Corpus Linguistics and Psycholinguistics

Stefan Th Gries[a,b], [a] University of California, Santa Barbara, CA, United States and [b] Justus Liebig University Giessen, Giessen, Germany

Abstract

This paper provides a brief survey of the use of corpus-linguistic data and methods/statistics in the domain of psycholinguistics.

Key Points

- This article defines the notion of a corpus.
- It exemplifies how corpora can serve as primary data for psycholinguistic studies.
- It discusses a variety of corpus-based quantitative measures and how they can inform psycholinguistic research.
- It discusses advantages and challenges of conducting psycholinguistic research with corpora.

Introduction

The notion of “a corpus” is a radial/prototype category. At its center are corpora that

- consist of many machine-readable Unicode text files;
- are meant to be representative for a particular kind of speaker, register, genre, topic variety, or language, which means the corpus's sampling scheme represents the variability of the population it is meant to represent;
- contain contextualized data from natural communicative settings, which means the data in the corpus were produced not for the corpus and that their production was untainted by their collection.

That means corpora like the British National Corpus (BNC) or, to a lesser extent, the Corpus of Contemporary American English (COCA) are fairly prototypical examples, whereas collections of essays written by language learners are less prototypical because (i) writing on an assigned topic under time pressure is not the most natural setting and (ii) such data of course do not cover a wide variety of speakers, registers, or genres. Note that my definition of corpus here does not include what might better be called databases, e.g., example collections (such as speech error collections).

With some simplification, corpus linguistics can inform, or contribute to, psycholinguistic studies in two main ways. First, the psycholinguistic study is based on the analysis of examples from corpora, i.e. the corpus data are the primary data of the study. Second, the much more frequent case—because most psycholinguistic work is experimental in nature, see below—is experimental psycholinguistic studies using quantitative information of various kinds from corpora as predictors, control variables, confounds, moderators, or response variables. In the next section, I briefly mention a few example studies of the first type, before I then turn to the kinds of quantitative measures corpus data can provide and how they relate to many notions that psycholinguists are often interested in.

Corpora as Primary Data

While generally in the minority, there are a variety of psycholinguistic studies in which corpora constitute the primary data. For example, there is by now a healthy body of research on **structural priming** that is corpus-based. The maybe earliest such study is actually sociolinguistic in nature: **Sankoff and Laberge (1978)** explore the degree to which speakers choose three pronominal forms in Montreal French in a way that is comparable to runs tests per speaker, and the find that speakers indeed stick to a form more often than expected by chance. More influential from a psycholinguistic perspective was **Estival (1985)**, who found robust priming effects for the English voice alternation even when confounds such as discourse effects or lexical repetitions and others were partial out. A second wave of corpus-based priming research was launched by **Gries (2005)** and **Szmrecsanyi (2006)**. The former studied the dative alternation and particle placement in the British Component of the International Corpus of English and, using linear discriminant analyses with multiple predictors, finds robust priming effects for both alternations, but also that priming effects are confounded by the verbs' preferences for specific constructions. In fact, such lexically specific effects can be extremely predictive: Exploring the alternation of the *will* versus the *going-to* future, **Gries (2016)** finds that 87.7% of all future choices can be predicted correctly just by predicting the most frequent future choice per lexical verb. Szmrecsanyi studies a wider variety of alternations and, using binary logistic regression, also finds robust and lexically-specific priming effects, but also observes that structural priming can be strengthened by other lexical items (as when *more* not used in an analytic comparative still primes analytic comparatives).

Another psycholinguistic area of research sometimes relying on corpora is that of **disfluencies**. One influential study is **Clark and Fox Tree (2002)**, who study the filled pauses *uh* and *um* in a variety of spoken corpora—most notably, the London-Lund corpus (≈170K words), but also smaller, less prototypical corpora—and advance the filler-as-word hypothesis, according to which *uh*'s and *um* 's meanings are to announce minor or major delays in speaking respectively. They then report supporting evidence based on, among other things, the lengths of the filled pauses and their positions as well as delays before and after them. Another example is Rühlemann et al. (2013), who compare four different kinds of pauses (short vs. long and filled vs. unfilled) pauses in the Narrative Corpus, a corpus of narratives (≈150K words) extracted from the spoken-conversation part of the BNC XML. They confirm previous results suggesting that most pauses—but not long silent ones—are more frequent in conversational narratives than in general conversation, and they identify a variety of preferred positions and collocates of pauses, several of which are related to production planning, They, too, infer from their findings that pauses should not be considered disfluencies but instead offer “immense potential for the study of speech and cognition” (p. 87).

Many other applications of corpora as primary data are also possible (and some are mentioned in the following section, too), but given space constraints, let us now turn the more frequent use of corpora, namely as a source of quantitative metrics quantifying various dimensions of psycholinguistic interest.

