31 New Technologies and Advances in Statistical Analysis in Recent Decades

STEFAN T. GRIES

Introduction

For most of the twentieth century, much of (theoretical) linguistics was predominantly generative in nature and with that theoretical orientation came a methodological predominance of judgments of acceptability or grammaticality: Speakers of a language would claim that a certain linguistic expression was acceptable/grammatical or not (in which case, that expression was starred) and, from that, linguists would infer some theoretical consequences. However, insightful critique (e.g., Labov, 1975) made very clear how problematic such judgment data could be, and one of the consequences was that firm commitment to binary grammaticality soon became softened and, without much empirical validation, sentences were then also prefixed with one or more question marks to indicate uncertainty or variability of their grammaticality, which at least introduced some recognition of gradience, or probability, into theoretical discussions. Usage-based linguistics (UBL), on the other hand, incorporated probability as a theoretical notion right from the start, making frequency not just an arbitrary performance phenomenon but a crucial component of the inner workings of mental grammar. This chapter will discuss why and how statistical methods have become increasingly more prominent in UBL studies. More specifically, the next section on the conceptual background discusses the "why" and outlines how the very nature of UBL—its commitment to certain kinds of data and its central notions and mechanisms—virtually requires a certain amount of statistical methodology. After that, the next section discusses examples of statistical applications we find in current UBL work in approximately ascending order of complexity: (1) frequencies of (co-)occurrence and association measures; (2) predictive modeling (e.g., regressions and other approaches); and (3) exploratory methods (e.g., cluster or correspondence analyses). Finally, each of these three areas is revisited with an eye to discussing necessary next steps that UBL could benefit from.

The Handbook of Usage-Based Linguistics, First Edition. Edited by Manuel Díaz-Campos and Sonia Balasch.

Background

If one adopts the current usage-based theoretical perspective on language/linguistics, one also, minimally implicitly, adopts a perspective that essentially has to be statistical in nature simply because most, if not all, of contemporary UBL involves a methodological commitment to study language and develop linguistic theory on the basis of naturally-occurring language use, i.e., corpora, which has of course been a methodological commitment shared in a variety of other research areas, such as (typological) discourse-functional linguistics (see work by Givón, 1979, 1992a, 1992b), variationist sociolinguistics (e.g., Cedergren & Sankoff, 1974), or general corpus linguistics (e.g., Biber, 1993; Leech & Fallon, 1992) even before linguists started using the term *usage-based* (which Bybee & Beckner, 2015, p. 953 attribute to Langacker, 1987). However, UBL also unavoidably involves statistical approaches for more general and more theoretical ways, namely via

1. a great many central concepts or notions of the theory that are probabilistic/statistical in nature;
2. models and/or stipulated mechanisms of a kind that ultimately require the use of statistical techniques for their study, validation, and confirmation.

To appreciate this connection of UBL to statistical methods, consider Bybee and Beckner's (2010) excellent overview of UBL, which reviews most essential notions/mechanisms of UBL (highlighted in bold below).

- They motivate UBL's assumption of **domain-general cognitive processes**, quoting "Elman and Bates (1997: 1180) writ[ing] that 'language evolved through **quantitative changes** in social perceptual, and cognitive abilities, including **statistical learning**" (p. 954, my emphasis).
- They discuss the role that **repetition** plays in UBL both in how it reinforces the **entrenchment** of units as well as their **chunking** into greater units (2010, p. 955). This, of course, relates the UBL statistically and corpus-linguistically to token/type frequency counts (per corpus (part)), arguably the simplest statistic to be computed from corpora:
  - token frequency per corpus is supposed to be causally related to entrenchment (see Baayen et al., 2016; Bybee & Thompson, 1997; Langacker, 1987; Schmid, 2010), which in turn is supposed to be causally related to matters of ease/speed of lexical access, age of acquisition, resistance to or acceleration of, grammatical change, and many other (psycho)linguistic effects (see Ellis, 2002);
  - token frequencies per corpus part involves the corpus-linguistic notion of dispersion, which quantifies how evenly distributed occurrences of something are in a corpus (see Gries, 2008, 2020), which is similarly relevant to learning and processing (see Adelman, et al., 2006; Ambridge, et al., 2006; Baayen, 2010; Gries, 2022b);
  - type frequency, by contrast, is supposed to be causally related to matters of productivity, acquisition, and grammaticalization (Bybee & Beckner, 2015, pp. 966–967).

- They speak about "**learning** when two (or more) events tend to co-occur, or when one event tends to predict another" (2010, p. 955), which establishes a clear connection to the following:
  - statistically, conditional probabilities and associative learning algorithms, but also to simplest cases of regression modeling, namely monofactorial statistics such as Pearson's *r* (the simplest case of linear regression modeling) or Chi-squared tests (relatable to the simplest case of generalized linear/logistic regression modeling); see Ellis (2006), especially p. 8: "Learners FIGURE language out: their task is, in essence, to learn the probability distribution P(interpretation | cue, context), the probability of an interpretation given a formal cue in a particular context, a mapping from form to meaning conditioned by context";
  - corpus-linguistically, measures of association that—typically, at least—quantify how much two linguistic elements, or a linguistic and an extra-linguistic element, tend to occur together.
- They discuss **categorization** (the process underlying category formation), which requires that the (human) categorizer compares the **similarity** of a to-be-categorized (linguistic) stimulus to previous stimuli on multiple dimensions to determine what category on which level of categorization a stimulus should be associated with. Naturally, this relates to **gradience** (the fact that, because similarity is a continuous metric, category boundaries are often difficult to distinguish). Statistically, this again involves metrics from simple conditional probabilities (as in measurements of cue validity in early prototype theory (e.g., Rosch, 1978) or within the Competition Model (e.g., Zhao & Fan, 2021)) to more complex tools like similarity-based methods (e.g., cluster analyses) or dimension-reduction methods (e.g., factor/principal component analyses).
- They argue that similarities among words (and, ultimately, constructions) are captured in **networks** (which are studied with statistical methods from network analysis).
- They claim that, because of how gradience and variability are built into UBL, it can also accommodate (synchronic and diachronic) **variation and change** (e.g., grammaticalization), which have often been statistically modeled with regression or other kinds of predictive modeling techniques (e.g., variable rules in variationist sociolinguistics).

