Frequency-Dependent Regularization in Mandarin Elastic Word Length

Skyler Jove Reese (mwreese@ucdavis.edu)[1]
Zoey Liu (liu.ying@ufl.edu)[2]
Masoud Jasbi (jasbi@ucdavis.edu)[1]
Emily Morgan (eimorgan@ucdavis.edu)[1]

[1] Department of Linguistics, University of California – Davis, Davis, California, United States
[2] Department of Linguistics, University of Florida, Gainesville, Florida, United States

Abstract

In English binomial expressions, "bread and butter" is preferred over "butter and bread". Morgan & Levy (2015) show that for these types of expressions, frequently used phrases tend to have stronger, more extreme preferences. In contrast, there is roughly an equal preference for "bishops and rooks" versus "rooks and bishops", a much less common pairing. This paper extends this research to the concept of Mandarin elastic word length, a phenomenon in which most Mandarin words have long and short forms. We find evidence for frequency-dependent regularization in the elastic length of Noun-Noun compounds in Chinese, demonstrating that frequency-dependent regularization extends to structures with more than two alternations and to languages other than English.

Keywords: frequency-dependent regularization, Mandarin Chinese, elastic word length, hierarchical Bayesian modeling, corpus analysis

Introduction

An important feature of human language is the presence of alternations—having multiple ways to communicate the same concept. At the phrasal level there are binomials like "butter and bread" vs "bread and butter". Syntactically, we have argument structure alternations like "I gave her a card" vs "I gave a card to her". With lexical items there are examples like "fridge" vs "refrigerator". This is a concept which applies to all language varieties in some way or another, but languages vary in what kinds of alternations they allow and on what levels of structure.

This paper will ask whether preference strength is correlated with frequency. For example, Morgan & Levy (2015) found that words that occur together in binomials more often (e.g. "bread", "butter") tend to have stronger preferences for one ordering (e.g. "bread and butter" vs "butter and bread")—in other words, their preferences are more extreme. But words which occur together in binomials less often tend to have more balanced preferences (e.g. "bishops and rooks" or "rooks and bishops"). This phenomenon is known as frequency-dependent regularization. It is an interesting discovery because it could just as well have been that case that learners develop balanced preferences at high frequencies, but that is not what we see. Note that "regularization" here is being used to refer to preference strength for a particular alternation. This should not be conflated with "regularization" in the sense of the replacement of irregular forms for regular ones which we find in language acquisition and language change.

Evidence for frequency-dependent regularization has been found in the case of English binomials ("bread and butter" vs "butter and bread") (Morgan & Levy, 2015) and for the English dative alternation ("gave her a book" vs "gave a book to her") (Liu & Morgan, 2020) but it has not been investigated in any language other than English. This leads us to wonder whether this is something common to all languages or simply unique to English. Additionally, this phenomenon has only been demonstrated with linguistic constructions that only allow two structural alternatives. Our primary research question therefore is whether we can find frequency-dependent regularization in the elastic length of Mandarin Chinese noun-noun compounds. Inquiry of this nature could extend evidence for frequency-dependent regularization to a language other than English and to structures with more than two alternations.

To address our research question, we use elastic length in Mandarin Chinese as the test case. Most Mandarin lexical items have two relatively synonymous alternations—a monosyllabic form and a disyllabic one. One example of elastic length is the word for 'coal', 煤炭(mei2 tan4), which can also be expressed simply as 煤(mei2). These two forms are interchangeable in a variety of contexts with minimal or negligible differences in meaning. For example, a store where coal is sold can be referred to either as a 煤炭店(mei2 tan4 dian4) or as a 煤店(mei2 dian4). However, interchangeability in all contexts is not a requirement for the 'elastic length' determination. For example, a coal mine is not called a 煤炭矿(mei2 tan4 kuang4) and can only be called a 煤矿(mei2 kuang4) (Duanmu, 2012).

