Broad consistency but cross-linguistic variation in the structure of information in sentences

Anonymous CogSci submission

Abstract

Optimal coding theories of language predict that speakers should keep the amount of information in their utterances relatively uniform under the constraints imposed by their language. But how much of a role do these constraints provide, and does it vary across languages? We find a consistent non-uniform shape which characterizes both spoken and written sentences of English but is tempered by predictive context. We then show that other languages are also characterized by consistent but non-English shaped curves related to their typological features, but that sufficient context produces more uniform shapes across languages. Thus, producers of language appear to structure their utterances in similar near-uniform ways despite varying linguistic constraints.

Keywords: information theory; efficient communication; language typology; computational modeling

We use language for a variety of purposes like greeting friends, making records, and signaling group identity. These purposes all share a common goal: Transmitting information that changes the mental state of the listener (Austin, 1975). For this reason, language can be thought of as a code, one that allows speakers to turn their intended meaning into a message that can be transmitted to a listener, and subsequently converted by the listener back into an approximation of the intended meaning (Shannon, 1948). How should we expect this code to be structured?

If language has evolved to be a code for information transmission, its structure should reflect this process of optimization (Anderson & Milson, 1989). The optimal code would have to work with two competing pressures: (1) For listeners to easily and successfully decode messages sent by the speaker, and (2) For speakers to easily code their messages and transmit them with minimal effort and error. A fundamental constraint on both of these processes is the linear order of spoken language–sounds are produced one at a time and each is unavailable perceptually once it is no longer being produced.

Humans accommodate this linear order constraint through incremental processing. People process speech continuously as it arrives, predicting upcoming words and building expectations about the likely meaning of utterances in real-time rather than at their conclusion (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Since prediction errors can lead to severe processing costs and difficulty integrating new information on the part of listeners, speakers should seek to minimize the chance of prediction errors when choosing what to say. However, the cost of producing more predictable utterances and thus minimizing the likelihood of errors is using more words. Therefore the optimal strategy for speakers seeking to minimize their production costs is to produce utterances that are just at the prediction capacity of listeners without exceeding this capacity (Genzel & Charniak, 2002). In other words, speakers should consistently transmit information as close to the listener's fastest decoding rate as possible.

This Uniform Information Density hypothesis has found support at a variety of levels of language from the structure of individual words to the syntactic structure of utterances (see Gibson et al., 2019 for a review). Further, speakers make online word choices that smooth out the information in their utterances (Jaeger & Levy, 2007). While speakers can make bottom-up choices such as controlling which of several near-synonyms they produce, they cannot control the grammatical properties of their language. Properties like canonical word order impose top-down constraints on how speakers can structure what they say. While speakers may produce utterances as uniform in information density as their languages will allow, these top-down constraints may create significant and unique variation across languages.

How significant are a language's top-down constraints on speakers? Yu, Cong, Liang, & Liu (2016) analyzed how the information in words of English sentences of a fixed length varies with their order in the sentence (e.g. first word, second word, etc). They found a surprising non-linear shape, and argued that this shape may arise from top-down grammatical constraints on language. We build on these ideas, asking (1) Whether this shape depends on listener's predictive models, (2) Whether this shape varies across linguistic contexts, and (3) Whether this shape is broadly characteristic of a diverse set of languages or varies predictably from language to language. We find that languages are characterized by highly-reliable but cross-linguistically variable information structures that co-vary with top-down linguistic features. Listeners' predictive coding flattens these shapes across languages, in accord with predictions of the Uniform Information Density hypothesis.

Study 1: Information in Written English

An influential early test of the Uniform Information Density hypothesis was performed by Genzel & Charniak (2002), who analyzed the amount of information in successive sen-

tences of the same text. They found that the amount of information increased across sentences when each was considered in isolation. They reasoned that since all prior sentences provide the context for reading each new sentences, the amount of total information (context + paragraph) was overall constant for human readers.