Corpora Providing Quantitative Information

As mentioned above, corpus data are more commonly used in psycholinguistics as a source of quantitative metrics that are computed from the distributional patterns in corpus data and that operationalize psycholinguistic dimensions of interest. Table 1 provides an overview of various corpus-linguistic statistics that are regularly computed and what they describe or operationalize.

The simplest corpus statistic is that of **token frequencies** of occurrence, which answers the question how often something (a form, a lemma, an expression, etc.) is attested in a corpus or in a relevant part of a corpus (either on its own or with something else). Frequencies are seen as a cause and an operationalization of degree of cognitive entrenchment and, thus, familiarity/commonness, which in turn is correlated with age and speed of acquisition/learning, routinization (and thus grammaticalization), phonetic reduction in speech, and ease and speed of (e.g., lexical) access as manifested in experimentally-obtained reaction/naming times. Frequencies are usually provided in a normalized form (e.g., as frequency per million words) or, maybe more usefully, van Heuven et al.'s (2014) Zipfscale, a scale whose values are both normalized and logged.

Another important kind of frequency is **type frequency**, which answers the question how many different things are attested with or in slots of something, e.g., how many different verbs are attested in the verb slot of the English ditransitive or how many different nouns are attested with a certain suffix representing some case and number combination. Such values, or values derived from them (such as the frequencies of hapaxes in a construction), are positively correlated with the degree of productivity of the linguistic element in question, and arguably high type frequencies in a constructional slot are also positively correlated with the degree of schematicity of the construction: Many more verbs are used with the *going-to* future than with the semantically more restricted ditransitive, for instance.

Crucially, while (token) frequency has probably been the single most predictor in psycholinguistics, evidence is mounting that commonness of element, but also entrenchment as measured by reaction times, maybe be better predicted by **dispersion**, i.e., a statistic that quantifies how evenly an element is distributed over the parts of a corpus and, therefore, how regularly one might be exposed to an element, which in turn is also correlated with recency of exposure. While there is still some debate as to how dispersion is best measured, studies such as Baayen (2010b) or Gries (2022, Gries, 2024, and more indicate that dispersion is important and predictive and suggest that the (causal) role of frequency might have been overestimated for decades.

Table 1 Psycholinguistic notions and corpus-linguistic ways to explore them.

Corpus method/statistic(s)	*Psycholinguistic notion*	*Brief description*
Token frequency of (co-)occurrence	Entrenchment	Cognitive process by which a linguistic pattern is established as a cognitive routine through repetition
Type frequency of (co-)occurrence	Generalization, schemati-city, category formation, & productivity	Cognitive process by which categories are formed through generalization
Dispersion, concordancing	Commonness, recency, & regularity of encounter	The property of something occurring regularly in many contexts so one's chances of encountering something are high
Co-occurrence (collocation, colligation, collostruction), frequency, conditional probabilities, & association measures	Contingency	The degree to which one thing (a form, a feature of a use, a meaning) increases the probability of another thing
	Surprisal	The degree to which one thing is surprising, given the presence or absence of other things
Entropy	Uncertainty	The degree to which many things are equally likely or to which something has no strong preference
Frequency of (co-)occurrence, relative entropy (Kullback-Leibler divergence)	Prototypicality	Degree to which an expression is a central, best, or most representative member of a category (and may serve to anchor/define it)

The second most important corpus-linguist statistic after frequency is the class of **association measures** (AMs), which quantify (in various ways) contingency, i.e., how much one element increases the probability of another element. If one is concerned with the association between words, this is referred to as *collocation*, if one is concerned with the association between words and more abstract—e.g., syntactic—constructions, this is referred to as *collostruction*. Two main different classes of AMs have been most widely used: (i) significance-based measures that are very sensitive to co-occurrence frequency and corpus size such as the log-likelihood statistic G^2, $p_{\text{Fisher-Yates exact}}$, and the *t*-score and (ii) less frequency-sensitive heuristic and effect-size measures such as Pointwise Mutual Information (*PMI*), the logged odds, conditional probabilities and their differences (ΔP), and the information-theoretic measure of the Kullback-Leibler divergence (*KLD*); see **Pecina (2010)**, **Gries, 2024**. Since contingency is the very foundation of associative learning, the establishment of connections between forms, meanings, contexts, etc., AMs are among the most important statistics corpora can provide to measure collocations and collostructions, but also the association of forms to meanings that underlies the acquisition of all linguistic signs; even the most modern current large language models are, in a sense, just very very powerful ways of quantifying contingency.

