¹ Speakers of diverse languages structure their utterances for efficient communication

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

³ [1] Carnegie Mellon University

⁴ [2] University of Chicago

⁵ Author Note

⁷ Correspondence concerning this article should be addressed to Josef Klafka, 5000

⁸ Forbes Ave. E-mail: klafka@cmu.edu

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 in their utterances, surveying numerous diverse languages. We find that speakers produce frequent and informative words at regular parts of their utterances, depending on language they use, which is predictable in part from the features and genealogy of their language. This 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 the information distribution in the language they understand.

Keywords: keywords

Word count: X

Speakers of diverse languages structure their utterances for efficient communication

Introduction

One of the defining features of human language is its power to transmit information. We use language for a variety of different proximal tasks tasks: greeting friends, taking notes, signalling group identities. But all share a common unifying purpose: changing the mental state of the listener or reader of our communications (Austin, 1975). Language can thus naturally thought of as a code that allows speakers to turn their intended meanings into a form that can be transmitted through the air (or paper, or electrons) and subsequently converted by listeners back into an approximation of the intended meaning (Shannon, 1948).

Beyond its utility as a metaphor, this coding perspective on language is powerful because it affords a framework for rational analysis: If language has evolved to be an optimal code for information transmission, how should this code be structured (Anderson & Milson, 1989)? The optimal code would have to optimize two competing pressures: (1) a pressure for listeners to easily and successfully decode messages sent by the speaker, and (2) a pressure for speakers to easily code their messages and transmit them with minimal effort and error. A fundamental constriaint on both of these processes is the serial nature of human language: sounds are produced one at a time and each is unavailable perceptually once it is no longer being produced.

The solution that languages have arrived on is incremental processing: Listeners 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). Because prediction errors can lead to severe processing costs and difficulty integrating new information, an optimal coding strategy is for speakers seeking to minimize their prodction costs is to produce utterances that are are just at the prediction capacity of listeners without exceeding this capacity (Aylett & Turk, 2004;

48 Genzel & Charniak, 2002).

49 One salient constraint on this transmission processes is the serial nature of language:

50 This coding perspective is powerful because it lends itself naturally to a rat

51 Languages can thus be thought of as a code that can be used to transform speaker's
52 intented meaning into a

53 • Coding/decoding and information transfer

54 • language as prediction

55 • how should the code be structured? optimization questions like this have a had * a lot
56 power across different parts of language: zipf, kemp & regier, etc etc. noisy channel
57 stuff. (or does this go in discussion)

58 • don't under or overutilize channel, stay near limit – constant entropy rate

59 • evidence for this at multiple levels from individiual lexical choice *(info/information
60 theory paper), inclusion of non-obligatory aspects of language (filled pauses, that),
61 across multiple utterances in a discourse

62 • what about over the course of an individual utterance–perhaps the strongest
63 constraints from language itself?

64 • yu et al. found this puzzling effect.

65 • we replicate that in a different kind of analysis and show that it extends to trigrams
66 etc???

67 • BUT variation cross-linguistically. why? constraints from language.

68 We use language to communicate information. Although we often need to say things
69 we have never said before, we are easily and quickly able to produce novel utterances with
70 next to no prior planning. Our listeners incrementally process each word we say in
71 utterances they may have never heard before, and predict which words will come next within
72 fractions of a second of hearing the beginning of the word (for evidence see e.g. Kutas &

Federmeier, 2011). Encoding new and unpredictable information into our utterances

surprises our listeners, and presents problems for communication. Listeners may struggle to

decode unpredictable words, which can cause difficulty decoding the following words as well.

Efficient and effective communication relies on interlocutors being able to process novel

information without being surprised by new information in unpredictable words. On the

other hand, to make best use of the capacity of the communication channel, speakers should

encode as much information as possible into each word they say.

Due to their differences in distributional frequency, words vary in how much

information content they contain (Shannon, 1948). Frequency is a meaningful predictor of

word length: the more common the word, the shorter the word tends to be (Zipf, 1949).

Over and above frequency, contextual information content predicts word length, due to

speakers minimizing effort in the amount of information they encode in speech (Piantadosi,

Tily, & Gibson, 2011). Speakers tend to smooth out their information to a more uniform

rate within utterances that allows for easier decoding on the listener side (Jaeger & Levy,

2007). Across utterances, context allows speakers to produce more informative and

surprising sentences later in communication. The later a sentence occurs in a paragraph, the

the informative it is and the more unusual and surprising words it contains (Genzel &

Charniak, 2002). From the level of individual words to sentences in paragraphs, information

content shapes and is shaped by the context it appears in.

The pressure for efficient communication influences language at all timescales, from

online speech production to evolution. Mahowald, Fedorenko, Piantadosi, and Gibson (2013)

found that speakers tend to use shorter, more efficient word forms in predictive contexts,

minimizing communicative cost without sacrificing intelligibility. Predictability and

ambiguity shape the distribution of color terms in languages (Gibson et al., 2017) as well as

word lengths (Piantadosi et al., 2011) across languages, and speakers are able to

deterministically decode words despite ambiguity in word forms (Piantadosi, Tily, & Gibson,

2012).

How do frequency and predictive context determine how we distribution information at the level of utterances? We do not speak only in phrases or paragraphs, but conversationally we speak in utterances and in writing we produce sentences as a coherent unit of a complete thought. Languages possess very different grammatical and morphological structures–the same thought can be encoded in at least as many ways as there are different languages. The average rate of information speakers transfer in syllables differs from language to language (Aylett & Turk, 2004). How do the different structures seen in different languages, with different rates of information encoding, affect how speakers structure information in production?