Yu et al. (2016) applied this same logic to analysis of the information in individual sentences, computing the entropy of each successive word in an utterance. The first word of each sentence tended to contain little information, while words in the middle of sentences each contained roughly the same amount of information as one another, and the final word of each sentence contained much more information than any other word. They found the same distribution across sentence lengths, from sentences with 15 words to sentences with 45 words. They took this as evidence against the Uniform Information Density Hypothesis as, unlike Genzel and Charniak's (2002) results, information plateaued in the middle of sentences. We replicate their analysis here, and build an additional model to bring their analysis more in line with Genzel and Charniak's (2002) methods. Finally, we also develop a method for averaging the curves for sentences of different lengths together to provide a single typical information structure.

Data

Following Yu et al. (2016), we selected the British National Corpus (BNC) for analysis (British National Corpus Consortium, 2007). The British National Corpus is ˜100 million word corpus consisting of spoken (10%) and written (90%) English from the late 20th Century.

Pre-processing

We began with the XML version of the corpus, and used the `justTheWords.xsl` script provided along with the corpus to produce a text file with one sentence of the corpus on each line. Compound words (like "can't") were combined, and all words were converted to lowercase before analysis. This produced a corpus of just over six million utterance of varying lengths. From these, we excluded utterances that were too short to allow for reasonable estimation of information shape (fewer than 5 words), and utterances that were unusually long (more than 45 words). This exclusion left us with 89.83% of the utterances (Fig 1.

Estimating information

To estimate how information is distributed across utterances, we computed the lexical surprisal of each word under two different models (Levy, 2008; Shannon, 1948). Intuitively, the surprisal of a word is a measure of how unexpected it would be to read that word, and thus how much information it contains. First, following Yu et al. (2016), we estimated a unigram model which considers each word independently: $\text{surprisal}(\text{word}) = -\log P(\text{word})$. This unigram surprisal measure is a direct transformation of the word's frequency and thus less frequent words are more surprising.

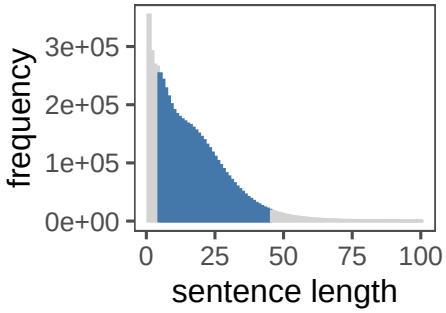

Figure 1: The distribution of sentence lengths in the British National Corpus. We analyzed sentences of length 5-45 (colored).

Second, we estimated a trigram model in which the surprisal of a given word (w_i) encodes how unexpected it is to read it after reading the prior two words (w_{i-1} and w_{i-2}): $\text{surprisal}(w_i) = -logP(w_i|w_{i-1}, w_{i-2})$. This metric encodes the idea that words that are low frequency in isolation (e.g. "meatballs") may become much less surprising in certain contexts (e.g. "spaghetti and meatballs") but more surprising in others (e.g. "coffee with meatballs"). The difficulty of correctly estimating these probabilities from a corpus grows combinatorically with the number of prior words, and in practice trigram models perform well as an approximation (see e.g. Chen & Goodman, 1999; Smith & Levy, 2013).

Model details We estimated the surprisal for each word type in the British National Corpus using the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013). Each utterance was padded with a special start-of-sentence word "$\langle s \rangle$" and end of sentence word "$\langle /s \rangle$". Trigram estimates did not cross sentence boundaries, so for example the surprisal of the second word in an utterances was estimated as $(\text{surprisal}(w_2) = -P(w_2|w_i, \langle s \rangle)$. Naïve trigram models will underestimate the surprisal of words in low-frequency trigrams (e.g. if the word "meatballs" appears only once in the corpus following exactly the words "spaghetti and", it is perfectly predictable from its prior two words). To avoid this underestimation, we used modified Kneser-Ney smoothing: this method discounts all ngram frequency counts and interpolates lower-order ngrams into the calculations. These lower-order ngrams are weighted according to the number of distinct contexts they occur as a continuation (see Chen & Goodman, 1999).

