Availability, informatively and bustiness: Why average corpus measures are an inaccurate guide to surprisal in language

Sihan Chen (sihanc@mit.edu)
Department of Brain and Cognitive Sciences, MIT
43 Vassar St. Cambridge, MA 02139 USA

Edward Gibson (egibson@mit.edu)
Department of Brain and Cognitive Sciences, MIT
43 Vassar St, Cambridge, MA 02139 USA

Michael Ramscar (michael.ramscar@uni-tuebingen.de)
Department of Psychology, University of Tübingen
Schleichstraße 4, 72076 Tübingen, Germany

Abstract

It has been proposed that Chinese classifiers facilitate efficient communication by reducing the noun uncertainty in context. Although recent evidence has undermined this proposal, it was obtained using the common method of equating noun occurrence probabilities with corpus frequencies. This method implicity assumes words occur uniformly across contexts, yet this is inconsistent with empirical findings showing word distributions to be bursty. We hypothesized that if language users are sensitive to burstiness, and if classifiers provide information about upcoming nouns, this information will be less important in reducing uncertainty about noun after their first mention. We show that classifier usage provides more information at earlier mentions of nouns and and less information at later mentions, and that the actual classifier distribution appears inconsistent with previous proposals. These results support the idea that classifiers facilitate efficient communication and indicate that language users representations of lexical probabilities in context are dynamic.

Keywords: information theory; lexical processing; corpus analysis; burstiness; dynamic language modeling

Introduction

Although it is often claimed that language makes humans unique, it remains the case that centuries of contemplating its nature has yielded little consensus as to what 'language' actually is, let alone how human communication processes actually work, or are realized in our brains. In recent years, many researchers have adopted tools from information theory (Gibson et al., 2019) in order to explain how languages enable the communication of information. Information theory models communication as an abstract process involving a sender, a receiver, and a channel, and perhaps critically, formally defines what information is and how it can be quantified in order to specify how communication across can be optimized and the effects of channel noise alleviated.

The information provided by communicative events – e.g. symbols occuring in a code – is defined by their probabilities of occurrence, with higher probabilities defined as being less informative and smaller probabilities more informative (Hartley, 1928). The theory then shows how although physical channels impose limits on the amount of information that can be transmitted across them at a given time (channel capacity), if information is defined in this way, communication can be optimized if the probabilities of communicative events occurring and their encodings are arranged so as to keep the average number of code-symbols occurring at any point in communication close to the channel's capacity (Shannon, 1948).

Table 1: The relative rates of occurence of 8 most frequent and the 2 least frequent English letters in Google ngrams (Norvig, 2013).

LETTER	OCCURRENCE RATE	ENCODING
E	12.49%	I
T	9.28%	O
A	8.04%	II
O	7.64%	OI
I	7.57%	IO
N	7.23%	OO
S	6.51%	III
R	6.28%	IIO
...	...	...
Q	0.12%	OOOO1
Z	0.09%	OOOOO

This can be explained more intuitively if we consider the rates at which English letters occur in the Google ngrams dataset (Norvig, 2013), the most and least frequent of which are shown in Table 1. Given the way lexical information was defined above above, it follows that E will communicate the least information of any English letter (it occurs most often, and so is most predictable), and Q and Z the most. If a binary encoding of these words were to employ 1 symbol to encode E, and 4 symbols each to encode Q and Z (see Table 1), then it follows that given this distribution, the average number of symbols needed to communicate the orthographic forms of English words will be much lower than if they were encoded using the same number of symbols. The highly skewed cod-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ing probabilities will also serve to ensure that the average amount of information communicated at any point in time will be much closer to a channel's capacity limit than if English letters were distributed across words with more similar frequencies, or (in the worst case from a coding perspective) if they were all equally frequent (when all of the efficiency benefits that a skewed distribution brings will be lost).

