LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

CHAPTER 10
Statistical Measures for Usage-Based Linguistics

Stefan Th. Gries and Nick C. Ellis
University of California, Santa Barbara and University of Michigan, Ann Arbor

The advent of usage-/exemplar-based approaches has resulted in a major change in the theoretical landscape of linguistics, but also in the range of methodologies that are brought to bear on the study of language acquisition/learning, structure, and use. In particular, methods from corpus linguistics are now frequently used to study distributional characteristics of linguistics units and what they reveal about cognitive and psycholinguistic processes. This paper surveys a range of psycholinguistic notions that are becoming ever more important in theoretical and cognitive linguistics—for example, frequency, entrenchment, dispersion, contingency, surprisal, Zipfian distributions—and current corpus-linguistic approaches toward exploring these notions and their roles for linguistic cognition.

Keywords corpus data; psycholinguistics; associative learning; frequency; dispersion; contingency/association; surprisal

Usage-Based Approaches: Psycholinguistics and Corpus Analysis

Usage-based approaches see language as a large repertoire of symbolic constructions. These are form–meaning mappings that relate particular patterns of lexical, morphological, syntactic and/or prosodic form with particular semantic, pragmatic, and discourse functions (Bates & MacWhinney, 1989; Goldberg, 2006; Robinson & Ellis, 2008; Tomasello, 2003; Trousdale & Hoffmann, 2013). These allow communication because they are conventionalized in the speech community. People learn them from engaging in communication, the "interpersonal communicative and cognitive processes that

We thank Matt O'Donnell and Adam Kilgarriff for helpful reactions to a prior draft.

Correspondence concerning this article should be addressed to: Stefan Th. Gries, Department of Linguistics, University of California, Santa Barbara, Santa Barbara, CA 93106-3100. E-mail: stgries@linguistics.ucsb.edu

DOI: 10.1111/lang.12119

everywhere and always shape language" (Slobin, 1997). Repeated experience results in their becoming entrenched as language knowledge in the learner's mind.

Constructionist accounts thus investigate processes of language acquisition that involve the distributional analysis of the language stream and the parallel analysis of contingent cognitive and perceptual activity, with abstract constructions being learned from the conspiracy of concrete exemplars of usage following statistical learning mechanisms relating input and learner cognition (Rebuschat & Williams, 2012). Psychological analyses of these learning mechanisms are informed by the literature on the associative learning of cue-outcome contingencies, where the usual determinants include: factors relating to the form such as frequency and salience; factors relating to the functional interpretation such as significance in the comprehension of the overall utterance, prototypicality, generality, and redundancy; factors relating to the contingency of form and function; and factors relating to learner attention, such as automaticity, transfer, overshadowing, and blocking (Ellis, 2002, 2003, 2006, 2008). These various psycholinguistic factors conspire in the acquisition and use of any linguistic construction. Research into language and language acquisition therefore requires the measurement of these factors.

From its very beginnings, psychological research has recognized three major experiential factors that affect cognition: frequency, recency, and context of usage (e.g., Anderson, 2000; Bartlett, 1932/1967; Ebbinghaus, 1885). "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" (Ellis, 2006, p. 8). But assessing these probabilities is nontrivial, because constructions are nested and overlap at various levels (morphology within lexis within grammar); because sequential elements are memorized as wholes at (and sometimes crossing) different levels; because there are parallel, associated, symbiotic, thought-sound strands that are being chunked—language form, perceptual representations, motoric representations, . . . , the whole gamut of cognition—and because there is no one direction of growth—there is continuing interplay between top-down and bottom-up processes and between memorized structures and more open constructions: "Language, as a complex, hierarchical, behavioral structure with a lengthy course of development . . . is rich in sequential dependencies: syllables and formulaic phrases before phonemes and features . . . , holophrases before words, words before simple sentences, simple sentences before lexical categories, lexical categories before complex sentences, and so on" (Studdert-Kennedy, 1991, p. 10).

Constructions develop hierarchically by repeated cycles of differentiation and integration. Recent developments in corpus and cognitive linguistics are addressing these issues of operationalization and measurement with increasing sophistication (Baayen, 2008, 2010; Gries, 2009, 2013; Gries & Divjak, 2012). This paper summarizes relevant factors and how these can be operationalized and explored on the basis of corpus data.

Psycholinguistic Desiderata and Corpus-Linguistic Responses

Frequency

The most fundamental factor that drives learning is the frequency of repetition in usage. This determines whether learners are likely to experience a construction and, if so, how strongly it is entrenched, accessible, and its processing automatized.

Sampling

Language learners are more likely to experience more frequent usage events. They have limited exposure to the target language but are posed with the task of estimating how linguistic constructions work from an input sample that is incomplete, uncertain, and noisy. Native-like fluency, idiomaticity, and selection presents another level of difficulty again. For a good fit, every utterance has to be chosen from a wide range of possible expressions to be appropriate for that idea, for that speaker and register, for that place/context, and for that time. And again, learners can only estimate this from their finite experience. Like other estimation problems, successful determination of the population characteristics is a matter of statistical sampling, description, and inference.

Entrenchment

Learning, memory, and perception are all affected by frequency of usage: the more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The power law of learning (Anderson, 1982; Ellis & Schmidt, 1998; Newell, 1990) describes the relationships between practice and performance in the acquisition of a wide range of cognitive skills—the greater the practice, the greater the performance, although effects of practice are largest at early stages of learning, thereafter diminishing and eventually reaching an asymptote. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed. The more times we experience conjunctions of features, the more they become associated in our minds, the more these subsequently affect perception and categorization in

the sense that we perceive and process them as a chunk; so a stimulus becomes associated to a context and we become more likely to perceive it in that context.

Fifty years of psycholinguistic research has demonstrated language processing to be exquisitely sensitive to usage frequency at all levels of language representation: phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax (Ellis, 2002). Language knowledge involves statistical knowledge, so humans learn more easily and process more fluently high frequency forms and regular patterns that are exemplified by many types and that have few competitors. Psycholinguistic perspectives thus hold that language learning is the associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is a key determinant of this kind of acquisition because "rules" of language, at all levels of analysis from phonology, through syntax, to discourse, are structural regularities that emerge from learners' lifetime analysis of the distributional characteristics of the language input.

Counting Frequencies in Corpora

Frequencies of occurrence and frequencies of co-occurrence constitute the most basic corpus-linguistic data. In fact, one somewhat reductionist view of corpus data would be that corpora typically have actually nothing more to offer than frequencies of (co-)occurrence of character strings and that anything else (usage-based) linguists are interested in—morphemes, words, constructions, meaning, information structure, function—needs to be operationalized in terms of frequencies of (co-)occurrence. Thus, linguistic data from corpora can be ranked in terms of how (in)directly a particular object of interest is reflected by corpus-based frequencies. On such a scale, frequency per se and the way it contributes to, or more carefully "is correlated with," entrenchment is the simplest corpus-based information and is typically provided in the form of tabular frequency lists of word forms, lemmas, *n*-grams (interrupted or contiguous sequences of words), and so on. While seemingly straightforward, it is worth noting that even this simplest of corpus-linguistic methods can require careful consideration of at least two kinds of aspects.

