Chinese words shorten in more predictive contexts

Yanting Li, **Gregory Scontras**, and **Richard Futrell**
Department of Language Science
University of California, Irvine
{yantil5, g.scontras, rfutrell}@uci.edu

Abstract

In Mandarin Chinese, abbreviation happens commonly to compound words across different syntactic categories. What is the motivation behind this shortening of words? This paper presents an investigation of this phenomenon from an information-theoretic point of view. A corpus study was conducted to measure the average amount of information contained in the full (long) form and the abbreviated (short) form of words given certain contexts. The amount of information was then compared between the long and short forms of a word, revealing that the short one usually contains less information, and therefore is more likely to be used in more predictive contexts. This result indicates that speakers of Chinese can choose to use shorter words when the context is more predictive, in accordance with considerations of efficiency.

Introduction

People use language to transfer information. In the past two decades, language scientists have been researching the (in)efficiency of natural languages at achieving this goal at various levels, including phonology (Goldsmith, 2002; Cohen Priva, 2015), morphology (Dye, Milin, Futrell, & Ramscar, 2018; Rathi, Hahn, & Futrell, 2022), the lexicon (Piantadosi, Tily, & Gibson, 2011; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021; Mahowald, Dautriche, Braginsky, & Gibson, 2022; Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho, 2017), syntax (Levy & Jaeger, 2007; Futrell, Mahowald, & Gibson, 2015; Hahn, Jurafsky, & Futrell, 2020; Gibson et al., 2019; Ferreira, 2008; Kurumada & Jaeger, 2015), and more. Among them, many researchers proposed different ways to predict word length in language use, with the idea that longer words require more effort to produce. As one of the earliest attempts at answering this question, Zipf (1949) proposed using word frequency as a predictor. Half a century later, Piantadosi et al. (2011) showed that average information content of a word is more effective than frequency at predicting word length: more informative words are longer. This finding suggested that human lexicons are coded in an information-theoretically efficient way to convey meaning: more informative content gets encoded more robustly. Mahowald et al. (2013) looked into the use of English nouns specifically and found that speakers choose to use shorter words in more predictive contexts, which provided more evidence supporting the claim made by Piantadosi et al. (2011): the less information content a word holds, the shorter it is likely to be. Kachakeche, Scontras, and Futrell (2022) conducted similar research on vowel dropping phenomenon in Arabizi, an emerging writing system for Arabic speakers using Roman characters; the authors found similar results supporting the claim. Here we ask whether similar pressures apply in Mandarin Chinese.

In Chinese, each character represents approximately one morpheme. Many words are elastic, meaning that they can be expressed either by a short form or a long form (Duanmu, 1997; Duanmu & Dong, 2016). Li, van Deemter, Paperno, and Fan (2019) applied the method of Mahowald et al. (2013) and conducted a corpus study on such words. Specifically, they investigated 442 monosyllabic–disyllabic word pairs (e.g., 虎 *hǔ* vs. 老虎 *lǎohǔ* 'tiger') and showed that the short (monosyllabic) forms carry less information than the long (disyllabic) forms. While this result suggests that Chinese word length may be influenced by information content, prosody is another factor that influences the choice between monosyllabic–disyllabic word pairs (Duanmu, Feng, Dong, & Zhang, 2018; Duanmu, 2012; Dong, 2015). One way to minimize the potential influence of prosody is by examining non-monosyllabic word pairs instead.

Table 1 contains some examples of non-monosyllabic word pairs. The full forms consist of more than two units, and the abbreviations are all disyllabic, created by selecting one character from each unit in the full form. Unlike English, Chinese does not have an alphabet, nor does it insert spaces in between words in writing, so it does not have a consistent way of shortening phrases by generating acronyms like *NBA* (National Basketball Association) or *WTO* (World Trade Organization) in English. In addition, abbreviation can be ambiguous because a morpheme may have more than one meaning, so that different words containing overlapping characters may be abbreviated to the same form. For instance, 文管 *wénguǎn* can be the abbreviation for both 文化管理 *wénhuà guǎnlǐ* "cultural management" and 文物管理 *wénwù guǎnlǐ* "cultural relic management". The good news is that such ambiguity can often be resolved given the appropriate context. When the intended meaning is clear given the context, speakers should be able to use a less effortful abbreviation, rather than the full form, without worrying about confusing the listener.