Natural languages have been shown to exhibit a lot of the same properties as information theoretic codes: for example, just as frequencies of letters in Table 1 appear to vary systematically, so too do the frequencies of all words across languages (Estoup, 1912; Zipf, 2013). Further, words that occur more often, and are thus more predictable, not only tend to be shorter in line with our earlier discussion of coding (Piantadosi, Tily, & Gibson, 2011), they are also more likely to be phonetically reduced in production (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009). These (and other) similarities between information theoretic codes and natural languages have prompted researchers to propose that information theory might not only be a good model for understanding human communication (Gibson et al., 2019), but also, more specifically, that languages are efficient in the way that information theoretic codes are, in that they serve to keep the average rate of information in communication constant ('smooth'), optimizing communication in much the same way as information theory proposes (Aylett & Turk, 2004; Jaeger & Levy, 2006).

In its the explicitly probabilistic characterization of communication, information theory also serves to highlight some hitherto unappreciated challenges that face any theory of human communication that seeks to explain how language is learned and used in terms of their underlying process, which appear to be predictive in their nature (Schrimpf et al., 2021). In particular, it serves to highlight the potential problems that nouns pose to any probabilistic account of human communication. Nouns (both common and proper) comprise by far the dominant number of lexical types in most languages, and as a result this means both that are learned continuously across the lifespan (no speaker will ever learn all of the nouns of any modern language), and that nouns dominate the long tails of low frequency types found in distributions of words and lexical combinations (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). Indeed, the distributional properties of nouns appears to guarantee that unless their prediction is somehow supported systematically in grammars, nouns will inevitably be by far the least predictable part of speech, such that the average amount of information communicated over time must inevitably be subject to huge peaks wherever a noun occurs (Dye, Milin, Futrell, & Ramscar, 2018). All of which raises some questions: do natural languages contain systems that help to manage the uncertainty associated with nouns; or are they in fact far less efficient than has often been claimed.

With regards the first question, it has been shown, first, that as compared to English articles (which are typically not marked for grammatical gender), German articles serve to make nouns more predictable in context, such that the German noun class system supports the use of more informative nouns after articles than English (suggesting a functional role for an aspect of language that has tended to confound traditional, logical theories) (Dye, Milin, Futrell, & Ramscar, 2017); and, second, that English makes more use of prenominal adjectives than German (Dye et al., 2018). These findings suggest that English and German do contain devices to smooth the information associated with noun phrases, but that they employ different means to achieve a similar end. At which point another, different mechanism for managing the uncertainty associated with nouns can be added to the mix, because a similar functional role has been proposed for Chinese classifiers (words that link numerals and nouns together): it has been suggested Chinese classifiers also serve to provide information that makes nouns more predictable in context (Klein, Carlson, Li, Jaeger, & Tanenhaus, 2012).

However, recent findings have cast serious doubt on this last proposal. First, Zhan and Levy (2018) have claimed that as opposed to the kind of continuous, skewed distributions associated with information theory, Chinese classifiers in fact comprise a set of specific classifiers, and a general classifier *ge*. The general classifier *ge* can be paired with virtually every noun, whereas specific classifiers can only proceed specific subsets of nouns, based on the properties of the entities they refer to. For example, specific classifier *pian* usually proceeds nouns referring to objects that are thin and flat, such as leaves and papers. Second, the evidence provided by a study that sought to directly test whether a function of classifiers is to reduce uncertainty in noun phrases in Mandarin Chinese suggest that they do not in fact do this. Rather, in a series of corpus analyses Zhan and Levy (2018) revealed that Mandarin speakers tend to use the general classifier when they use low-frequency nouns, and that the use of specific classifiers tends to be reserved for higher frequency nouns. Given that this pattern of observations runs counter to any obvious information-theoretic account, Zhan and Levy (2018) suggest that the way Chinese classifiers are used supports an availability account of classifier usage instead. (Zhan & Levy, 2018) propose that given that nouns must be 'retrieved' in order for any specific classifier information associated with them to be accessed, and given that lower frequency nouns will be harder to 'retrieve' than higher frequency nouns, it would appear to follow that their results best support an account of classifier use in which speakers default to using the general classifier whenever an upcoming noun is hard to retrieve from memory (i.e., unavailable).