First, counting tokens such as words requires an (often implicit) process of tokenization, that is, decisions as to how the units to be counted are delimited. In some languages, whitespace is a useful delimiter, but some languages do not use whitespace to delimit, say, words (Mandarin Chinese is a case in point) so a tokenizer is needed to break up sequences of Chinese characters into words and different tokenizers can yield different results. Even in languages that do

use whitespace (e.g., English), there may be strings one would want to consider words even though they contain whitespace; examples include proper names and titles (e.g., *Barack Obama* and *Attorney General*), compounds (*corpus linguistics*), and multiword units (e.g., *according to*, *in spite of*, or *on the one hand*). In addition, tokenization can be complicated by other characters (how many words are *1960* or *Peter's dog*?) or spelling inconsistencies (e.g., *armchair linguist* vs. *armchair-linguist*). Practically, this means that it is often a good idea to explore an inventory of all characters that are attested in a corpus before deciding on how to tokenize a corpus.

Second, aggregate token frequencies for a complete corpus can be very misleading since they may obscure the fact that tokens may exhibit very uneven distributions in a corpus, a distributional characteristic called *dispersion*, which is important both psycholinguistically and corpus-linguistically/statistically.

Dispersion

While frequency provides an overall estimate of whether learners are likely to experience a construction, there is another dimension relevant to learning: dispersion, that is, how regularly they experience a construction: Some constructions are equally distributed throughout language and will thus be experienced somewhat regularly, others are found aggregated or clumped in particular contexts or in bursts of time and may, therefore, only be encountered rarely, but then frequently in these contexts. In other words, frequency answers the question "how often does *x* happen?" whereas dispersion asks "in how many contexts will you encounter *x* at all?"

Sampling Discourse Contexts

Language users are more likely to experience constructions that are widely or evenly distributed in time or place. When they do so, contextual dispersion indicates that a construction is broadly conventionalized, temporal dispersion shares out recency effects.

Sampling Linguistic Contexts: Type and Token Frequency

Token frequency counts how often a particular form appears in the input. Type frequency, on the other hand, refers to the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection or a syntactic construction specifying the relation among words. For example, the regular English past tense *-ed* has a very high type frequency because it applies to thousands of different types of verbs, whereas the vowel change exemplified in *swam* and *rang* has much lower type

frequency; thus, in a sense, type frequency is a kind of dispersion. The productivity of phonological, morphological, and syntactic patterns is a function of type rather than token frequency (Bybee & Hopper, 2001). This is because: (a) the more lexical items that are heard in a certain position in a construction, the less likely it is that the construction is associated with a particular lexical item and the more likely it is that a general category is formed over the items that occur in that position; (b) the more items the category must cover, the more general are its criterial features and the more likely it is to extend to new items; and (c) high type frequency ensures that a construction is used frequently and widely, thus strengthening its representational schema and making it more accessible for further use with new items (Bybee & Thompson, 2000). In contrast, high token frequency promotes the entrenchment or conservation of irregular forms and idioms; irregular forms only survive because they are high frequency.

The overall frequency of a construction compounds type and token frequencies, whereas it is type frequency (dispersion over different linguistic contexts) that is most potent in fluency and productivity of processing (Baayen, 2010). These factors are central to theoretical debates on linguistic processing and the nature of abstraction in language regarding exemplar-based versus abstract prototype representations, phraseology and the idiom principle versus open rule-driven construction, and the richness of exemplar memories and their associations versus more abstract connectionist learning mechanisms that tune the feature regularities but lose exemplar detail (Pierrehumbert, 2006). Metrics of dispersion over different linguistic contexts are therefore key to these inquiries.

Measuring Dispersion and Type Frequency in Corpora

Because virtually all corpus-linguistic data are based on frequencies, the fact that very similar or even identical frequencies of tokens can come with very different degrees of dispersion in a corpus makes the exploration of dispersion information virtually indispensable. This fact is exemplified in Figure 1. Both panels represent the frequency of words (logged to the base of 10) on the *x*-axis and the dispersion metric *DP* (cf. Gries, 2008) on the *y*-axis. *DP* is very straightforward to compute: (i) for each part of the relevant corpus, compute its size s_i in percent of the whole corpus; (ii) also, for each part of the corpus, compute how much of a token it contains in percent of all instances of the token t_i; and (iii) compute and sum up the absolute pairwise differences $|s_i\text{-}t_i|$, and divide the sum by 2. Thus, *DP* falls between 0 and approximately 1 and low and high values reflect equal and unequal dispersion respectively. While

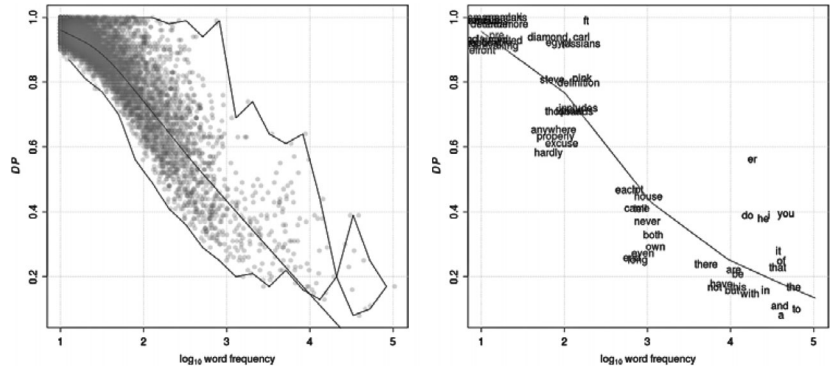

Figure 1 The relation between (logged) frequency (on the *x*-axes) and *DP* (on the *y*-axes): all words in the BNC sampler with a frequency ≥10 (left panel), 68 words from different frequency bins (right panel).

there is the expected overall negative correlation between token frequency and dispersion (indicated by the solid-line smoother)—infrequent tokens cannot be highly dispersed, frequent ones are likely to be highly dispersed—there is a large amount of diverse dispersion results for intermediately frequent words. The left panel shows, for example, that especially in the frequency range of 2–3.5, words with very similar frequencies can vary enormously with regard to their dispersion; in the right panel, this is exemplified more concretely: words such as *hardly* and *diamond*, for instance, have nearly the exact same frequency but are distributed very differently.

Because especially in psycholinguistics word frequency is often used as a predictor or a control variable, results like these show that considering dispersion is just as important, or even more important for such purposes (cf. Gries, 2010, for how dispersion measures can be better correlated with reaction time data than the usual frequency data).