Yu, Cong, Liang, and Liu (2016) proposed a novel method for quantifying the amount of information typically found at each part of English utterances. Treating each word position by itself, Yu et al. (2016) calculated the Shannon entropy (a measure of uncertainty) over the words of each position, producing an estimate for the amount of uncertainty reduced once the listener hears the words used in a given position, related to the Shannon concept of information as a reduction in uncertainty (Shannon, 1948). They found a distinctive three-step shape to the distribution of information in English written sentences.

In this paper, we expand on this body of prior work in a number of novel ways. We replicate the results from Yu et al. (2016) with a different metric, one intimately tied to incremental word processing. We find the same distribution of information based on word frequency in English adult speech as in English adult writing, as well as in English child speech. We also apply a contextual version of the same metric, and find that in communication, listeners are able to decode information from the utterances they hear at a constant rate past the initial words. We expand the study of information density to the largest set of languages yet considered, and incorporate contextual and typological information. We find that each language has a characteristic, robust frequency-based

information distribution tied to the language's morphology, phonology and syntax. Speakers of a language will tend to distribute information according to that language's distribution of information in their utterances. As soon as a child can form multi-word utterances in a given language, they will produce utterances obeying the characteristic distribution of information in a given language.

Methods

Shannon (1948) defined information as "the reduction in uncertainty about one variable given that the value of another variable is known". We use a metric proposed by Shannon and applied by Levy (2008): lexical surprisal, which defines the information in word based on the ratio of possible continuations of the sentence after to before the word is seen. Equivalently, we can compute surprisal with the predictability of the word, 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.

$$surprisal(word) = -logP(word)$$

The surprisal of a word is also correlated with the processing cost of a word, with evidence from eye-tracking (Smith & Levy, 2013) and ERP (Frank, Otten, Galli, & Vigliocco, 2015) studies, among other sources. Considered without context, the surprisal of an individual word is inversely proportional to the frequency of that word, so that the less often a person has seen a word, the more information that word holds. 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's 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).

₁₄₈   However, when reading or listening, people don't just consider each word as an isolated

₁₄₉   linguistic signal. Instead, listeners use the words they have already heard to predict and

₁₅₀   decode the word they are currently processing. Following this incremental processing

₁₅₁   paradigm, we can also condition the surprisal of a word in its context. In the formula below,

₁₅₂   widenotes the word currently being read or heard, while wi-1denotes the first word before the

₁₅₃   current word, wi-2denotes the second word before the current word, and so on.

₁₅₄

$$surprisal(w_i|w_{i-1}w_{i-2}...) = -logP(w_i|w_{i-1}w_{i-2}...)$$

$$= -log\frac{P(w_i, w_{i-1}w_{i-2}, ...)}{P(w_{i-1}w_{i-2}...)}$$

₁₅₅   When we use a word or two of context in our surprisal calculations, then the set of

₁₅₆   reasonable final items in our ngrams is greatly restricted. "Flower" may contain less

₁₅₇   information than "azalea" when we consider the words independently of their context, but

₁₅₈   with context this can be reversed. Flower appears in a variety of contexts, and so the

₁₅₉   information content of a word like "flower" in a particular context may be higher than

₁₆₀   "azalea". If you only have azaleas in your garden, then hearing someone say "in that garden,

₁₆₁   look at the flowers" may be higher surprisal for you: you expect them to say "azalea". This

₁₆₂   prediction does not need many words before to work out, for example in "I take my coffee

₁₆₃   with cream and sugar". When you hear "cream and", you automatically predict "sugar" or

₁₆₄   maybe "honey", but there are few possible continuations with even those two words. Hearing

₁₆₅   "I" restricts the next word to a verb, or possibly an adverb, and since you have just heard

₁₆₆   the speaker refer to themselves in the first person singular, your set of possible completions is

₁₆₇   significantly restricted.

₁₆₈   To find the distribution of information within a corpus, we calculate the surprisal or

₁₆₉   contextual surprisal for each unique word or sequence of words within the corpus. We take

₁₇₀   all utterances of a given length, and for each word position in utterances of that length, we

171 compute the average of the surprisals for all of the non-unique words that occur in that

172 position, conditioned or not conditioned on context. By computing these averages for each

173 word position in an utterance, we compute a low-dimensional approximation to the average

174 distribution of information in the corpus. With the surprisal metric, we base the information

175 contained in each word on how often the word is encountered in its context in the corpus. As

176 long as the corpus is representative of the language or population we study, then the

177 distribution of information is approximated for that language or population as a whole.

178   The flexibility of the surprisal metric we employ in this paper allows us to calculate the

179 anticipated information for an individual utterance, as most work with the metric has done

180 in the past. Averaging together the surprisal values for a word position within utterances is

181 actually a step further than prior work, and indicates the tendencies speakers gravitate

182 towards instead of examining individual stimuli in psycholinguistic experiments.

183   The frequency-based surprisal metric gives us an idea of when in their utterances

184 speakers say frequent i.e. independently information-rich words. The context-based surprisal

185 metric show us how speakers tend to distribute the information in utterances relative to

186 real-time processing in communication. We expect a priori that our frequency-based

187 surprisal curve will be flat. No one part of the sentence will on average have words that are

188 more frequent than another across utterance lengths. Similarly, we expect that there will be

189 a small smoothing effect for our contextual surprisal metric.