Elasticity of NN Compounds

Mandarin two-word compounds can appear in four alternations with a long or short first word followed by a long or short second word. These length patterns are referred to with 1s (short) and 2s (long) as "1+1", "1+2", "2+1" and "2+2". An example of an item which appears in all four forms can be seen below in Table 1. Because of the presence of these alternations, we have the opportunity to look for frequency-dependent preference extremity. In simpler terms, we can investigate whether nouns which get used more often have stronger preference strength for one form over the another. We can consider two types of frequencies for these

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.*

compounds. The first is the *overall frequency* of the compound in any of its four alternations. This metric tells us how often the compound is used generally. *Relative frequency*, on the other hand, refers to the individual frequency of one alternation type. This metric tells us about the regularization of a given compound. In this paper, we ask whether overall frequency predicts regularization.

Table 1: Four Length Patterns for a NN Compound

Length Pattern	Alternation	Gloss
1+1	简字	"simple" + "character"
1+2	简汉字	"simple" + "Chinese character"
2+1	简体字	"simple form" + "character"
1+2	简体汉字	"simple form" + "Chinese character"

Researchers have looked both at Noun-Noun (NN) compounds and Verb-Object (VO) compounds, which exhibit different phonological well-formedness constraints due to the headedness of each compound (Duanmu, 2012; Qin & Duanmu 2017; 2019; 2021). This study will focus on NN compounds. With the second noun being the head of a NN compound and Mandarin nouns receiving stress on the initial syllable, 1+2 length patterns are phonologically ill-formed (Duanmu, 2007). This theory is supported by a corpus study using the Lancaster Corpus of Mandarin Chinese (Duanmu, 2012).

On the other hand, acceptability judgements for similar items show a dispreference for 1+1 compounds which is just as strong as 1+2 and which cannot be accounted for by the phonological well-formedness theories (Qin & Duanmu, 2017; 2019). While 1+1 compounds might be phonologically well-formed, it appears that speakers and listeners have a strong preference against them. Intuitively, this might be due to the fact that short compounds are highly homophonous and present a great deal of ambiguity.

Qin & Duanmu's acceptability judgment studies also discovered significant effects for generative constraints beyond phonological well-formedness. Among the constraints which are supported by the strongest evidence are boundedness (whether the short form is a bound or free morpheme) and homography (whether the noun's sense is the most prominent sense among words that the short form can represent). Short forms are claimed to be dispreferred when they are bound morphemes or lower frequency homographs. Above and beyond the influence of phonological, morphological, and semantic constraints, we can ask whether higher overall frequency of a compound predicts stronger preferences for a given length pattern.

In this work, we look specifically at NN compounds in Mandarin that alternate between all four forms and examine whether there is frequency-dependent regularization among them. Evidence of frequency-dependent regularization in a language other than English and in a phenomenon which is uniquely Chinese could expand the scope of this theoretical paradigm. It is also beneficial to demonstrate that this concept applies to a situation in which speakers have more than two options for alternations; here they have four.

Present Study

In conversation with previous work, this paper seeks to answer the following question: do Mandarin NN compounds show frequency-dependent regularization?

To answer this question, we perform two analyses. The first analysis implements a linear regression between the overall frequency of a compound and the preference strength (relative frequency) of length patterns, computed as the information entropy of the length pattern distribution. The second analysis fits a hierarchical Bayesian model which additionally accounts for phonological, morphological, and semantic constraints which have been claimed to affect elastic length of compounds. This hierarchical model also includes a component which models preference strength, this time as a latent model parameter and allows us to model the strength of this effect above and beyond the effects of generative constraints.

Methodology

Both of our analyses rely on a corpus analysis of NN compounds. We extracted NN compound types from the Lancaster Corpus of Mandarin Chinese (McEnery & Xiao, 2004), and then extracted token frequencies from Simplified Chinese Google Books (Michel et al., 2011).