If this relationship between contextual predictability and

Word type	Full form	Abbreviation	English meaning
N + N	上海博物馆	上博	Shanghai Museum
Adj + N	高速铁路	高铁	high-speed railway
V + N	断绝外交关系	断交	sever diplomatic relations
V + V	禁止使用	禁用	forbid to use
Adv + V	完全封锁	完封	completely lockdown

Table 1: Examples of abbreviations in Chinese. All examples are chosen from the materials used in this study.

shortening holds, then the finding of the corpus study in Mahowald et al. (2013) should generalize to Chinese: the unabbreviated (long) form of a word should contain more information than its abbreviated (short) counterpart in the contexts where it is used. As information content of a word can be measured by its surprisal—that is, the negative log probability of the word given context (Hale, 2001; Levy, 2008; Hale, 2016)—we can test whether word length in Chinese abbreviation alternations can also be predicted by surprisal by measuring the average surprisal of the short and long forms of a word in a corpus and comparing them (Piantadosi et al., 2011; Mahowald et al., 2013).

Research questions

To see whether considerations of efficiency influence abbreviation decisions in Chinese, we explore whether speakers of Chinese tend to use the abbreviated (short) form of a word more than its full (long) form when the context makes the word more predictable. To investigate this question, the predictability of both forms represented by their average surprisals will be calculated and compared. Our hypothesis is that long forms are less predictable; in other words, long forms contain more information than their short counterparts.

Method

In order to get the data that is necessary for answering the above research question, we need short and long word pairs that can be used interchangeably to express a single concept, as well as a large Chinese corpus where those word pairs appear in context. In addition, a Chinese language model is needed to calculate the surprisal of those short and long forms given the contexts they appear in. In our data analysis, we will use the average estimated surprisal of the concept as the independent variable, its word length (short vs. long) as the categorical dependent variable, and see if surprisal can be used to predict word length.

Materials

Word pairs The short and long word pairs come from the Chinese abbreviation dataset built by Y. Zhang and Sun (2018). The dataset contains over 7,000 pairs of Chinese words and their corresponding abbreviations. Only a portion of the pairs were kept after a screening process, which will be described in more detail below.

Corpus The corpus used for this study is Chinese Gigaword Fifth Edition (Parker, Graff, Chen, Kong, & Maeda, 2011), a comprehensive collection of newswire texts in Chinese. Among the eight distinct sources of newswire in the corpus, seven use simplified Chinese characters, and one uses traditional Chinese characters. We excluded the traditional characters (i.e., Central News Agency, Taiwan) because our language model is trained on simplified Chinese data. Apart from the difference in orthography, vocabulary for certain concepts or entities used by news agencies in Taiwan might be different from the seven other news agencies from mainland China.

Language model and tokenizer The language model used in this study is the Chinese Pre-trained Language Model (CPM), developed by Z. Zhang et al. (2021). We used the CPM model for the tokenization of texts and the calculation of surprisals. Neural language models such as this one deliver surprisal values that are a psychologically relevant for human language processing (Goodkind & Bicknell, 2018; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). The models operate over tokenized versions of text; in the case of the CPM model, Chinese characters get mapped to numerical tokens (e.g. 中国 *zhōngguó* "China" gets mapped to the token 98).

Procedure

Generating a preliminary dataset A text file was generated by combining all simplified Chinese news from the corpus. For each piece of news, only the main body was kept while metadata such as headline, date, and time were discarded, and the whole piece was stored as one line. We then searched through the text file for words that are either the short or long form of a word pair from the Y. Zhang and Sun list. Whenever a target word is encountered, an entry was created and saved, including 1) the target word, 2) its word form (i.e., short or long), 3) pre-context (i.e., everything preceding the word within the piece of news), 4) post-context (i.e., everything following the word within the piece of news), and 5) line number. More than 30 million entries were created.

Screening word pairs Based on the preliminary dataset, we analyzed the frequency of the short and long forms for all word pairs. Based on these frequencies, a pair was discarded if:

- either of its short or long form appeared no more than 10 times in the corpus, indicating a lack of productivity, or
- the ratio between the frequency of its short and long form in the corpus was smaller than 0.1 or larger than 10. This ensures that neither form is overwhelmingly dominant, which may mean that the short and long forms are not true alternatives.