As for type frequency, this is a statistic that is usually computed from frequency lists (as when one determines all verbs beginning with *under*-), but probably more often from concordance displays that show the linguistic element in question in its immediate context. As discussed, in the case of morphemes or constructions, the type frequency of an element is the number of different types that the element co-occurs with, for example, the number of different nouns to which a particular suffix attaches or the number of different verbs that occur in a slot of a particular construction. While this statistic is easy to obtain, it is again not necessarily informative enough because the type frequency per

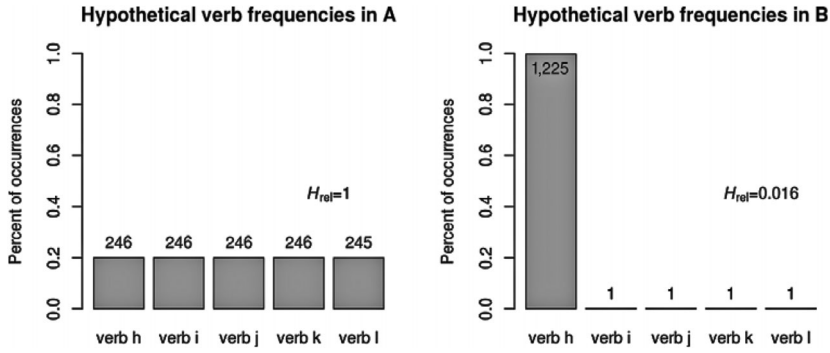

Figure 2 Type-token frequency distributions for constructions *A* and *B* in a hypothetical data set.

se does not also reflect the frequency distribution of the types. For instance, two constructions *A* and *B* may have identical token frequencies in a corpus (e.g., 1,229) and identical type frequency of verbs entering into them, say, 5, but these may still be distributed very differently, as is exemplified in Figure 2.

A measure to quantify the very different frequency distributions is relative entropy H_{rel}, a measure of uncertainty that approximates 1 as distributions become more even (as in the left panel) and that approximates 0 as distributions become more uneven and, thus, more predictable (as in the right panel). The Zipfian distributions that are so omnipresent in corpus-linguistic data typically give rise to small entropy values; see also below. In sum, both dispersion and (relative) entropy are useful but as yet underutilized corpus statistics that should be considered more often in corpus-linguistic approaches to both cognitive/usage-based linguistics as well as psycholinguistics.

Contingency

Form–Function Contingency

Psychological research into associative learning has long recognized that while frequency of form is important, so too is contingency of mapping (Shanks, 1995). Cues with multiple interpretations are ambiguous and so hard to resolve; cue-outcome associations of high contingency are reliable and readily processed. Consider how, in the learning of the category of birds, while eyes and wings are equally frequently experienced features in the exemplars, it is wings that are distinctive in differentiating birds from other animals. Wings are important features to learning the category of birds because they are reliably associated with class membership while being absent from outsiders. Raw

frequency of occurrence is therefore less important than the contingency between cue and interpretation. Reliability of form–function mapping is a driving force of all associative learning, to the degree that the field of its study has become known as "contingency learning." These factors are central to the Competition Model (MacWhinney, 1987, 1997, 2001) and to other models of construction learning as the rational learning of form-function contingencies (Ellis, 2006; Xu & Tennenbaum, 2007).

Context and Form-Form Contingency

Associative learning over the language stream allows language users to "find structure in time" (Elman, 1990) and thus to make predictions. The words that they are likely to hear next, the most likely senses of these words, the linguistic constructions they are most likely to utter next, the syllables they are likely to hear next, the graphemes they are likely to read next, the interpretations that are most relevant, and the rest of what's coming (next) across all levels of language representation, are made readily available to them by their language processing systems. Their unconscious language representation systems are adaptively tuned to predict the linguistic constructions that are most likely to be relevant in the ongoing discourse context, optimally preparing them for comprehension and production. As a field of research, the rational analysis of cognition is guided by the principle that human psychology can be understood in terms of the operation of a mechanism that is optimally adapted to its environment in the sense that the behavior of the mechanism is as efficient as it conceivably could be, given the structure of the problem space and the cue–interpretation mappings it must solve (Anderson, 1989). These factors are at the core of language processing, small and large, from collocations (Gries, 2013), to collostructions (Gries & Stefanowitsch, 2004; see below) to formulas (Ellis, 2012), parsing sentences (Hale, 2011), understanding sentences (MacDonald & Seidenberg, 2006), and reading passages of texts (Demberg & Keller, 2008).

Measuring Contingency in Corpus Linguistics

Quantifying contingency has a long tradition in corpus linguistics. The perhaps most fundamental assumption underlying nearly all corpus-linguistic research is that similarity in distribution, of which co-occurrence is the most frequent kind in corpus research, reflects similarity of meaning or function. Thus, over the last decades a large variety of measures of contingency—so-called association measures – have been developed (cf. Pecina, 2009 for a recent overview). The vast majority of these measures are all based on a 2×2 co-occurrence table of the kind exemplified in Table 1. In this kind of table, the two

Table 1 Schematic co-occurrence table of token frequencies for association measures

Observed frequencies	Element *y*	Other elements	Totals
Element x	*a*	*b*	*a*+*b*
Other elements	*c*	*d*	*c*+*d*
Totals	*a*+*c*	*b*+*d*	$a+b+c+d = N$

linguistic elements *x* and *y* whose mutual (dis)preference for co-occurrence is quantified—these can be words, constructions, other patterns—are listed in the rows and columns, respectively, and the four cells of the table list frequencies of co-occurrence in the corpus in question; the central frequency is *a*, which is the co-occurrence frequency of *x* and *y*.

Most association measures require that one computes the *expected* frequencies *a*, *b*, *c*, and *d* that would result from *x* and *y* co-occurring together as often as would be expected from their marginal totals (*a*+*b* and *a*+*c*) as well as the corpus size *N*. The following measures are among the most widely used ones:

(1) a. pointwise $MI = log_2 \frac{a}{a_{expected}}$

b. $z = \frac{a - a_{expected}}{\sqrt{a_{expected}}}$

c. $t = \frac{a - a_{expected}}{\sqrt{a}}$

d. $G^2 = 2 \cdot \sum_{1}^{4} obs \cdot log \frac{obs}{exp}$

e. $-\log 10\, p_{\text{Fisher-Yates exact test}}$

Arguably, (1e) is among the most useful measures because it is based on the hypergeometric distribution, which means (i) quantifying the association between *x* and *y* is treated as a sampling-from-an-urn-(the corpus)-with-replacement problem and (ii) the measure is not computed on the basis of any distributional assumptions such as normality. Precisely because of the fact that (1e) involves an exact test, which could involve the computations of theoretically hundreds of thousands of probabilities for just one pair of elements *x* and *y*, the log-likelihood statistic in (1d) is often used as a reasonable approximation. In addition, since some measures have well-known statistical characteristics—mutual information (*MI*) is known to inflate with low expected frequencies (i.e., rare combinations) and *t* is known to prefer frequent co-occurrences—researchers sometimes compute more than one association measure.