Averaging curves To develop a characteristic information curve for sentences in the corpus, we needed to aggregate sentences that varied dramatically in length (Fig **??**A). We used Dynamic Time Warping Barycenter Averaging (DBA), an algorithm for finding the average of sequences that share and underlying pattern but vary in length (Petitjean, Ketterlin, & Gançarski, 2011). DBA inverts standard dynamic time warping, discovering a latent invariant template from a set of sequences.

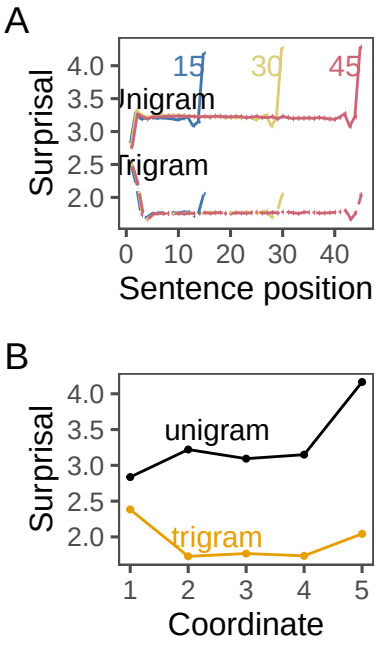

Figure 2: (A) Surprisal by sentence position of length 15, 30, and 45 sentences in the British National Corpus under unigram and trigram surprisal models. Error bars indicate 95% confidence intervals (tiny due to sample size). (B) Characteristic information curves produced by the DBA algorithm averaging over all sentence lengths in each corpus.

We used DBA to discover the short sequence of surprisal values that characterized the surprisal curves common to sentences of varying sentence lengths. We first averaged individual sentences of the same length together and then applied the DBA algorithm to this set of average sequences. DBA requires a parameter specifying the length of the template sequence. We chose a 5-length template sequence based on our inspection of the curves from the British National Corpus as well as the curves we found in subsequent studies. However, the results of this and the following studies were robust to other choices of this length parameter (7 and 10).

Results and Discussion

We began by replicating Yu et al.'s (2016) analyses, examining the surprisal of words in sentence of length 15, 30, and 45 estimated by our unigram model. In line with their computations, we found a reliably non-linear shape in sentences of all 3 lengths, with the information in each word rising for the first two words, plateauing in the middle of sentences, dipping in pen-ultimate position, and rising steeply on the final word (Fig. 2A).

In comparison, under the trigram model we observed 3 major changes. First, each word contained significantly less information. This is to be expected as the knowing two prior words makes it much easier to predict the next word. Second, the fall and peak at the ends of utterances was still observable, but much less pronounced. Finally, the first word of each sentence was now much more surprising than the rest of the words in the sentence, because the model had only the start of sentence token $\langle s \rangle$ to use as context. Thus, the trigram model likely overestimates the information for humans reading the first word. Together, these results suggest that Yu et al. (2016) overestimated the non-uniformity of information in sentences. Nonetheless, the final words of utterances do consistently contain more information than the other words.

Fig. 2B shows the barycenter produced by DBA. The algorithm correctly recovers both the initial and final rise in information under the unigram model, and the initial fall and smaller final rise in the trigram model. We take this as evidence that (1) these shapes are characteristic of all lengths, and (2) that DBA effectively recovers characteristic information structure. In sum, the results of Study 1 suggest that sentences of written English have a characteristic non-uniform information structure, with information rising at the ends of sentences. This structure is more pronounced when each word is considered in isolation, but some of the structure remains even when each word is considered in context. Is this structure unique to written English, or does it characterize spoken English as well? In Study 2, we apply this same analysis to two corpora of spoken English–the first of adults speaking to other adults, and the second of adults and children speaking to each other.

Study 2: Information in Spoken English

Spoken language is different from written language in several respects. First, the speed at which it can be processed is constrained by the speed at which it is produced. Second, speech occurs in a multimodal environment, providing listeners information from a variety of sources beyond the words conveyed (e.g. prosody, gesture, world context). Finally, the both words and sentence structures tend to be simpler in spoken language than written language as they must be produced and processed in real-time (Christiansen & Chater, 2016). Thus, sentences of spoken English may have different information curves than sentences of written English.

