N-GRAM FREQUENCY EFFECTS ON SPEECH PRODUCTION IN MANDARIN CHINESE

Ching Chu Sun, Peter Hendrix

Eberhard Karls Universität Tübingen
ching-chu.sun@uni-tuebingen.de

ABSTRACT

We report the results of two studies that investigate frequency effects of random multi-word sequences in speech production. First, we collected acoustic durations for random three-word sequences in a reading aloud experiment in Mandarin Chinese. Second, we extracted acoustic durations of random word trigrams from a spoken corpus of Taiwan Mandarin [6]. We analyzed both data sets using quantile generalized additive mixed-effect regression models (QGAMMS) [5]. For both the experimental data and the spontaneous speech data, we observed significant effects of trigram frequency on acoustic durations that existed over and above the effects of the components unigram and bigram frequencies. The robust effects of *n*-gram frequency indicate that the acoustic signal in speech production is influenced by the combinatorial properties of words, even when the word sequences under investigation do not form fixed expressions or lexical bundles.

Keywords: speech production, trigram frequency, combinatorial properties, Mandarin Chinese

1. INTRODUCTION

Word frequency effects are among the most well-documented effects in the psycholinguistic literature. Higher frequency words typically lead to shorter response times in behavioral experiments. Frequency effects, however, are not limited to the word level. Recent studies documented frequency effects of multi-word sequences in the lexical decision [3] and reading aloud task [12, 8]. Frequency effects on acoustic durations have been reported as well, both at the word level [4, 7, 9] and at the phrase level [12, 1, 2] - with shorter acoustic durations for more frequent words and phrases.

Here, we present an investigation of the duration of word *n*-grams in both elicited speech and spontaneous speech. Below, we first report the results of a reading aloud experiment. Next, we take a closer look at pronunciation durations of multi-word sequences in a data set extracted from a corpus of spontaneous speech: the Taiwan Mandarin corpus [6]. Whereas previous work focused on frequency effects for carefully selected phrases that form well-defined lexical units, the word *n*-grams in the current study are random word trigrams without a predefined linguistic structure or a well-established lexical status.

2. READING ALOUD EXPERIMENT

2.1. Methods

2.1.1. Participants

Thirty participants took part in the experiment. All participants were native speakers of Mandarin Chinese that were born in mainland China. Their mean age was 25.86 (sd = 4.16). Twenty-two participants were female, whereas eight were male.

2.1.2. Materials

The stimuli for the experiment were sequences of three words. We henceforth refer to these three-word sequences as word trigrams and to the words in a three-word sequence as word $n-2$, word $n-1$, and word *n*. From the Chinese Lexical Database (CLD) [11], we selected 400 unique two-character words to serve as the middle word in the word trigrams. We refer to these words as the base words. For each base word, we extracted a random three-word sequence in which the base word appeared as word $n-1$ from a 466 million word corpus of Mandarin Chinese: the Simplified Chinese Corpus of Webpages (SCCoW) [10]). To avoid the inclusion of spurious items, word trigrams were considered for selection if and only if their frequency in SCCoW was equal to or greater than 10.

2.1.3. Design

We randomized the order of the 400 word trigrams between participants. The response variable is the acoustic duration of the pronunciations in the reading aloud experiment. The predictors of primary interest are the frequencies of the word unigrams, bigrams, and trigrams. These frequencies were extracted through Google searches that were restricted to documents in simplified Chinese from mainland China.

The frequency counts for word unigrams, word bigrams and word trigrams are highly correlated. It is therefore difficult to establish the independent contribution of the frequency of multiword sequences on the basis of the raw frequency counts. To obtain more independent frequency measures, we entered the (log-transformed) Google frequency counts into a principal components analysis with varimax rotation. The loadings of the frequency measures on their corresponding rotated components were all greater than 0.800 (mean: 0.933). The largest loading of a non-target frequency count on a rotated component was 0.437. The rotated components (henceforth RCs) thus provide measures of unigram, bigram, and trigram frequencies that are, to a large extent, independent. The rotated components for the frequency unigram, bigram, and trigram frequency counts were entered into the analysis as predictors.