A final important characteristic of UBL is its recognition that **adult grammars still change** (Bybee & Beckner, 2015, p. 976), which makes accounting for **individual variation** essential (see "sin 5" in Dąbrowska, 2016). The statistical consequence of, among other things, recognizing individual differences is the move toward more sophisticated regression modeling techniques, which help to separate, for instance, speaker-specific idiosyncrasies from more general trends likely to be characteristic of the wider population that is studied.

Thus, nearly all explanatory notions and mechanisms of UBL are strongly related with all kinds of statistical approaches, and maybe especially with statistical approaches that apply to, or take as input, corpus data from natural language usage. In the following sections, I will discuss selected and hopefully representative studies exemplifying how many of these notions and mechanisms are targeted with statistical methods.

Current approaches

Frequency of (co-)occurrence and association

As mentioned above, the two most basic statistical methods involve frequency. UBL studies interested in entrenchment often discuss token frequencies, i.e., how often something occurs in a corpus/text. These can be divided into absolute frequencies (i.e., raw counts) and relative frequencies (i.e., probabilities where raw counts are normalized against the size of a corpus (measuring context-free entrenchment) or against the frequency of a co-occurring element (measuring contextual entrenchment by expressing frequency as a probability). By contrast, studies interested in category formation and productivity often discuss type frequencies, i.e., how many different elements occur in a slot/with something else because, e.g., the larger the type frequency in a slot of a potential construction *C*, the more likely speakers are to form a more schematic category over instances of *C*. For example, Quochi (2016) looks at type and token frequencies of the radial-category family of Italian light-verb constructions and their L1 acquisition in CHILDES (MacWhinney, 1991). She explores ≈2100 instances of *fare* ("do") + noun constructions from children and adults in terms of the nouns/noun categories they occur with and the type-token-ratios of verb-related nouns. Tracking new types over time, she finds that *fare* + nouns derived from verbs by suffixation appear to be rote-learned rather than creatively produced. The time course of acquisition Quochi observes is one where children first pick up on the most frequent uses, then develop a more abstract schema, which becomes generalized to intransitive actions.

The overall importance attributed to frequency notwithstanding, the exact nature of the cognitive/psycholinguistic mechanisms is often hard to pin down, as are the best ways of measuring frequencies and their effects. For instance, De Vogelaer (2012) finds that standardization in the gender systems of different dialects of Dutch affects high-frequency items (with clearest results emerging from spoken data) whereas re-semanticization affects low-frequency items (with clearest results emerging from age-of-acquisition data and frequencies from acquisition corpora). Similarly, while high token frequencies have often been used to motivate phonological reduction (Bybee & Beckner 2015, pp. 964–966), Raymond and Brown (2012) show the picture is more complex: Their study of initial fricative reduction in a corpus of New Mexican Spanish controls for many contextual factors appearing to co-determine reduction and shows that, with such controls in place, "there was no influence on reduction in the complete dataset or the subsets tested of preceding phone frequency or *s*-word frequency" (p. 48) and that this "suggests that speakers are sensitive to how often a word occurs in environments that encourage reduction, but not measurably to non-contextual probabilistic measures of use" (p. 49). Thus, while frequency is a cornerstone of UBL, the time may have come to re-evaluate its primacy and how much it is a real cause or just correlated with causes.

The second basic statistical method involves the association of elements, i.e., the degree to which, typically, two elements are attracted to each other such as forms with other forms or forms with meanings (i.e., constructions), which is central to many aspects of associative learning. Within general corpus linguistics, studies mostly focus on collocation, i.e., the preferred co-occurrence of two words (e.g., *tea* going with *strong*, not with the near synonym *powerful*) and, in general corpus linguistics, this is usually studied using association measures (AMs) such as the log-likelihood value G^2, (pointwise) Mutual Information (*MI*), and/or the *t*-score (see, e.g., Evert, 2009; Gries, 2022a). In UBL, on the other hand, there is more work on the association of words and more schematic constructions; the family of methods called collostructional analysis (Gries & Stefanowitsch, 2004a, 2004b; Stefanowitsch & Gries, 2003, 2005) has shown how especially the semantic pole argument structure in constructions is reflected by the words strongly attracted to them. Three main methods are used:

- collexeme analysis: measuring how much words are attracted to, or repelled by, a syntactically defined slot in a construction (e.g., the verb slot in the ditransitive construction or the noun slot in the N-*waiting-to-happen* construction);
- (multiple) distinctive collexeme analysis: measuring which slot of two or more functionally similar constructions a word (dis)prefers to occur in (e.g., the verb slot in the two constructions making up the dative alternation);
- covarying collexeme analysis: measuring how much elements in two slots of one construction (dis)like to co-occur (e.g., the two verb slots in the *into*-causative, i.e., in V DO_{NP} *into* V-*ing*).

All three methods involve computing an AM—the *p*-value of a Fisher-Yates exact test or the log-likelihood value G^2 have been used most often—from 2×2 co-occurrence tables that provide the frequencies with which one element (as opposed to all others) co-occurs and does not co-occur with another element (or all others). For instance, for a collexeme analysis of the *as*-predicative, one might quantify the co-occurrence of *regard* with, or the occurrence of *regard* within, the *as*-predicative (V DO_{NP} *as* XP as in *The Minbari regarded the Shadows as a powerful enemy*) based on Table 31.1.