Applications of association measures are numerous but, for a long time, they were nearly exclusively applied to collocations, that is, co-occurrences where both elements *x* and *y* are words. For example, researchers would use association

measures to identify the words $y_{1\text{-}m}$ that are most strongly attracted to a word *x*; a particularly frequent application involves determining the collocates that distinguish best between each member $x_{1\text{-}n}$ of a set of *n* near synonyms. For example, Gries (2003) showed how this approach helps distinguish notoriously difficult synonyms such as *alphabetic*/*alphabetical* or *botanic*/*botanical* by virtue of the nouns each word of a pair prefers to co-occur with.

In the last 10 years, a family of methods called *collostructional analysis*—a blend of *collocation* and *constructional*—has become quite popular. This approach is based on the assumption—independently arrived at in cognitive/usage-based linguistics and corpus linguistics—that there is no real qualitative difference between lexical items and grammatical patterns, from which it follows that one can simply replace, say, word *x* in Table 1 by a grammatical pattern and then quantify which words $y_{1\text{-}n}$ "like to co-occur" with/in that grammatical pattern. In one of the first studies, Stefanowitsch and Gries (2003) showed how the verbs that are most strongly attracted to constructions are precisely those that convey the central senses of the (often polysemous) constructions. For example, the verbs in (2) and (3) are those that are most strongly attracted to the ditransitive V NP_{REC} NP_{PAT} construction and the *into*-causative V NP_{PAT} into V-*ing* construction, respectively; manual analysis as well as computationally more advanced methods (see below) reveal that these verbs involve concrete and metaphorical transfer scenarios as well as trickery/force respectively.

(2) *give*, *tell*, *send*, *offer*, *show*, *cost*, *teach*, *award*, *allow*, *lend*, . . .
(3) *trick*, *fool*, *coerce*, *force*, *mislead*, *bully*, *deceive*, *con*, *pressurize*, *provoke*, . . .

Additional members of the family of collostructional analysis have been developed to, for instance, compare two or more constructions in terms of the words that are attracted to them most (cf. Gries & Stefanowitsch, 2004), which can be useful to study many of the syntactic alternations that have been studied in linguistics such as the dative alternation (*John gave Mary the book* vs. *John gave the book to Mary*), particle placement (*John picked up the book* vs. *John picked the book up*), *will*-future versus *going-to* future versus *shall*, and so on.

If, as we argued above, contingency information was really more relevant than mere frequency of occurrence, then it should be possible to show this by comparing predictions made on the basis of frequency to predictions made on the basis of contingency/association strength. Gries, Hampe, and Schönefeld (2005, 2010) study the *as*-predicative exemplified in (4) using collostructional

analysis and then test whether subjects' behavior in a sentence-completion task and a self-paced reading task is better predicted by frequency of co-occurrence (conditional probability) or association strength ($-\log10\, p_{\text{Fisher-Yates exact test}}$).

(4) a. V NP_{DO} *as* XP
b. John regards Mary as a good friend.
c. John saw Mary as intruding on his turf.

In both experiments, they find that the effect of association strength is significant (in one-tailed tests) and much stronger than that of frequency: Subjects are more likely to complete a sentence fragment with an *as*-predicative when the verb in the prompt was not just frequent in the *as*-predicative but actually attracted to it; similarly, subjects were faster to read the words following *as* when the verb in the sentence was predictive for the *as*-predicative. Similarly encouraging results were obtained by Ellis and Ferreira-Junior (2009), who show that measures of association strength such as p_{FYE} (and others, see below) are highly correlated with learner uptake of verb use in constructions and more so than frequency measures alone.

In spite of the many studies that have used association measures to quantify contingency, there have been few attempts to improve how contingency is quantified. Two problems are particularly pressing. First, nearly all association measures neither include the type frequencies of *x* and *y* in their computation nor the type-token distributions (or [relative] entropies, see above) because the type frequencies are just conflated in the two token frequencies *b* and *c*. Thus, no association measure at this point can distinguish the two hypothetical scenarios represented in Figure 3, in which one may be interested in quantifying the association of construction *A* and verb *h*. In both cases, A is attested 1,229 times with 5 different verb types, of which the verb of interest, *h*, accounts for 500. All existing association measures would return the same value for the association of *A* and *h* although a linguist appreciating the notion of contingency/predictiveness may prefer a measure that can also indicate that, in the left panel, another verb may be more strongly attracted to *A* than in the right panel, where *h* is highly predictive of *A*. There is one measure that has been devised to at least take type frequency into consideration—Daudaravičius and Marcinkevičienė's (2004) lexical gravity *G*—but even this one would not be able to differentiate the two panels in Figure 3 since they involve the same type frequency (5) and only differ in their entropy.

In the absence of easily recoverable frequency distributions of, say, constructions from parsed corpora, this kind of improvement will of course be

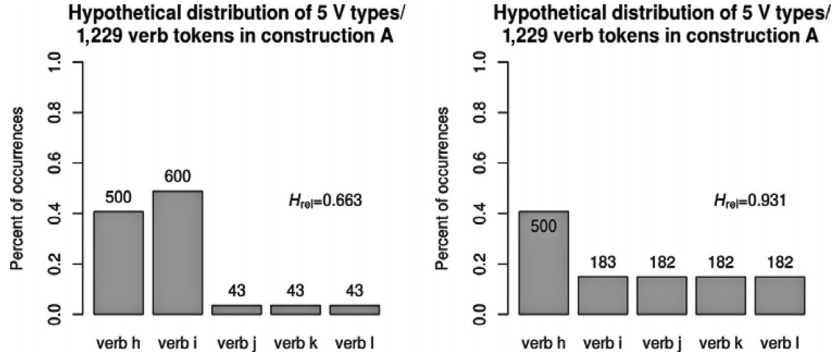

Figure 3 Type-token frequency distributions for constructions *A* and *B* in a hypothetical data set.

very hard to come by; studies like Roland, Dick, and Elman (2007) provide important first steps toward this goal.

A second problem of nearly all association measures is their bidirectionality: they quantify the mutual association of two elements even though, from the perspective of psycholinguistics or the psychology of learning, associations need not be mutual, or equally strong in both directions (just like perceptions of similarity are often not symmetric; cf. Tversky 1977). While there have been some attempts at introducing directional association measures based on ranked collocational strengths (cf. Michelbacher, Evert, & Schütze, 2011), the results have been mixed (in terms of how well they correlate with behavioral data, how well they can separate some very strongly attracted collocations, and in terms of the computational effort the proposed measures require). The currently most promising approach is the measure ΔP from the associative learning literature as introduced into corpus linguistics by Ellis (2007). ΔP is a measure that can be straightforwardly computed from a table such as Table 1 as shown in (5), that is, as simple differences of proportions:

(5) a. $\Delta P_{y|x} = \frac{a}{a+b} - \frac{c}{c+d}$
b. $\Delta P_{x|y} = \frac{a}{a+c} - \frac{b}{b+d}$

When applied to two-word units in the spoken component of the British National Corpus (cf. Gries, 2013a), this measure is very successful at identifying the directional association of two-word units that traditional measures flag as mutually associated. For instance, (6a) lists two-word units in which the

first word is much more predictive of the second one than vice versa, and (6b) exemplifies the opposite kind of cases.

