Choosing between Long and Short Word Forms in Chinese

Lin Li[♠△★], Kees van Deemter[♠♡], Denis Paperno[◇], Jinyu Fan[♣], Zewangkuanzhuo[△]
[♠]Department of Information and Computing Sciences, Utrecht University
[△]Department of Computer Science, Qinghai Normal University
[★]CMLI Research Center, Minzu University of China
[♡]Department of Computing Science, University of Aberdeen
[◇]Department of Languages, Literature and Communication, Utrecht University
[♣]Department of Physics and Information , Qinghai Normal University
{l.li1, c.j.vandeemter, d.paperno}@uu.nl
{2104439819, 2497138704}@qq.com

Abstract

Between 80% and 90% of all Chinese words have long and short form such as 老虎/虎 *(lao-hu/hu , tiger)* (Duanmu, 2013). Consequently, the choice between long and short forms is a key problem for lexical choice across NLP and NLG in Chinese.

Following on from earlier work on abbreviations in English (Mahowald et al., 2013), we bring a probabilistic perspective to word length choice, using both a behavioural and a corpus-based approach. Thus, we hypothesise that, in Chinese, short forms are likelier in supportive than in neutral contexts. Our corpus and behavioral study supported this hypothesis, but a closer analysis revealed striking differences between different types of Chinese words.

1 Introduction

Choosing words is an important task for both Natural Language Generation (Gatt and Krahmer, 2018; Reiter et al., 2005; Stede, 1994; Polguère, 2000; Wanner, 1996) and systems that perform summarisation and machine translation. Many Chinese words can be expressed by either a short form or a long form. These words are known as *elastic* words (Guo, 1938; Duanmu, 2013; Qin and Duanmu, 2017). For instance, one Chinese sentence in Table 1[1] contains three elastic words, 学习/学*(xue-xi/xue, to study)*, 养殖/养*(yang-zhi/yang, to cultivate)*, and 蜜蜂/蜂*(mi-feng/feng, bee)*. Such long and short form pairs are interchangeable in some contexts with little difference in meaning. Choosing between a long and a short form of the same word is an important problem for any NLG system that produces Chinese text.

Various theories have sought to address the issue of elastic words. Speech rate theory (Guo, 1938) proposed that speakers control their speech rate by alternating between long and short forms. And processing need theory (Pan, 1997) suggests (without offering computational detail) that the choice between long and short forms can help to decrease information density, thus aiding listeners' understanding.

他	想	学习	养殖	蜜蜂。
ta	*xiang*	*xue-xi*	*yang-zhi*	*mi-feng*
He	*want*	*learn*	*cultivate*	*bee.*
他	想	学习	养	蜜蜂。
他	想	学习	养	蜂。
他	想	学	养	蜂。
他	想	学	养	蜜蜂。
他	想	学	养殖	蜜蜂。
*他	想	学	养殖	蜂。
*他	想	学习	养殖	蜂。

He wants to lean how to cultivate bee.

Table 1: Example of an ordinary sentence with elastic words.

In this paper, we adopt a similar perspective, informed by work in information theory (Shannon, 1948). The information content of a word depends on its context, that is, a word conveys more information content in unpredictable contexts than in predictive ones. More specifically, we follow the study design of Mahowald et al. (2013), who test the hypothesis that short forms should convey less information than their longer counterparts, given that language is designed to be information-theoretically optimal.

2 Related work

Our work in Chinese builds on previous studies on other languages (Mahowald et al., 2013; Piantadosi et al., 2011; Willems et al., 2015; Seyfarth, 2014; Jaeger and Buz, 2017; Lewis and Frank, 2016). Piantadosi et al. (2011) investigated

[1] This sentence contains 3 elastic words, which has 8 varied-length sentences. * indicates degraded sentences.

Proceedings of The 12th International Conference on Natural Language Generation, pages 34–39,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. ©2019 Association for Computational Linguistics

the correlation between word length and information content across 11 languages (not including Chinese) and suggested that information content (specifically, surprisal) is a good predictor of word length.