We first extracted all NN compounds from the Lancaster Corpus of Mandarin Chinese. We operationalized NN compounds as adjacent items marked as either "n" or "ng" (common nouns), a stricter formulation than Duanmu (2012) which also includes "f", "fg", "s" (location, direction, and space words). These were removed from our analysis because they largely represent single syllable words without disyllabic alternations. Our initial extraction resulted in a list of 39,324 compounds.

Human Annotation

Using the list of NN compounds, any compound occurring four or more times in the corpus was subjected to human annotation to ensure the quality of the parsing. Proper nouns, multisyllabic words, and adjacent nouns which did not form a compound in context were removed, leaving us with 1,501 usable items.

From this list, the items were subjected to a second round of annotation, this time providing disyllabic forms for monosyllabic words and monosyllabic forms for disyllabic words. Two new annotators were used for this portion, each of them providing the missing length alternations or noting the lack thereof. Strikingly, one of the annotators was extraordinarily more conservative than the other. Out of our 1,501 potential items, Annotator 1 claimed that 1,309 of them contained a noun without elastic length. Contrarily, Annotator 2 only marked 17 items as not having elastic length. Of those 17 items, only 10 of them were in agreement with Annotator 1.

The extremely low inter-annotator agreement is not without precedent, given that in every experimental study on elastic word length, participants have demonstrated a very high degree of variance in their acceptability judgements.

In an attempt to reconcile these differences, we provided each annotator with the other's annotations which contained all four length pattern alternations. The annotators were then asked to rate each item with binary acceptability. All items which were assessed as acceptable by both annotators were included in a new meta-annotated dataset with 977 items.

Ngram Extraction

We obtained token frequencies (n-gram counts) from the Simplified Chinese Google Books Corpus (Michel et al., 2011). We used the R package *ngramr* (Carmody, 2023) to obtain the ngram counts for each alternation by item across the time period from 1900 to 2019. This period was chosen because prior to the twentieth century, modern Mandarin Chinese as it is known today was not the most prevalent written form. Classical Chinese dominated as the written form of all Chinese languages prior to the twentieth century.

The data set used in the analyses to follow includes items either attested to or rated as acceptable by both annotators. Bigrams were collected for alternations for all items. We define the **overall frequency** of an item as the summed frequency of its four alternations. This provides a measure of frequency of the item that is not confounded with the preference for a particular alternation.

Finally, items which did not have at least one attestation of each of the four length patterns were removed in order to focus our inquiry on fully elastic compounds. The complete dataset used for analysis contained 683 items.

Generative Constraint Annotation

The data was then further annotated for two generative constraints, namely boundedness and homography. Boundedness is the property of a given noun such that its short (monosyllabic) form is a bound morpheme. Boundedness was determined from the presence or absence of an isolated short form appearing in the Lancaster Corpus of Mandarin Chinese. At least one appearance of the short form as an individual token classified the noun as unbound, a score of zero. An absence of any appearances classified the noun as bound, a score of one. Each of the two nouns in a given compound were given a binary boundedness score. Homography is a pervasive phenomenon in Mandarin Chinese whereby a single grapheme is used to represent a variety of senses or discrete lexical items which can be disambiguated using their long (disyllabic) forms. Typically, one sense of a grapheme is the primary or most prominent sense of the character. We can think of this for example with the English word 'run' which has a variety of senses—to run a race, to run a program, to run a school, to run the numbers, etc. The primary sense of the word 'run' is the sense of 'to run a race'.

In order to determine which sense of a Chinese character is the primary sense, we consulted the Contemporary Chinese Dictionary 7th Edition (Chinese Academy of Social Sciences, 2016), which is the foremost authority on standard Mandarin lexicography. The primary sense of a character was operationalized as the sense which is listed first in the Contemporary Chinese Dictionary under the entry for the short form. Annotations for homography were also done on a binary basis. Nouns were given a homography score of zero if they held the primary sense of the short form used to represent them. Nouns whose sense was not in line with the primary sense of the character used as its short form were given a homography score of one. Again, each constituent noun in a NN compound was given its own score for both boundedness and homography.

