Word Frequency Effects on Homophonous Words in Mandarin Chinese

Ethan Sherr-Ziarko

University of Oxford
Faculty of Linguistics, Philology, and Phonetics

Abstract

Previous work on the effects of word usage frequency on production has found a significant correlation between the usage frequency and duration of the word. This study examines usage frequency effects on the production of homophonous words in a corpus of Mandarin Chinese, seeking to determine the validity of previous results cross-linguistically. Analysis of the corpus reveals a similar pattern to that found in spoken English, but with additional differences at the high and low end of the usage frequency spectrum.

1. Introduction

Recent phonetic research has shown that word usage frequency affects pronunciation ([4],[8]): frequent words tend to shorten, and this phenomenon has lead Bybee to propose that lemma frequency should be considered part of the mental representation of words, as it appears to have an effect on their production ([3]).

This idea is not without controversy. Newmeyer ([15]) argues that the effects of word frequency can be explained away by the fact that in general repetition of any motor function – not only language – will result in increased routinization and increase the speed with which the action can be performed. This argument might hold water when examining words with both very different frequencies and phonological forms – such as, for example, the very frequent word *about* versus the infrequent word *impale*.

Phonological frequency refers to the frequency of the particular sound of a word within the language – for example, in English *in* and *inn* are canonically pronounced using the same set of phonemes in the same order, and hence share the same phonological frequency. Previous theories and models ([13],[14]) of word production have placed great importance on the phonological frequency of words, arguing that it should be considered the main variable causing any observed frequency-based effects in a language's phonetics.

This leads to the idea of **frequency inheritance**, the theory that words with identical phonological representations will 'inherit' the frequency values of all other words with identical form in the language (i.e. homophones) and therefore should all behave as if they had the same frequency. If true this would support the idea that frequency effects on pronunciation are simply a product of the repetition and routinization of a given phonological form.

Lemma frequency refers to the frequency of a word's semantic representation rather than its phonological structure, meaning that there is a lemma frequency representation for every single word and morpheme in a given language. This means that while *in* and *inn* may be pronounced in nearly identical fashion, they would have very different lemma frequencies for the purposes of representation in the mental lexicon. If the theory of frequency inheritance is true, we would not expect lemma frequency to matter, but many recent studies ([1],[5],[6],[8]) have pointed to lemma frequency as a greater motivating factor than simple phonological form when considering frequency-based effects on language production.

To determine the more relevant representation to consider, and whether or not word frequency should be considered a part of the mental representation of words, it is very useful to study **homophonous words.** Since these words are ostensibly pronounced in exactly the same way, even a rather slight frequency dependent phonetic variation is interesting if it can be observed consistently, casting doubt on the value of phonological frequency as major component of linguistic representation while supporting the relevance of lemma frequency.

This study examined a corpus of recordings of spoken Mandarin Chinese, a language rich in homophones of varying frequencies, with the aim of testing whether results found in English ([8]) would also be present in a very different language.

1.1. Homophones and lemma frequency

Changes in duration have been shown in many cases ([1],[3],[12]) to be strong indicators of word-frequency, with high-frequency forms being shorter and low-frequency forms being longer, or over-articulated. Gahl set out to examine word frequency effects on homophone pairs independent of factors such as orthography, word function, speech rate, and morphological structure. She discovered significant durational effects based on lemma frequencies, with low-frequency words (like *thyme*) consistently

having significantly longer durations than their high-frequency pairs (eg *time*).

Other corpus-based studies of frequency effects on duration have not shown results as strong as Gahl's. Jurafsky ([11],[12]) and Bell et al. ([1]) examined the most frequent English function words and their lower-frequency homophone pairs (such as the preposition *to* vs. the infinitive marker *to*) and failed to find any significant effects when other factors such as speaking rate, position, and predictability were controlled for. This led Jurafsky to posit that lemma-frequency effects were not a significant predictor of word duration, but Gahl's results to the contrary indicate that this is more likely an issue of the stimuli examined (function words) not behaving identically to content words in terms of their routines of access and pronunciation.