Mahowald et al. (2013) hypothesised that *when an English word w has a long and a short form, then the choice between the two is affected by the extent to which the occurrence of w is "predictable": in contexts where w has high probability, the short form is more preferred than in contexts where w has low probability*. They used two approaches to testing this idea: a corpus study and a behavioural one. Both these studies made use of word pairs w_1, w_2 , where w_1 and w_2 have nearly identical meanings but different length; for instance, *math/mathematics*, *chimp/chimpanzee*, *dorm/dormitory*, etc.

Corpus Study *Surprisal* (Shannon, 1948) quantifies the information content conveyed by a word in a given context. Surprisal values of 22 word pairs were obtained from Google N-gram corpus of English (Mahowald et al., 2013). The mean surprisal for long forms was significantly higher than short forms, confirming the authors' expectations. For all 22 word pairs, the long forms of 18 word pairs showed higher surprisal than short forms.

Behavioural Study Mahowald et al. (2013)'s behavioural study asked participants to complete a sentence whose final word w was missing. Participants were offered a choice between the long and short form of w. Half the context sentences described a situation that made w highly probable (*supportive contexts*), the other half made w not very probable (*neutral contexts*).

A validation experiment ensured that suitable contexts had been chosen as supportive and neural contexts: Participants were asked to complete a sentence by filling in one word. The validation experiment showed that in supportive contexts, the target word (i.e., either its long or its short form) was chosen 33 times as often as in neutral contexts, suggesting that suitable contexts had been chosen.

For their main behavioural study, the authors hypothesised that in supportive contexts, a larger proportion of short forms is chosen than in neutral contexts. This hypothesis was confirmed, with short forms being 11% more frequent in supportive than in neutral contexts.

We decided to replicate the two studies by (Mahowald et al., 2013) as precisely as possible, but focusing on Chinese instead, to see how the choice between long and short word forms might differ across the two languages.

3 Experiments and results

We planed to find out whether the findings of Mahowald et al. (2013) in English can be replicated for Chinese. Our main goal was to test whether predictiveness of context affects the form choice for elastic words. We focus on *nominal* elastic words because these are particularly plentiful, and particularly varied in terms of their form (Duanmu, 2013). As our elastic words, we used the list of elastic words in the appendix of Dong (2015), focusing on those cases in which (1) the part-of-speech is Noun, and (2) the short form is a free morpheme (i.e., a morpheme that can occur as a complete word).

3.1 Corpus study

In the corpus study, the predictiveness of a context was assessed as a continuous measure using surprisal. The surprisal of w is high if it's very unpredictable in a given context. For instance, consider
我的包坏了，我要去买新包。
wo-de-bao-huai-le, wo-yao-qu-mai-xin-bao.
My bag is torn out, I want to buy a new bag.
The last word 包 *(bao, bag)* is highly predictable, so its surprisal would be very small in this context.

We made use of the simplified Chinese corpora of Google Ngram corpus[2]. And we focused on the 442 word pairs from Dong (2015) that meet the criteria above. First, we trained a trigram language model on Google Ngram corpus. Then for each word occurrence, we calculated the surprisal using a trigram language model to estimate word occurrence probability. Surprisal of the occurrence of w is defined as

$$-\frac{1}{N}\sum_{i=1}^{n}\log P(W = w|C = c_i),$$

where N is the total frequency of word w and c_i indicates the ith occurrence context of word w in the corpus. We tested whether short forms tend to have lower mean surprisal than the corresponding long forms using Student's t-test.

[2] https://books.google.com/ngrams

3.2 Behavioural study

In the behavioral study, context predictiveness was a binary variable. For each word, two contexts, supportive and neutral, were presented to experimental participants. In our work, a context is *supportive* if it makes the target word has a very high probability of occurrence; a context is *neutral* if does not.

Following Mahowald et al. (2013), we recruited two disjoint groups of native speakers of Chinese. Each group contained 52 undergraduates at Qinghai Normal University. One group of speakers was given the main behavioural study and the second group was presented with a validation experiment (described below).

