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

Josef Klafka[1] & Daniel Yurovsky[1,2]

[1] Carnegie Mellon University

[2] University of Chicago

Author Note

Correspondence concerning this article should be addressed to Josef Klafka, 5000 Forbes Ave. E-mail: jklafka@andrew.cmu.edu

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 do these constraints influence information structure, and how does this influence 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; communication; efficiency; syntax; typology; language development; computational modeling

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

Introduction

One of the defining features of human language is its power to transmit information. 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 our listener (Austin, 1975). For this reason, we can describe language as a cryptographic 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 as 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 to a listener 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 meaning of an utterance in real time rather than at its conclusion (Kutas & Federmeier, 2011; Pickering & Garrod, 2013; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This solution creates new guidance for speakers. 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 most efficient strategy is 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 transmission of information, with the optimal rate of information transfer as close to the listener's fastest decoding rate as possible. The hypothesis that speakers follow this optimal strategy is known as the *Uniform Information Density* hypothesis.

Using information theory, a mathematical framework for formalizing predictability, researchers have tested and confirmed this optimal coding prediction across several levels and contexts in language production. For example, Genzel and Charniak (2002) provided a clever indirect test of Uniform Information Density across sentences in a paragraph. They showed that the predictability of successive sentences, when analyzed in isolation, decreases, as would be expected if readers use prior sentences to predict the content of future sentences. Thus, based on the increasing amount of context, they found that total predictability remains constant. At the level of individual words, Mahowald, Fedorenko, Piantadosi, and Gibson (2013) showed that speakers use shorter alternatives of more predictable words, maximizing the amount of information in each word while minimizing the time spent on those words.

Other research has suggested that efficient encoding impacts how speakers structure units between words and sentences. The inclusion of complementizers in relative clauses (Jaeger & Levy, 2007) and the use of contractions (Frank & Jaeger, 2008) are two situations in sentence formation in which speakers can omit or reduce words to communicate more efficiently and maximize use of the communication channel without exceeding the listener's capacity.

How languages evolve is shaped by efficient communication as well. Piantadosi, Tily, and Gibson (2011) showed that more easily predictable words in a language may tend to become shorter over time, maximizing the amount of information transmitted over the communication channel at every second by speakers in each language. Semantic categories of words across languages can also evolve to be structured efficiently. Categories such as kinship terms (Kemp & Regier, 2012) maintain a trade-off between informativeness and complexity. Structure in langauge evolves from a trade-off between efficient and learnable encoding on the one hand and an expressive and descriptive lexicon on the other (Kirby, Tamariz, Cornish, & Smith, 2015). Languages may come to efficiently describe the particular environment in which they are spoken over the course of evolution: features of the world that are relevant to speakers become part of a language, while irrelevant features are disregarded (Perfors & Navarro, 2014).

However, despite this literature using the predictive coding model of language, one level has not yet been studied in depth: how speakers structure each individual utterances. This level may show the strongest effects of variation between languages. 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 of a language, 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 determining how its speakers structure their speech? Yu, Cong, Liang, and 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 in the English 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.

Methods

We measure information structure within languages, using a universal information metric proposed for the study of information transmission more generally by Shannon (1948) and applied to words specifically by Levy (2008): lexical surprisal. We can compute surprisal with the predictability of the word based on previously heard or seen words in its context, as in the formula below. The surprisal of a word is inversely proportional to the predictability of a word, such that less common and less predictable words carry more information. For example, “flower” has less information than “azalea” because “flower” is much more common than “azalea”. Though the two words have the same length in number of letters, it is more difficult to process “azalea” when reading it here than when reading “flower”. Frequency is intimately tied information content in words, with much of the differences between words frequencies being explained by information content cross-linguistically (Piantadosi et al., 2011). The surprisal of a word is also correlated with the processing cost of a word, shown by evidence from e.g. eye-tracking (Smith & Levy, 2013) and ERP (Frank, Otten, Galli, & Vigliocco, 2015) studies.

However, when reading or listening, people don’t just consider each word as an isolated linguistic signal. Listeners use the words they have already heard to predict and decode the word they are currently hearing. Following this incremental processing paradigm, we can also condition the surprisal of a word in its context. Ideally, we would like to measure the predictability of each word in an utterance using all of the information available to that word. For example, in an utterance of twenty words, we would like to use the previous 19 words of context to predict the 20th word. However, we would need to train on a corpus of many trillion word tokens to predict with this amount of context. Regardless of computational constraints, we want to directly compare how predictable each word is regardless of its position in an utterance.