2. The Question and Hypothesis

This experiment examined whether or not there is any variation in pronunciation among homophonemic words in Mandarin Chinese – meaning words that are canonically constructed of the same phonemes. Mandarin has both a relatively small sound inventory, and a relatively great number of homophones. Although the dearth of phonemes in complementary distribution in Mandarin is partially overcome by its system of tones, there remains a huge number of phonemically identical words. Most monosyllabic words in the language have homophones, and often there are over a dozen for a given word.

This aspect of its phonology makes Mandarin an interesting language to study when considering the effects of word frequency on pronunciation. Frequency inheritance effects have been shown to not be particularly strong in Mandarin ([5]), but there has been no work examining whether lemma-frequency effects can be observed.

The research question addressed is whether or not the lemma frequency of a word affects its pronunciation relative to its homophone pairs? To address this question, word duration seems the most salient property to examine, as it has been shown to vary significantly based on word frequency in other languages. The hypothesis is that when examining a controlled set of speech data in a corpus, mean word duration of homophone pairs will vary inversely with lemma frequency.

3. Experimental Design and Issues

Since the corpus-based study [8] produced significant results, where similar lab-based studies had failed to reveal an effect ([9],[6],[11]), it was decided to use a large corpus to test the research question. The corpus used was the Mandarin Hub4 Broadcast News Corpus [10], which consists of 14740 recordings (roughly 30 hours of speech) of Mandarin broadcast news from 1997, by 27 different speakers. The data was recorded at 16kHz, and is largely clear and free of interference, although certain recordings have music overlaying the speech which causes occasional intelligibility issues. The recordings in the corpus have been force-aligned by Jiahong Yuan ([10]), yielding Praat ([2]) text grids which contain transcripts in Chinese characters, duration, and tone information for each segment in the recording.

The corpus is not without limitations. Firstly, due to its relatively small size, there are few extremely infrequent words, and generally only one or two tokens of those that do appear. This is a common problem in corpus-based studies, and is largely unavoidable. A second potential issue is that, given that the recordings are of news broadcasts, the speech data is most likely scripted (i.e. read from teleprompters), not spontaneous.

A number of criteria were used to select data to be examined. To avoid potential issues with lemma frequency conflicting with morpheme frequency in compound words, only monosyllabic tokens were considered. Among those, Chinese characters with multiple possible pronunciations were eliminated from consideration. The reason for this was largely a practical one – the available word-frequency data for Mandarin Chinese is based upon character frequency rather than solely lemma-frequency, due to the fact that it comes from corpora of written Mandarin. Therefore, although frequency information is available for a word such as 和 hé (“together”/ “also”), the same character can also be phonetically realized as hè, huó, or huò, all of which have separate lemmas. Finally, as function words may behave differently than content words in production ([1], [12]) only content words (nouns, verbs, adjectives, and adverbs) were examined.

Word frequency data was based on data compiled from the Internet corpus of Mandarin Chinese, and the Lancaster Corpus of Mandarin Chinese ([15]). Frequency values were given in parts per million. Although these lists are based on written data rather than spoken, the sheer quantity of data (over 280 million words) should provide a good estimate of relative frequency in speech.

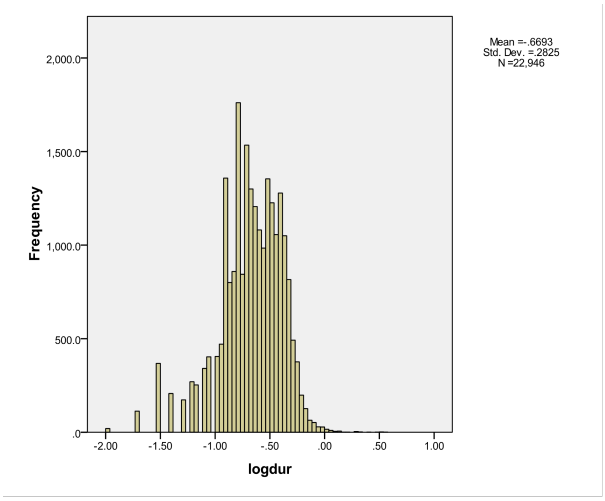
Automated scripts were created to compile from transcripts of the recordings a list of all valid words to be examined in the corpus. In total, the corpus contains 1650 instances of monosyllabic words, among which there are 491 total phonological forms with at least one homophone pair. From those pairs a

number were eliminated due to not meeting the criteria of the character having only one phonetic realization, or due to being function words, leaving a total of **322** pinyin homophone pairs, among which there are a total of **936** different lemmas.