However, in common with almost all other studies in the field (e.g., (Dye et al., 2018; Bell et al., 2009; Piantadosi et al., 2011), [1], in Zhan and Levy (2018) analyses, the uncer-

[1] A range of estimation measures are typically employed in estimating lexical probabilities for the purposes of information theoretic analyses of language, for example: the *prior probabilities* of words (i.e., their frequencies in a corpus); the *joint probabilities of lexical sequences* (estimated from sequence frequencies in a corpus); and the *conditional probabilities* between words (the conditional prob-

tainty associated with nouns was estimated by calculations based on average corpus frequencies. Given that these estimates are based on counts of how many times a word (or ngram) appears in a corpus – which is an aggregate of texts that contain a wide range of contexts – they implicitly assume that words occur across different contexts in a relatively uniform manner. This assumption is, however, inconsistent with the fact that occurrence of words in context is bursty (Katz, 1996; Altmann, Pierrehumbert, & Motter, 2009; Slone, Abney, Smith, & Yu, 2023): words, especially low frequency words, are likely to occur multiple times in a few texts, but never in most others. For example, the word 'asystole' occurs frequently in medical texts related to cardiac arrest, but will almost never be found in any other types of text. Thus an important implication of burstiness is that when most words first appear in a text, the chance that they will reappear is much, much higher than the overall likelihood of their appearing in any particular random text. This in turn indicates that the actual likelihoods with which words occur across context must be dynamic, and since estimates made by averaging the occurrence across all contexts in a corpus cannot capture this kind of dynamicity, it further indicates that in some contexts, word occurrence probabilities based on average frequencies will be far from correct. Moreover, it follows that if language users are sensitive to this property in their use of linguistic codes, then after encountering a specific word in a context, their expectations about encountering that same word again should increase, which would suggest that in many contexts, the probabilities estimated by aggregating across contexts (corpus frequencies) will be poor predictors of their communicative behavior.

Importantly, there is evidence that language users may in fact be sensitive to the bursty nature of word occurences: when referring to a novel object, speakers tend to shorten their reference phrases as further reference to an object is made (Krauss & Weinheimer, 1964); similarly, if a word has already been mentioned in context, speakers will tend to articulate more rapidly, and in a more phonetically reduced manner, when it recurs (Aylett & Turk, 2004; Bortfeld & Morgan, 2010; Tippenhauer, Fourakis, Watson, & Lew-Williams, 2020; Shi, Gu, & Vigliocco, 2022). Given this, it is worth considering what the implication of burstiness – and of language users potential sensitivity to it – are when it comes to the way we might expect Chinese classifiers to be used in text and discourse. If classifiers provide information about the upcoming nouns, as an information-theoretic account of language predicts, then this information will serve a more important role in reducing the high uncertainty of a noun when it is initially introduced. However, if language users are sensitive to burstiness, it will play a less important role in reducing the now lower uncertainty of the same noun afterwards, since it has now become more predictable as a result of its already having occurred in that context. In other words, if language users are sensitive to burstiness, and they have a choice of which classifier to employ, we would expect then to use less informative classifiers as they reuse nouns in any given context. By contrast, and interestingly, the availability account proposed by Zhan and Levy (2018) makes the opposite prediction: if language users have to retrieve nouns in order to access their classifiers, and if they default to the gerneral classifier when they fail to retrieve nouns, then given that most nouns are low frequency and low frequency nouns will be harder to retrieve, we should expect the classifiers they use to become more informative when nouns are reused in context.

These two accounts thus provide a potentially fruitful setting in which to examine – and ideally illustrate – the dynamic nature of human source codes. In the following study, we sought to test the predictions laid out above – while hopefully illustrating the dynamicity of noun probabilities in context – by examining whether the information provided by classifiers does in fact decrease when nouns are reused in context.