We furthermore entered a number of control variables into the statistical analysis. First, we included the initial phoneme and final phoneme of the word trigram and the tone of each word in the word trigram as predictors. In addition, the total number of strokes in the orthographic form of a word trigram and the total number of phonemes in the pronunciation of a word trigram were included as measures of visual complexity and phonological length. All control variables were extracted from the Chinese Lexical Database (CLD [11]).

2.1.4. Procedure

The experiment was carried out in a soundproof booth. Prior to each trial a fixation mark was shown at the location of the first character of the word trigram. Next, the word trigram was presented in white KaiTi 80 point font on a black background. The word trigram remained on the screen for 3,000 milliseconds. After each stimulus, a blank screen appeared for 1,000 ms, followed by the fixation mark for the next trial. The experimental items were preceded by 10 practice items. Each experimental session had a duration of about 60 minutes, including setup and a 5 minute break halfway through the experiment.

2.2. Analysis

We removed predictor outliers that were further than 3 standard deviations from the predictor mean prior to analysis. We initially carried out a non-linear mixed-effect regression using a generalized additive mixed-effect model (GAMM) [13, 14]. The residuals of this model, however, violated the normality assumption. Normality of the residuals is not an assumption of quantile regression models. We therefore opted for a quantile regression of the median with a quantile generalized-additive mixed-effect model (QGAMM) [5] for the analysis of the acoustic durations in the experimental data.

The random effects of participant and item were modelled through random intercepts. We used parametric terms to model the effects of lexical tone and the total number of phonemes in the word trigram. The effects of the rotated components representing word unigram, word bigram and word trigram frequencies, as well as the effect of the total number of strokes in the word trigram were modelled through smooth terms. We restricted all smooth terms to fourth order non-linearities.

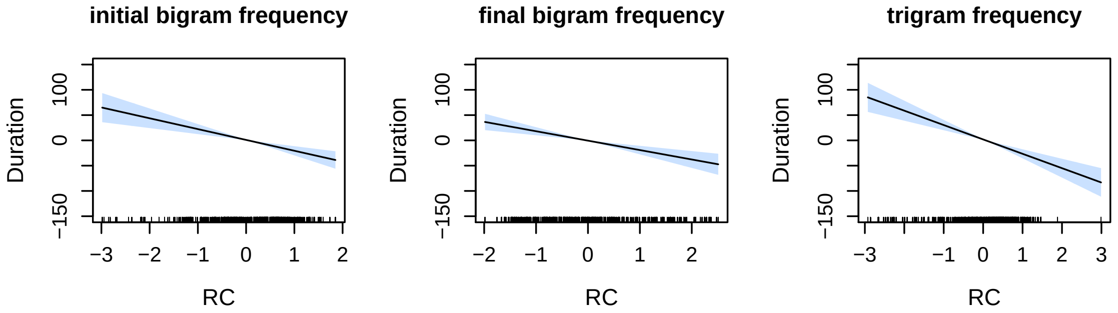
2.3. Results

The QGAMM fit to the pronunciation durations revealed significant random effects of participant (χ^2 = 13713.826, $p < 0.001$) and item (χ^2 = 6814.935, $p < 0.001$), as well as of the initial phoneme of the trigram (χ^2 = 842.526, $p = 0.006$). We furthermore observed a significant effect of the lexical tone of the final word of the trigram, of the total number of phonemes in the pronunciation of the word trigram (z = 20.183, $p < 0.001$), and of the total number of strokes in the orthographic form of a trigram (χ^2 = 10.505, $p = 0.001$).

Here, however, we are primarily interested in the effects of (the rotated component that correspond to) the unigram, bigram, and trigram frequency counts. The unigram frequency of the first (χ^2 = 14.895, p = 0.001), the second (χ^2 = 19.547, $p = 0.001$), and the third (χ^2 = 22.685, $p < 0.001$) word in a word trigram all had a significant effect on the acoustic duration of the pronunciation of a trigram, with shorter durations for word trigram that contained high frequency words.

We observed significant effects of the frequency of the word-initial (χ^2 = = 20.347, $p < 0.001$) and word-final (χ^2 = 20.618, $p < 0.001$) bigram as well. Finally, the QGAMM analysis revealed a significant effect of trigram frequency (χ^2 = 35.272, $p < 0.001$) over and above the effects of the frequency of the component unigrams and bigrams. Interestingly, the

Figure 1: Results of the QGAMM fitted to the acoustic durations in the experimental data. Plotted are partial effects of the frequency of the trigram-initial bigram (left panel), the trigram-final bigram (middle panel) and the word trigram (right panel).



effect size of the effect of trigram frequency is larger than the effect size of the frequency of the component unigrams and bigrams. The effects of bigram and trigram frequency for the experimental data are presented in Figure 1. As can be seen in Figure 1 the bigram and trigram frequency effects are linear in nature.