The language young children hear is further different from the language adults speak to each other. Child-directed speech tends to simpler than adult-directed speech on a number of dimensions including the lengths and prosodic contours of utterances, the diversity of words, and the complexity of syntactic structures (Snow, 1972). The speech produced by young children is even more distinct from adult-adult speech, replete with simplifications and modifications imposed by their developing knowledge of both the lexicon and grammar (Clark, 2009). In Study 2, we ask whether spoken English–produced both by adults and children– has the same information structure as written English.

Data

To estimate the information in utterances of adult-adult spoken English, we used the Santa Barbara Corpus of Spoken American English, $\sim$ 250,000 word corpus of recordings of naturally occurring spoken interactions from diverse regions

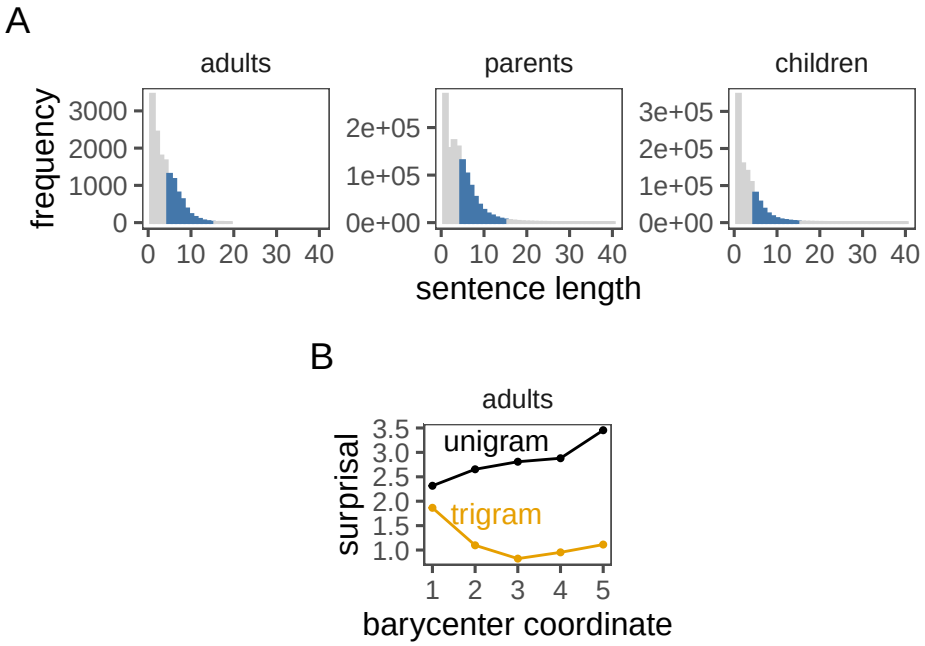

Figure 3: (A) The distribution of sentence lengths in the spoken English corpora: Adults in Santa Barbara, and parents and children in CHILDES. We analyzed sentences of length 5-15 (colored). (B) Charateristic surprisal curves for these corpora.

of the United States (Du Bois, Chafe, Meyer, Thompson, & Martey, 2000). For parent-child interactions, we used all of the North American English corpora in the Child Language Data Exchange System (CHILDES) hosted hosted childes-db (MacWhinney, 2000; Sanchez et al., 2019). We selected for analyses all $\sim$ 1 million utterances produced by children (mostly under the age of five), and $\sim$ 1.7 million utterances produced by the parents of these children.

Data Processing

All pre-processing and modeling details were identical to Study 1 except for the selection of sentences for analysis. Because the utterances in both the Santa Barbara Corpus and CHILDES were significantly shorter than the sentences in the British National Corpus, we analyzed all utterances of at least 5 and most 15 words (see Fig. 3A). Models were estimated separately for each of the 3 corpora.

Results and Discussion

The information curves found in adults-adult utterances were quite similar to those of parent-child utterances and child-parent utterances (Fig. 3B). Under the unigram model, information rose steeply in the beginnings of utterances, was relatively flatter in the middle of utterances, and the rose even more steeply at the ends. Under the trigram model, the first parts words of sentences contained the most information, information was relatively constant in the middle of utterances, and then rose slightly again at the ends.