4. Data Analysis and Discussion

22946 total utterances containing **322** valid pairs were examined. This is an extremely large sample, meaning that even small correlations will likely be considered significant by statistical tests. For this reason, it is important to examine the data carefully in order to avoid a type I error. Figure 4.1 is a histogram showing the distribution of $\log_{10}$ duration among all tokens of the experiment.

Figure 4.1; Histogram of logDuration.



logDuration appears to be normally distributed, but with a greater volume of tokens falling on the left side of the curve, which indicates a shorter duration. In spite of the log transformation, this figure lends a small amount of initial support to the hypothesis that more frequent words will be pronounced with shorter durations, based on the fact that the shorter durations appear to be more frequent. The mean syllable duration of all tokens taken from the corpus was 257.10ms, with a standard deviation of 155.9ms. The lowest frequency word examined had a frequency of 0.7 parts per million, while the most frequent was 11879.45 parts per million.

As there were **27** different speakers whose tokens were taken from the corpus, it is necessary to confirm that there are no significant differences between the speakers which could have undue influence on any statistical tests. The frequencies and durations of all speakers' utterances fall in a similar range, and hence should not cause issues for the experiment, but an issue arises when examining the differences in standard deviations of duration between speakers. Although small, incidental differences would likely cancel each other out when subsumed into an overall mean, an extreme outlier could cause issues in statistical analysis if they caused hundreds of outliers to be introduced into the experiment.

Most of the speakers have roughly similar standard deviations of duration, but one speaker (labelled ZHH in the corpus) has one that is roughly double that of all the other speakers. This could indicate inaccuracies in the durational measurements in the corpus for that speaker, or that speech rate for that speaker varied to such a degree that durational measurements of their utterances would not be consistent enough to use in statistical tests. Examination of that speaker's recordings showed that the majority of their recordings are overlayed with music, causing interference in the spectrograms and waveforms, and resulting in a number of inaccurate durational measurements. Fortunately only **551** of the total **22946** tokens examined were by that speaker and so this speaker's utterances were dropped from the data in order to avoid any potential confounds.

With potential confounds accounted for, we can begin to examine the entirety of the data. As this experiment examined the relationship between duration and frequency, statistical comparisons between the two were made. A Pearson correlation between frequency and duration among all tokens was significant at the $p<.01$ level, but the actual correlation of 0.139 is not particularly high, with the significance likely due to the extremely large sample size. A scatterplot of all the data also does not reveal any immediately obvious correlation, and while a univariate ANOVA indicates that there is significant interaction between duration and frequency at the $p<.001$ level, but again due to the huge sample size, an ANOVA shows every possible factor as highly significant.

In order to obtain a meaningful statistical result, it is necessary to be somewhat more selective with the data. The mean durations of all tokens of each utterance were taken, and compiled into a separate data set. From each homophone group (ranging from two to eleven homophones) two words were selected to be representative of the group. The words selected all followed the same criteria: for the first word of the pair, the word with the highest frequency in the group was always taken. For the second word, the lowest frequency word in the group with at least five utterances in the corpus was chosen. In cases where no word among the homophone group had five utterances in the corpus, the lowest frequency word in the group was chosen. The mean durations of the low frequency homophones were compared with their high frequency pairs in a paired sample t-test.

Table 1 shows the descriptive statistics of the pairs. With a sample size of **322** word pairs, the t-test showed the difference in means to be statistically significant at the p<.002 level.

Table 1; Descriptive statistics of low-frequency and high frequency pairs.

Paired Samples Statistics

		Mean	N	stdDev	stdError
Pair 1	lowFreq	253.46	322	108.70	6.05
	highFreq	231.36	322	70.39	3.92

The mean duration of the low-frequency forms was 22.1ms greater than that of the high-frequency forms, which follows the pattern found in [8].