Methods

Corpus data preprocessing

Following Zhan and Levy (2018), we based our analysis of classifier use on the SogouCS corpus (Zhang, 2021), a compilation of short Chinese news articles from the website Sohu News. We mainly followed the preprocessing procedures described in Zhan and Levy (2018): we first filtered out all the texts not related to the content, such as author credits, datelines, and advertisements. Since we were also interested in the number of times each noun occurred in each article, it was important to distinguish individual articles in the corpus. To this end, author credits were employed as an end-of-article indicator. We next removed files that contain texts that are not news articles (e.g. lists of essay prompts, lists of names, lists of definition, transcripts, interviews). Then, within the remaining texts, we only included complete sentences, those ending with a period, a question mark, or an exclamation mark in our analysis set. The resulting files were then fed (one sentence per line, and one article per file) into the Stanford CoreNLP toolkit (Manning et al., 2014) and all the numeral-classifier-noun (NCN) trigrams, where the head noun has a `nummod` dependency, and the numeral has a `mark:clf` dependency with a classifier were extracted from the parser output. For the purposes of our analysis, we only included numerals ranging from one to five and classifiers from a list on Wikipedia[2]. Given that the distribution of numerals follows Benford's Law (Benford, 1938) which describes the nonlinear distribution of numerals (which is similar to the distribution of color adjectives in Table 1, with 1 being the most frequent numeral, 2 the the 2nd most frequent numeral, etc.), this meant that we could expect to capture around half of the classifiers in the corpus while keeping analysis manageable.

ability of a target word *w* given a previous word *wi-1)* in a corpus). What all of them have in common is that they are estimated from counts that average across contexts and thus implicitly assume that words and ngrams tend to occur uniformly across contexts.

[2]classifiers listed under the "Classifiers Proper" section `https://en.wikipedia.org/wiki/List_of_Chinese_classifiers`

We also removed NCN trigrams that are not actually numeral-classifier-noun constructions (e.g. "两面针" referrs to a plant, despite being parsed as a trigram). Since not all articles had author credits, each file could potentially contain more than one article. To increase the chance that each file contains only one article, we only included the first ten mentions of each noun in each file (visual inspection of our data indicated that the chances of the same noun being repeated more than a few times in a short news article were exceedingly small, and that articles containing more than this had simply evaded our pre-screening for lists etc.).

Analysis

At the end of the processing phase, our final dataset comprised 1583309 instances of 81531 NCN trigrams, from 557207 files. For the purpose of this analysis, we will represent these as trigrams (a,c,n) satisfying $(a,c,n) \in \{(a,c,n) \mid a \in A, c \in C, n \in N\}$, where $A = \{1,2,3,4,5\}$, C is the set of classifiers, and N is the set of nouns in the dataset. Each appearance of a noun can be uniquely identified by two numbers: the article and the n^{th} mention in the article. Therefore, we can further represent each appearance of a noun n by n_{ij}, where $i \in I$ refers to the article, and $j \in J$ refers to the index of the number of times it has been mentioned.

We operationalized the informativity of classifiers as the **diversity** of classifier usage given a numeral and the j^{th} appearance of nouns within the same text. Here we conditioned the diversity of classifiers on each numeral separately because in the same way that classifiers provide information about upcoming nouns, numerals provide information about upcoming classifiers and nouns. For instance, the noun '月亮(Moon)' is unlikely to follow any numeral greater than one, since in Mandarin, the word only refers to the Moon, and not satellites of other planets. According to the information-theoretic account, if users are sensitive to burstiness (such that when a noun occurs, they increase their estimates of its likelihood of reoccurring), then classifiers will play a less important role in reducing the uncertainty of upcoming nouns at later appearances of an article. Hence, we should expect **a lower diversity** of classifiers, because the need for classifiers to help reduce the uncertainty about upcoming nouns will be reduced. In contrast, according to the availability account, nouns that have already been mentioned will be more available when they reappear later, and because the theory proposes that nouns must be retrieved for classifiers to be accessed (Zhan & Levy, 2018), collectively, the range of classifiers available to speakers ought to be greater at subsequent mentions of nouns as opposed to at first mention, such that on this account, we should expect **a higher diversity** of classifiers at subsequent as opposed to first mention. Since the predictions of both accounts are independent of the information provided by numerals, we should broadly expect the patterns predicted by each account to be observed after each numeral.