Analysis

Linear Regression

Our linear regression model uses the log of the overall frequency of an item to predict the information entropy contained within its length pattern distribution.

Information entropy (commonly abbreviated H) is a common measure used to determine how uniform or skewed a distribution is. High entropy is a feature of more uniform distributions, while low entropy is characteristic of strong preference. Thus, we predicted that an increase in log frequency will correspond to a decrease in entropy. In simple terms, we expect these metrics to be negatively correlated across NN item types.

A linear regression model of the form "H ~ 1 + LogFreq" was fit using the *brms* package in R (Bürkner, 2021). We found a significant effect of log overall frequency ($\bar{\beta} = -.12$; $\text{CrI} = [-.11, -.14]$), which initial evidence for frequency-dependent regularization. This sets the stage for our subsequent model which further tests for frequency-dependent regularization while controlling for generative constraints on elastic length.

Hierarchical Bayesian Model

Frequency-dependent regularization is a phenomenon which we find in alternations which already have some degree of preference due to a variety of factors. Factors which have been demonstrated to be relevant for the elasticity of Mandarin NN compounds are phonological well-formedness (Duanmu, 2007), boundedness, and homography (Qin & Duanmu, 2017; 2019). In order to determine the relative explanatory power of these factors in determining length pattern alternation distributions, alongside the contribution of frequency-dependent regularization, we design a hierarchical Bayesian model. The flexibility of our hierarchical model allows us to take into account the modulating effect of overall frequency alongside a straightforward regression with effects for generative constraints.

The core of the model is a Dirichlet-Multinomial distribution which is a multicategory extension of the Beta-Binomial distribution as implemented in Morgan & Levy

(2015) to demonstrate frequency-dependent regularization in binomial ordering preferences. A multinomial distribution (Equation 1) puts priors over a random variable (D) which can have one of four outcomes (1+1, 1+2, 2+1 and 2+2). It has three parameters, written as the vector $\hat{\pi}$, each of which is the likelihood of a given outcome, with the fourth outcome being completely determined by the other three: $\pi_{2+2} = 1 - (\pi_{1+1} + \pi_{1+2} + \pi_{2+1})$.

$$\mathrm{D} \sim \mathrm{Multinomial}(\hat{\pi}) \tag{1}$$

The Dirichlet distribution then serves as the conjugate prior of the Multinomial distribution with three expected mean parameters ($\hat{\mu}$), one for each of the outcomes with the fourth outcome's mean being determined by the other three $\mu_{2+2} = 1 - (\mu_{1+1} + \mu_{1+2} + \mu_{2+1})$ and one concentration parameter (ν) which modulates how tightly clustered or dispersed the multinomial distribution samples are. Ultimately, a lower ν value causes a sparser multinomial distribution which results in stronger preferences. Higher ν values then result in weaker preferences, so we anticipate overall frequency to correlate negatively with this ν parameter.

$$\hat{\pi} \sim Dirichlet(\hat{\mu}, \nu) \tag{2}$$

There are two components of the model which feed directly into the parameters of the Dirichlet distribution. The first one is a regression (Equation 3) which predicts the concentration parameter as a function of overall frequency of the item (F_i). With the Dirichlet distribution serving as the prior for our multinomial distribution, its concentration parameter influences the skewedness of the resulting multinomial distribution. If the slope of this regression (m) is significantly negative, it will indicate frequency-dependent regularization.

$$\nu = e^{m*F_i+b} \tag{3}$$

The second component is a multinomial regression (Equation 4) which predicts the expected means using intercepts for each of the four length patterns, with the intercept for β_{2+2} remaining zero. The intercepts here encode our phonological well-formedness constraints by serving as a baseline proportion for each length type alternation. Slopes for boundedness (β_b) and homography (β_h) are also included in this regression. All slopes and intercepts are represented in Figure 1 by the vector $\hat{\beta}$. Slopes are multiplied by values of the generative annotations for boundedness (b_1, b_2) and homography (h_1, h_2) represented in Figure 1 as the design matrix $\hat{X}$.