(6) a. *upside down, according to, volte face, ipso facto, instead of, insomuch as*
b. *of course, for example, per annum, de facto, at least, in situ*

In sum, the field of corpus-linguistic research on contingency/association is a lively one. Unfortunately, its two most pressing problems—type-token distributions and directionality—are currently only addressed with methods that can handle only one of these at the same time; it remains to be hoped that newly developed tools will soon address both problems at the same time in a way that jibes well with behavioral data.

Surprisal

Language learners do not consciously tally any of the above-mentioned corpus-based statistics. The frequency tuning under consideration here is computed by the learner's system automatically during language usage. The statistics are implicitly learned and implicitly stored (Ellis, 2002); learners do not have conscious access to them. Nevertheless, every moment of language cognition is informed by these data, as language learners use their model of usage to understand the actual usage of the moment as well as to update their model and to predict where it is going next.

There is considerable psychological research on human cognition and its dissociable, complementary systems for implicit and explicit learning and memory (Ellis, 2007, 2015; Rebuschat, 2015). Implicit learning is acquisition of knowledge about the underlying structure of a complex stimulus environment by a process that takes place naturally, simply, and without conscious operations. Explicit learning is a more conscious operation where the individual makes and tests hypotheses in a search for structure. Much of the time, language processing, like walking, runs successfully using automatized, implicit processes. We only think about walking when it goes wrong, when we stumble, and conscious processes are called in to deal with the unexpected. We might learn from that episode where the uneven patch of sidewalk is, so that we don't fall again. Similarly, when language processing falters and we do not understand, we call the multimodal resources of consciousness to help deal with the novelty. Processing becomes deliberate and slow as we think things through. This one-off act of conscious processing too can seed the acquisition of novel explicit form–meaning associations (Ellis, 2005). It allows us to consolidate new constructions as episodic fast-mapped cross-modal associations

(Carey & Bartlett, 1978). These representations are then also available as units of implicit learning in subsequent processing. Broadly, it is not until a representation has been noticed and consolidated that the strength of that representation can thereafter be tuned implicitly during subsequent processing (Ellis, 2006). Thus the role of noticing and consciousness in language learning (Ellis, 1994; Schmidt, 1994).

Contemporary learning theory holds that learning is driven by prediction errors: that we learn more from the surprise that comes when our predictions are incorrect than when our predictions are confirmed (Clark, 2013; Rescorla & Wagner, 1972; Rumelhart, Hinton, & Williams, 1986; Wills, 2009), and there is increasing evidence for surprisal-driven language processing and acquisition (Dell & Chang, 2014; Demberg & Keller, 2008; Jaeger & Snider, 2013; Pickering & Garrod, 2013; Smith & Levy, 2013). For example, Demberg and Keller (2008) analyze a large corpus of eye-movements recorded while people read text to demonstrate that measures of surprisal account for the costs in reading time that result when the current word is not predicted by the preceding context. Surprisal can be seen as an information-theoretic interpretation of probability. It is computed as shown in (7).

(7) $\text{surprisal} = -\log_2 p$

The probability in question can be unconditional or conditional probabilities of occurrence of different kinds of linguistic elements of any degree of complexity. The simplest possible case would be the unconditional probability (i.e., relative frequency) of, say, a word in a corpus. A slightly more complex example would be a simple forward transitional probability such as the probability of the word *y* directly following the word *x*, or a conditional probability such as the probability of a particular verb given a construction. More complex applications include the conditional probability of a word given several previous words in the same sentence or, to include a syntactic example, the conditional probability of a particular parse tree given all previous words in a sentence (as in, say, Demberg & Keller, 2008).

Whatever the exact nature of the (conditional) probability, equation (7) shows that surprisal derives from conditional probabilities, which means it, too, can in fact be computed from Table 1, namely as $-\log_2 {}^{a}/_{a+b}$ or $-\log_2 {}^{a}/_{a+c}$, and, as Figure 4 clearly shows, surprisal is therefore inversely related to probability and thus also very strongly correlated with ΔP.

In usage-based linguistics, surprisal has been studied in particular in studies of structural priming, for example, when Jaeger and Snider (2008) show that surprising structures—for example, when a verb that is strongly attracted to

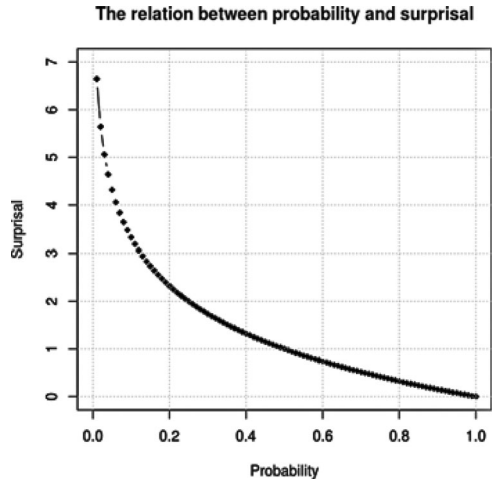

Figure 4 The relationship between probability (on the *x*-axis) and surprisal (on the *y*-axis).

the ditransitive is used in the prepositional dative—prime more strongly than nonsurprising structures. Whichever way surprisal is computed, it is a useful addition to the corpus-linguistic tool kit and may ultimately also be viewed as a good operationalization of the notoriously tricky notion of salience.

The complementary psychological systems of implicit, expectation-driven, automatic cognition as opposed to explicit, conscious processing are paralleled in these complementary corpus statistics measuring predictability in context vs. surprisal. Contemporary corpus pattern analysis also focusses upon their tension. Hanks (2009, p. 64) talks of norms and exploitations as the *Linguistic Double Helix*:

> Much of both the power and the flexibility of natural language is derived from the interaction between two systems of rules for using words: a primary system that governs normal, conventional usage and a secondary system that governs the exploitation of normal usage.

The *Theory of Norms and Exploitations* (TNE, Hanks, 2013) is a lexically based, corpus-driven theoretical approach to how words go together in collocational patterns and constructions to make meanings. He emphasizes that the approach rests on the availability of new forms of evidence (corpora, the Internet) and the development of new methods of statistical analysis and inferencing. Partington (2011), in his analysis of the role of surprisal in irony, demonstrates that the reversal of customary collocational patterns (e.g., *tidings of great joy,*

overwhelmed) drives phrasal irony (*tidings of great horror, underwhelmed*). Similarly, Suslov (1992) shows how humor and jokes are based on surprisal that is pleasurable: we enjoy being led down the garden path of a predictable parse path, and then have it violated by the joke-teller.

Zipf's Law and Construction Learning

Zipf's law states that in human language, the frequency of words decreases as a power function of their rank in the frequency table. If p_f is the proportion of words whose frequency in a given language sample is f, then $p \approx \alpha f^{-1/s}$ with $s \approx 1$. Zipf (1949) showed this scaling relation holds across a wide variety of language samples. Subsequent research has shown that many language events (e.g., frequencies of phoneme and letter strings, of words, of grammatical constructs, of formulaic phrases, etc.) across scales of analysis follow this law (Ferrer i Cancho & Solé, 2001, 2003).