190   **Written and Spoken English have the same information distribution**

191   We first turn to working with written English in the British National Corpus (BNC;

192 Leech, 1992). The BNC is a collection of spoken and written records (90% written) from the

193 turn of the century, intended to be a representative sample of British English. Using their

194 word entropy metric without context, Yu et al. found a distinctive three-step distribution for

195 information in written English sentences in the corpus. The first word tended to contain

little information. While the middle words of sentences each had more information than the first word, information was flat and non-increasing across the middle of sentences. The final word, they found, contained the most information out of any in the sentence, with a noticeable spike in information. They found the same distribution across sentence lengths, from sentences with 15 words to sentences with 45 words.

We first replicate the Yu et al. (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.

See in Figure @ref(fig:bnc_unigrams) a qualitative comparison of the information trajectory they found and the information trajectory we found. It is easily observable that the two information trajectories are identical, and we have found the same distribution of information in the BNC as Yu et al. (2016) did. The first words in written English sentences tend to contain little information, while the final words tend to contain a great deal of information.
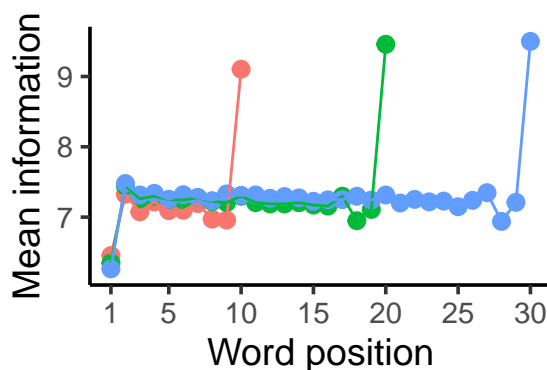

Figure 1. (#fig:bnc_unigrams)BNC frequency-based information curves

However, this only characterizes written English sentences. If this information distribution is characteristic more generally of English, then we will find the same

²¹⁴ distribution for spoken English. We use the Switchboard corpus of spoken English telephone

²¹⁵ conversations between adults to compute the distribution of information across spoken

²¹⁶ utterances in English. The sentences in the BNC, compared to the conversations in

²¹⁷ Switchboard, feature much longer sentences with larger vocabularies and more complex

²¹⁸ sentence structures. We split the conversations in Switchboard not by sentences, as in the

²¹⁹ BNC, but by turns.

²²⁰   For Switchboard we find the same information trajectory as we did for the BNC, and

²²¹ we find the same distribution of information regardless of utterance length. The longer the

²²² utterance, the more pronounced the information distribution shape becomes. A qualitative

²²³ comparison of the two trajectories is in Figure @ref(fig:switchboard_unigrams).
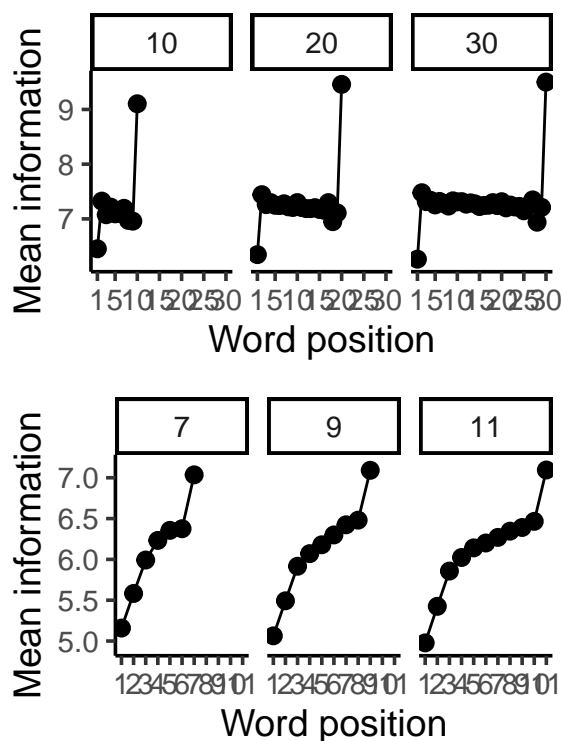
Figure 2. (#fig:switchboard_unigrams)Switchboard frequency-based information curves

²²⁴   We have found a unique average distribution of information that appears to

²²⁵ characterize the English language as a whole, regardless of medium (spoken or written) or

²²⁶ utterance length. This distribution indicates that in English, the words we speak or write at

the beginnings of utterances have little information, while the words we speak or write at the ends of utterances have a lot of information. The words in the middle of utterances have a middling amount of information, without the increasing trend in information from word to word that we might expect from (Genzel & Charniak, 2002).

What about context? So far we've only discussed the frequency-based metric, considering words on their own without any explicit incorporation of prior context. As previously discussed, listeners decode information and process what they hear incrementally, using prior heard words to ease the comprehension process. We now include one word of context (bigrams) and then two words of context (trigrams) for each word in our measurements. We observe a flattening effect of context across both modalities and all speaker populations. 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 curves for the BNC and Switchboard are in Figures @ref(fig:switchboard_bigrams) and @ref(fig:switchboard_trigrams).

Speakers produce information at a more or less constant rate, avoiding peaks or troughs in their information distribution, except at the beginnings of utterances, where listeners may not have any context to predict what the speaker is going to say. Speakers of English tend towards a characteristic and uneven distribution of word information based on frequencies within their utterances. Their interlocutors, however, once they have a word or two of context, decode information at a more or less constant and optimal rate.