Later work showed that collostructional attraction was correlated with priming effects in corpora (Gries, 2005; Szmrecsanyi, 2006) and experiments (Gries & Wulff, 2009) but also with experimental results such as sentence-completion experiments and self-paced reading (Gries, Hampe, & Schönefeld, 2005, 2010), and acceptability judgments (Backus & Mos, 2011). More recent studies have extended the method and/or combined it with other kinds of data. Perek (2014), for instance, involves an extension of collexeme analysis of verbs occurring in the conative construction (e.g., *John kicked at Mary*). Based on fictional prose from the British National Corpus, he finds that even collexemes most strongly attracted to the conative exhibit a considerable range of verbs/verb classes while many collostructional studies of similar constructions resulted in semantically much more homogeneous verb classes; the best example is probably the strong representation of transfer-related verbs in the ditransitive (see Stefanowitsch & Gries, 2003). Based on Croft's (2002) insightful critique of the notion

Table 31.1 A hypothetical co-occurrence table for *regard* in the *as*-predicative

	as-*predicative*	*Not* as-*predicative*	*Sum*
regard	100	1400	1500
Not *regard*	900	149100	150000
Sum	1000	150500	151500

of postulating constructional polysemy when all/most that motivates that notion is the occurrence of different verbs in a construction, Perek then does separate collexeme analyses on "sub-constructions" of the conative as defined by classes of verb senses (e.g., of cutting, pulling, or striking) and finds that, once the resolution of the analysis is increased this way, the verbs preferred in the "sub-constructions" do indeed reflect their distinct notable semantic features.

Other recent applications and extensions include the following:

- Bernolet and Colleman (2016) demonstrate that all collostructional studies would benefit from taking polysemy more seriously than they have so far by showing that collostructional attraction should not be measured on the level of form alone, but on the level of form-sense pairings.
- Ellis et al. (2016) study verb-argument construction in native and learner language and correlate the results of collexeme analyses with many other data (e.g., frequencies and network analyses).
- Hoffmann et al. (2019) apply a co-varying collexeme analysis to the comparative correlative constructions (e.g., *the more, the merrier*) and show that the two slots of the construction usually contain the same kind of grammatical/syntactic material, indicating that one's account of the construction should not attempt to treat the construction's slots as independent.
- Flach (2020) revisits Gries, Hampe, and Schönefeld's question—what predicts experimental results better, frequency or association?—with data on *gonna/wanna/gotta* contraction and shows that measures of contingency/association consistently outperform mere co-occurrence frequency.

Thus, while AMs are not complex, they are nonetheless instructive as a first proxy of association/contingency relations within the speech community studied (with the corpus or in the experiment) they provide; proposals on how to improve such methods will be made below. However, the above is not to imply that all studies of co-occurrence in UBL are collostructional in nature or involve AMs. For example, Huang, Wible, and Ko (2012) study how differences in transitional probability make the last word of a phrase (e.g., *fact*) faster to read when it is part of a multi-word expression (e.g., *as a matter of fact*) or not (e.g., *whether this is a fact*). L1 and L2 speakers of English were presented with multi-word expressions and other phrases ending in the same word, and Huang et al. used eye-tracking to measure fixation probabilities, first-fixation durations, and gaze durations. They report that (the more predictable) words in multi-word expressions have significantly lower fixation probabilities and shorter first-fixation as well as gaze durations; the results of a second, follow-up experiment show that training changes the results for the L2 learner by making the final word of a multi-word expression more predictable.

Predictive modeling

As mentioned in the preceding section, frequencies and AMs are useful, but only as a first step, if only for the fact that they are monofactorial. On their own, they do not include other predictors/determinants of whatever linguistic choice one is interested in, which means they cannot shed too much light on most complex cognitive/psycholinguistic phenomena, in particular variation and change such as on why speakers make certain choices and how those might change in the short term (maybe due to priming) and in the long term (maybe due to grammaticalization). For this, UBL, like all other areas in linguistics, requires methods that can accommodate multifactorial relationships—potentially many causes affecting, or at least correlating with, usually one outcome. Thus, many UBL practitioners have turned to predictive modeling, particularly regression modeling. Such studies nearly always involve supervised learning: a predictive modeling technique tries to "learn" from some data set which predictors explain (most of) the variability of some response (which is typically binary, sometimes numeric, and (too) rarely ordinal or categorical). Examples of the former include studies of alternation phenomena on many levels of linguistic analysis, such as morphemic, lexical, or syntactic alternations. Two of the earliest examples in cognitive/usage-based linguistics are Gries (2003a, 2003b), studying the alternations of particle placement (*Riker gave back the phaser* vs. *Riker gave the phaser back*) and the dative alternation (*Riker gave Picard the phaser* vs. *Riker gave the phaser to Picard*) on the basis of concordances that were annotated for around 20 predictors from many levels of linguistic analysis (phonology, morphology, syntax, semantics, and discourse pragmatics). In each case, he applied a linear discriminant analysis to determine (1) which factors are most predictive for the constructional choices speakers make and (2) what, therefore, appears to be the most prototypical instantiations of these constructions.

Later work on such variability in alternations typically uses binary logistic regression modeling and is, thus, statistically more advanced, but otherwise similar in spirit to this earlier work. Sokolova et al. (2012), for instance, study the Russian locative alternation in the 98m-word Russian National Corpus. Using a version of the Behavioral Profiling approach, they annotate 1,920 examples of the locative alternation with *gruzit'* "load" (non-passives and passive participles) for (1) the presence or absence of three prefixes (none vs. *na*, *za*, or *po*); (2) whether the construction is used in a reduced form or not; and (3) whether the verb is used in the participial form or not. Model selection indicates that the three predictors and one of their pairwise interactions are significantly and predictively correlated with the choice of construction in the locative alternation. One particularly interesting aspect of this analysis is the strong effect of the first predictor, given how its effect goes against the often-made assumption that the prefixes are semantically empty. If they were, how could they have so much predictive power?

A similar example is De Vaere et al. (2021), who study German *geben* ("give") in 1,301 occurrences of two alternating ditransitive constructions in the German Reference Corpus (DeReKo), which were annotated for 20 morphosyntactic, semantic, and pragmatic factors. A logistic regression model that, laudably, includes curvature for numeric predictors (to avoid the often-implicit assumption of straight-line effects only), protection against overfitting with penalization, and a bias-corrected C-index shows that the main meaning of *geben* is not so much "literal transfer from one person to another" (as in *give* or *hand*) but a more general "transfer" meaning.