3. SPONTANEOUS SPEECH DATA

3.1. Methods

3.2. Participants

The spontaneous speech data were extracted from a 20-hour corpus of spontaneous speech in Mandarin, the Taiwan Mandarin corpus [6]. This corpus contains spontaneous speech about a self-selected topic for 55 native speakers of Mandarin Chinese from Taiwan. Thirty of these speakers were female, whereas 25 were male.

3.2.1. Materials

As was the case for the experimental data, the stimuli in the study of spontaneous speech were word trigrams. We selected 136 two-character words to serve as the final word in the word trigrams. For these 136 words, we extracted the duration of 20 random word trigram tokens in which the base word appeared as the final word from the Taiwan Mandarin corpus. The total number of trigram tokens in the data set of spontaneous speech, therefore, was $2,720$.

3.2.2. Design

As was the case for the experimental data, the response variable for the spontaneous speech data was the acoustic duration of the pronunciations of the word trigram tokens selected from the Taiwan Mandarin corpus. As before, our main interest is in the effects of the frequency of word unigrams and word bigrams in a word trigram, as well as in the frequency of the word trigram itself. We collected these frequency measures through Google searches that were restricted to documents from Taiwan that were written in traditional Chinese.

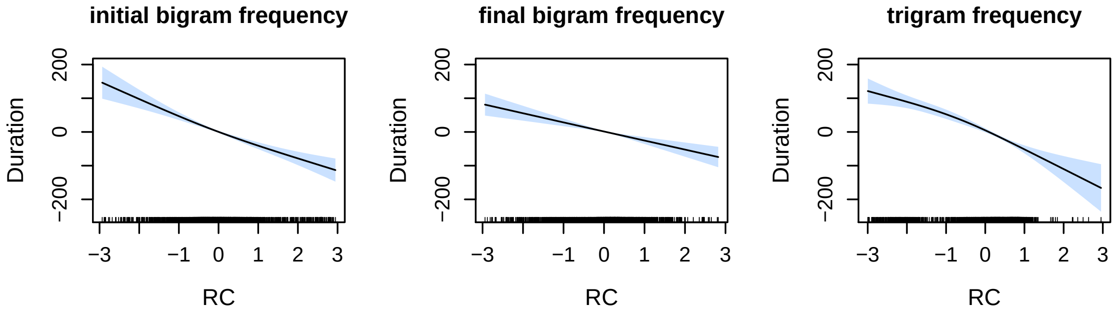
To be able to tease apart the effect of unigram, bigram, and trigram frequencies we again applied a principal components analysis with varimax rotation to the (log-transformed) Google frequency counts. The loadings of the frequency measures on their corresponding rotated components were all greater than 0.815 (mean: 0.910). The largest loading of a non-target frequency count on a rotated components was 0.344. As was the case for experimental data set, the rotated components (RCs) therefore provide measures of unigram, bigram, and trigram frequencies that are to a large extent independent.

In addition to the rotated components for the unigram, bigram, and trigram frequency counts, we entered a number of control variables as predictors into the analysis. The set of control variables was identical to the set of control variables used for the experimental data and consistent of the initial and final phoneme of the word trigram, the tone of each word in the word trigram, the total number of strokes in the orthographic form of the trigram, and the total number of phonemes in the pronunciation of a word trigram.

3.3. Analysis

As was the case for the experimental data, we analyzed the spontaneous speech data with a quantile regression of the median in a generalized-additive mixed-effect model (QGAMM). As before, we mod-

Figure 2: Results of the QGAMM fitted to the acoustic durations in spontaneous speech. Plotted are partial effects of the frequency of the trigram-initial bigram (left panel), the trigram-final bigram (middle panel) and the word trigram (right panel).



elled the random effect of participant through a random intercept, the effects of lexical tone and the total number of phonemes in the pronunciation of a word trigram through parametric terms, and the effect of the total number of strokes in the orthographic form of a word trigram as well as the effects of the rotated components representing word unigram, word bigram, and word trigram frequencies through smooth terms that were restricted to fourth-order non-linearities. Predictor outliers further than 3 standard deviations from the predictor mean were removed prior to analysis.