Main experiment As target (elastic) words, we chose 42 word pairs from Dong (2015)'s list. To avoid complications that could be caused by abstract nouns, we ensured that all target words describe physical objects.[3] We constructed supportive and neutral contexts for each of these words. Like Mahowald et al. (2013) before us, we ensured that the length of the supportive context and neutral context was always approximately the same. A comprehension question followed each item. For instance, a (supportive) context sentence might be
大蒜素能抗癌，所以我妈做菜时会放很多....
da-suan-neng-kang-ai,suo-yi-wo-ma-zuo-cai-shi-hui-fang-hen-duo...
Garlic allicin is helpful for anti-cancer, so my mom cooks dishes with ...
where participants could fill in the dots by choosing between 大蒜*(da-suan, garlic)* and 蒜*(suan, garlic)*.
This was followed by a question saying
大蒜素是不是能抗癌?
da-suan-su-shi-bu-shi-neng-kang-ai?
Is Garlic allicin helpful for anti-cancer?

The follow up questions were included to make sure the participants had read and understood the sentences.

Validation experiment In our work, we performed a validation experiment analogous to that of Mahowald et al. (2013). Its goal was to ensure that the target word is more likely in the context we had constructed as supportive. Indeed, in supportive contexts, the target word was chosen 38 times as often as in neutral contexts; this figure is quite similar to Mahowald et al. (2013) and shows that our choice of supportive and neutral sentence contexts was appropriate.

3.3 Results

The results of the two studies confirm our initial hypothesis. In the corpus study, we found that the mean surprisal for long forms (11.78) was higher than that for short forms (10.87). In 324 (73.30%) of the 442 cases, long forms showed significant higher mean surprisal than its short counterpart. The difference between mean surprisal of short and long forms is statistically significant (t=11.66, p<0.01 by paired-sample *t*-test), confirming our hypothesis.

In the behavioural study, we found that the short form was more often chosen in supportive context (51.73%) than in neutral context (48.27%); the difference is significant under a paired-samples *t*-test ($t = 3.04$, $p < 0.05$).

Our two studies about Chinese suggest that the probability of occurrence of a word (more precisely: of a word pair) in a given context, as defined in two very different types of study, affects the choice between the long and the short version of the word, in accordance with our hypotheses. On the other hand, this effect was not as strong as it had been in the work of Mahowald et al. (2013).

4 Distinguishing between types of words

We analysed our data by distinguishing between different types of elastic words. We classified elastic words into seven categories, according to the relation between two morphemes in the long form. The following notation for different types of morphemes is used in characterising types of elastic words. Before explaining what these categories are, we define a few linguistic terms.

Affixes are bound and functional morphemes (Liao, 2014), denoted by 0 in what follows; they are members of a closed set (Wang et al., 2015), for instance, 老 *in* 老虎*(lao-hu, tiger)*, 头 *in* 骨头*(gu-tou, bone)*, etc. **Pseudo-affixes** are very similar to affixes, but they are not a member of the closed set; they are free morphemes that can be used separately (i.e., as a separate word) but also as part of a multi-syllabic word, in which case they lose their original meaning. Pseudo-affixes will be denoted by $0'$. For example, 蜂蜜*(feng-mi, honey)* starts with the pseudo-affix 蜂*(feng, bee)*, followed by the short form 蜜*(mi, honey)* (Note

[3] All 7 categories of elastic words discussed in section 4 were represented by 6 long-short word pairs.

that 蜂*(bee)* can also be used separately, which means ’bee’. A **free morpheme** can occur as an independent word and can also combine with other morphemes to form a new word. Free morphemes will be denoted by the symbols X, X', and Y such as 树*(shu, tree) and* 井*(jing, well)*. The word 谎话*(huang-hua, lie)*, by contrast, starts with the word 谎*(lie)*, followed by the pseudo-affix 话*(hua, talk)* (Note that 话*(talk)* can be used separately.)