We therefore use a simplifying *Markov assumption*: we condition our next predictions on a fixed-size context window instead of all preceding words. Although these models may seem to use an inconsequential amount of context when predicting the next word, bigram and trigram models introduce a great deal of improvement over unigram models across tasks (Chen & Goodman, 1999). Models which incorporate more than two words of context have issues with overfitting to the corpus and only predicting observed sequences, often generalizing poorly.

When we use a word or two of context in our surprisal calculations, then the set of reasonable final items in our ngrams is greatly restricted. "Flower" may contain less information than "azalea" when we consider the words independently of their context, but with context this can be reversed. Flower appears in a variety of contexts, and so the information content of a word like "flower" in a particular context may be higher than "azalea". If you only have azaleas in your garden, then hearing someone say "in that garden, look at the flowers" may be higher surprisal for you: you expect them to say "azalea". This prediction does not require many words for context. For example, in the sentence "I take my coffee with cream and sugar", when hearing "cream and", a listener might automatically predict "sugar", but there are few possible continuations with even the two words "cream and". Hearing "I" restricts the next word to a verb, or possibly an adverb, and since the listener has heard the speaker refer to themselves in the first person singular, their set of possible completions is significantly restricted.

We train two types of ngram language models independently on a corpus. One of our models is frequency-based: we do not incorporate context into our surprisal calculations. To incorporate context into our models, we train bigram and trigram language models, which incorporate one and two words of context for each processed word, respectively. The frequency-based surprisal metric gives us an idea of when in their utterances speakers say frequent i.e. independently information-rich words. The context-based surprisal metric show us how speakers tend to distribute the information in utterances relative to real-time processing in communication. We expect a priori that our frequency-based surprisal curve will be flat. No one part of the sentence will on average have words that are more frequent than another across utterance lengths. Similarly, we expect that there will be a small smoothing effect for our contextual surprisal metric such that the word in each position of an utterance is more predictable than its frequency-based counterpart.

Estimating information

To estimate how information is distributed across utterances, we computed the lexical surprisal of each word under two different models. 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. Simply the less often a person has seen a word, the more information that word holds.

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 a corpus using the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013). Each utterance was padded with a special start-of-sentence token "$\langle s \rangle$" and end of sentence token "$\langle / s \rangle$". Trigram estimates did not cross sentence boundaries, so for example the surprisal of the second word in an utterances was estimated as 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 as implemented in the KenLM toolkit (Heafield et al., 2013). Briefly, this smoothing technique discounts all ngram frequency counts, which reduces 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). For a thorough explanation of modified Kneser-Ney smoothing, see Chen and Goodman (1999).

Aggregating 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.

We used DBA to discover the short sequence of surprisal values that characterized the

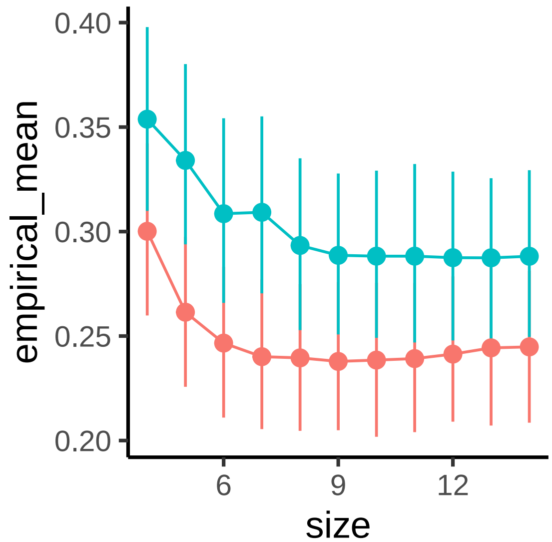


Figure 1. The final cost for the EM algorithm in fitting each size of barycenter

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.

Optimizing the size hyperparameter for barycenter averaging. We give the intuition for how we chose size X. How we find the barycenter produces a cost: the further the barycenter is from each data point, the higher the cost. We tried every size of barycenter in between 4 and 15 coordinates and found that 10 afforded the smallest size of the barycenter with the lowest cost.