Such approaches have also been applied to research on diachronic change. Shank et al. (2014) study the realization vs. omission of *that* after *I think* in a stratified sample of ≈5,800 instances in corpus data spanning the time period from 1560 to 2012. They annotated those instances for 26 predictors involving features of the corpus (file) as well as features regarding the matrix and the complement clause; the clause-based features involved, among others, person, tense, polarity as well as the length of material between the two clauses. A stepwise regression analysis revealed a variety of effects, in particular some interactions involving the predictor TimePeriod. For instance, over time, *that* realization became less likely in spoken, but more likely in written data. Similarly, the effect of the length of the complement subject or the harmony of polarity between matrix and complement clause are not constant across time.

Given the above-mentioned recognition of the theoretical importance of individual variation and its statistical corollaries, most such studies now involve generalized linear mixed-effects modeling, a kind of modeling that can account for (1) the repeated-measurements structure of most experimental and observational data (where we collect multiple data points from each subject/speaker and/or for each stimulus/lexical item) and (2) the hierarchical nature of corpus data (where multiple uses of one speaker are nested into one file, which is nested into one register, which is nested into one mode, etc.), which can help disentangle speaker or register-specific effects from overall effects. A UBL study of learner language that involves an advanced version of this is Lester (2019), who studied the realization/omission of *that* as a relativizer (e.g., *Bester hated the way that/- telepaths were treated*). 800 relative clauses with/without *that* (40% from native speakers, 60% from German and Spanish learners) were retrieved from two corpora and annotated for 13 variables (including what would normally be the response variable, i.e., *that*-realization) including task type, semantic predictors, structural/complexity predictors, priming and disfluencies, etc. He then fitted a generalized additive mixed model (GAMM, a model that can handle speaker-specific idiosyncrasies, but also curvature of numeric predictors) on the native speaker data, cross-validated it (with a bootstrap, see Egbert & Plonsky, 2020), applied it to the learner data, and then computed how much the actual learner choices deviate from the imputed NS choices, which became the response variable in a second GAMM. That model resulted in several significant linear and non-linear predictors. For example, all learners overused *that* for subject, predicate-nominal, and direct-object roles of the relative-clause heads, but the Spanish learners performed more nativelike than the German ones, and self-priming effects differed between the German and the Spanish learners. More generally, the data did not support the study's initial expectation that learners would follow the same processing-based strategy (of producing *that* in complex contexts). Instead, learners under-produced *that* in structurally complex contexts and when production was difficult. This study is a nice example of how applying more advanced statistical methods to offline observational data can still shed light even on the interplay of domain-general processing characteristics, linguistic predictors, and group (learners vs. native speakers) as well as individual speaker differences.

In the more recent past, a few first studies adopted predictive modeling methods from the domain of machine learning, in particular, classification trees and random forests, which, while completely different in their conceptual underpinnings (see Efron, 2020, or, in a more linguistic setting, Chambaz & Desagulier, 2016), are often applicable even to data that pose problems for regression analysis (e.g., by being skewed/imbalanced or violating the assumptions of regression models) and which often trump regression analyses in terms of predictive power. One recent application is Fonteyn and Nini (2020), a study of whether gerunds are used with *of* (e.g., *eating of* *meat*) or not (*eating meat*) based on ≈14,000 instances from the EMMA (Early Modern Multiloquent Authors) corpus that was annotated for the response variable (whether of is used or not) and for three predictors:

- determiner use: bare/none vs. possessive (e.g., *their eating of that bread*), *the* (e.g., *the stopping of the play*), *a* (e.g., *a fulfilling of rites*), demonstrative (e.g., *this introducing of God's name*), and quantification (e.g., *every mentioning of the name*);
- function of the gerund: subject vs. object vs. subject complement, and three groups of prepositions: frequent ones, temporal, and other;
- verb type: lexical vs. light vs. possessive *have*.

In addition, they added the speaker producing the sentence (as a kind of random effect with 19 levels) as well as their age (in years) and generation (as a ternary factor) and the genre of the text in which the gerund appeared (18 genre labels). A conditional inference forest and a conditional inference tree indicate that the language-internal predictor of determiner is by far the most important one (especially with its levels *bare/no determiner* and *the*), and that is true across nearly all individual speakers, but less important predictors vary a lot more between speakers, which "challenge[s] the common belief that all constraints on grammatical variation are shared by all individuals in a community" (p. 302) and, thus, again supports the UBL's assumption of the importance of individual variation.

Another study, which combined regression modeling and random forests (and corpus and experimental data), is Azazil's (2020) study of frequency effects in the L2 acquisition of the catenative verb construction (e.g., *He enjoys smoking / to smoke*) by German learners of English. A sentence-completion experiment with advanced learners aimed at determining whether learners' sentence completions were correlated with (1) the frequency of the matrix verb (above, *enjoy*), (2) the frequency of the matrix verb in the construction, and (3) the proportion of uses of the matrix verb in the construction relative to all its uses (i.e., a collostructional kind of question); matrix verbs had either a *to-* or *-ing* construction bias. To single out just one result from Experiment 1, a random forest shows that the third collostructional kind of measure had the highest variable importance when it came to predicting a variable representing the target-likeness of the learners' completions. A later mixed-effects model "confirmed" the higher importance of that predictor compared to the more general/less context-bound frequency measures.

In sum, the kinds of predictive modeling methods summarized here offer UBL practitioners techniques that kill many birds with one stone: they can tackle many kinds of variation, change, and acquisition questions by including domain-general predictors, frequency effects, linguistic and other contextual predictors while still controlling for individual or lexical differences. This is an area of active research and lively discussion of the advantages and disadvantages of different methods. For instance, alternation studies such as Baayen (2011) and Baayen et al. (2013) compare a variety of different classifiers (the above kinds of regression and tree-based models, but also memory-based learning, naïve discriminative learning, and support vector machines) in terms of their predictive power and their potential "mappability" on actual cognitive processes.

Exploratory tools

The final major strand in statistical applications involves exploratory methods, i.e., methods that are not hypothesis-testing, but hypothesis-generating, in nature; the most prominent examples of such methods in UBL studies are cluster, factor/principal components, and correspondence analyses. While different in nature, these kinds of methods can help researchers see (similarity-based) structures in often highly multi-dimensional and noisy data sets to, for example, decide how many senses of a lexical item to distinguish, how many subsenses of a construction to postulate, what linguistically meaningfully different temporal stages to distinguish in a diachronic study, etc.