A pair would also be discarded if:

- Either of its short or long form is shared with other pairs. Otherwise it would be difficult to tell which concept the word is representing without manually going through the data.
- Its short form has other meanings that are not shared with the long form. For example, 乔峰 *qiáofēng* is the short form of 乔戈里峰 *qiáogēlǐfēng*, which is the name of K2, the second highest mountain on Earth (Wikipedia contributors, 2023). Meanwhile, 乔峰 is also the name of a famous fictional character. Same as the previous point, there is no

way to tell which role the word is serving unless the contexts are manually checked, which is labor intensive and may introduce bias.

- Its short form is a substring of its long form. For example, 国足 *guózú* is the short form of 中国足球队 *zhōngguó zúqiú duì* "Chinese soccer team". However, since the short form exactly matches the 2nd and 3rd character of the long form (underlined), whenever a long form is encountered, the short form is inevitably encountered. In contrast, the word pair of 人口所 *rénkǒusuǒ* and 人口研究所 *rénkǒu yánjiūsuǒ* "Institute of Population Research" is acceptable, since the short form matches the 1st, 2nd and 5th character of the long form (underlined), not a continuous substring.

Another case where a word pair was discarded is if either its short of long form has Unicode-related tokenization artefacts.

After the screening, 1418 word pairs were left.

Cleaning the preliminary dataset Since the word pairs are downsized, the preliminary dataset was downsized accordingly. Entries with target words in the discarded word pairs were deleted.

In addition, entries with empty pre-context and/or post-context were deleted, as the CPM language model only calculates surprisals starting from the second token. For an entry with empty pre-context, the target word will appear at the very beginning, so CPM will not be able to calculate its surprisal.

We also wanted to make sure that the token(s) of the target word stay(s) the same when the word is tokenized in context and independently out of context. For example, when tokenized independently, the token for 国企 *guóqǐ* "state-owned enterprise" has token ID 8038. When it appears in the phrase 中国企业 *zhōngguó qǐyè* "Chinese enterprise", the tokenization of the target word following its pre-context changes to

中国	企
98	5390

where the independent token of 国企 (i.e., 8038) does not show up, suggesting that the concept of "state-owned enterprise" is not in the entry. Such entries with inconsistent tokenization were then discarded.

Selecting a sample dataset for calculation As the size of the remaining dataset was still fairly large (≈20GB), a sample was selected for the ease of computation. From the remaining word pairs, 100 were randomly selected with one criterion: their short and long forms should each have no fewer than 100 occurrences in the remaining dataset. Then, for each word, 50 entries were randomly selected. A total of 10,000 entries were selected.

For each entry, the surprisal of the concept will be calculated, meaning adding up the surprisal of the short form and the surprisal of the long form given the same pre-context. With this consideration in mind, the tokenization of the pre-context should stay the same when substituting the target word with its alternative form. If the pre-context tokenization changed, the entry was substituted with a new entry (if available). After applying this criterion, 9992 entries were kept.

Calculating the surprisal of the concepts For each specific entry, the information available includes: 1) the target word w, 2) the word form of the target word (*short* or *long*), and 3) the context c, which was limited to a maximum of 200 characters (punctuation included) preceding the target word. Next, for each datapoint, we need to compute the surprisal of the *concept*—the shared meaning of the long and short form—given the context. The probability of a concept given a context c is given by

$$P(\text{concept} \mid c) = P(w_{\text{short}} \mid c) + P(w_{\text{long}} \mid c) \tag{1}$$

which is summing up the probability of the short form given the context c and the probability of the long form given the same context. The probabilities of words given contexts are generated by the CPM language model. If the word is split into $K > 1$ tokens $w_1, \ldots, w_K$, the probability of the word is then calculated as:

$$P(w \mid c) = \prod_{k=1}^{K} P(w_k \mid c, w_1, \ldots, w_{k-1}).$$

Using these probabilities, we calculated the surprisal of concepts given contexts, $-\log P(\text{concept} \mid c)$.