We set the β_{2+2} intercept at zero to serve as a baseline by which the other intercepts are compared. Because the 2+2 alternation consists of two disyllabic nouns, neither effect of boundedness nor homography apply, as they are effects of the short form of a noun. Thus, the initial regression value before normalization for the 2+2 length pattern ends up being a value of 1 as detailed in Equation 4d.

(4)

$$y_{1+1} = e^{-\beta_{1+1} + \beta_b*b_1 + \beta_b*b_2 + \beta_h*h_1 + \beta_h*h_2} \quad \text{a)}$$

$$y_{1+2} = e^{-\beta_{1+2} + \beta_b*b_1 + \beta_h*h_1} \quad \text{b)}$$

$$y_{2+1} = e^{-\beta_{2+1} + \beta_b*b_2 + \beta_h*h_2} \quad \text{c)}$$

$$y_{2+2} = e^{-0} = 1 \quad \text{d)}$$

$$y_{tot} = y_{1+1} + y_{1+2} + y_{2+1} + y_{2+2} \quad \text{e)}$$

$$\mu_{1+1} = \frac{y_{1+1}}{y_{tot}} \quad \text{f)}$$

$$\mu_{1+2} = \frac{y_{2+1}}{y_{tot}} \quad \text{g)}$$

$$\mu_{2+1} = \frac{y_{2+1}}{y_{tot}} \quad \text{h)}$$

Figure 1 illustrates the relationships between various probability distributions and parameters in our model. Using Stan integration with R (Stan Development Team, 2019), we sampled the posteriors of each latent model parameter in 4 chains each with 10000 iterations, half burn-in. All models converged with an $\hat{R} < 1.001$.

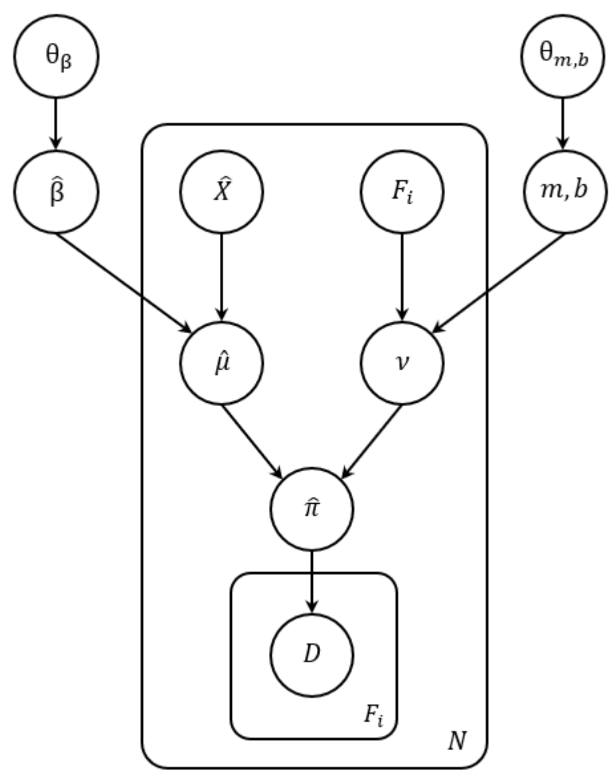

Figure 1: Graphical representation of our hierarchical Bayesian model architecture. $\hat{X}$, F_i, and D are all data points. θ_β and $\theta_{m,b}$ are uninformative normal priors centered at zero with a standard deviation of ten. $\hat{\beta}$, $\hat{\mu}$, ν, m, and b are all latent model parameters.

Results

To answer the question about whether we find frequency-dependent regularization in length alternation preferences of NN compounds in Mandarin we look to (1) the linear regression results and (2) the regression predicting ν in our hierarchical model.