Unfortunately, we cannot compare amount of information across corpora–surprisal is highly correlated with corpus size (e.g. there is less information in adults' speech in Santa Barbara than in children's speech in CHILDES). However, we can compare the shapes of these curves both to each-other and to the written English sentences in Study 1 2B. All of these curves appeared to share their important qualitative features, including the sharp rise at the end under the unigram model and the attenuation of this rise under the trigram model. There are small differences–such as the flatter shape in the middle of written sentences than spoken utterances, but this difference is pronounced in the utterances of the Santa Barbara corpus relative to utterances of parents in CHILDES, suggesting that it may be partly a function of corpus size.

Thus, English–both written and spoken, both produced by adults and by children-appears to have a characteristic shape. Are the features of this shape features of English, or features of language more broadly? In Study 3 we apply this technique to a diverse set of written languages of different families to ask whether these structures vary cross-linguistically.

Study 3: Cross-linguistic variation in information

Data

To measure cross-linguistic variation in the structure of information across sentences, we constructed a corpus of Wikipedia articles from all languages with at least 10,000 articles. This resulted a set of 152 languages from 16 families. We then used two measures of lexical similarity to investigate cross-linguistic variation in information curves.

To target lexical differences between languages, we used the 40-item Swadesh word list consisting of basic concepts appearing in all langauges (Swadesh, 1955; Wichmann et al.,

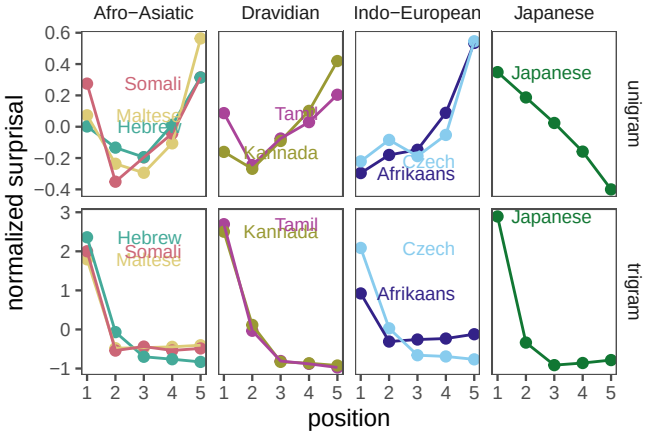

Figure 4: Characteristic information curves (centered) for a sample of languages from Wikipedia

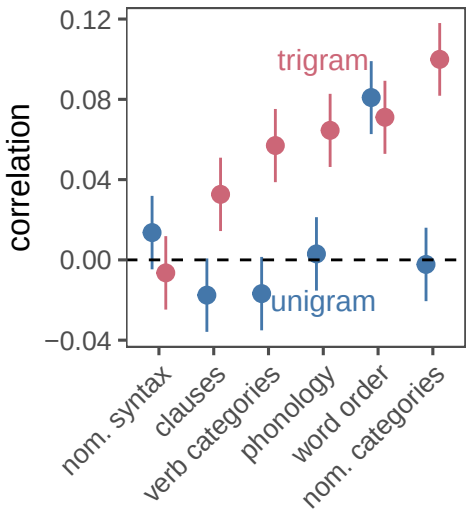

Figure 5: Pairwise correlations between languages' centered information curves and the number of linguistic features they share of each type. Error bars indicate 95% CIs.

2016). The more similar two languages' words are, the more similar the two languages are. To understand the relationship between information curves and typological features, we used the World Atlas of Language Structures, a collection of morphological, syntactic, phonological, and other features of linguistic structure (WALS; Dryer & Haspelmath, 2013). As WALS is a compiled database from dozens of papers from different authors, most features for most languages are fairly sparse. We use a iterative imputation algorithm for categorical data Multiple Imputation Multiple Correspondence Analysis to fill in the missing features (MIMCA; Audigier, Husson, & Josse, 2017).