English CDS and child speech

We found that English speakers and writers used the same robust and distinctive distribution of information within each utterance, regardless of the number of words in their utterances. To determine if this distribution truly characterizes all speakers of the English language as a whole, we wanted to examine speech from English-speaking children who are
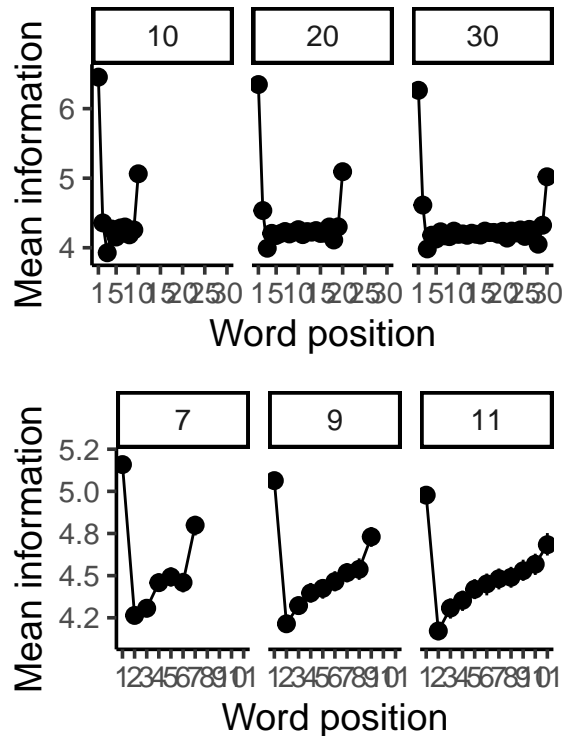
Figure 3. (#fig:switchboard_bigrams)Switchboard bigram context-based information curves

producing their very first multi-word utterances. We hypothesize the three-step distribution of information we found for English will characterize child speech and child-directed speech.

We use the Providence corpus from CHILDES (Evans & Demuth, 2012; MacWhinney, 2000), which consists of about 650000 utterances from parent-child conversations for six monolingual English-speaking children talking with the parents. We obtained the Providence corpus using the childesr frontend to the childes-db database (Sanchez et al., 2019). The child age range was between 11 months and 4 years. Sessions recorded in the home. Parents said a majority of the utterances, but children take up an increasing share of the conversation as they grow older and are actually able to speak full utterances. The utterances in the Providence corpus are on average significantly shorter than those in the BNC and Switchboard; over 90% of the utterances in the Providence corpus are at most 10 words long, with the median at around 6 words. Similar to our treatment of the Switchboard corpus, we split the Providence corpus up by conversational turns and separate utterances.
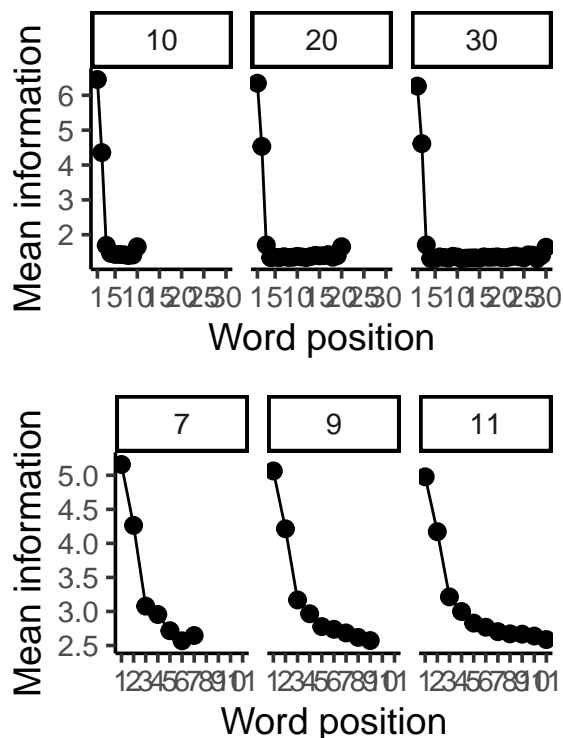
Figure 4. (#fig:switchboard_trigrams)Switchboard trigram context-based information curves

We observe the same distinctive distribution of information for parents and children in the Providence corpus as we did for adults in the BNC and Switchboard. The distribution of information we found at the level of individual words in English, therefore, characterizes the English language as a whole and not only adult utterances, not only written utterances. See @ref(fig:providence_unigrams).

What about context in the Providence corpus? When incorporating one or two words of predictive context, we observe the same trend as in the Switchboard and BNC corpora. Beyond the first couple of words, once your interlocutor has enough context to predict with some accuracy what you will say next, then you decode information from their speech stream at a constant and optimal rate. This applies to parents and children speaking to one another, as well as adults speaking and writing to one another. See @ref(fig:providence_bigrams) and @ref(fig:providence_trigrams).
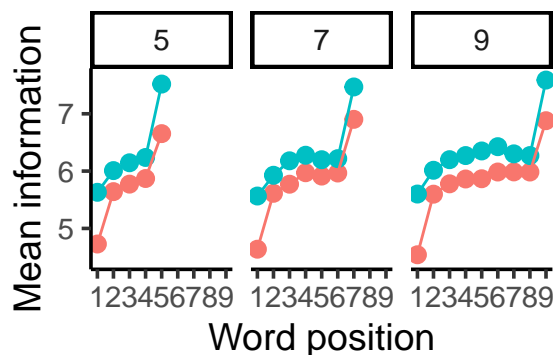
Figure 5. (#fig:providence_unigrams)Providence frequency-based information curves. Lines around each point indicate 95% confidence intervals computed with non-parametric bootstrap

Child speech and CDS across languages

So far, we have only looked at the distribution of information in words in English, both with and without context. We have examined child speech, child-directed speech and adult-directed speech, as well as writing samples selected to be representative of British English as a whole. But this only captures the picture for English.

