Top-down language features predict little variation in information structure across languages

Anonymous CogSci submission

Abstract

What role does communicative efficiency play in how we organize our utterances? In this paper, we present a novel method of examining how much information speakers in a given language communicate in each word, surveying numerous diverse languages. We find that speakers produce frequent and informative words at regular parts of their utterances, depending on language they use. The information distribution for each language is derived in part from the top-down features of the language, but largely derives from bottom-up choices made by speakers during communication. Each robust information distribution characterizes both spoken and written communication, and emerges in children's earliest utterances. However, in real-time communication, in-context word predictability allows listeners to process information at a constant, optimal rate, regardless of information distribution.

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

Introduction

We use language for a variety of purposes like greeting friends, making records, and signaling group identity. But, these purposes all share a common goal: Transmitting information that changes the mental state of the listener or reader (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 or reader, 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 accomodate 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 (Kutas & Federmeier, 2011; Pickering & Garrod, 2013; 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 prediction errors. However, the cost of producing more predictable utterances is using more words. Thus, 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 (Aylett & Turk, 2004; Genzel & Charniak, 2002). In other words, speakers should maintain a constant rate information of 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 (Jaeger & Levy, 2007; Piantadosi, Tily, & Gibson, 2011; see Gibson et al., 2019 for a review). Further, speakers make lexical choices that smooth out the information in their utterances (Jaeger & Levy, 2007; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). However, while speakers can control which of several near-synonyms they produce, or whether to produce an optional complementizer like "that," they cannot control the grammatical properties of their native language like canonical word order that impose top-down constraints on the structure of utterances. While speakers may produce utterances as uniform in information density as their languages will allow, these top-down constraints may impose significant variation.

How significant are these top-down constraints? One previous paper analyzed the information content in English sentences and found a surprising three-step shape where information first rises, then plateus, and then sharply rises again at the ends of sentences (Yu, Cong, Liang, & Liu, 2016). 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-linguisticly variable structures that co-vary with typological features, but that predictive coding flattens these shapes across languages, in accord with predictions of the Uniform Information Density hypothesis.

Study 1: Information in Written English

The Uniform Information Density Hypothesis predicts that people should structure their utterances so that the amount of information in each unit of language remains constant. An influentual early test of this idea was performed by Genzel & Charniak (2002) who analyzed the amount of information in successive paragraphs of the same text. They found that the amount of information increased across paragraphs when each was considered in isolation. They reasoned that since all prior paragraphs provide the context for reading each new paragraph, the amount of total information (context + paragraph) was 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. Surprisingly, they found found a distinctive three-step distribution for information in a corpus of written English. The first word of each sentence tended to contain little information; words in the middle of sentences each contained roughly the same amount of information, nut 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 plateud in the middle of sentences rather than increasing as it did at the beginnings and ends.

We replicate their analysis here, bringing it more in line with Genzel and Charniak's (2002) methods. While Yu et al. (2016) considered only the information in each word, we build two different models. The first, replicating their analysis, considers the surprisal of each word read in isolation. The second, following Genzel & Charniak (2002), is a trigram model which considers the surprisal of each word having read the prior two words. We take this trigram model as a better proxy for human reader's processing of these sentences. We also develop a method for averaging the curves for sentences of different lenghts together to provide a single typical information curve signature.

Corpus

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 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 lenghts. 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.

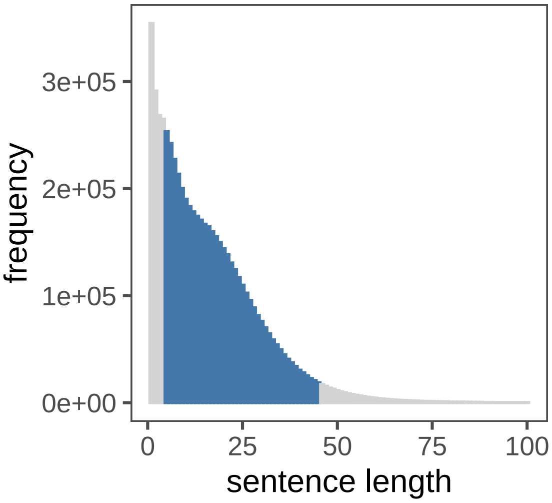

Figure 1: Sentence length distribution in the British National Corpus. Lengths included in analysis are dark.