The next notion, **surprisal**, can be seen as (inversely) related to contingency and association because, in a sense, it quantifies the opposite of association. While high association scores indicate that, given some element, another element is highly expected, high surprisal scores indicate that, given some element, another element is highly surprising. While many different AMs have been developed, surprisal is usually computed as the negative log of a conditional probability -log *p*(element B | element A): If such a probability of B given A is low, the negative log will return a higher positive number than if the probability of B given A is high. In a study of priming of the voice alternation, where corpus data also provided the primary data as discussed in the previous section, **Jaeger and Snider (2008)** show that priming from the priming verb to the target verb is stronger if the priming verb appears in a surprising construction, i.e., a construction that it usually disprefers.

Another information-theoretic statistic is **entropy**, which quantifies the unorderedness, uncertainty, or unpredictability of a distribution: Entropy is high when a set of alternatives—e.g., words or constructions given a certain meaning to be communicated—consists of many nearly equally likely alternatives and it is particularly low when one alternative accounts for most occurrences of the set. **Linzen and Jaeger (2015)**, for example, study expectation-based processing and show that the entropy of potential parse completions is correlated with how quickly subjects read subject-verb sentence beginnings that could be completed with either a direct object or a sentential complement. Specifically, reading times are longer when the entropy in the disambiguation region after verbs is reduced, but their results show that surprisal is also an important predictor of reading times. Finally, **Lester & Moscoso del Prado Martín (2017)** provide some truly fascinating entropy-related findings. For example, they find that entropies of syntactic distributions of nouns—i.e., how many different syntactic contexts does a noun occur in and how evenly so?—has an impact on how fast speakers read that noun even when it is presented without context: "nouns that provide more possibilities for expanding their phrasal nodes [i.e., with a higher entropy] could be accessed more quickly because their selection does not immediately commit the speaker to any particular phrasal structure" (p. 2589).

The final measure to be discussed here is the **relative entropy**, i.e. the *KLD* mentioned above in the context of association measures. While the *KLD* can be used for dispersion, association, keyness, and many other scenarios (see **(Gries, 2024)**), two applications are particularly interesting. The first one is discussed in **McDonald and Shillcock (2001)**, who use the *KLD* as a measure of contextual distinctiveness, i.e. the degree to which a word affects its linguistic context. Specifically, they measure how much the frequency distribution of words around a node word of interest diverges from the overall frequency distribution of words, and they show that this measure does not simply follow from existing measures but explains reaction times to words better than (logged) frequency does. A second application of relative entropy has been proposed in **Milin et al. (2009)**, who show that individual Serbian nouns from a given inflectional class are recognized faster if the frequency distribution of their inflectional forms does not diverge much from the frequency distribution of inflectional forms of the whole class. **Baayen et al. (2011)** and **Lester (2017)** conclude from this that the frequency distribution of the whole inflectional class is its prototype, the default expectation of the processor for the class, and **Gries (2018)** then shows that such lexical-distribution prototypes for members of an alternation can strongly boost the performance of predictive models in alternation studies.

Challenges and Advantages of Corpus Data

Given the overall predominance of experimental work over corpus-based work in psycholinguistics, it is useful to briefly consider advantages and disadvantages of the latter, especially since it was argued in the past that corpus data may not have much to contribute to psycholinguistic studies (of, say, priming, see **Pickering & Branigan, 1999**).

Advantages of Corpus Data

The maybe biggest advantage of corpus data is their potential for results with a **higher ecological validity**: Experimental studies are usually conducted in highly controlled conditions and based on carefully elicited/selected and often decontextualized data that might not at all be representative of anything. In contrast, as an authentic/spontaneous and highly contextualized type of data, corpus data can come with a higher level of ecological validity. A second, related advantage is that corpora help **avoid imbalanced input distributions** that experiments often entail. Such controlled (i.e., well balanced) designs of experiments often mean that participants are exposed to unrepresentative distributions of the investigated linguistic elements, which can be problematic given that learning or habituation effects, but now based on *un*representative input, can be observed even over a small numbers of experimental stimuli. A final advantage is that corpora often force researchers to face **unexpected data**. It is not at all uncommon that corpus data reveal a wider variety of expressions or uses of expressions than analysts might have considered or even expected to be possible. For example, it has often been assumed that *donate* cannot be used ditransitively, but **Stefanowitsch (2007)** finds multiple instances of such uses in web-scraped that are even consistent enough to give rise to what he calls a DONATE frame. Thus, corpora can reveal patterns in data that might challenge old assumptions and, thus, spark new research.