We now turn to a small number of typologically diverse languages, and conduct the same analysis, using a monolingual adult-child speech corpora from CHILDES (MacWhinney, 2000) to compare the results from these languages directly to our results from the English Providence corpus. We use corpora for Spanish, German, French, Mandarin and Japanese. Similarly to the English Providence corpus, all of the following corpora consist mainly of shorter utterances: most utterances in the corpora are under 10 words long. For Spanish, we use the corpus of conversations between primary school-aged children between 6 and 9 years old and the researchers, in which the children were asked to recite stories in their native Spanish (Shiro, 1996). The students came from working class public schools and upper class private schools. For German, we use the Wagner corpus (Wagner, 1985). This corpus consists of a collection of mini-corpora collected from socio-economically diverse families, with children aged from 1;0 to 14;0, mainly on the younger end of that range. The corpus mainly consists of parent-child conversations between the children and their primary
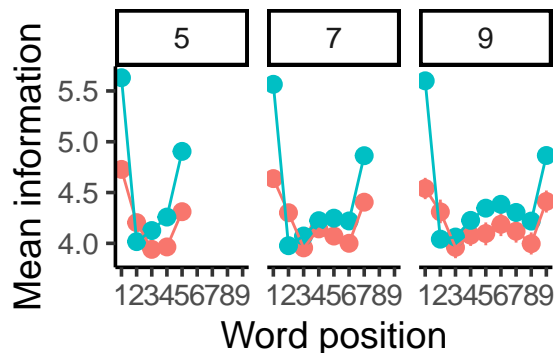
Figure 6. (#fig:providence_bigrams)Providence bigram context-based information curves

caregivers. For French, we use the Palasis corpus (Palasis, 2009), a longitudinal study of children between the ages of 2;5 and 4;0 from the same kindergarten class. The conversations in this corpus take place between the children and other children in the same class, and between the children and the investigator. For Mandarin Chinese, we use the Zhou corpus of dinner conversations from the Shanghai area (Li & Zhou, 2015). The children in this corpus are 5 or 6 years old, and half come from working class families (the other half from upper class families). For Japanese, we use the Okayama corpus (Shirai, 2001). This corpus consists of conversations between children between the ages of 2;2 and 4;2 and their mothers.

We observe a distinct and characteristic frequency-based information trajectory emerge for each language, robust across each utterance length within each language. We see the same distribution of information for both parents and children in each language, with the child's curve normally below the parent's curve, likely due to the parent speaking a larger share of the utterances in the corpus and using a larger vocabulary than the child. We see the opposite amplitude trend in the Shiro corpus, where the children speak more than the adult investigators as the children are the ones telling the stories. We include the frequency-based information curve from the Providence corpus for comparison.

English, Spanish, French and German feature similar information curves shapes, with slight variations. The German information curve features lower information for longer
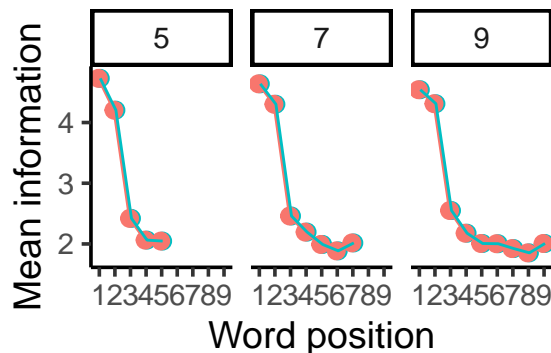
Figure 7. (#fig:providence_trigrams)Providence trigram context-based information curves

towards the beginnings of utterances, possibly due to the grammatical restriction that the second word in German utterances must be a verb (V2). Spanish features a larger spike in the amount of information in the final words of utterances. For Japanese and Mandarin, we observe completely different frequency-based information curve trajectories. The Japanese frequency-based information curve trajectory begins high and finishes low, the mirror image of the European language information curves. The Mandarin curve begins low and finishes low, but features high information in the middle of utterances. We hypothesize this may be due to both Japanese and Mandarin speakers typically ending their utterances with particles, which contain very little information on their own. The penultimate noun with the speakers pair the particles are where the information lies.

For the bigram and trigram information curves, we see the same contextual smoothing effect as for all the corpora we worked with in English. While the frequency-based information curves may depend based on the language, the contextual information curves show the same trajectory cross-linguistically. Using more than two words of context is difficult for parent-child speech corpora because the utterances are so short on average (less than 10 words). We hypothesize that the frequency-based information curves may vary based on the genealogy and typology of the languages in question, but this does not extend to the information curves with two words of context in particular.
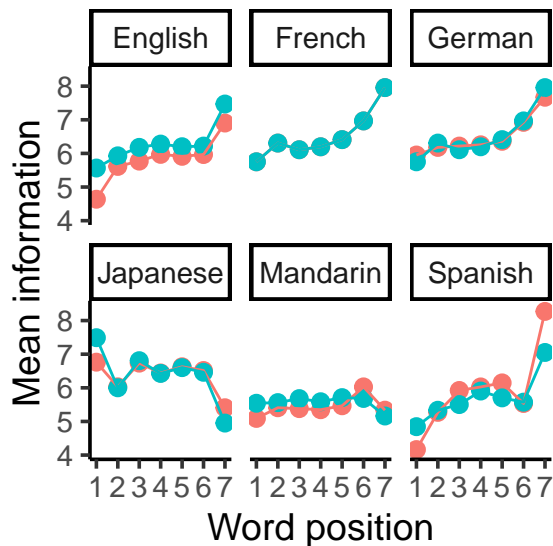
Figure 8. CHILDES frequency-based information curves