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, asking how unexpected that word would be in the absence of any context: $\text{surprisal}(\text{word}) = -\log P(\text{word})$. This unigram surprisal measure is a direct transformation of the word's frequency and thus less frequent words are more surprising.

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"). In principle, we would like to encode the surprisal of a word given all of the prior sentential context ($\text{surprisal}(w_i) = -P(w_i|w_{i-1}w_{i-2}...w_1)$. However, the difficulty of correctly estimating these probabilities from a corpus grow 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 ($\mathrm{surprisal}(w_2) = -P(w_2|w_i, \langle s\rangle)$).

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 which discounts all ngram frequency counts–reducing the impact of rare ngrams on probability calculations–and interpolates lower-order ngrams into the calcuations. These lower-order ngrams are weighted according to the number of distinct contexts they occur as a continuation (e.g. “Francisco” may be a common word in a corpus, but likely only occurs after “San” as in “San Francisco”, so it receives a lower weighting; 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 **??**). 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 is an extension of standard Dynamic Time Warping, which searches for an invariant template in shifted and stretched instances of that template [e.g. all instances of the vowel ‘a’ in acoustic recordings of a speaker’s productions even if the instances vary in their duration and intensity]. DBA inverts this idea, discovering a latent invariant template from a set of sequences.

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 5 as the length of the 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

We replicate Yu et al.’s (2016) result using the surprisal metric in place of the entropy metric. We use the frequency-based or “contextless” surprisal metric, which determines the average distribution of information based on word frequencies in a corpus. A priori we expect that the frequency-based metric will produce a flat distribution of information across word positions in the BNC. We find the same frequency-based information trajectory as Yu et al. with little information in the first words of utterances and the most information in the final word, see Figure **??**fig:bnc_raw).

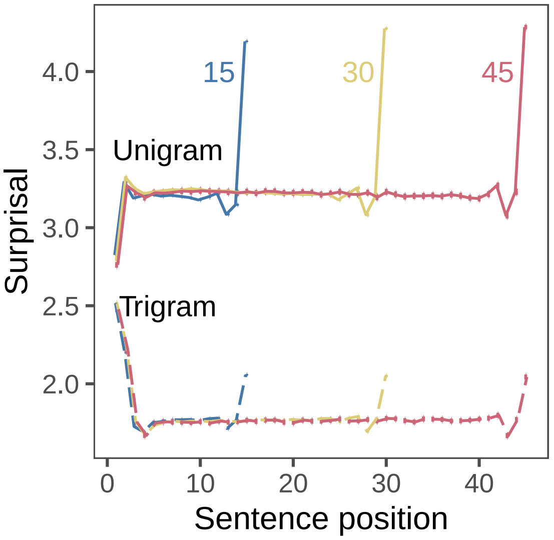

Figure 2: Surprisal by sentence position in the British National Corpus. Error bars indicate 95% confidence intervals (sometimes invisible because of precision)

We now include two words of context (trigrams) for each word in our measurements. We observe a flattening effect of context for both spoken and written English. After the first word or two, where the listener does not have access to prior context, then they decode information at a flat and more or less uniform rate. The contextual information curve for the BNC is in Figure @ref(fig:bnctrigrams).

We now show that spoken English from adults, parents and children all follows the same characteristic frequency-based and context-based information distributions as written English. English as a whole is characterized by the three-step frequency-based distribution, and context enables the listener or reader to process English at a nearly constant rate.

Experiment 2: English in Other contexts

Spoken language and written language diverse in a number of respects. Speech in general is produced and therefore can only be processed at a slower rate than the written word. Speech occurs in a multimodal environment, along with other features such as prosody and visual information such as gesture and information from the environment. The words in speech are often different from the set of words writers draw from in a particular language: writers do not have to obey the same cost of production or the now-or-never bottleneck in comprehension [CITE CHRISTIANSEN and CHATER]. See Figure [WHICH FIGURE] for the distribution of word lengths in the Santa Barbara corpus of spoken conversations between adult strangers (Du Bois, Chafe, Meyer, Thompson, & Martey, 2000). Compared to Figure