We use the implementation of DBA in the Python package tslearn (Tavenard et al., 2017), which fits the barycenter to a time-series dataset through the expectation-maximization algorithm (EM; Moon, 1996). DBA in this implementation allows us to specify the size of the barycenter. In order to choose the optimal size for the barycenters, we computed the final cost from the EM algorithm for each barycenter size and chose the barycenter size which minimized the average final cost. Time costs were negligible for computing a larger barycenter.

Application to the corpus. Once we have fitted our language model, we can compute the surprisal of a continuation by simply taking the negative log-probability of that word's ngram probability. To find the average information for a given position in a corpus, we take all utterances of a given length, and for each word position in utterances of that length, we compute the average of the surprisals for all of the non-unique words that occur in that position, conditioned or not conditioned on context. By computing these averages for each word position in an utterance, we compute a low-dimensional approximation to the average distribution of information in the corpus. With the surprisal metric, we base the information contained in each word on how often the word is encountered in its context in the corpus. As long as the corpus is representative of the language or population we study, then the distribution of information is approximated for that language or population as a whole.

Study 1: The Shape of Information in Written English

Genzel and Charniak (2002) performed an influential early test of the Uniform Information Density hypothesis. They analyzed the amount of information in successive sentences 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 that the reader decoded within each sentence was constant overall.

Yu et al. (2016) applied this same logic to each word within sentences, computing the entropy over each successive word position in an utterance. The entropy metric Yu et al. (2016) used is the average surprisal of all the words in a given word position, computed only within that position, and weighted by how many times each word occurs in that position. Their formula is given by $\mathrm{H}(X) = -\sum_w P(w \in X) \log P(w)$, where X is a word position (e.g. first words, fifth words, final words) and w is a word occuring in position X.

The first word of each sentence tended to contain little information, while words in the

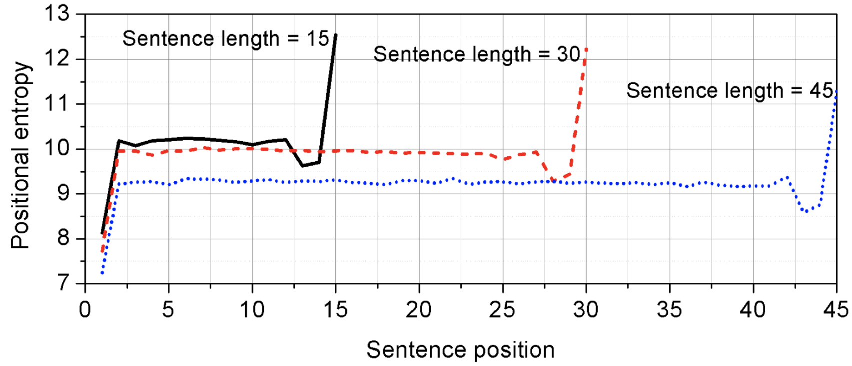


Figure 2. A caption

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.

UID would predict a result similar to Genzel and Charniak's (2002) results: a smooth and monotonically increasing affine function of word position. Yu et al. (2016) interpreted their uneven information curve as evidence against the Uniform Information Density Hypothesis as, unlike Genzel and Charniak's (2002) results, information plateaued in the middle of sentences. See Fig 2.

We replicate the analysis from Yu et al. (2016) 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 (Leech, 1992). The BNC is an approximately 100 million word corpus consisting of mainly of

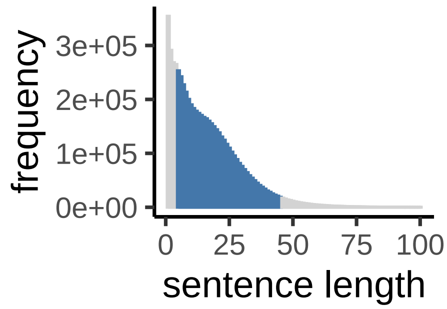


Figure 3. The distribution of sentence lengths in the British National Corpus. We analyzed sentences of length 5-45 (colored).

written (90%) with some spoken transcriptions (10%) of English collected by researchers at Oxford University in the 1980s and 1990s. The BNC is intended to be representative of British English at the end of the 20th century, and contains a wide variety of genres (e.g. newspaper articles, pamphlets, fiction novels, academic papers). Yu et al. (2016) only used the written portion of the BNC, although we include the much smaller spoken portion as well.