With our simple linear regression model examining the relationship between the information entropy (skewedness) of the alternation distribution and overall frequency of the compound, we found a significant effect of overall frequency in predicting entropy in the expected direction ($\bar{\beta} = -.12$; $\mathrm{CrI} = [-.11, -.14]$). This negative relationship provides evidence for frequency-dependent regularization in the elastic length of Mandarin NN compounds.

Within our hierarchical Bayesian model, we examine the regression component which directly influences our concentration parameter (ν). This parameter serves as our metric of frequency-dependent regularization. If we find that higher overall frequency value corresponds to lower ν values, this means for higher frequency items, the distribution over alternations is less uniform and more skewed.

The slope parameter from our exponential regression (Equation 3) is significantly different from zero in the negative direction (Table 2). This provides further evidence in favor of frequency-dependent regularization above and beyond the effects of the generative constraints of phonological well-formedness, boundedness, and homography.

Table 2: Slope and Intercept Estimates for Concentration (ν) Exponential Regression

	mean	sd	95% CI	N_{eff}
m	-.03553	0.0027	[-.0697, -.0009]	4255
b	1.7986	0.0302	[1.3936, 2.1939]	4519

This relationship between ν and overall frequency is visualized in Figure 2. We can see that as the log of overall frequency increases, the predicted ν value decreases.

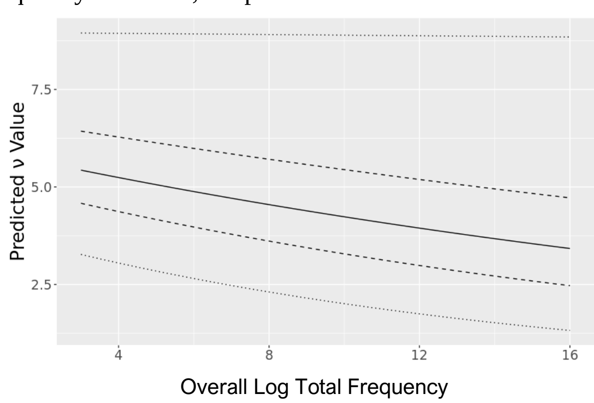

Figure 2: The relationship between overall frequency (logarithmic) and the predicted concentration parameter value; the horizontal axis represents the log of overall frequency for a given NN compound and the vertical axis represents the predicted ν value from our exponential regression (Equation 3). The solid line is the mean prediction with the dashed lines showing a 50% credible interval and the dotted lines showing a 95% credible interval.

Our multinomial regression included intercept components for each of our length patterns (holding the intercept for the 2+2 pattern as 0) and slope components for boundedness (β_b) and homography (β_h). Our intercepts all had values significantly greater than zero ($\beta_{1+1} = 2.27$; $\mathrm{CrI} = [2.13, 2.41]$), ($\beta_{1+2} = 2.01$; $\mathrm{CrI} = [1.89, 2.12]$), ($\beta_{2+1} = 1.70$; $\mathrm{CrI} = [1.60\ 1.81]$), demonstrating that all three other length patterns were less preferred than 2+2, with 1+1 being the least preferred and 2+1 being most preferred after 2+2.

Neither slope parameter had an effect significantly different from zero ($\beta_b = -.03$; $\mathrm{CrI} = [-.22, .18]$), ($\beta_h = .06$; $\mathrm{CrI} = [-.07, .14]$), so we cannot make any claim about the effects of boundedness or homography on the relative preference strength of any length patterns. It is worth noting, however, that 529 out of 683 items contained two free morphemes, thus this dataset was not well-suited to examining the effect of boundedness.

In addition to examining the posterior distributions of our parameters, we also plot actual entropy of each item alongside predicted entropy as a function of our predicted ν value for each item by overall frequency. This allows us to visualize how well our hierarchical model is capturing the relationship between frequency and regularization in our dataset. Predicted entropy in this instance was calculated by sampling 10,000 multinomial distributions from a Dirichlet distribution with a particular ν value (and with $\hat{\mu}$ determined by the intercepts in $\hat{\beta}$) and calculating the average entropy across all of the sampled distributions.