Next, for each concept, we calculate its average surprisal when appearing as the short form and when appearing as the long form. If a context precedes the short form, it is labeled as c^{short}. If it precedes the long one, it is then labeled as c^{long}. Adapted from the equation proposed by Piantadosi et al. (2011) and used in Mahowald et al. (2013), we can calculate the average surprisal of a given concept in the contexts for short and long forms as:

$$-\frac{1}{N}\sum_{n=1}^{N} \log P(\text{concept} \mid c_n^{\text{short}}) \tag{2}$$

$$-\frac{1}{M}\sum_{m=1}^{M} \log P(\text{concept} \mid c_m^{\text{long}}), \tag{3}$$

given a corpus of N short forms for the given concept, and M long forms for the same concept.

Data analysis Our hypothesis predicts that long forms will be used in less predictive contexts, so the average estimated surprisal of a concept should be higher when it appears as the long form. To see whether the hypothesis is supported by our data, we first calculated the average estimated concept surprisal according to the form (short vs. long) it appears as. We will use *average* concept surprisal as short (Eq. 2) and *average* concept surprisal as long (Eq. 3) to refer to these values.

In Figure 1, we plot these two values for each concept, and connected the two values either with a red line or a blue one. A red line indicates that the *average concept surprisal*

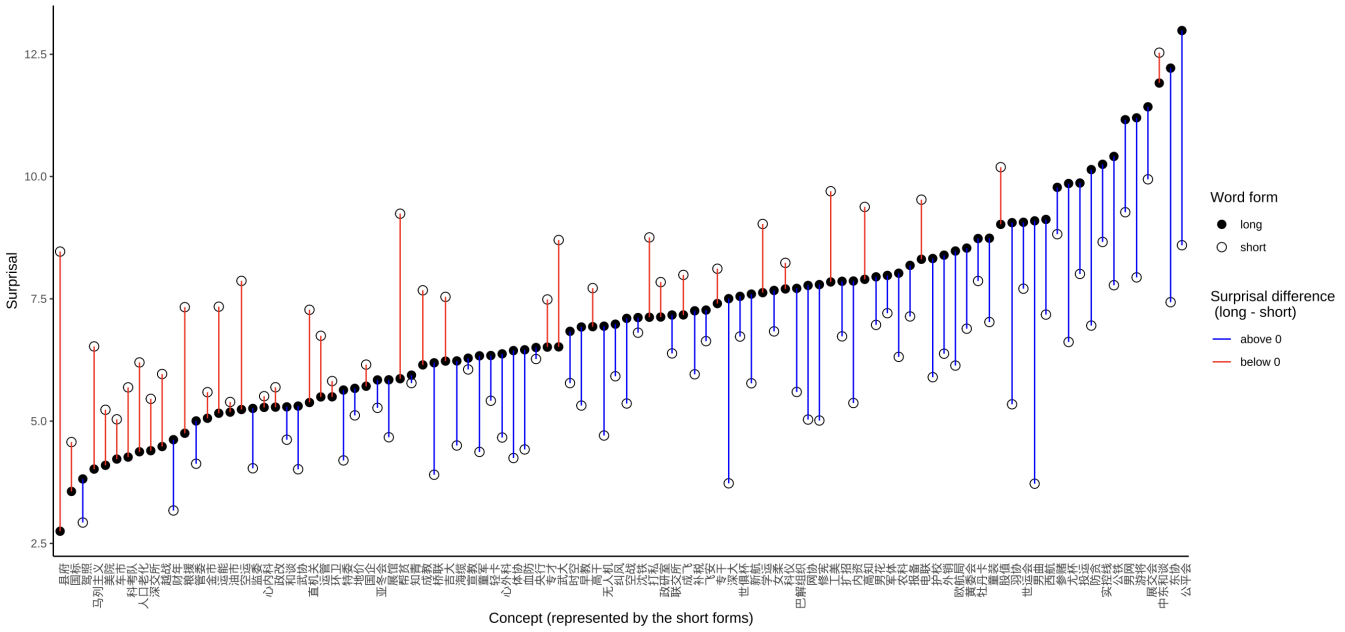

Figure 1: Average estimated surprisal of both long and short form for each of the 100 concepts investigated. For each concept, there is a solid circle representing the long form and a hollow circle representing the short form. Each pair of circles are connected by a solid line showing the difference between the surprisal of the two forms (long − short). If the difference is negative, the line is colored red. Otherwise it is colored blue.