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 3).

Due to the size of the corpus, we do not include it along with our submission, but we include the scripts and directions in the copy of our GitHub repository.

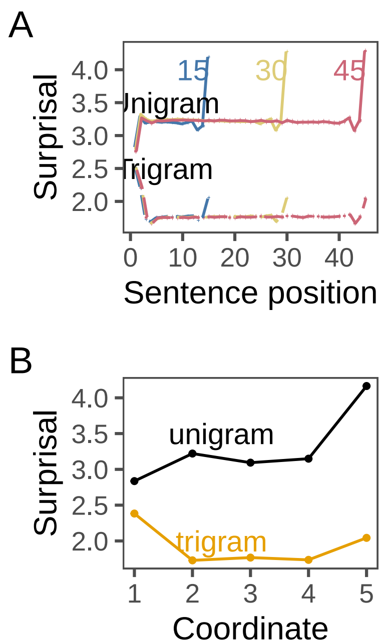


Figure 4. (A) Surprisal by sentence position of length 15, 30, and 45 sentences in the British National Corpus under unigram and trigram surprisal models. Error bars indicate 95% confidence intervals (tiny due to sample size). (B) Characteristic information curves produced by the DBA algorithm averaging over all sentence lengths in each corpus.

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. 4A). Qualitatively, we found the same shape in utterances of all other lengths we sampled, from utterances with 5 words to utterances with 44 words.

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. 4B shows the barycenters produced by the dynamic time warping barycenter averaging algorithm. 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, a $\sim$ 250,000 word corpus of recordings of naturally occurring spoken interactions from diverse regions 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 through the childes-db interface (MacWhinney, 2000; Sanchez et al., 2019). We selected for analysis 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. 8A). 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. 8B). 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), which rules out directly comparing the surprisal values between the BNC, the SBC and English CHILDES. However, we can compare the shapes of these curves both to each-other and to the written English sentences in Study 1 **??**B. 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: Language structures and large-scale data analysis

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 (Swadesh, 1955), retrieved from the ASJP database (Wichmann et al., 2016). The Swadesh list is designed to include near-universal words that target basic cognitive concepts, and are useful in determining the genealogical similarities and differences between languages. Qualitatively, the more similar two languages' words are, the more similar the two languages are. To quantify this intuition, we computed the average normalized Levenshtein distance (LDN; Holman et al., 2008) over all items on the Swadesh list between pairs of languages.

To more rigorously described the top-down typological similarites and differences between languages, we used data from the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013). The WALS database has data for 144 typological features in thousands of languages from across the world. These features describe aspects of morphology, syntax, phonology, etymology and semantics–in short the features describe the structures in each language.

We used Multiple Imputation Multiple Correspondence Analysis to fill in the missing data for the features we selected using statistical imputation (MIMCA; Audigier, Husson, & Josse, 2017). MIMCA begins with mean imputation, converts the categorical WALS features into a numerical contingency table with dummy coding, then repeatedly performs principle components analysis and reconstructs the contingency table. As WALS is a compiled database from dozens of papers from different authors, most of the features and languages are fairly sparse. Even limiting ourselves to the 152 language corpora we pulled from Wikipedia and 122 features from WALS, there are tens of thousands of individual possible data values, fewer than half of which were already computed for those languages in the WALS database.

Data processing

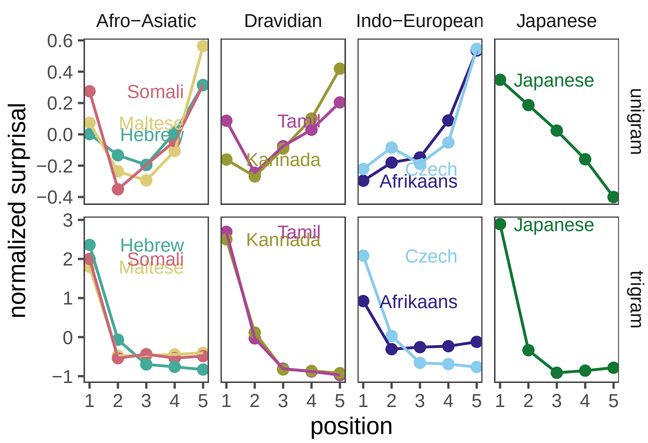


Figure 5. Characteristic information curves (centered) for a sample of languages from Wikipedia