Research by Goldberg (2006), Ellis and Ferreira-Junior (2009), Ellis and O'Donnell (2012), and Ellis, O'Donnell, and Römer (2012) shows that verb argument constructions are (1) Zipfian in their verb type-token constituency in usage, (2) selective in their verb form occupancy, and (3) coherent in their semantics, with a network structure involving prototypical nodes of high betweenness centrality and a degree distribution that is also Zipfian. Psychological theory relating to the statistical learning of categories suggests that learning is promoted, as here, when one or a few lead types at the semantic center of the construction account for a large proportion of the tokens. These robust patterns of usage might therefore facilitate processes of syntactic and semantic bootstrapping.

Zipfian distributions are also characterized by a low entropy because of how the most frequent elements in a distribution reduce the uncertainty, and increase the predictability, of the distribution. In a learning experiment of Goldberg, Casenhiser, and Sethuraman (2004), subjects heard the same number of novel verbs (type frequency: 5), but with two different distributions of 16 tokens, a balanced condition of 4-4-4-2-2 (with a relative entropy of $H_{\text{rel}} = 0.97$), and a skewed lower-variance condition of 8-2-2-2-2 ($H_{\text{rel}} = 0.86$). The distribution that was learned significantly better was the one that was more Zipfian and had the lower entropy, providing further evidence for the psycholinguistic relevance of Zipfian distribution and the notion of entropy.

Semantic Network Analysis

Constructions map linguistic forms to meanings. One of the greatest challenges in usage-based research is how to quantify relevant aspects of meaning, for example, for verb-argument constructions (VAC):

– *prototypicality*: For each verb type occupying a VAC, how prototypical is it of the VAC?
– *semantic cohesion*: For each VAC, how semantically cohesive are its verb exemplars?
– *polysemy*: Are there are one or several meaning groups associated with a VAC form and can we identify these semantic communities?

Analysis of construction meanings typically rests on human classification, as illustrated so well in the ground-breaking corpus linguistic work on the meanings of English Verb Pattern Grammar (Francis, Hunston, & Manning, 1996). But we can go some way toward quantifying these analyses, and this will become increasingly important as we pursue replicable research to scale in large corpora. O'Donnell and Ellis applied methods of network science to these goals (O'Donnell, Ellis, Corden, Considine, & Römer, 2015; Römer, O'Donnell, & Ellis, 2014).

Consider the *into*-causative VAC (as in *He tricked me into employing him*) described here. Wulff, Stefanowitsch, and Gries (2007) present a comparison of the verbs that occupy this construction in corpora of American and British English using distinctive collexeme analysis. They take the verbs that are statistically associated with this VAC in the two corpora, qualitatively group them into meaning groups, and show a predominance of verbal persuasion verbs in the cause predicate slot of the American English data as opposed to the predominance of physical force verbs in the cause predicate slot of the British English data. Their qualitative methods for identifying the semantic classes were clearly described:

> First, the three authors classified the distinctive collexemes separately. The resulting three classifications and semantic classes were then checked for consistency. Verbs and classes which had not been used by all three authors were finally re-classified on the condition that finally a maximum number of distinctive collexemes be captured by a minimum number of semantic classes. The resulting classes are verbs denoting communication (e.g. *talk*), negative emotion (e.g. *terrify*), physical force (e.g. *push*), stimulation (e.g. *prompt*), threatening (e.g. *blackmail*), and trickery (e.g. *bamboozle*). (p. 273)

This pattern was discussed on the Corpora list (www.hit.uib.no/corpora/ November 20, 2013) and Kilgarriff (Kilgarriff, Rychly, Smrz, & Tugwel, 2004) posted the types of verb that occupy the pattern in 113436 hits in the enTenTen12 corpus (a 12 billion word corpus of web crawled English texts

collected in 2012, http://www.sketchengine.co.uk). Following the methods described in O'Donnell et al. (2015), we took these verb types and built a semantic network using WordNet, a distribution-free semantic database based upon psycholinguistic theory (Miller, 2009). WordNet places verbs into a hierarchical network organized into 559 distinct root synonym sets (synsets such as *move1* expressing translational movement, *move2* movement without displacement, etc.), which then split into over 13,700 verb synsets. Verbs are linked in the hierarchy according to relations such as hypernym [verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to *perceive* is an hypernym of to *listen*], and hyponym [verb Y is a hyponym of the verb X if the activity Y is doing X in some manner (to *lisp* is a hyponym of to *talk*)]. Algorithms to determine the semantic similarity between WordNet synsets have been developed that consider the distance between the conceptual categories of words and their hierarchical structure in WordNet (Pedersen, Patwardhan, & Michelizzi, 2004). We compared the verb types occupying the *into*-causative pairwise on the WordNet Path Similarity measure as implemented in the Natural Language Tool Kit (Bird, Loper, & Klein, 2009), which ranges from 0 (no similarity) to 1 (items in the same synset). We then built a semantic network in which the nodes represent verb types and the edges strong semantic similarity. Standard measures of network density, average clustering, degree centrality, transitivity, and so on, were then used to assess the cohesion of the semantic network (de Nooy, Mrvar, & Batagelj, 2010). We also applied the Louvain algorithm for the detection of communities within the network representing different semantic sets (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

Figure 5 shows the semantic network for verb occupying the *into*-causative VAC built using these methods, with 7 differently colored communities identified using the Louvain algorithm. In these networks, related concepts are closer together. The more connected nodes at the center of the network, like *make*, *stimulate*, *force*, and *persuade*, are depicted larger to reflect their higher degree. For each node we have measures of degree, betweenness centrality, and so on. There are 57 nodes connected in the network by 130 edges. The cohesion metrics for the network as a whole include network density 0.081, average clustering of 0.451, a degree assortativity of 0.068, transitivity 0.364, degree centrality 0.212, and betweenness centrality 0.228, and a modularity score, which reflects the degree to which there are emergent communities, of 0.491. We have colored the communities following the same scheme we used above when describing the qualitative results of Wulff et al. (2007). There are clear parallels, and community membership seems to make sense. For example, the [*deceive*] community [*deceive, fool, delude, dupe, kid, trick, hoodwink*] is