Child language input before children learn to read [AT WHAT AGE] is entirely spoken, and children’s main interlocutors are their parents. In particular, the language input children receive may differ in terms of the vocabulary, rate

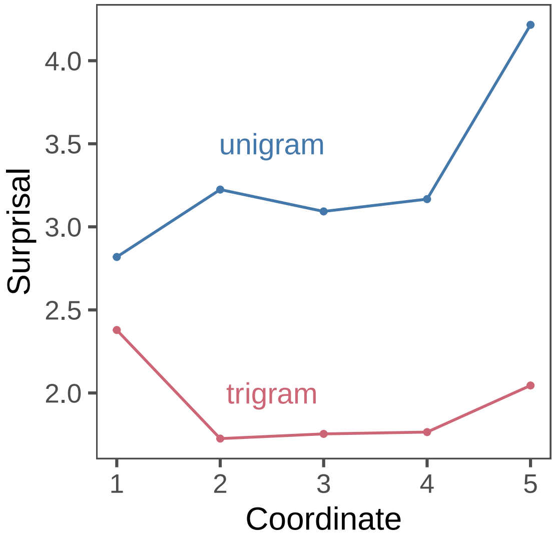

Figure 3: Charateristic surprisal curves for the British National Corpus

of speech, length of words and syntax from adult speech [CDS CITATION]. In North American English, speech to children is often simplified [CDS CITATION]. In return, child speech is simplified: children typically do not speak their first multiword utterances until their second birthday [IS THIS RIGHT?]. Much complex children's language input actually comes from children's books [CITE Jessica Montag]. See Figure [WHICH FIGURE] for a typical distribution of word lengths in child-directed and child speech. The vast majority of utterances are only a few words long.

In this context, we will show that despite the differences in utterance length, vocabulary and grammatical complexity between spoken and written English, we observe the same information trajectory in both modalities. For spoken English, we use the Santa Barbara corpus of spoken English telephone conversations (Du Bois et al., 2000). We use the North American English collection of corpora from CHILDES (MacWhinney, 2000), which includes `1.07` million child utterances mainly before the age of five, and `1.7` million parent child-directed utterances. We apply a similar cleaning procedure to the one we used for the BNC. We lowercase and remove punctuation from all utterances in the corpora. To obtain the CHILDES utterances, we use childesr [childesr citation], a frontend in the R programming language [CITE R core team].

We observe the exact same frequency-based distribution for our information curves in adult conversation, parent speech to children and children's speech as in spoken English. This indicates that the three-step frequency-based distribution in English holds for all modalities and ages in the language. Children, from their first multiword utterances, tend to produce words according to the information distribution we have found, and continue following that distribution when they grow up and speak to other adults. Final words in English are important in child-directed speech [CITE ASLIN PAPER].

We observe that the contextual distribution is likewise similar to the distribution we found for the BNC: high and then immediately flattening out once the listener has a word or two of context to predict the word their interlocutor will produce next.

This only captures the picture for English. In our next experiment, we find unique frequency-based distributions for each language, which are determined in part by top-down features in the language such as word order. We will show a similar contextual smoothing effect, where readers and listeners in each language are able to decode information at a constant rate with only a couple of words of context.

Experiment 3: Large-scale data and linguistic features

We pulled corpora for 159 diverse languages from Wikipedia, an online general knowledge repository with separate repositories for each of hundreds of languages. We filtered our selection of language corpora on Wikipedia to those which had at least $10,000$ articles. We used the Wikiextractor tool from GitHub [WIKIEXTRACTOR CITATION] to retrieve the dumps of entire Wikipedia archives, then lowercased, removed punctuation and split the corpora into sentences. We trained and tested a model on each language's Wikipedia corpus independently, constructing unigram and trigram surprisals for each language separately. We create frequency-based and contextual surprisal curves for each corpus. The distribution of sentence lengths in Wikipedia corpora tend to resemble the distribution of sentence lengths in the BNC (another written corpus).

See Figure [Diff languages plot] for examples of frequency-based languages from several different language families. We see that the frequency-based curve for German resembles the English curve we found in Experiments 1 and 2 with a three-step distribution, while the Japanese curve looks very different from all the other curves. The Slavic language curves (for Serbian, Russian and Slovak) resemble each other.

Regardless of the shape of their respective unigram curves, the trigram curves for all the languages look the same as the English curve.