as short is higher than the *average concept surprisal as long*, whereas a blue line indicates the opposite. Among the 100 concepts sampled, 63 have a blue line, indicating that the majority of the long forms tend to have higher surprisal than their short counterpart, which means that the long forms tend to contain more information in general. An unpaired t-test was conducted for each concept to see whether there is a significant difference between the average estimated surprisal of the short and long form at the level of individual concept. Among the 100 pairs, 35 pairs had reliably positive difference scores (at $p < .05$), suggesting that their long forms contain more information on average than their short forms. In the opposite direction, 14 pairs had reliably negative difference scores, suggesting that their short forms contain more information on average.

Table 2 reports the results of a mixed-effects logistic regression model that was fitted to see whether word form (short vs. long) can be predicted by surprisal of the concept, with random intercepts and slopes for concept. The average surprisal for the long forms is 7.09, higher than that of the short form at 6.51 ($\beta = 0.030$, $z = 2.399$, $p < .05$), so the effect of surprisal on determining word length is significant. A paired t-test conducted on the long and short forms also indicated a significant difference in their average estimated surprisal ($t = 3.1109$, $p < .05$).

To answer our research question, we can say with confidence that, for Chinese word pairs, the full (long) form generally contains more information than the abbreviated (short) form.

	Estimate	Std. Error	z value	p value
(Intercept)	−0.217	0.085	−2.546	0.011
Surprisal	0.030	0.013	2.399	0.016

Table 2: Summary of fixed effects from a mixed-effects logistic regression predicting whether a form is long vs. short as a function of its surprisal in context.

Discussion and future work

In this paper, we applied the method in Piantadosi et al. (2011) and Mahowald et al. (2013) to investigate whether the full (long) form of a Chinese word contains more information than its abbreviated (short) form. We hypothesized that the full forms contain more information, and the results of our corpus study provided evidence supporting this hypothesis. In other words, when the context is more predictive, speakers are likely to choose the shorter word to maximize efficiency.

While this trend holds generally of our data, there were some notable exceptions of concepts that behaved in the opposite direction of our prediction. After a close examination of the news materials, we saw that the concept with the most negative difference score (i.e., with a long form that on average contained less information than its short form) is 县人民政府 *xiànrénmín zhèngfǔ* "People's government of the county" and 县府 *xiànfǔ*. Interestingly, these alternating forms appear in very different contexts in our data. The short form usually appears in pieces quoting news from Taiwan,

where the long form does not, which suggests that these two forms are not in fact true alternatives. As discussed earlier, vocabulary might have different meanings in Taiwan, which is exactly the case for 县府, since its corresponding full form should be be 县政府 *xiànzhèngfǔ* "government of the county" instead. Verifying the news content in such detail is difficult for a corpus study, so this word pair was kept in our analysis. Future work will attempt to filter the news content in more detail.

As the current experiment was run on a restricted dataset, the next step will be to run it on the full dataset and see if the results remain consistent. One difference between the current study and the corpus study of Mahowald et al. (2013) is that the word pairs under investigation here are not limited to nouns whereas Mahowald et al. (2013) chose to look at nouns only to avoid the potential effect of syntactic category on surprisal and word length. As shown in Table 1, abbreviation in Chinese can happen to different combinations of syntactic categories. With around 1,400 word pairs, it will be interesting to look at groups of these word pairs according to their syntactic categories and see if the pressure of shortening words in predictive context is different across the groups.

Another difference is that both Mahowald et al. (2013) and Li et al. (2019) trained a trigram model for the calculation of surprisal while in this study a neural language model was used to calculate surprisal with a context length of up to 200 characters. Our corpus contains news articles and, according to the writing convention, abbreviations can only be used when their full form is introduced beforehand. Because of this convention, the longer the context, the more likely the full form is included in the context of the short form, making the short form more predictable. We are interested in the potential impact of context length on the results, namely whether similar results would obtain when context length is shortened so the long form is less likely to appear in the context of the short form. Future work will aim to find out whether context length affects the predictability of the short form, and also whether context size affects our conclusion that long forms contain more information.

One additional point to explore is whether the amount a word shortens (i.e., the difference in number of characters between long and short forms) correlates with the average information change between long and short contexts. If disyllabic is the most popular designated length for abbreviation, one potential hypothesis to be tested is that longer words lose more information content when they are reduced to this ideal length.