Language structure and large-scale data analysis: methods and preprocessing

To make a claim about language as a whole and languages on a larger scale, we needed to use larger corpora and a much larger number of languages. We pulled corpora for 159 diverse languages from Wikipedia, each of which had at least 10000 articles on the knowledge base. We split each article up into sentences; the variance in sentence lengths for Wikipedia was significantly larger than for the CHILDES corpora we used in the previous section. Most sentences in Wikipedia contained between 10 and 30 words, unlike the CHILDES corpora which mainly contained utterances with under 10 words. We excluded the small fraction of utterances with more than 50 words since they were small in number and, from manual inspection, uncharacteristic of typical written sentences.

How do we analyze more than 40 different surprisal curves for each language, adding up to several thousand surprisal curves total? We used two different strategies, which yielded identical results upon analysis. Each strategy gave us a five-dimensional vector space embedding for each language in a Wikipedia "slope space". For the first strategy, we split each sentence length by number of words into fifths, and computed surprisal values for the closest word position to each quintile. We then computed the slopes between the surprisal
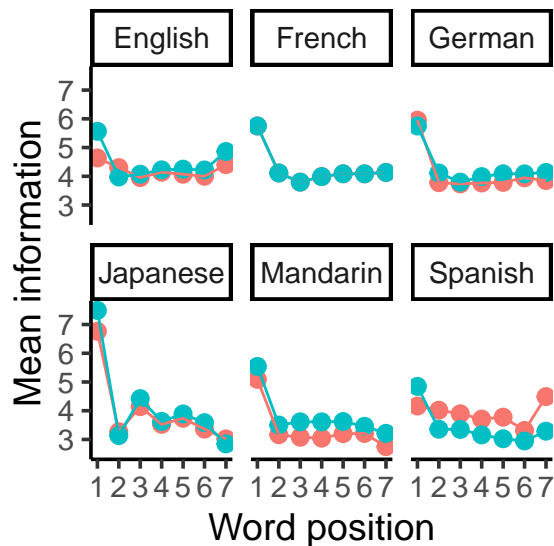
Figure 9. CHILDES bigram context-based information curves

values at neighboring quintiles, yielding five slope values for each curve. For the second strategy, we split each sentence length by number of words into sections based on those areas of the surprisal curves that had seemed most important before: between the first and second word; between the second and third word; between the third word and third-to-last word; between the third-to-last word and the second to last word; and between the second-to-last word and the last word. We then similarly computed surprisal values at each of these positions, and similarly computed slopes between the surprisal values at each position, giving us another five slope values for each language summarizing the surprisal curves. Illustrations of these two strategies are in Figure 11

We computed unigram, bigram and trigram values for each language, and grouped the contextual and frequency-based values for each treatment.

To more rigorously described the differences between languages, we used data from the World Atlas of Language Structures (WALS; Dryer & Haspelmath, 2013). WALS has data for 144 typological features in 2569 languages around the world. These features describe aspects of morphology, syntax, phonology, etymology and semantics–in short the features
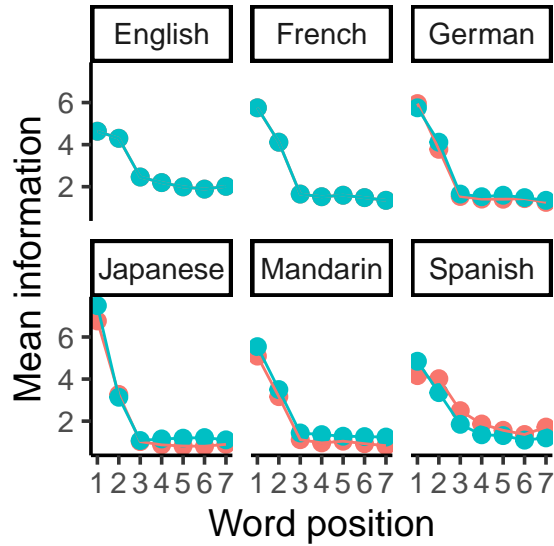
Figure 10. CHILDES trigram context-based information curves

describe 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. Even limiting ourselves to the 159 language corpora we pulled from Wikipedia and 122 features from WALS, there are nearly 20000 individual possible feature values, fewer than half of which were already inputted for those languages in the WALS database.

We used the Multiple Imputation Multiple Correspondence Analysis algorithm (MIMCA; Audigier, Husson, & Josse, 2017) to fill in the missing data using statistical imputation. MIMCA essentially uses mean imputation to begin with the missing values, then performs repeatedly performs principle components analysis on and reconstructs the contingency table formed from observations in categorical variables. By the end of this we had frequency-based and contextual information curves for the 159 language corpora pulled from Wikipedia, along with 122 typological features for each language.