To compare languages more rigorously, we used two databases of language similarity features. To target lexical differences between languages, we used the 40-item Swadesh list (Swadesh, 1955), retrieved from the ASJP database (Wichmann et al., 2016). The Swadesh list is a well-known method for comparing lexical similarity between languages, by quantifying the similarity between the words on the list for pairs of languages, and is often used to compare genetic relationships between languages. We computed the average normalized Levenshtein distance, a string edit distance measure (LDN; Holman et al., 2008) between each pair of our Wikipedia languages.

We split the WALS features by type of feature: the nomina-

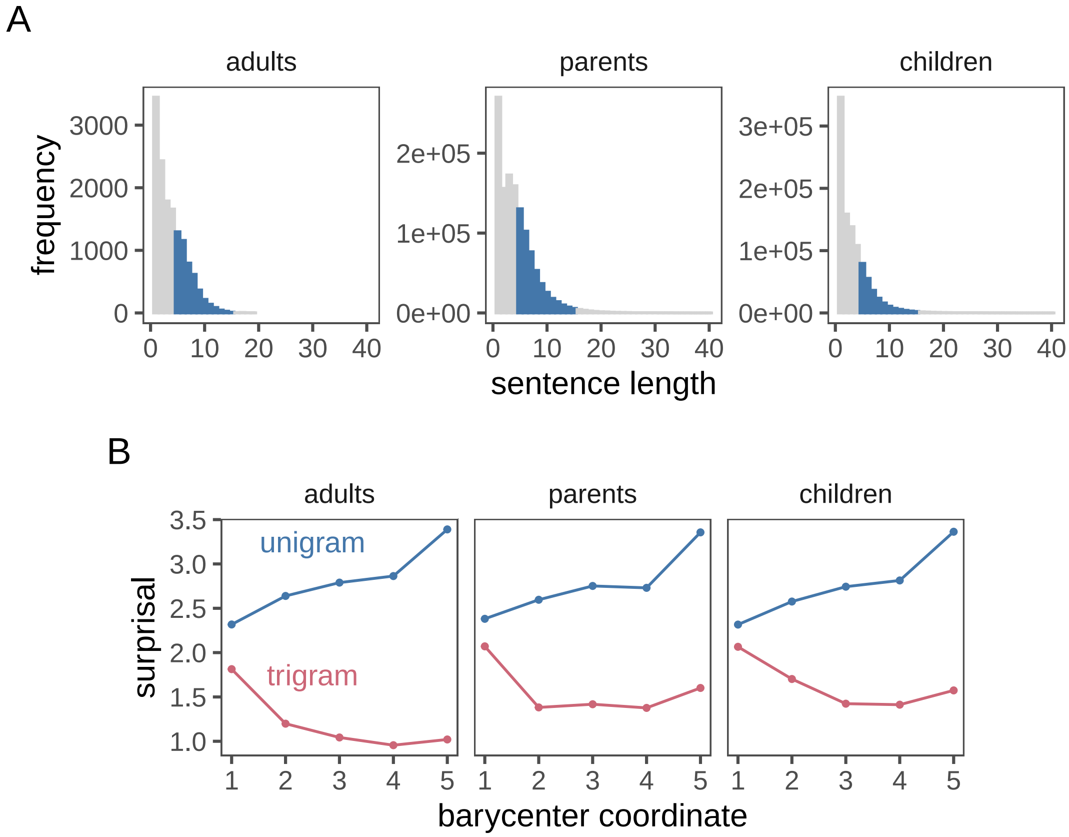

Figure 4: (A) Sentence length distributions in the spoken language Corpora: Adults in Santa Barbara, and Parents and children in CHILDES. (B) Charateristic surprisal curves for these corpora.