Using the terminology above, Duanmu (2015) distinguished the following categories:[4] (1) X-$0X$, (2) X-$0'X$, (3) X-$X0$, (4) X-$X0'$, (5) X-XX, (6) X-$X'X/XX'$. We add another category, namely (7) X-YX/XY, where the long form adds a free morpheme Y that has a different meaning from X. Most of the long and short form pairs of category (7) have hyponymy relations such as 手表/表 *(shou-biao/biao, watch) and* 台灯/灯 (tai-deng/deng, lamp). Moreover, they have identical meaning in some contexts, thus are interchangeable in these contexts. We take this category into consideration because X-YX/XY is in generally consistent with the definition of elastic words and is very common in Chinese.

Post-hoc analysis of our corpus study Fig.1 shows average surprisal of long and short word forms investigated in our corpus study grouped by category. One category X-XX showed a reverse trend with our hypothesis. If the category X-XX is ignored, then in the words of remaining 6 categories together, the surprisal for long forms (11.83) is much higher than that for short forms (10.51). In 322 (74.02%) of the 435 cases, long forms showed higher mean surprisal than its short counterpart (t=11.78, p<0.01).

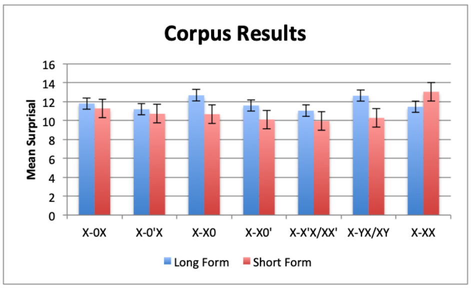

Figure 1: The y-axis is the mean surprisal of the long and short form. The x-axis shows the 7 categories of long and short forms. The blue bar indicates the long form, and the red bar is the short form.

Post-hoc analysis of our behavioural study Fig.2 shows that the frequency of short forms in supportive and neutral contexts. Using the classification above, 5 of 7 categories indicated results that are consistent with our prediction. For the categories X-$0X$ and X-XX we found the reverse, suggesting that the short form is *more* probable in a neutral context than in a supportive context.

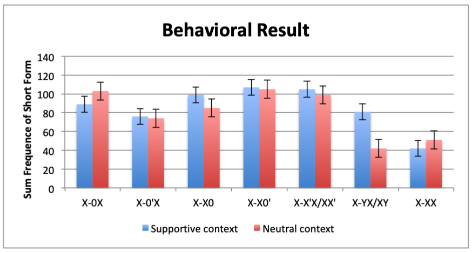

Figure 2: The y-axis indicates the frequency of the short form being chosen by speakers in supportive vs. neutral contexts. The x-axis shows the category of elastic word.

This post-hoc analysis suggests that a revised definition of elastic words in Chinese might be considered. In particular, it might be better not to adopt the same strategy to choose the long and short form of category X-XX. Not only because the words in this category show contrary results with all other categories in both studies, but they arguably play a different role in the language: whereas the long forms of elastic words in the other categories may be seen as adding more information in an extra morpheme (thus aiding the hearer’s understanding), words in the X-XX category merely reduplicate the same morpheme.

5 Discussion

Lexical choice – in automatic summarisation and machine translation as well as NLG – is not just mapping concept into words, because the choice of words depends on linguistic context. We investigated the correlation between the degree to which a concept (or word pair) is predictable, on the one hand, and the length of the word on the other. We hypothesized that speakers use short forms more often in supportive contexts than in neutral contexts. The results of our two studies were consistent with this hypothesis. And the result confirmed the idea that the information-theoretic factors (i.e., factors related to the probability of expressing a given concept (or word pair)) that influence the choice between long and short forms of elastic words in Chinese.

[4] Dashes denote separation between long and short forms; slashes denote alternatives.