However, the WALS features describe specific structural differences between languages, while our surprisal metric is lexical. To target lexical differences between languages, we computed the average normalized Levenshtein distance (LDN; Holman et al., 2008) over the
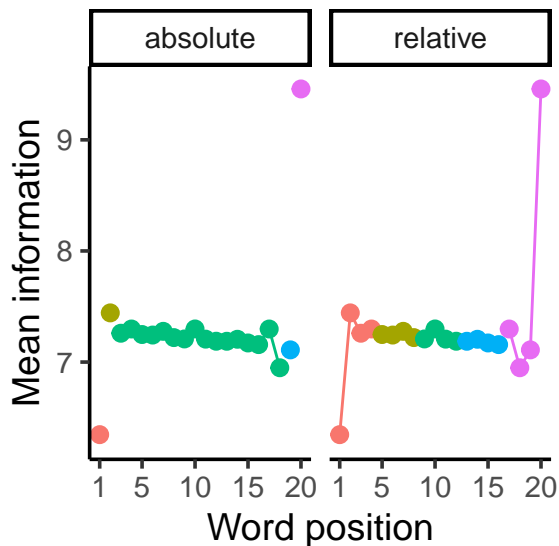
Figure 11. Illustration of slope treatments for Wikipedia information curves: relative on top and absolute on bottom

377 40 item Swadesh list (Swadesh, 1955), retrieved from the ASJP database (Wichmann et al.,

378 2016). The Swadesh list is designed to include near-universal words that target basic

379 cognitive concepts, and are useful in determining the genealogical similarities and differences

380 between languages. The results of classifying languages using the Swadesh list and LDN are

381 correlated with those using WALS features, but the Swadesh list and LDN do not suffer from

382 the same sparsity problem as WALS (Holman et al., 2008).

383   **Language structures and large-scale data analysis: results**

384   We ran a hierarchical clustering algorithm on the unigram information curves using the

385 hclust package from the R stats core library (Team & others, 2013). We used the complete

386 linkage algorithm for hierarchical clustering with distances between information curves in

387 language computed using cosine distance between their embeddings in the slope space. The

388 complete linkage algorithm at every step pairs each language or cluster of languages with its

389 closest neighboring language or cluster. A sample from the dendrogram is shown in Figure

390 **??**, and the full tree can be constructed using the code in our GitHub repository. The

391 language family-like structure can be seen at first glance, although the dendrogram does not

exactly replicate language genealogy for all 159 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 the entire story.
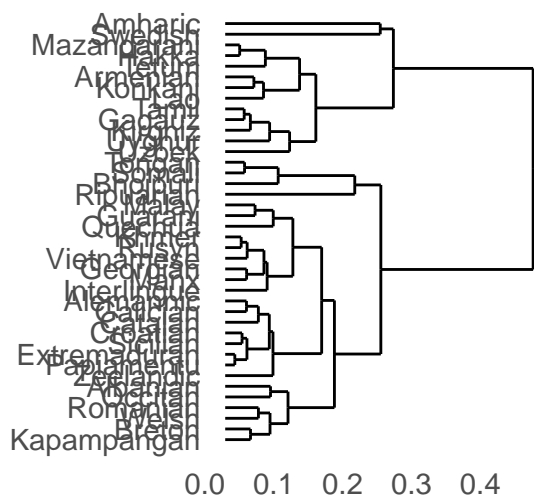
A sample of the contextual information curves (two words of context) are plotted in Figure @ref(fig:wiki_trigrams), and all trigram information curves for the languages we used follow the same pattern. The first few words in utterances for each language are surprising, but with even a couple words of predictive context for each word, the amount of information in each word drops off and speakers produce information at a constant, optimal rate.
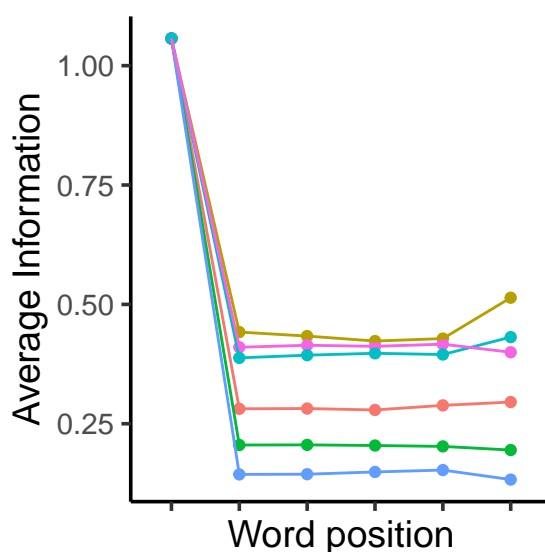

Figure 12. (#fig:wiki_trigrams)Some trigram information curves from the Wikipedia data

First we examined the effects of individual typological features on the shapes of the unigram information curves. We ran logistic regressions using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2014), checking whether the cosine distance between two languages' embeddings in the slope space played a role in determining if they had the same value for a given WALS feature. Individual WALS features do not necessarily have ordinal values. Some, such as the "Number of Cases" feature, are easy to quantify and order. Others are more difficult. For example, how does one order "relative clauses appear after the nouns they modify", "relative clauses appear before the nouns they modify" and "free order of relative clauses and nouns"? We chose the identify relation to avoid deciding on the basis of individual features. We found that 100 out of the 120 features from WALS we examined was statistically significant ($p < .001$) in determining whether two languages had the same shape to their unigram information distributions. The results for some important features are in Figure @ref(fig:linear_models), and the rest of the results are in the Appendices.

