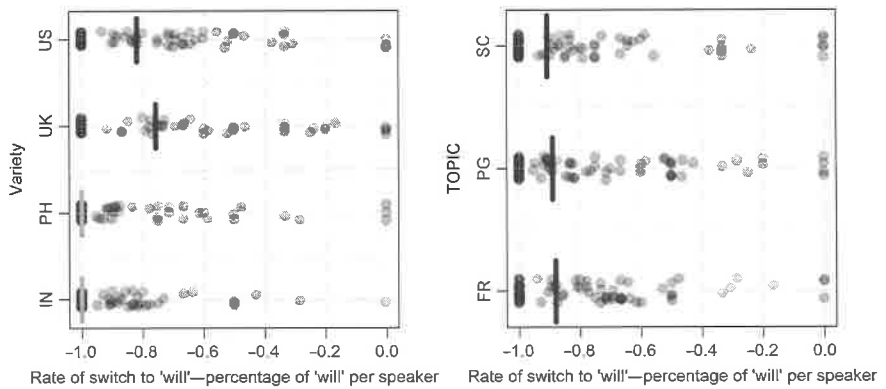

Figure 7.4 Switch rates to *will* minus percentages of use of *will* per file/speaker by variety (left) and by topic (right)

in speakers' minds.[4] 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.[5] 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 $^{1884}/_{2329}$ = 80.9% of all instances of *will*, *shall*, and *going to* correctly just on the basis of what the speaker did the 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) given 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 $p < 10–6$). Third, this approach cannot easily accommodate multiple kinds of random effects at the same time.

3 This extension uses exact binomial tests to test for each lexical item whether its occurrences with each of the constructions are more or less frequent than expected from the constructions' frequencies in the corpus and is implemented in Gries (2014), see <http://tinyurl.com/collostructions> for details and examples.

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

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

Estival, D. (1985). Syntactic priming of the passive in English. *Text*, 5(1–2), 7–21.

Gries, S. (2014). Coll.analysis 3.5. A script for R to compute perform collostructional analyses (major update to handle larger corpora/frequencies). Accessed online at: http://tinyurl.com/collostructions

Gries, S. (2015a). Структурный прайминг: корпусное исследование и узуальные/экземплярные подходы/'Structural priming: A perspective from observational data and usage-/exemplar-based approaches'. In Andrej A. Kibrik, Alexey D. Koshelev, Aexander V. Kravchenko, Julia V. Mazurova & Olga V. Fedorova (Eds.), *Язык и мысль: Современная когнитивная лингвистика/'Language and Thought: Contemporary Cognitive Linguistics'* (pp. 721–754). Moscow: Languages of Slavic Culture.

Gries, S. (2015b). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125.

Gries, S. & Deshors, S. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, 9(1), 109–136.

Gries, S. & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97–129.

Hasselgård, H. & Johansson, S. (2012). Learner corpora and contrastive interlanguage analysis. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (Eds.), *A Taste for Corpora: In Honour of Sylviane Granger* (pp. 33–61). Amsterdam & Philadelphia: John Benjamins.

Hyland, K. & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6(2), 183–205.

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61(2), 6478–6672.

Mukherjee, J. & Gries, S. (2009). Collostructional nativisation in New Englishes. *English World-Wide*, 30(1), 27–51.

Mukherjee, J. & Hoffmann, S. (2006). Describing verb-complementational profiles of New Englishes: A pilot study of Indian Englishes. *English World-Wide*, 27(2), 147–173.

Sankoff, D., & Laberge, S. (1978). Statistical dependence among successive occurrences of a variable in discourse. *Linguistic Variation: Methods and Models*, 119–126.

Schenkein, J. (1980). A taxonomy for repeating action sequences in natural conversation. In Brian Butterworth (Ed.), *Language Production* (Vol. 1, pp. 21–47). London & New York: Academic Press.

Schneider, E. (2007). *Postcolonial Englishes: Varieties Around the World*. Cambridge: Cambridge University Press.

Stefanowitsch, A. & Gries, S. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1(1), 113–150.

Szmrecsanyi, B. (2006). *Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin & New York: Mouton de Gruyter.

Weiner, E. & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, 19(1), 29–58.