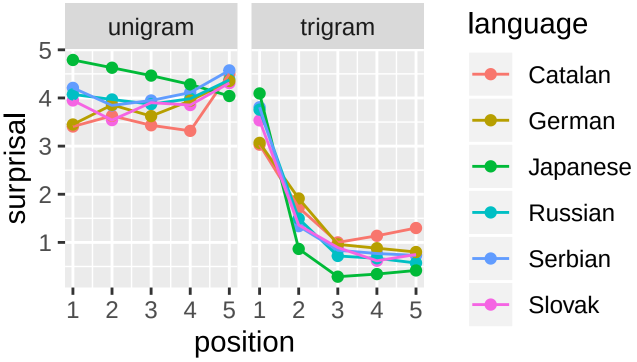

Figure 5: Wikipedia frequency-based and context-based curves for diverse languages

tive categories features describe aspects of morphology such as case systems and definite/indefinite articles; word order describes SVO as well as head-modifier word order; nominative syntax describes noun behavior such as possessives and adjectives acting as nouns; clauses describes phrasal and broader sentence syntax; and verb categories describe tense, mood and aspect as well as morphology on the verb.

As our surprisal metric is a lexical measure, we expect the Levenshtein distance to be high. To describe more structural relationships, we used the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013) to describe the morphology, syntax, phonology, etymology and semantics–in short the structures in each language. As WALS is a compiled database from dozens of papers from different authors, most of the features and languages are fairly sparse. We use a iterative imputation algorithm for categorical data Multiple Imputation Multiple Correspondence Analysis (MIMCA; Audigier, Husson, & Josse, 2017) to fill in the missing features.

We see that only word order shows a significant correlation with the shape of the frequency-based information curves in the wikipedia corpora. The rest of the features show virtually no correlation with the wikipedia corpora. For the context-based information curves, all categories of WALS features show correlations. However, across all languages, the context-based information curves are flat after the first word or two.

WHAT ABOUT THE SWADESH LIST?

By considering data from 159 diverse languages, we see that the frequency-based information curves display unique information distributions for each language. A small part of this cross-linguistic variation is explicable due to top-down features of the language in question, such as canonical word order. The majority of the variation for each language does

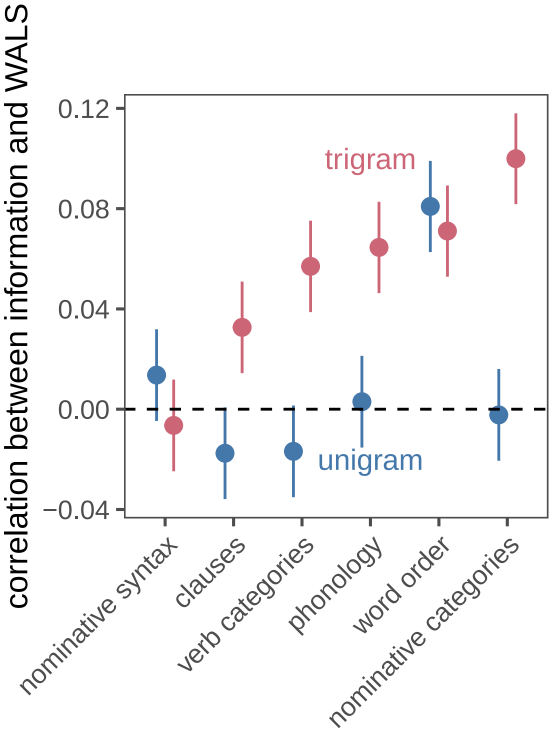

Figure 6: Correlations between Information curve shapes and WALS features.

not derive from top-down features, but we suspect instead arises from bottom-up choices made by the speakers in each language.

Conclusion

In this paper we have proposed a novel method for quantifying information structure in any language. Our method derives the unique distribution of information based on word frequency in the language, irrespective of whether the language in question is spoken or written. This information structure characterizes the earliest utterances in child speech all the way to complex sentences in knowledge base entries by adult writers. The shape of the information distribution in a language is correlated with the canonical word order in a language, but is mainly derived from the individual choices speakers make in communication in a language.

In communication, due to predictive processing, the variation in these unique frequency-based distributions are washed out. Instead, a uniform distribution of information emerges across languages, which may allow listeners to decode information quickly and at a nearly constant rate. Once they have enough context, listeners optimally decode information.

We average over the surprisal sequences of each length, which obscures most of the variation in information curve shapes. This may also mean that there's more bottom-up processing.

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.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

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.

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).

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., Bakker, D., & others. (2008). Advances in automated language classification. *Quantitative Investigations in Theoretical Linguistics*, 40–43.

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

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, *62*, 621–647.

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

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

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

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.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

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.

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*.