Evidence for efficiency in Chinese abbreviations

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

People use language to transfer information, and language scientists have been researching the (in)efficiency of natural languages at achieving this goal (Piantadosi et al., 2011; Gibson et al., 2019; Pimentel et al., 2021; Li et al., 2019). Such theories have been used to predict word length in language use, with the idea that longer words require more effort to produce. Piantadosi et al. (2011) and Mahowald et al. (2013) showed that average information content of a word is more effective than frequency at predicting word length: more informative words are longer. Here we ask whether similar pressures apply in Mandarin Chinese.

In Chinese, each character represents approximately one morpheme. Some long words in Chinese have predictable abbreviations, and different words containing overlapping characters may be abbreviated to the same form, thus creating ambiguity. However, such ambiguity can be resolved given the appropriate context. If the proposed relationship between contextual predictability and shortening holds, then the finding of Mahowald et al. (2013) should generalize to Chinese: the full (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—we can test whether word length in such abbreviation alternations can be predicted by surprisal by measuring the average surprisal of the short and long forms of a word in a corpus and comparing them.

Method The short and long word pairs, corpus, and language model used in this project are summarized in Table 1. Our data analysis uses the average estimated surprisal of the concept as the independent variable and its word length (short vs. long) as the categorical dependent variable.

We searched through the news corpus for either the short or long form of a word pair (i.e., a target word) from our materials. Whenever a target word is encountered, an entry including the target word and its context was saved. Word pairs and their entries were discarded if either form turned out to be unproductive, express more than one meaning, or generate Unicode-related tokenization artefacts. We also discarded entries if their target word was tokenized differently in isolation vs. in context, suggesting that the meaning of the target word did not actually appear in the entry. 1418 word pairs and around 6 million entries were kept. Among the remaining word pairs, 100 were randomly selected for ease of computation. Both forms of these 100 pairs have no fewer than 100 occurrences in the remaining dataset. 50 entries were randomly selected for each word for a total of 10,000 entries.

Each entry has the: 1) target word w, 2) form of the target word (*short* or *long*), and 3) context c, limited to the 200 characters preceding w. For each datapoint, we computed the surprisal of the *concept*—the shared meaning of the short and long form—given the context, $-\log P(\text{concept} \mid c)$. The probability of a concept given a context c is

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

where $P(w_{\text{short}} \mid c)$ and $P(w_{\text{long}} \mid c)$ are generated by the CPM language model. Next, we calculated the average surprisal of concepts when they appear as short forms and as long forms. We will use *concept surprisal as short* and *concept surprisal as long* to refer to these quantities.

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 its long form. To test our hypothesis, we subtracted *concept surprisal as short* from *concept surprisal as long*, and plotted the difference as shown in Figure 1. Each point represents a concept, and 63 of the 100 concepts sampled lie above 0, the predicted direction.

An unpaired t-test was conducted for each concept to compare the average estimated surprisal of

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 383-384.
Amherst, Massachusetts, June 15-17, 2023

Materials	Source	Remarks
Word pairs	Chinese abbreviation dataset (Zhang and Sun, 2018)	Over 7,000 pairs of Chinese words and their corresponding abbreviations
Corpus	Chinese Gigaword Fifth Edition (Parker et al., 2011)	5.3GB of news texts, only simplified Chinese ones are included
Language model	Chinese Pre-trained Language Model (CPM) (Zhang et al., 2021)	Used for the tokenization of texts and the calculation of surprisals

Table 1: Materials for the study.

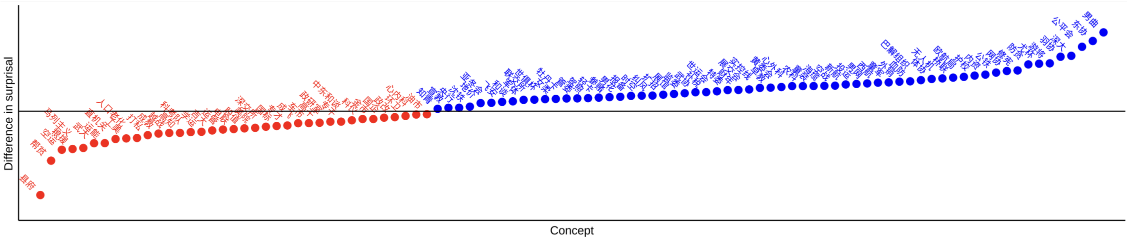

Figure 1: Difference in average estimated surprisal between the long and short form for each of the 100 concepts investigated. Concepts below the y = 0 line are colored red, whereas those above the line are colored blue.

the short and long form at the level of individual concept. 35 of the 100 pairs had reliably positive difference scores (at $p < .05$), suggesting that their long forms contain more information on average. In the opposite direction, 14 pairs had reliably negative difference scores, suggesting that their short forms contain more information on average.

To analyze the dataset altogether, a mixed-effects logistic regression model 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, significantly higher than that of the short form at 6.51 ($\beta = 0.030$, $z = 2.399$, $p < 0.05$). A paired t-test conducted on the long and short forms also indicated a significant difference in their average estimated surprisal in the predicted direction ($t = 3.1109$, $p < 0.05$).

Discussion The results of our corpus study provided evidence supporting our hypothesis: the full form of a Chinese word contains more information than its abbreviated form. In other words, when the context is more predictive, speakers are likely to choose the shorter word to maximize efficiency.

As the current experiment was run on a restricted dataset, the next step will be to run it on the full dataset—more alternating pairs in more contexts—to see if the results differ. We will also explore the possibility of relying on a word's backward predictability to predict its word length (the current analysis looks at forward predictability), and to explore whether the amount a word shortens correlates with the average information change between long and short contexts.

References

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

L. Li, K. van Deemter, D. Paperno, and J. Fan. 2019. Choosing between long and short word forms in mandarin. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 34–39.

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

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Chinese gigaword fifth edition.

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

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021. How (non-)optimal is the lexicon? In *Proceedings of NAACL-HLT*, pages 4426–4438, Online.

Y. Zhang and X. Sun. 2018. A chinese dataset with negative full forms for general abbreviation prediction. In *Proceedings of LREC*, pages 2065–2070, Miyazaki, Japan.

Z. Zhang, X. Han, H. Zhou, P. Ke, Y. Gu, D. Ye, Y. Qin, Y. Su, H. Ji, J. Guan, F. Qi, X. Wang, Y. Zheng, G. Zeng abd H. Cao, S. Chen, D. Li, Z. Sun, Z. Liu, M. Huang, W. Han, J. Tang, J. Li ad X. Zhu, and M. Sun. 2021. Cpm: A large-scale generative chinese pre-trained language model. *AI Open*, 2:93–99.