mportant Features	Exam
lative Clause and Noun	Noun–RC;
mber of Cases	No cases
bject, Object and Verb	SOV;
phological Imperative	Only singular; s
efinite Articles	Affix; dist
on of Case Affixes	Prefixes;

Figure 13. (#fig:linear_models)Some linear model results from Wikipedia and WALS features

We next compared how the cosine distance between two languages related to how many WALS features they had in common. r^2 value is .005734, which suggests that in aggregate there is not a correlation between how many WALS features languages have in common and the similarity of their frequency-based information curves. Figure @ref(fig:cosine_wals) displays the results. This result is surprising based on the significance of many WALS features in predicting the shapes of the frequency-based information curves,

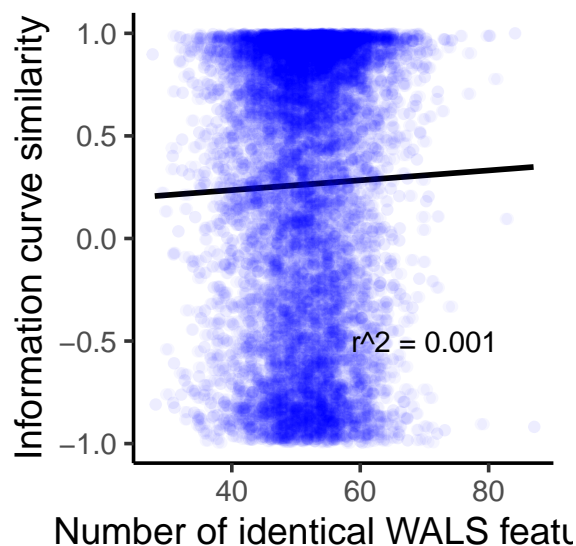₄₂₀ and we return to this result in the general discussion.

Figure 14. (#fig:cosine_wals)wals features vs cosine similarity

₄₂₁   For lexical features, we see a stronger correlation between the similarity of two

₄₂₂ languages in terms of their average LDN and the cosine distance between their information

₄₂₃ curves. Figure @ref(fig:cosine_ldn). We see a higher r^2 value here of .026, indicating that

₄₂₄ there is more correspondence between a language's lexical similarity to another language and

₄₂₅ their similarity in information curves.

₄₂₆   From these typological investigations, we conclude that the shape of a language's

₄₂₇ frequency-based information curve covaries with its typological and lexical similarity to other

₄₂₈ languages.

Discussion

₄₃₀   By considering the distribution of information at the level of utterances and sentences,

₄₃₁ we join together the information-theoretic work focusing on sub-word units and words, and

₄₃₂ that focusing on paragraphs. In doing so, we show that frequency and context-based metrics

₄₃₃ complement one another in studying efficiency and information in language. We directly link

₄₃₄ linguistic efficiency in a language to the genealogy and properties of that language, We
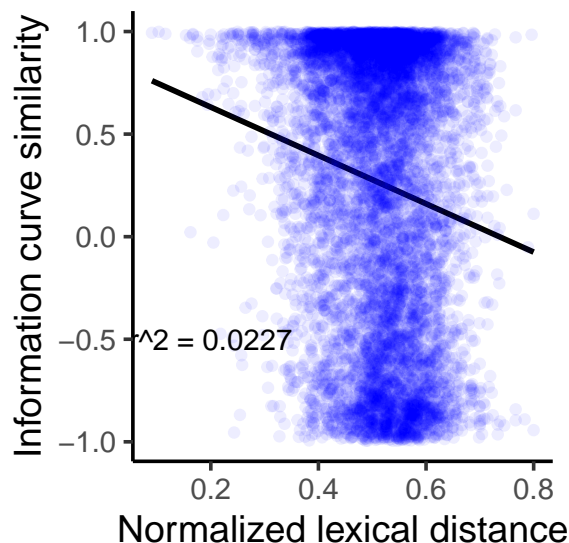
Figure 15. (#fig:cosine_ldn)ldn features vs cosine similarity

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. Also considering the developmental angle, observing language-specific information distributions arising as soon as children begin speaking in 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.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823.*

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from https://wals.info/

Evans, K. E., & Demuth, K. (2012). Individual differences in pronoun reversal: Evidence from two longitudinal case studies. *Journal of Child Language, 39*(1), 162–191.

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).

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., . . . Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings*

of the *National Academy of Sciences*, *114*(40), 10785–10790.

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.

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.

Li, H., & Zhou, J. (2015). *Study on dinner table talk of preschool children family in shanghai* (Master's thesis). East China Normal University, Shanghai, China.

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.

Palasis, K. (2009). *Syntaxe générative et acquisition: Le sujet dans le développement du système linguistique du jeune enfant* (PhD thesis). Nice.

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.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity

in language. *Cognition*, *122*(3), 280–291.

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

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.

Shirai, Y. (2001). The acquisition of causative morphology in japanese: A prototype account.
        East Asian Language Acquisition.

Shiro, M. L. de. (1996). Un estudio de las expresiones de modalidad en hablantes de dos
        culturas. *Boletı'n de Lingüı'stica*, (10), 43–60.

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.

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

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language*,
        12(2), 475–487.

524 Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., . . .

525   others. (2016). The asjp database. *Max Planck Institute for the Science of Human*

526   *History, Jena.*

527 Yu, S., Cong, J., Liang, J., & Liu, H. (2016). The distribution of information content in

528   english sentences. *arXiv Preprint arXiv:1609.07681.*

529 Zipf, G. K. (1949). Human behavior and the principle of least effort.