A limitation of the current study and the literature that inspired it is that predictability of a word is calculated based on its preceding context. However, the predictability of a word can also be affected by words following it, which may be related to the speaker's planning (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Harmon & Kapatsinski, 2021; Upadhye & Futrell, 2022). To further investigate the possibility of relying on a word's backward predictability to predict its word length, we can calculate the backward surprisal of the short and long forms using a language model trained backwards by reversing the word order in each sentence. We can then compare if using forward predictability, backward predictability, or both performs better at predicting word length. Together, this will allow us to know more about how elasticity of the Chinese language is utilized to achieve communicative efficiency.

References

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.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Dong, Y. (2015). *The prosody and morphology of elastic words in chinese: annotations and analyses* (Unpublished doctoral dissertation). University of Michigan.

Duanmu, S. (1997). Wordhood in chinese. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, *105*, 135–196.

Duanmu, S. (2012). Word-length preferences in chinese: a corpus study. *Journal of East Asian Linguistics*, *21*, 89–114.

Duanmu, S., & Dong, Y. (2016). Elastic words in chinese. In *The routledge encyclopedia of the chinese language* (pp. 490–506). Routledge.

Duanmu, S., Feng, S., Dong, Y., & Zhang, Y. (2018). A judgment study of length patterns in chinese: Prosody, last resort, and other factors. *Journal of Chinese Linguistics*, *46*(1), 42–68.

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.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

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.

Goldsmith, J. (2002). Probabilistic models of grammar: Phonology as information minimization. *Phonological Studies*, *5*, 21–46.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 10–18).

Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347–2353.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics (naacl)* (pp. 1–8). Association for Computational Linguistics.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, *128*(6), 1112.

Kachakeche, Z., Scontras, G., & Futrell, R. (2022). The efficiency of dropping vowels in romanised arabic script. In *Proceedings of the annual meeting of the cognitive science society* (pp. 1505–1511). Online: Cognitive Science Society. Retrieved from https://escholarship.org/uc/item/2m4141hs

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, *83*, 152–178.

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

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, *19*, 849–856.

Li, L., van Deemter, K., Paperno, D., & Fan, J. (2019). Choosing between long and short word forms in mandarin. In *Proceedings of the 12th international conference on natural language generation* (pp. 34–39).

Mahowald, K., Dautriche, I., Braginsky, M., & Gibson, T. (2022). Efficient communication and the organization of the lexicon. In A. Papafragou, J. Trueswell, & L. R. Gleitman (Eds.), *The oxford handbook of the mental lexicon.* Oxford: Oxford University 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.

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2011). *Chinese gigaword fifth edition.* Philadelphia: Linguistic Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/LDC2011T13 doi: https://doi.org/10.35111/102m-dr17

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.

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021, June). How (non-)optimal is the lexicon? In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4426–4438). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.350 doi: 10.18653/v1/2021.naacl-main.350

Rathi, N., Hahn, M., & Futrell, R. (2022). Explaining patterns of fusion in morphological paradigms using the memory-surprisal tradeoff. In *Proceedings of the annual meeting of the cognitive science society* (pp. 184–191). Online: Cognitive Science Society. Retrieved from https://escholarship.org/uc/item/0v03z6xb

Upadhye, S., & Futrell, R. (2022). Information-theoretic analysis of disfluencies in speech. In *Neurips 2022 workshop on information-theoretic principles in cognitive systems.* New Orleans. Retrieved from https://openreview.net/pdf?id=mlUfYl4ss_R

Wikipedia contributors. (2023). *K2 — Wikipedia, the free encyclopedia.* https://en.wikipedia.org/w/index.php?title=K2&oldid=1135546026. (Online; accessed 27-January-2023)

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings for the 42nd annual meeting of the cognitive science society* (pp. 1707–1713).

Zhang, Y., & Sun, X. (2018). A chinese dataset with negative full forms for general abbreviation prediction. In *Proceedings of the eleventh international conference on language resources and evaluation* (pp. 2065–2070). Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://doi.org/10.48550/arXiv.1712.06289

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., ... Sun, M. (2021). Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, *2*, 93–99.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.