We formulate classifier diversity as the entropy of classifiers, given a numeral and the j^{th} appearance of nouns within a text, aggregated across all the nouns, following the equation below.

$$H(C \mid a, j) = -\sum_{c} p(c \mid a, j) \cdot \log_2 p(c \mid a, j) \qquad (1)$$

The term $p(c \mid a, j)$ denotes the probability of classifier c being used at j^{th} mention after numeral a, which can be estimated by the Equation below:

$$p(c \mid a, j) = \frac{\#\{(a', c', n_{ij'}) \mid a' = a, c' = c, j' = j\}}{\#\{(a', c', n_{ij'}) \mid a' = a, j' = j\}} \qquad (2)$$

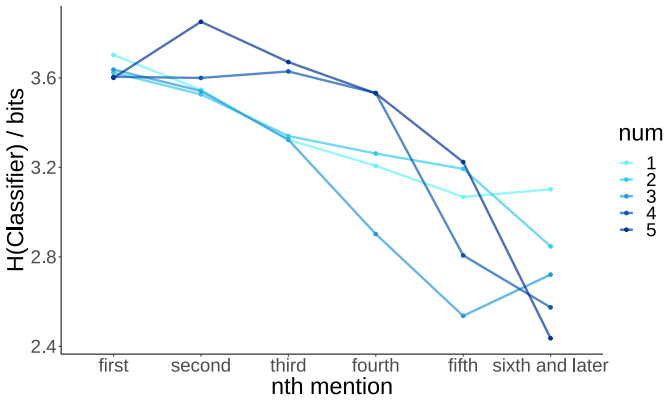

Figure 1: **The diversity of classifier (measured as conditional entropy) after a numeral as a function of mentions.** The entropy was calculated individually up to the fifth mention, and the entropy after the sixth mention is calculated as a whole. Each colored line represents a different numeral from one (light blue) to five (dark blue). As a noun is mentioned more and more times, classifier entropy goes down, meaning that classifiers become less important.

Results

Figure 1 shows the classifier diversity conditioned on each numeral, as a function of nth mentions. There is a clear decrease in classifier diversity as nouns recur more and more within a text. To quantify the trend, we conducted a mixed-effects linear regression on the entropy data, coding the number of mentions (1-5, and treating 6 and after as 6) as a fixed effect, adding random intercepts by numeral and random by-numeral slope of number of mentions. We found a significant negative slope for the number of mentions (β = -0.225, p = 0.005), suggesting on average, each time a noun is mentioned within the same text, the preceding classifier is 0.22 bits less diverse .

Next, to see if the default function that has been proposed for the the general classifer *ge* actually does produce usage patterns that run counter to the predictions of an information-theoretic account, we inspected the classifier distribution in our dataset and compared it to the distribution of prenominal words in other languages that have been proposed to serve as cues to reduce noun entropy: adjectives and articles (Dye

et al., 2017; Frantzi, 2022). We counted the occurrence frequency of English color adjectives from the Corpus of Contemporary American English (COCA)(Davies, 2008), and the occurrence of gendered articles in Greek from the Greek Web Corpus (elTenTen)(Jakubíček, Kilgarriff, Kovář, Rychlỳ, & Suchomel, 2013). The results are presented in Figure 2. The distribution of Mandarin Chinese classifiers (Figure 2a) is similar to those of Greek articles (Figure 2b) and English color adjectives (Figure 2c). Our results seem to contradict the notion of the general classifier *ge* being the 'default' classifier, in the following two senses. First, we'd expect a much higher occurrence frequency of *ge* if it were to be the default classifier. Instead, the distribution of Chinese classifiers resembles that of Greek articles and English color words. Based on these distributions, saying *ge* as the general classifier is the same as saying τηϛ as the general article in Greek or saying *white* is the general color adjective in English. Second, in our dataset, following (Zhan & Levy, 2018), we excluded nouns that were never paired with *ge* in the corpus, which already gave *ge* an advantage, but even with this, in some contexts (e.g. after fourth mentions and the numeral 4 in Figure 2a), *ge* is not even the most frequent classifier.

Discussion

It has been proposed that human languages are efficient communication systems in the sense that the uncertainty is smoothed out over messages in order to make the maximal use of the communication channel (Frank & Jaeger, 2008; Maurits, Navarro, & Perfors, 2010; Meister et al., 2021). As we noted in our introduction, the distribution of nouns pose a particular challenge to information theoretic account of human language processing. Thus although previous studies (Dye et al., 2018; Klein et al., 2012) have proposed that function words preceding nouns might serve as cues to reduce the uncertainty of the upcoming noun, a recent study (Zhan & Levy, 2018) on Chinese classifiers appears to undermine this specific proposal and, in finding evidence for a speaker availability account, raised questions about these other proposals.