Data processing

All processing was identical to Studies 1 and 2 except for the lengths of utterances chosen for analyses. To accommodate the variety of lengths across language corpora, we analyzed sentences of lengths 5 to 30.

For each pair of languages, we derived three pairwise similarity measures. To estimate the information structure similarity, we first centered each language's 5-point curve (since surprisal is highly correlate with corpus size), and then computed the cosine similarity between the two centered curves. To estimate Swadesh similarity, we computed the average normalized Levenshtein distance between each of the 40 words. Finally, to compare typological similarity, we added the number of features two languages shared.

Results and Discussion

Unlike the striking consistency across multiple English corpora, we found significant variability in the structure of information curves across languages estimated under the unigram model. Fig. 4 shows centered information curves for a sample of languages from several language families. Despite this variability under the unigram model, the shapes of information curves estimated under the trigram model were more similar cross-linguistically and to the shapes found in English: Sentences began with highly informative words (presumably due to lack of context) and then rapidly approached a uniform level of information. Consistent with this characterization, pairwise similarities in languages' information curves estimated under the unigram model were more correlated with their Swadesh distances (r = -0.11, t = -10.27, $p <$ .001) than their distances estimated under the trigram model ($r = 0.03$, $t = 2.33$, $p = .020$).

While the trigram information curves have a more consistent qualitative shape, there are differences between languages. The pairwise similarities between languages' trigram information curves were more correlated with the number of WALS features they shared ($r = 0.12$, $t = 13.35$, $p < .001$) than their similarities estimated under the trigram model ($r = 0.03$, $t = 2.72$,$p = .007$).

To understand which typological features contribute to these similarities, we split the WALS features by type, with categories such as nominative categories and nominative syntax describing morphology while word order describes subject-verb-object and head-modifier word orders. Fig. 5 shows the correlation between the similarity of information curves under both the unigram and trigram models and the number of features of each of these types two-languages shared. Under the unigram model, word order features appear to predict information curve similarity. In contrast, under the trigram model, all features types except for nominative syntax are reliably correlated with information curve similarity.

Taken together, these results suggest three broad conclusions. First, aspects of the history of languages–encoded in their lexicostatistics–structure the shapes of information in typical sentences. When each word in a sentence is considered alone, languages vary quite dramatically in how information is distributed across sentences. Second, a diverse set of typological features of languages are related to how information is structured for listeners who bring predictive

processing to language. These features appear to explain a small but reliable proportion of the variation in how uniformly information is distributed across utterances. Finally, despite this variation, two words of predictive context radically transform the structure of information in utterances, leading to significantly more uniformity in all languages irrespective of their typological structure. These analyses suggest that top-down constraints from language do play an important role in structuring speakers' utterances, but the speakers have tremendous power to choose efficient utterances within these constraints.

Conclusion

In this paper we have proposed a novel method for quantifying how information is typically distributed across the words of a sentence. We showed that English–whether written or spoken, whether produced by adults or children–has a prototypical information structure. We then showed that this information structure varies cross-linguistically in predictable ways, but also shares broad similarities. These results add to a growing body of research using information theory to explore the structure of language. While much of this work has considered the choices speakers make to structure their utterances within the constraints imposed by their linguistic codes, our work adds to this research program by investigating these constraints themselves. These results represent a small first step towards answering the question of how much these constraints shape speakers' productions, and how speakers interact with them.

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Audigier, V., Husson, F., & Josse, J. (2017). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, *27*(2), 501–518.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

British National Corpus Consortium. (2007). *British national corpus version 3 (BNC XML edition)*. Oxford: Oxford University Computing Services.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 199–206).

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 690–696).

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

MacWhinney, B. (2000). *The childes project: The database* (Vol. 2). Psychology Press.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678–693.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, 549–565.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, *21*(2), 121–137.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., ... others. (2016). The asjp database. *Max Planck Institute for the Science of Human History, Jena*.

Yu, S., Cong, J., Liang, J., & Liu, H. (2016). The distribution of information content in english sentences. *arXiv Preprint arXiv:1609.07681*.