An example of two kinds of cluster-analytic approaches is the pair of papers by Divjak and Gries (2006, 2008). In the former, they reported on the results of a behavioral profile analysis of ≈1,600 sentences featuring nine Russian verbs meaning "to try," which were annotated for 87 morphological, syntactic, and semantic features. Then, the frequency distributions of the features were analyzed with a hierarchical cluster analysis yielding three groups of near synonyms, which were interpreted based on the differences between and within clusters. The between-cluster differences can be summarized as follows (from Divjak & Gries, 2008, p. 193f.):

- a human is exhorted to undertake an attempt to move himself or others (rather than to undertake mental activities); often, these activities are negated;
- an inanimate subject undertakes repeated non-intense attempts to exercise physical motion; the actions are often uncontrollable and fail;
- an inanimate subject (concrete or abstract) attempts very intensely but in vain to perform what typically is a metaphorical extension of a physical action.

To validate these corpus findings, Divjak and Gries (2010) then analyzed the outcome of several sorting experiments with native speakers of Russian who sorted nine sentences that only differed in their verb meaning "to try" into groups based on their overall semantic similarity. Then, a score was computed to quantify the fit between the cluster analysis of the observational/corpus data and that of the experimental sorting data, which was then compared to the range of scores one might obtain from a null hypothesis distribution. The results show that the speakers' experimental sortings are very compatible with the corpus-based results; similarly supportive results were obtained from a comparison of the corpus-based clustering to an identical cluster analysis of the experimental data and a gap-filling task. This study is methodologically interesting in how cluster analyses from observational and experimental data are compared and evaluated.

One application of a correspondence analysis is by Delorge, Plevoets, and Colleman (2014). They studied the corpus frequencies with which dispossession verbs with *ont-* "away" occur in a variety of possessional transfer constructions. In a synchronic analysis, they found that verbs fall into clusters based on the constructions they (do not) "like" to occur in, with these clusters exhibiting clear patterns in terms of their semantics and their lexical profile. In an additional diachronic analysis, they found constructional specialization, showing how certain patterns solidify with time.

This area, too, is currently abuzz with new developments. For instance, there are now first applications of network analysis that directly target the kinds of network structures postulated in UBL (see Bybee & Beckner, 2015, pp. 959–961 and passim). A case in point is Ellis, Römer, and O'Donnell (2016), who developed semantic networks for verb-argument constructions (e.g., the V *about* N construction, the V *across* N construction, etc.) and, among other things, used a community-detection algorithm to identify semantically related coherent groups of verbs in these constructions. The results shed light on central UBL issues, such as the polysemy of constructions and their prototypical members. Similarly, Chen (2022) studied the network structure of ≈26,000 instances of Mandarin Chinese space particles in the constructional schema *zai* + NP + space particle in the 10m-word part of speech (PoS) tagged Sinica corpus.

Pending issues

Each of the above areas comes with a variety of issues that can be better addressed; I will discuss them in the same order as above.

Frequency of (co-)occurrence and association

While statistical methods, by definition, require frequencies as input, frequency need not be the seemingly all-powerful predictor that much work in usage-based linguistic has made it out to be. True, frequency data are: (1) relatively easy to obtain; (2) moderately to highly correlated with performance on many cognitive tasks such as acquisition, ease, and speed of processing; and (3) straightforwardly integratable into models/theories of the mental lexicon/construction. Yet, the fact that frequency of (co-)occurrence is highly correlated with cognitive tasks does not prove that it is also *causally* related to them. There is a small but growing body of literature reporting empirical findings that, minimally, undercut the centrality of frequency as a repetition counter and that indicates that other factors—recency (in its short-term form of priming and its long-term form of corpus dispersion), association/contingency, salience, context(ual) variability/distinctiveness)—are just as straightforward to integrate into our models/theories of the mental lexicon/constructicon and language production/comprehension but may be more powerful than frequency for various questions studied in UBL.

One such, much underappreciated study is that of McDonald and Shillcock (2001), who discussed many dimensions of lexical variation—frequency of occurrence, concreteness, context availability, age of acquisition, ambiguity—and their correlation with response time latencies, but, more importantly, then proposed a new dimension of lexical variation that is correlated with many of the above-mentioned ones but also contains additional information about words' lexical context. Their *contextual distinctiveness* (CD) "measures the amount of information conveyed by a word about its contexts of use" (p. 303) and is "derived from the distribution of words co-occurring with the word of interest, whereas Word Frequency (WF) is measured independently of this distribution" (p. 307). CD is correlated with observed log-transformed word frequency (r = -0.82), but its computation does not involve it directly because it is based on co-occurrence percentages. They show that CD accounted (marginally significantly) for variance in reaction times in a lexical decision task, even when word frequency and length were statistically controlled for ($r_{\text{part}} = 0.2$), whereas frequency did not when word length and CD were statistically controlled for ($r_{\text{part}} = -0.03$). They concluded "[w]ords that appear in relatively constrained (or distinctive) linguistic contexts have high [CD] scores and tend to attract longer lexical decision latencies" (p. 312). Other studies, such as Adelman, Brown, and Quesada (2006), Baayen (2010), and Gries (2010) all show that another factor—the dispersion of words in a corpus—can also have a higher degree of predictive power than the usual frequency measures. Yet other findings indicate that the frequency effect is also less monolithic than has been thought. See, for example, the works of Balota et al. (2001), Brysbaert, Mandera, and Keulers (2018:47), Diependaele, Lemhöfer, and Brysbaert (2013) and Rayner et al. (2006). Future work on the exact nature of the effect of frequency is therefore sorely needed; see Gries (2019a, Chapter 2, for more discussion).