We examined whether approximating noun occurrence probabilities from corpus statistics may be inadequate when it comes to understanding many aspects of human communication. Our principle concerns in this regard was that any measure that bases its lexical occurrence probability estimates on relative frequencies derived from a corpus must implicitly assume that nouns tend to appear in a uniform fashion across contexts, an assumption that is inconsistent with empirical facts that indicate that the distribution of nouns across contexts is uneven, i.e., bursty (Katz, 1996). The bursty nature of individual noun occurrence patterns means that the occurrence probabilities of nouns are in fact dynamic, such that that if a noun has already appeared in a context, this is a much (as in orders of magnitude) greater chance of it appearing again than there is of the same noun occurring in a random text. That is, because word occurrences are bursty, the base rate probability of a noun's occurrence can and usually will be

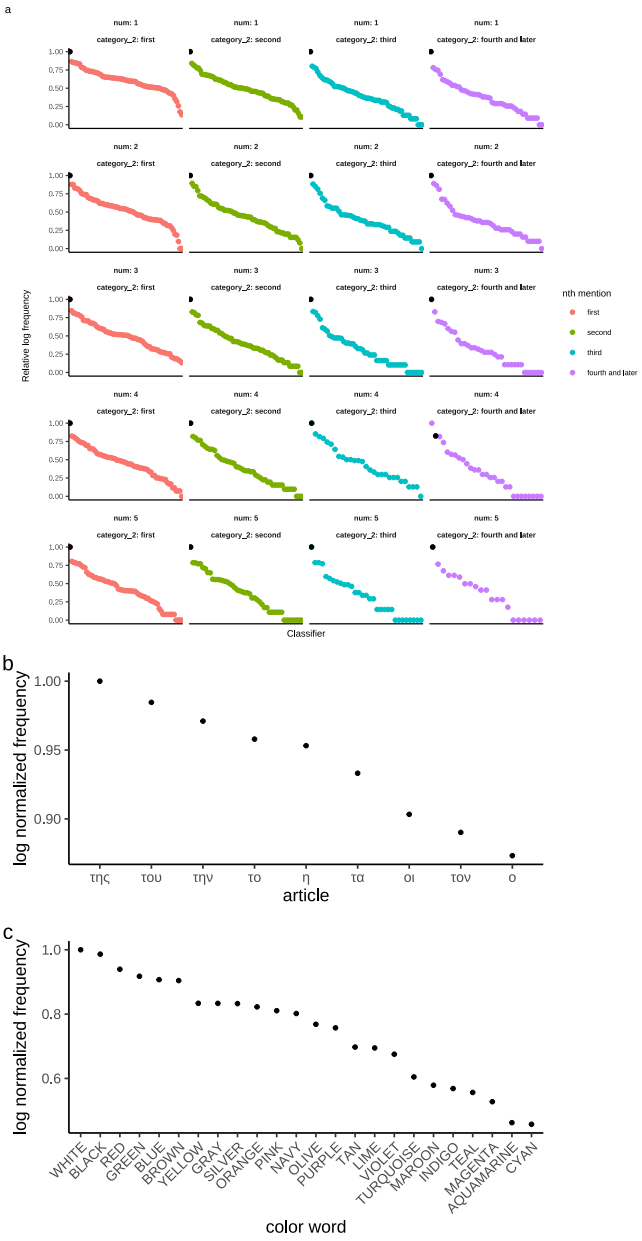

Figure 2: **Distribution of prenominal words in three languages,** ranked by their relative log frequency. **a)**Classifier distribution in our dataset extracted from SogouCS corpus (Zhang, 2021), faceted by numerals (rows) and j^{th} mention in an article (column). The general classifier *ge* is labeled by a black dot in each facet. **b)** Article distribution in the Greek Web Corpus (Jakubíček et al., 2013; Frantzi, 2022). **c)** Color word distribution from the Corpus of Contemporary American English (Davies, 2008).

massively lower than its re-occurrence probability in context.