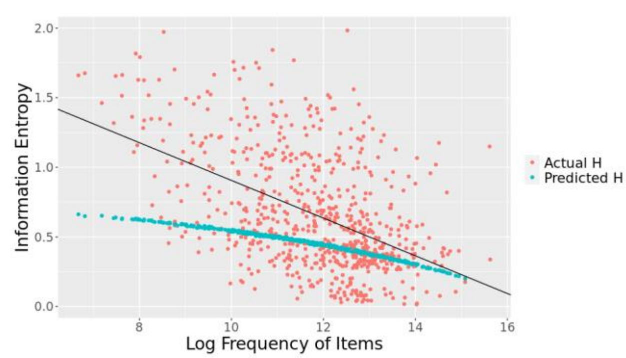

Figure 3: Actual and predicted entropy of length pattern distributions by overall frequency. Predicted entropy is calculated as a function of the relationship between overall frequency and ν in our hierarchical model ($\nu = e^{-0.355 * F_i + 1.80}$). The black line represents the line of best fit according to the actual data ($m = -.14, b = 2.26$).

We can see that the predicted ν-H relationship is much more clearly defined and dramatically less noisy than the actual relationship. We would expect if our prediction was capturing most of the variance that the predicted data would appear as a trendline over the actual data. This is clearly not the case, alerting us to the fact that there is still a sizeable amount of variance in this phenomenon that is yet to be explained.

It is also important to note that our model predicts the entropy of items with high overall frequency relatively well, with its performance decreasing as overall frequency decreases. This difference can be partially attributed to (1) the lack of variability in $\hat{\mu}$ values based on differing boundedness and homography scores (neither slope was significant) and (2) the high degree of skewedness in the $\hat{\mu}$ values calculated from the base intercepts $\hat{\mu} = [.0729,\ .0947,\ .1283, .7042]$. With a skewed set of $\hat{\mu}$ values, it would require extremely high ν values (greater than 1000) in order to predict entropy values as high as the actual entropy values that we see in the lowest frequency items. For this reason, high information entropy values which represent more uniform distributions are difficult to achieve by simply modulating our concentration parameter ν.

Discussion

This study looked for frequency-dependent regularization in the elastic length of Mandarin NN compounds. Evidence for such an effect is provided in two ways: first, by the significant slope predictor in a linear regression between overall frequency and information entropy. Second, we also found a significant slope predictor in a regression between overall frequency and the concentration parameter for a Dirichlet distribution in a hierarchical Bayesian model. This extends previous work on frequency-dependent regularization by demonstrating it in a language other than English, especially in a non-Indo-European language. It also extends the concept to include structures with more than two alternations.

Our initial regression model tells us that NN compounds with higher overall frequency have less uniform distributions. In most cases, the most common length pattern is the 2+2 pattern, but there are some exceptions to the rule.

However, this initial model does not account for the influence of generative constraints like phonological well-formedness, boundedness, or homography. Our second model has a more detailed architecture which does take these factors into account. Phonological well-formedness is accounted for by providing each of the four length patterns with an intercept parameter. Boundedness and homography of each noun are accounted for by slope parameters. All of these generative constraints and their parameters served as priors for a Multinomial-Dirichlet distribution which also utilized a concentration parameter to modulate frequency-dependent regularization.

The effects of phonological well-formedness constraints from our Bayesian model, did roughly match the theoretical predictions in that 1+2 << 2+1 << 2+2. 1+1 compounds, however, remain least preferred despite being perfectly acceptable phonologically. This is a phenomenon which has been demonstrated in previous acceptability judgment studies (Qin & Duanmu 2017; 2019) and still warrants an explanation.

One possible explanation is that short compounds are highly homographous and present a great deal of ambiguity. However, the slope parameter encoding the generative effect of homography on elastic length preference did not significantly differ from zero. This leads us to believe that the many senses of a single Chinese character are not the main source of speakers' dispreference for 1+1 NN compound length patterns. Additionally, boundedness of a noun did not seem to have a significant effect on length pattern preference. However, our dataset is dramatically skewed towards compounds with two free morpheme short forms which is typical of Chinese typology, so we cannot make strong claims here about the effects of boundedness.