When it comes to association/contingency, important future questions involve the degree to which association measures should conflate over various kinds of information. The most widely used measures/applications:

1. conflate frequency and association, which, depending on one's purpose, can be a feature or a bug. Measures such as p_{FYE}, G^2, t, or Dice react more to higher co-occurrence frequencies of the elements in question than their association (Gries, 2022a), which makes them easier to sort for quick heuristics, but harder to deal with for proper psycholinguistic purposes;
2. do not distinguish the direction of association. For instance, does a verb *V* attract a construction *C*, does *C* attract *V*, or do both attract each other? (See Gries, 2019b, for discussion.)
3. do not pay attention to corpus dispersion at all and, therefore, run the risk of inflating association scores for words that are severely underdispersed (e.g., because they might be specific to certain limited topics and/or registers); and
4. do not consider homonymy/polysemy much (see again Bernolet & Colleman, 2016).

To arrive at a more nuanced measurement of association (and one that is orthogonal to other factors) and to be able to integrate it into a theory (including a better understanding of the role of frequency) will require much additional empirical work involving both corpus and experimental data.

Predictive modeling

With regard to predictive modeling, much of what is needed is already underway with the field's slow move to mixed-effects modeling: We need techniques powerful enough to find every little bit of probabilistic structure in the data (given that humans are obviously so very good at that). This includes: (1) speaker- and lexical-specific effects to include the possibility of individual and lexical variation wherever possible; (2) more interactions (to make sure we can see when the effect of one predictor is contingent on another; and (3) more widespread attempts to deal with curvature in all studies involving cognitive effects, whose effects that are not necessarily captured with a straight line. While we often make the effects of predictors at least resemble a straight line with some transformation (e.g., logging for frequency effects), ultimately, more powerful/versatile methods would be more useful. Within a regression framework, these might include polynomials, (adaptive) splines, or generalized additive (mixed) models; in contexts where prediction is important, some machine-learning methods might also be appropriate (especially if we can then also interpret the effects they find).

In addition, it would be advantageous for UBL practitioners to be open to other statistical or machine-learning classifiers, including, but not limited to: (1) adaptive boosting, a classifier that has often been shown to outperform competing algorithms (see Hastie, et al., 2009, Chapter 10); (2) random forests enriched with elements from mixed-effects modeling such as speaker-specific effects (e.g., Fokkema et al., 2018); and (3) causal modeling, to move from mere correlational to causal reasoning (e.g., Larsson et al., 2021). All of these are promising and can move our understanding of phenomena to "the next level."

Exploratory tools

The main development to advance in this area is probably a push for using methods from the quickly evolving field of vector-space methods, both more traditional methods and more recent developments like GloVe, word2vec, BERT, and others. These methods are similar in that they are based on huge matrices with weighted co-occurrence information, e.g., co-occurrence frequencies of all pairs of words (within some context window such as *n* words, a sentence, or a complete document, in which the matrix is often referred to as a term-document matrix). Various kinds of transformations and dimension-reduction methods can then be applied to these matrices to represent the contexts, and, thus, the meanings/functions, of words using numeric vectors, which can then be compared to each other to assess semantic similarity.

Such methods are mathematically complex and often require data sets with sizes that are only available for high-resource languages (i.e., Indo-European languages), but, for those, they can shed light on many aspects of interest for UBL, including semantic similarity and categorization or network effects. As just one example of a more traditional vector-space semantic analysis, consider Perek and Hilpert's (2017) tweaking of Gries and Hilpert's Variability-based Neighbor Clustering (Gries & Hilpert, 2008) to work with vector-space representations to study the diachronic development of constructions (such as V *the hell out of* NP construction and the V POSS *way* PP construction) in data too big and noisy for manual analysis. For the former, new construction, their 1930s–2000s data from the Corpus of Historical American English reveal a slow and gradual expansion; for the latter, the data are noisier but are interpreted as a three-time-periods solution, with each period featuring somewhat distinctive verbs in the *way*-construction.

Uncertainty in the data

A final important issue that needs to be considered is the degree of uncertainty that accompanies many corpus-based measures that are not explored or quantified. Typically, corpus statistics are based on a researcher's sample, and the results are interpreted based on the quantitative findings. However, while dispersion measures provide some degree of "uncertainty" that a frequency or AM comes with in a corpus, the uncertainty based on the corpus sampling is usually left unexplored. This is less trivial than it sounds because a lot of times an argumentation/analysis rests on something being more or less frequent (in some context) than something else, or on some set of categories exhibiting a certain rank ordering of frequency, association, etc. However, if one's results are volatile (in the sense of "being extremely dependent on the exact sampling of the corpus"), researchers should hedge the strength of their conclusions, but it is this quantification of volatility that is usually missing.

To exemplify this, consider Figure 31.1, which plots gray words for some word types in the ICE-GB with their frequencies (on the *x*-axis) and their dispersions in the corpus (on the *y*-axis, high and low values mean clumpy and even dispersion respectively). It is understandable that some analysis might invoke the frequency differences among the different forms of GIVE or the difference between *coming* and *comes*. However, as we can see from the 95% data ellipses for both frequency and dispersion (derived from 1,000 bootstrapping samples of the corpus files), the frequencies and dispersions of *gives*, *giving*, and *gave* or of *comes* and *coming* actually overlap, so even if the observed ranking of these inflectional forms was 100% compatible with the theoretical predictions, these data might not constitute strong support for the analysis. Had the corpus been different, the results would change.

This is even more relevant to association data in collostructional studies of verb-argument constructions, constructional acquisition, grammaticalization, etc. Figure 31.2 plots verb types attested in the ditransitive in the ICE-GB with their frequencies (on the *x*-axis) and their association to the ditransitive (on the *y*-axis). If a researcher uses a measure of association that reflects frequency of occurrence, then the forms of *tell* "win out" over those of *give*, but more importantly in the present contexts are the ellipses. Some of the ellipses are quite big, indicating that the results from the corpus as a whole come with high degrees of uncertainty, but it is also striking that all ellipses are taller than they are wider, meaning that association data are much more volatile than frequency data. This is an observation that I have not seen discussed and that emphasizes the care we must exercise with our corpus-based co-occurrence measures.