Challenges of Corpus Data

The above notwithstanding, it is only fair to point out that corpus data also come with one major challenge: the fact that they are **noisy heterogeneous intercorrelated data**. Linguistic contexts vary considerably across the uses of a given linguistic element both within and across speakers and data distributions can be problematic because of the frequently highly imbalanced and Zipfian distributions: frequencies of elements on their own and with others decrease as a power function of their rank in the frequency table. On top of that, corpus data often present challenges in terms of (i) collinearity between predictors and (ii) the often necessary inclusion of many control variables to control statistically for what, in an experiment, would have been controlled for, or held constant by, experimental design. Addressing these kinds of problems can require complex statistical analyses that go much beyond the often straightforward balanced factorial designs that experimental psycholinguists are used to (see **Baayen, 2010a**), and with such analyses, corpus data arguably can address some of the concerns that have been raised again them.

Conclusion

While this overview could only be very selective, it has hopefully become clear that corpus data do have a lot to offer to various areas in contemporary psycholinguistics—like any method, they have advantages and disadvantages and striking the right balance between those when it comes to methodological choices is of course important. However, the wide range of distributional data that corpora provide as well as the wide range of distributional statistics computed from them have proven increasingly useful to the operationalization or exploration of many psycholinguistic notions; that alone, as well as the importance of validating experimental data with corpus data and vice versa (see **Gullberg et al., 2009**) will hopefully inspire more corpus-linguistic research in and for psycholinguistics.

Uncited References

Rühlemann, Bagoutdinov, & O'Donnell, 2013.

References

Baayen, R.H. (2010a). A real experiment is a factorial experiment? The Mental Lexicon, 5(1), 149–157.

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

Baayen, R.H., Milin, P., Dusica Filipović-Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychological Review, 118(3), 438–481.

Clark, H.H., & Fox Tree, J.E. (2002). Using *uh* and *um* in spontaneous speaking. Cognition, 84(1), 73–111.

Estival, D. (1985). Syntactic priming of the passive in English. Text—Interdisciplinary Journal for the Study of Discourse, 5, 7–22.

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

Gries, S.T. (2016). Variationist analysis: Variability due to random effects and autocorrelation. In Baker, P., & Egbert, J.A. (Eds.), Triangulating methodological approaches in corpus linguistic research (pp. 108–123). New York & London: Routledge.

Gries, S.T. (2018). The discriminatory power of lexical context for alternations: An information-theoretic exploration. Journal of Research Design and Statistics in Linguistics and Communication Science, 5(1–2), 78–106.

Gries, S.T. (2022). What do (most of) our dispersion measures measure (most)? Dispersion? Journal of 2nd Language Studies, 5(2), 171–205.

Gries S.T.. 2024 *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures* PhiladelphiaAmsterdam &

Gullberg, M., Indefrey, P., & Muysken, P. (2009). Research techniques for the study of code-switching. In Bullock, B.E., & Toribio, A.J. (Eds.), The cambridge handbook of linguistic code-switching (pp. 21–39). Cambridge: Cambridge University Press.

Jaeger, T.F., & Snider, N. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In Love, B.C., Kenneth McRae, K., & Sloutsky, V.M. (Eds.), Proceedings of the 30th cognitive science society conference (pp. 1061–1066).

Lester, N.A. (2017). The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition Ph.D. dissertation. Santa Barbara: University of California.

Lester, N.A., & Fermín Moscoso del Prado, M. (2017). Syntactic flexibility in the noun: Evidence from picture naming. In Papafragou, A., Grodner, D.J., Mirman, D., & Trueswell, J. (Eds.), Proceedings of the 38th annual conference of the cognitive science society (pp. 2585–2590).

Linzen, T., & Jaeger, P.F. (2015). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. Cognitive Science, 40(6), 1382–1411.

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), 295–323.

Milin, P., Dusica Filipović-Đurđević, D., & Fermín Moscoso del Prado, M. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. Journal of Memory and Language, 60(1), 50–64.

Pecina, P. (2010). Lexical association measures and collocation extraction. Language Resources and Evaluation, 44(1–2), 137–158.

Pickering, M.J., & Branigan, H.P. (1999). Syntactic priming in language production. Trends in Cognitive Sciences, 3(4), 136–141.

Rühlemann, C., Bagoutdinov, A., & O'Donnell, M.B. (2013). Windows on the mind: Pauses in conversational narrative. In Gilquin, G., & De Cock, S.

(Eds.), Errors and disfluencies in spoken corpora: Crosslinguistic perspectives (pp. 59–91). Amsterdam & Philadelphia: John Benjamins.

Sankoff, D., & Laberge, S. (1978). Statistical dependence among successive occurrences of a variable in discourse. In Sankoff, D., & Laberge, S. (Eds.), Linguistic variation: Methods and models (pp. 119–126). New York: Academic Press.

Stefanowitsch, A. (2007). Linguistics beyond grammaticality. Corpus Linguistics and Linguistic Theory, 3(1), 57–71.

Szmrecsanyi, B. (2006). Morphosyntactic persistence in spoken English: […]. Berlin & New York: Mouton de Gruyter.

van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology, 67(6), 1176–1190.