3.4. Results

The analysis of the pronunciations in the spontaneous speech data revealed a significant random effect of participant (χ^2 = 491.571, p < 0.001). In addition, we observed significant random effects of both the trigram-initial phoneme (χ^2 = 33.143, p = 0.001) and the trigram-final phoneme (χ^2 = 25.828, p = 0.001). Furthermore, we found a significant effect of the total number of phonemes in the pronunciation of a trigram (z = 12.384, p < 0.001).

The effects of the rotated components that encode the frequency of the first (χ^2 = 55.927, p < 0.001) and the second (χ^2 = 81.576, p < 0.001) word in a word trigram reached significance as well. We failed to observe a significant effect of the frequency of the final word in a word trigram (χ^2 = 0.024, p = 0.879).

Crucially, the effects of the (rotated components corresponding to) the frequencies of the word-initial bigram (χ^2 = 82.987, p < 0.001), the word-final bigram (χ^2 = 25.733, p < 0.001), and the trigram as a whole (χ^2 = 111.852, p < 0.001) were highly significant. The effects of bigram and trigram frequency for the spontaneous speech data are presented in Figure 2. As can be seen in Figure 2, the bigram and trigram frequency effects are linear or near-linear in nature. As was the case for the experimental data, the effect size of the effect of the frequency of the trigram as a whole is larger than the effect sizes of the frequencies of the components word unigrams and word bigrams.

4. DISCUSSION

Previous studies documented phrase frequency effects for the acoustic durations of pronunciation of word n-grams [12, 1, 2]. Acoustic durations, these studies found, are shorter for high frequency phrases as compared to low frequency phrases. Here, we examined the effects of the frequency of multi-word sequences on the acoustic durations in a reading aloud experiment and in spontaneous speech from the Taiwan Mandarin corpus [6]. For both data sets, we observed effects of the frequency of word trigrams on pronunciation durations that existed over and above the effects of the component word unigram and word bigram frequencies.

Whereas previous studies focused on acoustic durations of fixed expressions or well-defined lexical bundles, we imposed very few restrictions on the word trigrams selected for the studies reported here. Nonetheless, we observed robust effects of the frequency of the trigram as a whole in both elicited speech and spontaneous speech, that were highly significant across the predictor range. Furthermore, the effect of the frequency of the trigram as a whole was larger than the effect sizes of the unigram and bigram frequency effects. The combinatorial properties of words thus seem to play a pivotal role in the production of multi-word sequences, even when these sequences do not form fixed expressions or lexical bundles.

5. REFERENCES

[1] Arnon, I., Cohen Priva, U. 2013. More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech* 56, 349–371.

[2] Arnon, I., Cohen Priva, U. 2014. Time and again: The changing effect of word and multi-word frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon* 9, 377–400.

[3] Arnon, I., Snider, N. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62, 67–82.

[4] 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.

[5] Fasiolo, M., Goude, Y., Nedellec, R., Wood, S. N. 2018. *Fast Calibrated Additive Quantile Regression*. https://arxiv.org/abs/1707.03307v2.

[6] Fon, J. 2004. A preliminary construction of taiwan southern min spontaneous speech corpus (National Science Council [NSC-92-2411-H-003-050]). *Manuscript*.

[7] 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.

[8] Janssen, N., Barber, H. 2012. Phrase frequency effects in language production. *PLos one* 7, e33202.

[9] Pluymaekers, M., Ernestus, M., , Baayen, R. H. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118, 2561–2569.

[10] Shaoul, C., Sun, C. C., Ma, J. Q. 2016. The Simplified Chinese Corpus of Webpages (SCCoW). *Manuscript*.

[11] Sun, C. C., Hendrix, P., Ma, J. Q., Baayen, R. H. 2018. Chinese Lexical Database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods* 50, 2606–2629.

[12] Tremblay, A., Tucker, B. 2011. The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon* 6, 302–324.

[13] Wood, S. N. 2006. *Generalized Additive Models*. New York: Chapman & Hall/CRC.

[14] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1), 3–36.