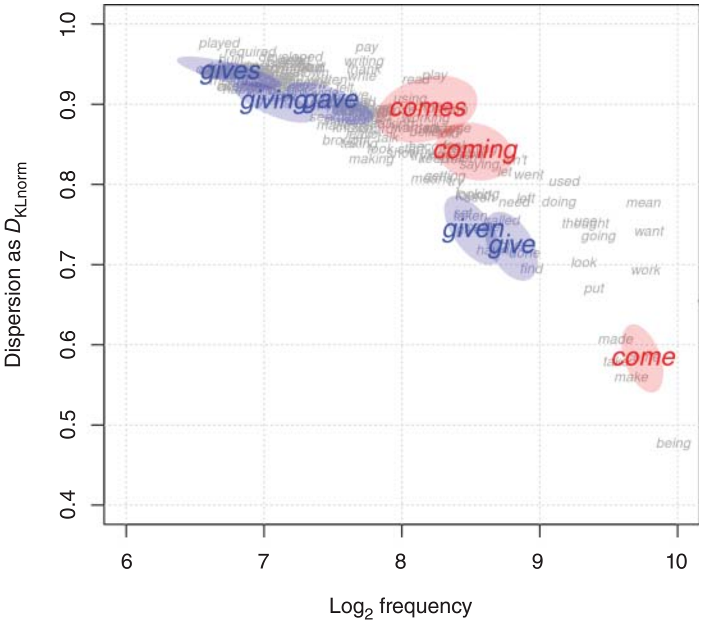

Figure 31.1 Frequency and dispersion (bootstrapped).

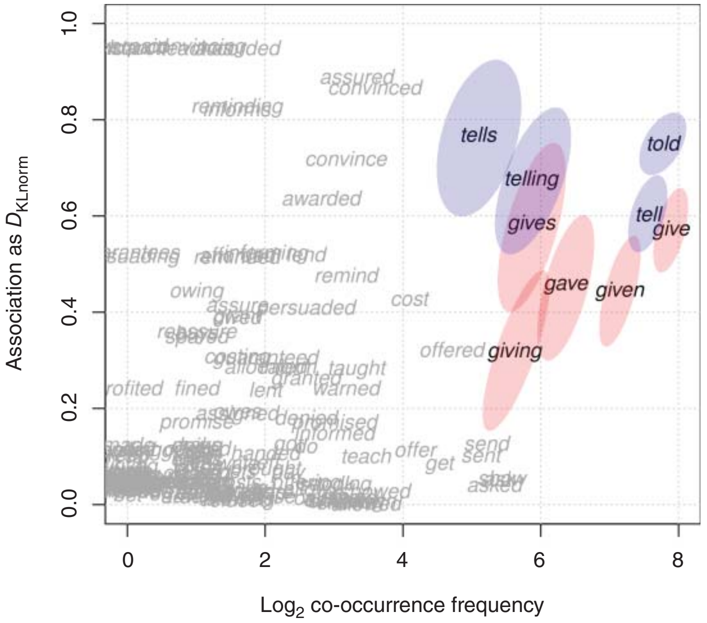

Figure 31.2 Frequency and association (bootstrapped).

Final remarks

Pending issues notwithstanding, it seems clear that UBL is undergoing a massive, but very positive development with regard to when and how statistical methods are used. Within the relatively short span of 15–20 years, the field has evolved from one that used frequency as a theoretical notion but did not tackle it with the correspondingly required methods, to one with rigorous quantitative analysis. Naturally, we all have to learn more, but the amount of progress is staggering and bodes well for the empirical rigor and, hopefully, the findings resulting from it.

REFERENCES

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Ambridge, B., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2)., 174–193.

Azazil, L. (2020). Frequency effects in the L2 acquisition of the catenative verb construction—evidence from experimental and corpus data. *Cognitive Linguistics*, *31*(3), 417–451.

Baayen, R. H. (2010). Demythologizing the word frequency effect: a discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian*

Journal of Applied Linguistics, *11*(2), 295–328.

Baayen, R. H., Endresen, A., Janda, L. A., A Makarova, A., & Nesset, T. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, *37*(3), 253–291.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220.

Backus, A., & Mos, M. B. J. (2011). Islands of (im)productivity in corpus data and acceptability judgments: Constructing two potentiality constructions in Dutch. In D. Schonefeld (Ed.), *Converging evidence: Methodological and theoretical issues for linguistic research* (pp. 165–192). Amsterdam: Benjamins.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition*, *20*(4), 639–647.

Bernolet, S., & Colleman, T. (2016). Sense-based and lexeme-based alternation biases in the Dutch dative alternation. In J. Yoon & S. T. Gries (Eds.), *Corpus-based approaches to Construction Grammar* (pp. 165–198). Amsterdam: John Benjamins.

Biber, D. (1993). Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, *19*(3), 531–538.

Brysbaert, M., Mandera, P., & Keulers, E. (2018). The word frequency effect in word processing: an updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Bybee, J. & Beckner, C. (2010). Usage-based theory. In H. Narrog & B. Heine (Eds.), *Handbook of linguistic analysis* (pp. 827–855). Oxford: Oxford University Press.

Bybee, J., & Thompson, S. A. (1997). Three frequency effects in syntax. *Berkeley Linguistics Society*, *23*, 65–85.

Cedergren, H. J., & Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, *50*(2), 333–355.

Chambaz, A., & Desagulier, G. (2016). Predicting is not explaining: Targeted learning of the dative alternation. *Journal of Causal Inference*, *4*(1), 1–30.

Chen, Alvin Cheng-Hsien. (2022). Words, constructions and corpora: Network representations of constructional semantics for Mandarin space particles. Corpus Linguistics and Linguistic Theory *18*(2). 209–235. https://doi.org/10.1515/cllt-2020-0012.

Croft, W. (2002). *Typology and universals*. Cambridge: Cambridge University Press.

Dąbrowska, E. (2016). Cognitive Linguistics' seven deadly sins. *Cognitive Linguistics*, *27*(4), 479–491.

Delorge, M., Plevoets, K., & Colleman, T. (2014). Competing 'transfer' constructions in Dutch: the case of *ont*-verbs. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 39–60). Amsterdam: John Benjamins.

De Vaere, Hilde, De Cuypere, Ludovic and Willems, Klaas. (2021). Alternating constructions with ditransitive geben in present-day German. Corpus Linguistics and Linguistic Theory *17*(1). 73–107.