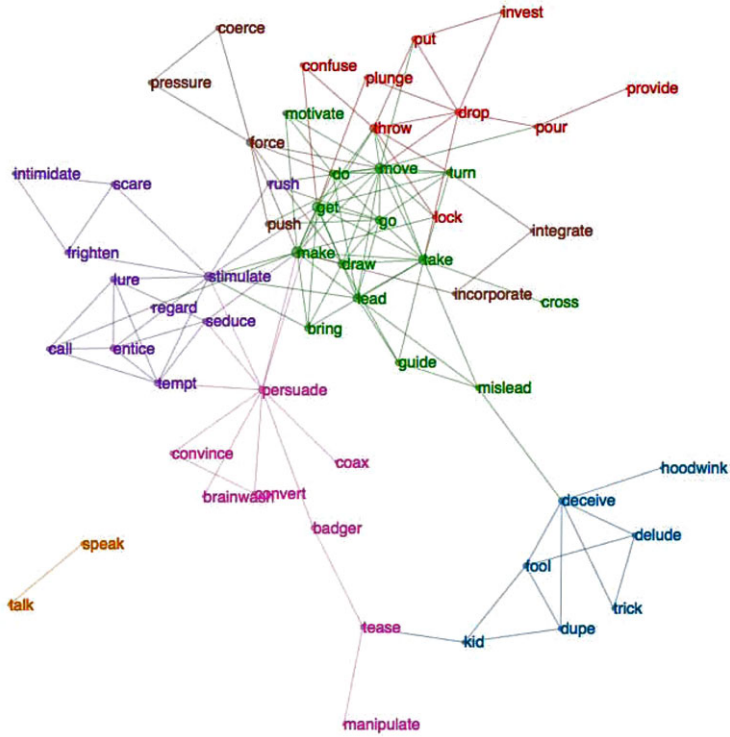

Figure 5 The semantic network for verbs occupying the *into*-causative VAC.

clearly separate from the [*force*] community [*force, push, coerce, incorporate, integrate, pressure*]. The [*persuade*] community is separated again [*persuade, tease, badger, convert, convince, brainwash, coax, manipulate*], and [*speak*, and *talk*] drift off into space on their own. Relating back to Kilgarriff's list of hits, the [*deceive*] community accounts for 44% of the total tokens, [*speak*], 17%, [*make*] 12%, [*throw*] 8%, [*stimulate*] 8%, [*force*] 6%, and [*persuade*] 4%.

These network science methods allow a variety of relevant metrics for semantics:

– *prototypicality*: The prototype as an idealized central description is the best example of the category, appropriately summarizing its most representative attributes. In network analysis, there are many available measures of centrality: degree centrality, closeness centrality, betweenness centrality, PageRank, and so on, each with its advantages and disadvantages (Newman, 2010). Historically first and conceptually simplest is degree centrality,

or degree, which is simply its connectivity in terms of the number of links incident upon a node. An alternative is betweenness centrality, which was developed to quantify the control of a human on the communication between other humans in a social network (Freeman, 1977). It is defined as the number of shortest paths from all nodes to all others that pass through that node. It is a more useful measure than degree of both the load and global importance of a node.

– *semantic cohesion*: In category learning, coherent categories, where exemplars are close to the prototype, are acquired faster than categories comprised of diverse exemplars. Graph theory also offers a number of alternatives for measuring network connectivity. The simplest is density, the number of edges in the network as a proportion of the number of possible edges linking those nodes. Other measures include average clustering, degree assortativity, transitivity, degree centrality, betweenness centrality, and closeness centrality (de Nooy et al., 2010; Newman, 2010).
– *polysemy and community detection*: A community within a graph or network is a group of nodes with dense connections to the other nodes in the group and sparser connections to other nodes that belong to a different community. Identification of communities has proven highly useful across a broad range of spheres to which network modeling can be applied, such as social networks, neural and gene networks. Analyses like those in Figure 5 suggest they might provide some traction in analyzing issues relating to issues of construction polysemy and homonymy. Nevertheless, there is a long way to go in properly analyzing the "hard problem" of construction semantics, which is just as hard as the hard problem of consciousness (Chalmers, 1995) in that we wish to understand how language prompts phenomenal experiences.

New developments like these network-/graph-based methods provide promising new avenues for exploring the functional side or pole of constructions—so far done largely manually or with simpler exploratory statistics such as cluster analyses—on the basis of the distributions of the formal side or pole of constructions. Given the scalability of these approaches, these are bound to take corpus-based studies in usage-based linguistics to new levels.

Conclusion

As we have argued above, speakers keep track of a wide array of co-occurrence information of both their language comprehension and production. It is becoming more and more obvious that this unconscious tracking of co-occurrence

statistics happens extremely early—in utero, in fact (cf. Moon, Lagercrantz, & Kuhl, 2012)—and also extremely fast. The latter has been demonstrated both in specific learning experiments with both children and adults but also in experiments that were not concerned with learning at all, but in which within-experiment learning had to be statistically controlled (cf. Gries & Wulff, 2009, for an example in L2 learning or Doğruöz & Gries, 2012, for an example in language contact situations). It is therefore imperative that both experimental and observational studies consider the speed and ubiquity of these learning processes alike—the unconscious pattern matcher in all of us hardly ever sleeps.

The processes and associations we describe here are all involved in every episode of language usage. Language processing is conditioned upon them all. So, for example, Ellis, O'Donnell, and Römer (2014) used free association and verbal fluency tasks to investigate VACs and the ways in which their processing is sensitive to these statistical patterns of usage (verb type-token frequency distribution, VAC-verb contingency, verb-VAC semantic prototypicality). In experiment one, 285 native speakers of English generated the first word that came to mind to fill the V slot in 40 sparse VAC frames such as 'he __ across the . . . ', 'it __ off the . . . ', and so on. In experiment two, 40 English speakers generated as many verbs that fit each frame as they could think of in a minute. For each VAC, they compared the results from the experiments with the corpus analyses of usage. For both experiments, multiple regression analyses predicting the frequencies of verb types generated for each VAC showed independent contributions of (i) verb frequency in the VAC, (ii) VAC-verb contingency, and (iii) verb prototypicality in terms of centrality within the VAC semantic network.

Future priorities concern both the range of corpus resources and statistical tools:

- We need more corpora, and more corpora representing diverse registers and with diverse layers of annotation—not just part-of-speech tagging, but syntactic parses, semantic as well as discourse annotation, and so on.
- We need more studies of the precise conditions when learning happens best and fastest, for example, how many high-frequency types in the Zipfian token distribution are best—1, 2, a few?—and what are the ideal distribution/dispersion conditions in which learning happens?
- We need more multivariate tools that include all the corpus statistics we can obtain—frequencies, dispersions, entropies, associations, and so on—but also new ones (such as the graph-based methods) that help us see the patterns in the structured but noisy mess that are corpora.

We hope that this agenda will lead to a stronger collaboration between usage-based theory on the one hand and corpus-linguistic practice on the other.

Final revised version accepted 16 September 2014

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–406.

Anderson, J. R. (1989). A rational analysis of human memory. In H. L. I. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 195–210). Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York: W.H. Freeman.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

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

Bartlett, F. C. (1932/1967). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–73). New York: Cambridge University Press.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media Inc.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics, P10008*.

Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, J., & Thompson, S. (2000). Three frequency effects in syntax. *Berkeley Linguistic Society*, *23*, 65–85.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, *15*, 17–29.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200–219.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204.

Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics, P09008*.