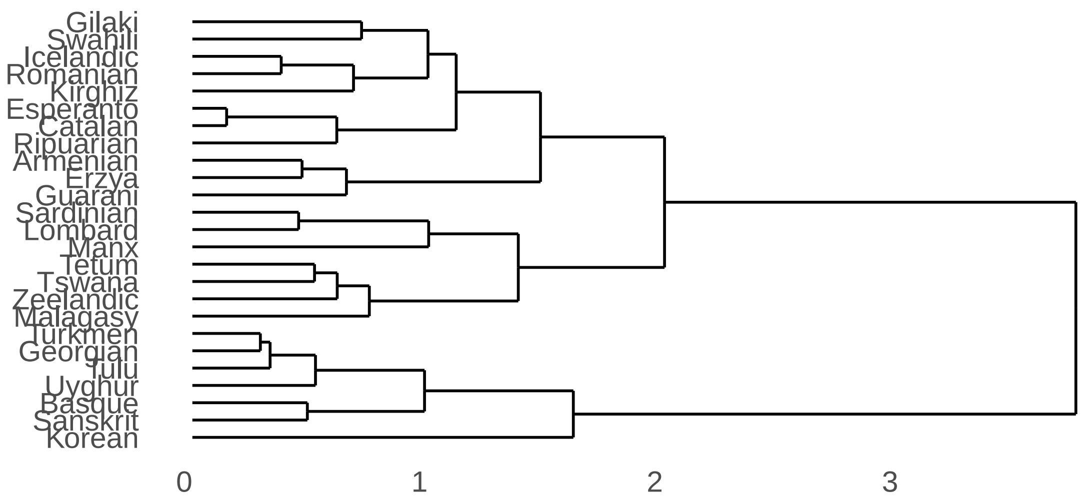
All processing was identical to Studies 1 and 2 except for the lengths of utterances chosen for analyses. To accommodate the variety of lengths across language corpora, we analyzed sentences of lengths 5 to 45.

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 barycenter 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 summed the number of WALS features each pair of languages shared the same value for.

Results and Analysis

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.

Hierarchical clustering. We ran a hierarchical clustering algorithm on the frequency-based information curves using the hclust package from the R stats core library (Team & others, 2013). We used the complete linkage algorithm for hierarchical clustering, with distances between information curves between languages computed using cosine distance between their embeddings in the slope space. The complete linkage algorithm at every step pairs each language or cluster of languages with its closest neighboring language or cluster. A sample from the dendrogram is shown in Figure **??**. From a quick glance, the unigram information curves appear to reproduce some of the genealogical relationships between languages, although the dendrogram does not exactly replicate language genealogy for all 152 languages. This suggests using a first-pass quantitative method that the information curves do correspond in some measure to language families, but language families do not explain all of the variation and relationships between frequency-based information curves.