While we do find that frequency-dependent regularization is playing a part in determining these length pattern preference distributions, there is much still to be learned about what factors contribute to these preferences. There remains a lot of variability in these distributions which cannot be accounted for by frequency-dependent regularization nor by phonological constraints.

It is also important to note that we restricted our investigation to NN compounds which are fully elastic, meaning that they have four grammatical length pattern alternations in order to be able to compare them to each other. Our results therefore do not reflect the elastic behavior of compounds which alternate between only three or two length patterns.

Finding evidence of frequency-dependent regularization in this Mandarin morphological four-way alternation has implications for the way we understand the relationship between overall frequency of a construction and its storage. Morgan & Levy (2015) claim that evidence of frequency-dependent regularization lends credibility to the hypothesis that generative constraints and item-specific knowledge have variable effects on preference strength. When processing novel or infrequent lexical items, interlocutors rely mostly on generative knowledge. Item-specific experience is then increasingly recruited as a function of overall frequency, until processing of highly frequent items relies almost entirely on item-specific storage rather than on generative knowledge. Morgan & Levy (2016) tested this using forced-choice and self-paced reading tasks to examine processing difficulty for binomial expressions. They found that the processing difficulty of novel binomials can be explained by generative constraints alone. The processing of high frequency binomials, however, relies primarily on direct experience with that particular binomial.

By demonstrating that frequency-dependent regularization applies to Mandarin NN compounds, we therefore demonstrate that the tradeoff between reliance on generative constraints and item-specific knowledge in processing is potentially a feature of human language generally rather than a unique quality of the English language.

References

Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software, 100(5)*, (pp.1-54).

Carmody, S. (2023). ngramr: Retrieve and Plot Google n-Gram Data. R package version 1.9.3.

Chinese Academy of Social Sciences (2016). 现代汉语词典 [Contemporary Chinese Dictionary] (7th ed). 商务印书馆有限公司 [*The Commercial Press*].

Duanmu, S. (2007). The phonology of standard Chinese (2nd ed). Oxford University Press.

Duanmu, S. (2012). Word-length preferences in Chinese: A corpus study. *Journal of East Asian Linguistics*, *21(1)*, (pp. 89–114).

Li, Y., Scontras, G., & Futrell, R. (2023). Chinese words shorten in more predictive contexts. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, (pp.1286-1291).

Liu, Z. & Morgan, E. (2020). Frequency-dependent Regularization in Constituent Ordering Preferences. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, (pp. 2990-2996).

McEnery, A., & Xiao, Z. (2004). The Lancaster Corpus of Mandarin Chinese: A Corpus for Monolingual and Contrastive Language Study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, (pp. 1175–1178).

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., The Google Books Team, Pickett, J. P., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331(6014)*, (pp. 176-182).

Morgan, E., & Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, (pp. 1649-1654).

Morgan E., Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, (pp. 384-402).

Qin, Z., & Duanmu, S. (2017). A judgment study of word-length preferences in Chinese NN compounds. *Lingua*, *198*, (pp. 1–21).

Qin, Z., & Duanmu, S. (2019). An acceptability judgment study of 1+1 NN compounds in Chinese. Lingua, *222*, (pp. 26–38).

Qin, Z., and Duanmu, S. (2021). 汉语动名定中复合词的词长搭配：一项基于语料库的量化研究 [Word-length preferences in Chinese VN compounds: a corpus study]. 世界汉语教学 [*Chinese Teaching in the World*] *35(3)*, (pp.450-464).

Stan Development Team. 2019. Stan Modeling Language Users Guide and Reference Manual, 2.26.

Acknowledgements

This research was supported by the UC Davis Society of Hellman Fellows Program and the UC Davis College of Letters and Science (to E.M.).