The results appear to indicate that lemma-frequency influences duration in Mandarin Chinese in a fashion similar to what was seen in English by Gahl [8]. This casts doubt on the viability of the theory of frequency inheritance in Mandarin Chinese. The statistical analyses appear to indicate fairly directly that lemma-frequency is a far more important factor than phonological frequency where usage-frequency based effects on the language are concerned. This indicates that Mandarin Chinese speakers do not necessarily access words in the mental lexicon mainly on the basis of phonemic structure (as argued in [7]), but rather that lemma categorization plays a more important role in the organization of the lexicon.

A somewhat surprising observation about the data is that although it does appear to hold that low-frequency words have longer durations while high-frequency are shorter, if examples on the extreme ends of the frequency scale are examined, the pattern actually reverses. If only extremely high frequency words (defined at words with a frequency of ≥100 parts per million, or among the 1000 most frequent words in the language) and extremely low frequency words (words with a frequency of ≤10 parts per million), then the mean duration is almost exactly opposite of what can be seen in Table 1. When examining the entirety of the data, those extremely high-frequency words have a mean duration of 266.05ms, while extremely low-frequency words have a mean duration of 219.67ms. This difference is significant at the p<.001 level, although the level of significance could again be related to the very large sample size of the extremely high-frequency words.

These contrasting results suggest a few possibilities. One is that it is simply a product of the Hub4 corpus itself, which is composed of scripted speech rather than spontaneous. It is possible that reading from a script could influence the pronunciation of extremely high or low-frequency words in unexpected ways. Alternatively, it is possible that this is a genuine phenomenon in Mandarin Chinese, and that although the pattern of higher lemma frequency resulting in shorter durations of utterances does appear to hold in general, words at the far ends of the frequency spectrum actually behave in dramatically different ways.

If this pattern were found to be consistent, it would lend support to Bybee's ([4]) theory of lemma-frequency as significant in language processing rather than to Newmeyer's ([15]) counter argument that such effects can be explained entirely by the routinization of motor function. If extremely high-frequency words in Mandarin were found to actually be pronounced with significantly greater durations than other words, it would indicate that lemma-frequency is having an effect that runs counter to the idea of increased routinization always resulting in increased performance speed. It would be an unexpected result based on previous findings regarding the relationship between lemma frequency and duration, but it is an interesting avenue for future investigation into the effects of lemma frequency cross-linguistically.

References

[1] Bell, A., D. Jurafsky, E. Fosler-Lussier, C.Girand, M. Gregory, & D. Gildea. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113: 1001–1024.

[2] Boersma, P. & D. Weenink. (2014). Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/

[3] Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press: Cambridge.

[4] Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language* 82: 711-733.

[5] Caramazza, A., A. Costa, M. Miozzo, & Y. Bi. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27: 1430–1450.

[6] Cohn, A., J. Brugman, C. Crawford, & A. Joseph. (2005). Lexical frequency effects and phonetic duration of

English homophones: An acoustic study. *Journal of the Acoustical Society of America* 118: 2036.
[7] Dell, G. (1990). Effects of frequency and vocabulary type on phonological speech errors. Language and Cognitive Processes 5: 313–349.
[8] Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84/3: 474-496.
[9] Guion, S. 1995. Word frequency effects among homonyms. *Texas Linguistic Forum* 35: 103–116.
[10] Huang, S., J. Liu, X. Wu, L.Wu, Y. Yan, Z. Qin, (1997). Mandarin Broadcast News Speech Corpus. https://catalog.ldc.upenn.edu/LDC98S73
[11] Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In eds R, Bod, J. Hay, and S. Jannedy, *Probabilistic Linguistics.* 39–95. Cambridge, MA: MIT Press.
[12] Jurafsky, D., A. Bell, & C. Girand. (2002). The role of the lemma in form variation. In eds. Carlos Gussenhoven and Natasha Warner, *Laboratory Phonology 7.* 1–34. Berlin: Mouton de Gruyter.
[13] Levelt, W., A. Roelofs & A. Meyer. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences 22: 1–75.
[14] Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. Journal of Speech, Language, and Hearing Research 44:778–792.
[15] Newmeyer, F. (2006). On Gahl and Garnsey on grammar and usage. *Language* 82: 399–404.
[16] Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, WaCk y! Working papers on the Web as Corpus. Gedit, Bologna.