De Vogelaer, G. (2012). Frequency, conservative gender systems, and the language-learning child: Changing systems of pronominal reference in Dutch. In S. T. Gries & D. S. Divjak (Eds.), *Frequency effects in language learning and processing* (pp.109–144). Berlin: De Gruyter.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863.

Divjak, D. S., & Gries, S. T. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, *2*(1), 23–60.

Divjak, D. S. & Gries, S. T. (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, *3*(2), 188–213.

Divjak, D. S., & Gries, S. T. (2009) Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In K. Dziwirek & B. Lewandowska-Tomaszczyk (Eds.), *Studies in cognitive corpus linguistics* (pp. 273–296). Frankfurt am Main: Peter Lang.

Efron, B. (2020. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115*(530), 636–655.

Egbert, J., & Plonsky, L. (2020). Bootstrapping techniques. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 592–610). Berlin: Springer.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of Construction Grammar. In *Language learning*, *66*(Suppl. 1, Language Learning Monograph Series). New York: John Wiley.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (vol. 2, pp. 1212–1248). Berlin: de Gruyter.

Flach, S. (2020). Schemas and the frequency/acceptability mismatch: Corpus distribution predicts sentence judgments. *Cognitive Linguistics*, *31*(4), 609–645.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034.

Fonteyn, L., & Nini, A. (2020). Individuality in syntactic variation: An investigation of the seventeenth-century gerund alternation. *Cognitive Linguistics*, *31*(2), 279–308.

Givón, T. (1979). *On understanding grammar*. Orlando, FL: Academic Press.

Givón, T. (1992a). The grammar of referential coherence as mental processing instructions *Linguistics*, *30*(1), 5–55.

Givón, T. (1992b). On interpreting text-distributional correlations: Some methodological issues. In D. Payne (Ed.), *Pragmatics of word order flexibility* (pp. 305–310). Amsterdam: John Benjamins.

Gries, S. T. (2003a). *Multifactorial analysis in corpus linguistics: A study of particle placement*. London: Continuum.

Gries, S. T. (2003b). Grammatical variation in English: A question of 'structure vs. function'? In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (pp. 155–173). Berlin: Mouton de Gruyter.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437.

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: Current studies, new directions* (pp. 197–212). Amsterdam: Rodopi.

Gries, S. T. (2012). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*, *18*(1), 137–165.

Gries, S. T. (2019a). *Ten lectures on corpus-linguistic approaches: Applications for usage-based and psycholinguistic research*. Leiden: Brill.

Gries, S. T. (2019b). 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, *24*(3), 385–412.

Gries, S. T. (2020). Corpus linguistics: Quantitative methods. In C. A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 340–344). Oxford: Wiley-Blackwell.

Gries, S. T. (2022a). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*.

Gries, S. T. (2022b). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies.*

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*(4), 635–676.

Gries, S. T., Hampe, B., & Schönefeld, D. (2010). Converging evidence II: More on the association of verbs and constructions. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 59–72). Stanford, CA: CSLI.

Gries, S. T., & Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora*, *3*(1), 59–81.

Gries, S. T., & Stefanowitsch, A. (2004a). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Gries, S. T., & Stefanowitsch, A. (2004b). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–236). Stanford, CA: CSLI.

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*. 163–186.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin: Springer.

Hoffmann, T., Horsch, J., & Brunner, T. (2019). The more data, the better: A usage-based account of the English comparative correlative construction. *Cognitive Linguistics*, *30*(1): 1–36. doi: 10.1515/cog-2018-0036.

Huang, P., Wible, D., & Ko, H. (2012). Frequency effects and transitional probabilities in L1 and L2 speakers' processing of multiword expressions. In S. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (vol. 1, pp. 145–176). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110274059.145.

Labov, W. (1975). Empirical foundations of linguistic theory. In R. Austerlitz (Ed.), *The scope of American linguistics* (pp. 77–133). Lisse: The Peter de Ridder Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar*, vol. I: *Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Larsson, T., Plonsky, L., & Hancock, G. R. (2021). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, *17*(3), 683–714.

Leech, G. N., & Fallon, R. (1992). Computer corpora—What do they tell us about culture? *ICAME Journal*, *16*, 29–50.

Lester, N. A. (2019). That's hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research*, *5*(1), 1–32.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295–322. https://doi.org/10.1177/00238309010440030101

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

Perek, F. (2014). Rethinking constructional polysemy: the case of the English conative construction. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy* (pp. 61–85). Amsterdam: John Benjamins.

Perek, F., & Hilpert, M. (2017). A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics*, *22*(4), 490–520.

Quochi, V. (2016). Development and representation of Italian light-fare constructions. In J. Yoon & S. T. Gries (Eds.), *Corpus-based approaches to construction grammar* (pp. 39–64). Amsterdam: John Benjamins.

Raymond, W. D., & Brown, E. L. (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In S. T. Gries & D. S. Divjak (Eds.), *Frequency effects in language learning and processing* (pp. 35–52). Berlin: De Gruyter.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmid, H.-J. (2010). Does frequency in the text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 101–133). Berlin: Mouton de Gruyter.

Shank, C., Plevoets, K., & Cuyckens, H. (2014). A diachronic corpus-based multivariate analysis of "I think that" vs. "I think zero". In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 279–303). Amsterdam: John Benjamins.

Sokolova, S., Lyashevskaya, O., & Janda, L. A. (2012). The locative alternation and the Russian 'empty' prefixes: A case study of the verb *gruzit* 'load'. In D. S. Divjak & S. T. Gries (Eds.), *Frequencyt effects in language representation* (pp. 51–85). Berlin: Mouton de Gruyter.

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

Stefanowitsch, A., & Gries, S. T. (2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, *1*(1), 1–43.

Szmrecsanyi, B. (2006). *Morphosyntactic persistence in spoken English*. Berlin: Mouton de Gruyter.

Zhao, H., & Fan, J. (2021). Modeling input factors in second language acquisition of the English article construction. *Frontiers in Psychology*, *12*(653258).