Daudaravičius, V., & Marcinkevičienė, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, *9*, 321–348.

de Nooy, W., Mrvar, A., & Batagelj, V. (2010). *Exploratory social network analysis with Pajek*. Cambridge, UK: Cambridge University Press.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Phil. Trans. R. Soc. B.*, *369*, 20120394. http://dx.doi.org/10.1098/rstb.2012.0394.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Doğruöz, A. S., & Gries, S.Th. (2012). Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics*, *10*, 401–426.

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology (H. A. Ruger & C. E. Bussenius (1913), Trans.)*. New York: Teachers College, Columbia University.

Ellis, N. C. (1994). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 211–282). San Diego, CA: Academic Press.

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*, 143–188.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 33–68). Oxford, UK: Blackwell.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*, 305–352.

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

Ellis, N. C. (2007). Implicit and explicit knowledge about language. In J. Cenoz (Ed.), *Knowledge about language (Vol. 6: Encyclopedia of Language and Education)*. Heidelberg, Germany: Springer Scientific.

Ellis, N. C. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned-attention, and the limited L2 endstate. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372–405). London: Routledge.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Ellis, N. C. (2015). Implicit AND explicit learning of language. In P. Rebuschat (Ed.), *Implicit and explicit learning of language*. Amsterdam: John Benjamins.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 187–220.

Ellis, N. C., & O'Donnell, M. B. (2012). Statistical construction learning: Does a Zipfian problem space ensure robust language learning? In P. Rebuschat & J. Williams (Eds.), *Statistical learning and language acquisition* (pp. 265–304). Berlin, Germany: Mouton de Gruyter.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2012). Usage-based language: Investigating the latent structures that underpin acquisition. *Currents in Language Learning*, *1*, 25–51.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality. *Cognitive Linguistics*, *25*, 55–98.

Ellis, N. C., & Schmidt, R. (1998). Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax. *Language & Cognitive Processes*, *13*, 307–336.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London, B.*, *268*, 2261–2265.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS*, *100*, 788–791.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1996). *Grammar patterns 1: Verbs. The COBUILD Series*. London: HarperCollins.

Freeman, L. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, *40*, 35–41.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science USA*, *99*, 7821–7826.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*, 289–316.

Gries, S. Th. (2003). Testing the sub-test: a collocational-overlap analysis of English *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics*, *8*, 31–61.

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

Gries, S. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction*. London, New York: Routledge.

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

Gries, S. Th. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*, 137–165.

Gries, S. Th., & Divjak, D. S. (Eds.). (2012). *Frequency effects in cognitive linguistics (Vol. 1): Statistical effects in learnability, processing and change*. Berlin, Germany: Mouton de Gruyter.

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

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

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

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science*, *35*, 399–443.

Hanks, P. (2009). The linguistic double helix: Norms and exploitations. In Hlaváčková, D., Horák, A. Osolsobě, K. & Rychlý, P. (Eds.), *After half a century of Slavonic Natural Language Processing (Festschrift for Karel Pala)* (pp. 63–80). Brno, Czech Republic: Masaryk University.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: MIT Press.

Jaeger, T. F., & Snider, N. E. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the Cognitive Science Society Conference* (pp. 1061–1066). Austin, TX: Cognitive Science Society.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*, 57–83.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwel, D. (2004). The sketch engine. *Proc EURALEX 2004*, Lorient, France., 105–116.

Kolb, P. (2008). *DISCO: A multilingual database of distributionally similar words*. In Proceedings of KONVENS-2008, Berlin, Germany.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 581–611). London: Elsevier.

MacWhinney, B. (1987). Applying the competition model to bilingualism. *Applied Psycholinguistics*, *8*, 315–327.

MacWhinney, B. (1997). Second language acquisition and the competition model. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 113–142). Mahwah, NJ: Erlbaum.

MacWhinney, B. (2001). The competition model: The input, the context, and the brain. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 69–90). New York: Cambridge University Press.

Michelbacher, L., Evert, S., & Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, *5*, 79–103.

Miller, G. A. (2009). *WordNet–About us*. Retrieved March 1, 2010, from http://wordnet.princeton.edu.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2012). Language experienced in utero affects vowel perception after birth: A two-country study. *Acta Paediatrica*, *102*, 156–160.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newman, M. E. J. (2010). *Networks: An introduction*. Oxford, UK: Oxford University Press.
O'Donnell, M. B., Ellis, N. C., Corden, G., Considine, L., & Römer, U. (2015). Using network science algorithms to explore the semantics of verb argument constructions in language usage, processing, and acquisition. Manuscript submitted for publication.
Partington, A. (2011). Phrasal irony: Its form, function, and exploitation. *Journal of Pragmatics*, *43*, 1786–1800.
Pecina, P. (2009). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*, 137–158.
Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). *WordNet::Similarity–Measuring the relatedness of concepts*. Paper presented at the Proceedings of Fifth Annual Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2004).
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329–347.
Pierrehumbert, J. (2006). The next toolkit. *Journal of phonetics*, *34*, 516–530.
Rebuschat, P. (Ed.). (2015). *Implicit and explicit learning of language*. Amsterdam: John Benjamins.
Rebuschat, P., & Williams, J. N. (Eds.). (2012). *Statistical learning and language acquisition*. Berlin, Germany: Mouton de Gruyter.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.
Robinson, P., & Ellis, N. C. (Eds.). (2008). *A handbook of cognitive linguistics and second language acquisition*. London: Routledge.
Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: a corpus analysis. *Journal of Memory and Language*, *57*, 348–379.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature Reviews Neuroscience*, *323*(6088), 533–536.
Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165–210). San Diego, CA: Academic Press.
Shanks, D. R. (1995). *The psychology of associative learning*. New York: Cambridge University Press.

Slobin, D. I. (1997). The origins of grammaticizable notions: Beyond the individual mind. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. *5*, pp. 265–323). Mahwah, NJ: Erlbaum.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

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

Studdert-Kennedy, M. (1991). Language development from an evolutionary perspective. In N. A. Krasnegor, D. M. Rumbaugh, R. L. Schiefelbusch, & M. Studdert-Kennedy (Eds.), *Biological and behavioral determinants of language development* (pp. 5–28). Mahwah, NJ: Erlbaum.

Suslov, I.M. (1992). Computer model of a "Sense of humour" II: Realization in neuronal networks. *Biophysics*, *37*, 249–258.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Boston: Harvard University Press.

Trousdale, G., & Hoffmann, T. (Eds.). (2013). *Oxford handbook of construction grammar*. Oxford, UK: Oxford University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Wills, A. J. (2009). Prediction errors and attention in the presence and absence of feedback. *Current Directions in Psychological Science*, *18*, 95–100.

Wulff, S., Stefanowitsch, A., & Gries, S. T. (2007). Brutal Brits and persuasive Americans: Variety-specific meaning construction in the into-causative. In G. Radden, K.-M. Köpcke, T. Berg, & P. Siemund (Eds.), *Aspects of meaning construction* (pp. 265–281). Amsterdam: John Benjamins.

Xu, F., & Tennenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.