To examine the implications of this we considered a potential function of Mandarin Chinese classifiers, namely that they serve to reduce the uncertainty associated with nouns in context. An earlier examination of this proposal appeared to find evidence against it, finding that on average Chinese speakers tended to use high frequency classifiers with most nouns most of the time, and, '[assuming] ... that a speaker must access a noun lemma in order to access its appropriate specific classifier' Zhan and Levy (2018) concluded that speakers must be defaulting to a high frequency classifier whenever they 'failed to retrieve' a noun lemma.

It is worth noting that the logic of Zhan and Levy (2018)'s analysis is that if Chinese classifiers serve to reduce the uncertainty associated with nouns, then they should always do this. As such, they assume that one should expect that Chinese speakers should tended to use lower frequency classifiers with most nouns most of the time. However, an information theoretic account of classifier use that takes account of burstiness makes far more nuanced predictions. It predicts that after a noun first appears in a context, prenominal function words such as classifiers will serve a less important role in reducing uncertainty of the noun and that at subsequent mention, speakers might either omit them, or where these function words are obligatory – as Chinese classifiers are – either produce them in reduced form (in speech), or in a less informative form if they are generating text.

It is notable that, if we were to consider what speakers do on average, teasing apart these two accounts from the data would be problematic. However, if we take burstiness into account, and consider their predictions in terms of mention, then the two accounts make opposite predictions. The information theoretic account predicts that Chinese speakers will use first more and then less diverse classifiers; the availability account predicts that Chinese speakers will use first less and then more diverse classifiers. We examined how Mandarin Chinese actually do use classifiers across context to tease apart these two accounts, operationalizing diversity as the conditional entropy of classifiers given a numeral at j^{th} mention. As predicted, we found that the diversity of classifiers given a numeral decreases as the number of mentions increases, suggesting not only that classifiers are indeed becoming less important as mention recur, but also highlighting how important it is for researcher to take the dynamic nature of occurrence probabilities into consideration when they seek to understand communicative behavior.

Although the burstiness of lexical distributions has long been understood (Katz, 1996), its potential influence has tended to be neglected in research that explicitly addresses prediction in language processing. Our results expose the limitations in current approaches to estimating lexical occurrence probabilities. Moreover, it follows that given that the occurrence of words in context is bursty (Katz, 1996; Altmann et al., 2009; Slone et al., 2023), all approaches that seek to estimate lexical occurrence probabilities from average relative frequencies taken from an collection of aggregated contexts will necessarily suffer the same limitations. Overall, they will tend to estimate the occurrence probabilities of high frequency words (whose distributions are more uniform across contexts) far more accurately than the occurrence probabilities of low frequency words (which are more bursty, Katz (1996)), because they will tend to massively underestimate the recurrence probabilities of low frequency words in context. Given the long tailed nature of lexical distributions (Estoup, 1912; Zipf, 2013), this means that they will tend to consistently make inaccurate estimates about the overwhelming majority of words they are applied to. Indeed, it is interesting to consider human intuitions about the distribution of items in codes from this perspective. If we assume that similar biases guide our intuitions, it may explain why researchers tend to overestimate the likelihood of classifiers like *ge*, assuming them to be "defaults," and why, as our results appear to indicate, that the more continuous relationship of more general to specific classifiers seems much harder to grasp intuitively.

At this point, it is incumbent on us to note one aspect of Zhan and Levy (2018)'s evidence that we have not addressed here. Following their corpus analysis, they conducted a behavioral experiment in which they asked participants to describe a series of stimuli, both under time pressure and not. Zhan et al. (2020) found that participants under time pressure were more likely to use general classifiers before low-frequency nouns than those who were not put under time pressure, and concluded that this too supported an availability based account. While we are currently in the process of beginning to examine the way that speakers actually use classifiers in spontaneous speech, we would note that a study that elicited spontaneous speech in order to examine whether language users employ adjectives to reduce the uncertainty associated with nouns, and if so, whether they are also sensitive to burstiness, found that English speaking participants were more likely to use adjectives to modify nouns at first mention, and that used significantly fewer adjectives on subsequent mentions (Kemper, Jenkins, Wonnacott, & Ramscar, 2024). Although this experiment was conducted among speakers of a different language, and there is a difference between describing a scene spontaneously and naming under time pressure, these results suggested that the latter might not be the most ecologically valid way of examining language production. In this paper, we have emphasized the challenge that the uncertainty associated with noun distributions poses to theories that assume grammars of languages have socially evolved to enable efficient communication, and we also showed how different grammatical mechanisms in various languages that appear to have socially evolved to alleviate this problem, but naming tasks given to participants under time pressure seem to preclude participants from making use of these socially evolved mechanisms that allow users to smoothly access nouns in context. We hope that future studies might serve to determine which of these conjectures is correct.