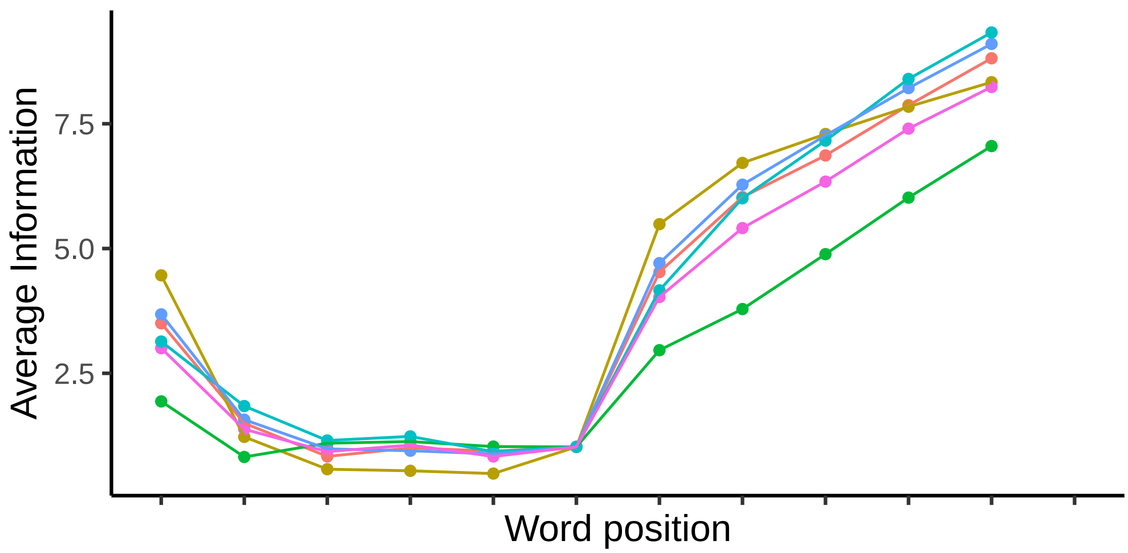
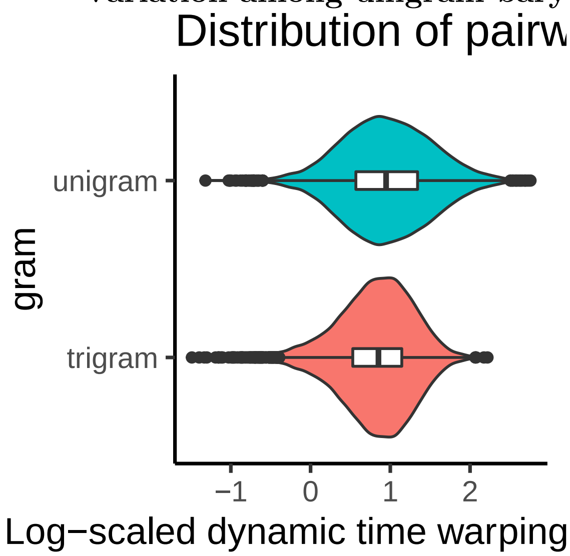


Figure 6. Some trigram information curves from the Wikipedia data

Qualitative analysis.

Variation among unigram barycenters versus trigram barycenters.



In this script, we quantify the average pairwise distance between unigram and trigram barycenters. This simple summary statistic (computed along with a non-parametric bootstrap) will give evidence that the unigram barycenters have more variation than the trigram barycenters.

We use the dynamic time warping distance metric to quantify the differences between the barycenters. The average unigram distance is 3.04, with a upper confidence interval bound of 3.07 and a lower confidence interval bound of 3.01. Those values are 2.50, 2.52 and 2.48 respectively for trigrams. Qualitatively, there appears to be a large difference in means.

We log-scale the distances and graph them with violin and boxplot to display the distribution. The violin plots for unigrams and trigrams actually appear similarly distributed, but unigrams has more large distance values (a fatter tail, in probability distribution terms) which allocates mass away from small distance values that dominate both trigrams and unigrams.

Global correlations. 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. 5 shows centered information curves for

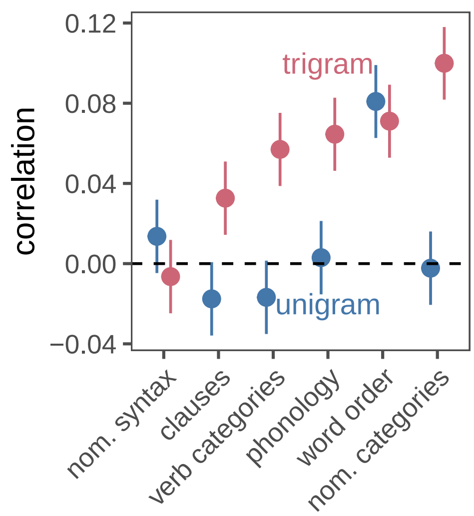


Figure 7. Pairwise correlations between languages' centered information curves and the number of linguistic features they share of each type. Error bars indicate 95% CIs.

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 ($r = 0.03$, $t = 2.72$,$p = .007$).

Correlations by type. 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. 7 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.

Mutual information between WALS features and barycenters. We use a final measure to quantify how much variation in WALS features explains variation in barycenters: mutual information between each coordinate of the barycenters and each WALS feature and feature group. We use the categorical-continuous measure derived and described in Ross (2014). How much does knowing the value of each WALS feature tell you about the value of each one of the barycenter coordinates? We reduce to a pairwise measure. The algorithm uses a variant of the k-nearest-neighbors regression algorithm. It essentially asks: how many points with the same WALS label are in a neighborhood around each language's e.g. first coordinate?

Discussion of results.

Results and plot.