We investigated the occurrence probability of long and short form in supportive and neutral contexts by means of Deep Neural Network language model[5]. The result shows that the probability of short form in supportive contexts (2.96%) is much higher than in neutral contexts(0.11%), which is consistent with our prediction (t=3.15, $p < 0.05$) again.

Our findings are broadly consistent with the view that the choice between long and short form of the same word in Chinese works the same way as the analogous choice in English. However, our post-hoc analysis suggests that this simple picture is complicated by differences between types of elastic words. In particular, the reduplicative X-XX category showed a trend that is opposite to what is seen in the other categories, with the longer version of the word *more* (not less) frequent than the shorter version in supportive contexts. Most reduplicative nouns are appellations that are used among family relatives, such as 妈妈/妈 (ma-ma/ma, mother) and 叔叔/叔(shu-shu/shu, uncle). This suggests that surprisal is not the only factor that influences the choice on which our study has focused.

Our research has obvious applications for lexicalization in Chinese NLG: Following findings in this work, a NLG system might implement the choice of word length based on its surprisal. However, this cannot be a universal strategy applied to the whole vocabulary without exception. At least the words of the X-XX category should have a different generation strategy. Furthermore, while we establish surprisal as a significant predictor of word form choice in Chinese, it might also be influenced by other factors, including, for instance, ambiguity avoidance (Khan et al., 2012), prosody (Duanmu et al., 2018; Duanmu, 2013), and style (Feng, 2016). We leave the investigation of these factors and their interaction to further study.

Acknowledgements

The first author of this paper received support from Qinghai Natural Science Foundation under Grant 2016-ZJ-931Q, Qinghai Major R&D Transformation Foundation under Grant 2019-GX-162, and National Natural Foundation under Grant 61862055, which is gratefully acknowledged.

[5] https://developer.baidu.com/platform/s91

References

Yan Dong. 2015. *The prosody and morphology of elastic words in Chinese: annotations and analyses*. Ph.D. thesis, University of Michigan.

San Duanmu. 2013. How many chinese words have elastic length. *Eastward flows the Great river: Festschrift in honor of Prof. William S.-Y. Wang on his 80th birthday*, pages 1–14.

San Duanmu. 2015. A study of elastic word length of monomorphemic nouns in chinese:'monosyllabic-only'nouns in the lexicon and in actual use.

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

Shengli Feng. 2016. Modern chinese: Written chinese. *The Routledge Encyclopedia of the Chinese Language*, pages 645–663.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Shaoyu Guo. 1938. 中国语词之弹性作用(the function of elastic word length in chinese). *Yen Ching Hsueh Pao*, 24:1–34.

T Florian Jaeger and Esteban Buz. 2017. Signal reduction and linguistic encoding. *Handbook of psycholinguistics*, pages 38–81.

Imtiaz H Khan, Kees van Deemter, and Graeme Ritchie. 2012. Managing ambiguity in reference generation: the role of surface structure. *Topics in Cognitive science*, 4(2):211–231.

Molly L Lewis and Michael C Frank. 2016. The length of words reflects their conceptual complexity. *Cognition*, 153:182–195.

Wei-Wen Roger Liao. 2014. *Morphology*, chapter 1. John Wiley Sons, Ltd.

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

Guowe Pan. 1997. 汉英语对比纲要(an outline comparison of chinese and english).

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

Alain Polguère. 2000. A "natural" lexicalization model for language generation. In *Proceedings of the Fourth Symposium on Natural Language Processing (SNLP2000). Chiangmai, Thailand*, pages 37–50. Citeseer.

Zuxuan Qin and San Duanmu. 2017. A judgment study of word-length preferences in chinese nn compounds. *Lingua*, 198:1–21.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Manfred Stede. 1994. Lexicalization in natural language generation: A survey. *Artificial Intelligence Review*, 8(4):309–336.

William S-Y. Wang, Chaofen Sun, and Jerome L. Packard. 2015. Morphology morphemes in chinese.

Leo Wanner. 1996. Lexical choice in text generation and machine translation. *Machine Translation*, 11(1-3):3–35.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. 2015. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.