References

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, *4*(11), e7678.

Aylett, M., & Turk, A. (2004, March). 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. Retrieved from http://dx.doi.org/10.1177/00238309040470010201 doi: 10.1177/00238309040470010201

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, 551–572.

Bortfeld, H., & Morgan, J. L. (2010, June). Is early word-form processing stress-full? how natural variability supports recognition. *Cognitive Psychology*, *60*(4), 241–266. Retrieved from http://dx.doi.org/10.1016/j.cogpsych.2010.01.002 doi: 10.1016/j.cogpsych.2010.01.002

Davies, M. (2008). *Corpus of contemporary american english.* Retrieved from https://www.english-corpora.org/coca/

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In *Perspectives on morphological organization* (pp. 212–239). Brill.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in cognitive science*, *10*(1), 209–224.

Estoup, J.-B. (1912). *Gammes sténographiques.* Institut sténographique.

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).

Frantzi, I. (2022). *The structure of greek gender classes and how they smooth signalling in noun phrases* [Bachelor Thesis].

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Hartley, R. V. L. (1928). Transmission of information 1. *Bell System technical journal*, *7*(3), 535–563.

Jaeger, T. F., & Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, *19*.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The tenten corpus family. In *7th international corpus linguistics conference cl* (pp. 125–127).

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Natural language engineering*, *2*(1), 15–59.

Kemper, S., Jenkins, H., Wonnacott, E., & Ramscar, M. (2024). Rethinking probabilities: Why corpus frequencies cannot capture speakers' dynamic linguistic behavior. *Paper in submission*.

Klein, N. M., Carlson, G. N., Li, R., Jaeger, T. F., & Tanenhaus, M. K. (2012, 09). Classifying and massifying incrementally in Chinese language comprehension. In *Count and Mass Across Languages.* Oxford University Press. doi: 10.1093/acprof:oso/9780199654277.003.0014

Krauss, R. M., & Weinheimer, S. (1964, January). Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, *1*(1–12), 113–114. Retrieved from http://dx.doi.org/10.3758/BF03342817 doi: 10.3758/bf03342817

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In K. Bontcheva & J. Zhu (Eds.), *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P14-5010 doi: 10.3115/v1/P14-5010

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 23). Curran Associates, Inc.

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. Retrieved from https://arxiv.org/abs/2109.11635 doi: 10.48550/ARXIV.2109.11635

Norvig, P. (2013). English letter frequency counts: Mayzner revisited. *Blogpost*. Retrieved from http://norvig.com/mayzner.htm

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.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5–42.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

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

https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shi, J., Gu, Y., & Vigliocco, G. (2022, December). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, *26*(4). Retrieved from http://dx.doi.org/10.1111/desc.13357 doi: 10.1111/desc.13357

Slone, L. K., Abney, D. H., Smith, L. B., & Yu, C. (2023). The temporal structure of parent talk to toddlers about objects. *Cognition*, *230*, 105266.

Tippenhauer, N., Fourakis, E. R., Watson, D. G., & Lew-Williams, C. (2020, November). The scope of audience design in child-directed speech: Parents' tailoring of word lengths for adult versus child listeners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(11), 2163–2178. Retrieved from http://dx.doi.org/10.1037/xlm0000939 doi: 10.1037/xlm0000939

Zhan, M., & Levy, R. P. (2018). Comparing theories of speaker choice using a model of classifier production in mandarin chinese..

Zhan, M., et al. (2020). *Investigating theories of speaker choice in a classifier language* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Zhang. (2021). *Sogou news corpus (sogou).* Zenodo. Retrieved from https://zenodo.org/record/5259056 doi: 10.5281/ZENODO.5259056

Zipf, G. K. (2013). *The psycho-biology of language: An introduction to dynamic philology*. Routledge.