Discussion

By considering the distribution of information at the level of utterances and sentences, we join together the information-theoretic work focusing on sub-word units and words, and that focusing on paragraphs. In doing so, we show that frequency and context-based metrics complement one another in studying efficiency and information in language. We directly link linguistic efficiency in a language to the genealogy and properties of that language. We provide evidence for a novel linguistic universal: low processing cost for listeners beyond the first words in utterances, driven by high average word predictability in conversation. With consideration to language acquisiton, we observe that children tend to distribute information in their utterances according to the their language's frequency-based information curve as soon as they form multi-word utterances.

Throughout this work we have averaged the surprisal values at each position. Averaging removes variation, which in turn may obscure trends in the data. As discussed in the methods section, the surprisal metric has historically been used for calculating the information and processing cost for individual utterances, and our use of the metric here is actually a step forward rather than a step back. Future work can investigate variation in how speakers distribute information in individual utterances.

Our Study 3 results from Wikipedia languages fall prey to Galton's problem: the languages are not drawn randomly from the set of all languages without regard to language family or any other kind of genealogy. Indo-European languages (which are more similar to each other) are overrepresented.

The WALS database we used to investigate typological variation in the information curves is overall sparse. We imputed well over 50% of the WALS features for most of our 159 languages, although all of the languages had at least 20 features evaluated in WALS. A large part of this is due to WALS being a collection of a number of different studies, instead of a systematic effort to catalogue variation across the world's languages. Additionally, WALS features are meant to describe specific microvariations in languages, not to provide a comprehensive typological representation of each language compared to each other language. This may be why the Swadesh list provided a higher correlation for describing the differences in information curves: Swadesh (1955) intended the list to allow researchers to more comprehensively compared and constrast lexical differences between languages. For our Wikipedia analysis, we also reduce all of a language's variation down to a five-dimensional vector. These information curve representations show a surprising amount of variation despite the degree of compression.

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.

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. Retrieved from https://wals.info/

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.*

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of* *the cognitive science society* (Vol. 30).

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

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).

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).

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

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.

Leech, G. N. (1992). 100 million words of english: The british national corpus (bnc).

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.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47–60.

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, *38*(4), 775–793.

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.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS One*, *9*(2).

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.

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., … Woods, E. (2017). Tslearn: A machine learning toolkit dedicated to time-series data.

Team, R. C., & others. (2013). R: A language and environment for statistical computing.

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*.

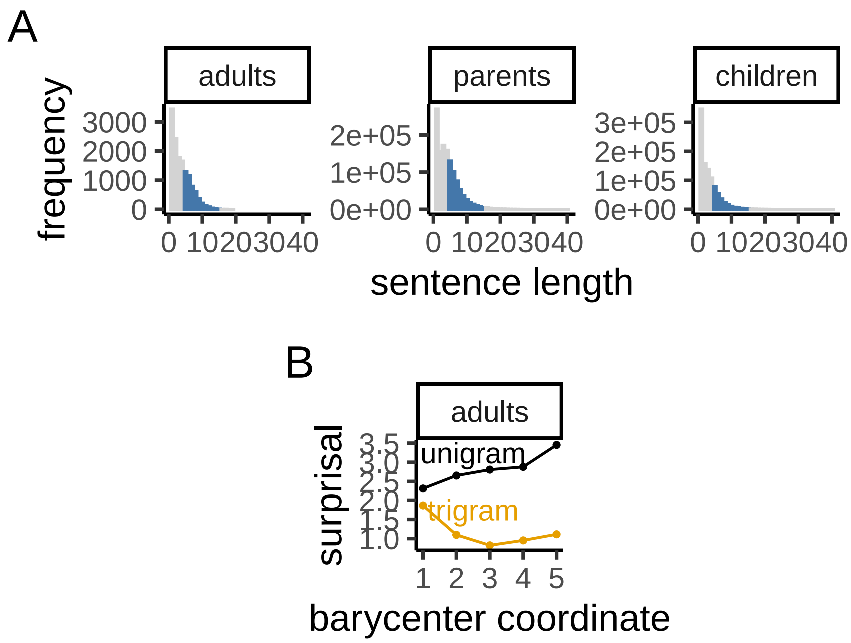


Figure 8. (A) The distribution of sentence lengths in the spoken English corpora: Adults in Santa Barbara, and parents and children in CHILDES. We analyzed sentences of length 5-15 (colored). (B) Charateristic surprisal curves for these corpora.