Tonal bilingualsim: the case of two related Chinese dialects

Wu, J.

6 Do Tonal Bilinguals Store the Unproduced Tonal Variant of Etymologically Related Words? [10]

Abstract

Tonal bilinguals of Jinan Mandarin and Standard Chinese (SC) produce different tonal variants of the same Jinan word. Usually all the variants are segmentally identical to the word's SC counterpart but only one of the variants is tonally identical to the word's SC counterpart (variant_id). The word-wise probability of variant_id varies between 0 and 1. Naming latency data were elicited for 400 Jinan words from 42 speakers to test whether speakers who produced the variant_id also store the unproduced variant which is not tonally identical to the SC counterpart (variant_ni). If the speakers who produced the variant_id do not store the unproduced variant_ni, the naming latency should only depend on the speaker's choice of variant_id (yes/ no) but not on the word-wise probability of variant_id. If the speakers who produced the variant_id do store the unproduced variant_ni, the naming latency should depend on the word-wise probability of the speaker's chosen variant. The latter was verified by our results.

6.1 Introduction

Etymologically related words are translation equivalents which are similar in sound. They are either inherited from the common ancestor as cognates or borrowed across languages as loan words.

The phonological similarity between a pair of etymologically related words varies along a continuum and etymologically related words can be practically distinguished as identical and non-identical. For instance, the Dutch 'computer' and the English 'computer' are more similar than the Dutch 'neus' and the English 'nose'. Experimental evidences suggest that etymologically related words with different degrees of phonological similarity may have different statuses in lexical representation and lexical access. For instance, only identical cognates showed cognate facilitation effect in eye-tracked reading, while non-identical cognates did not (Duyck, Assche, Drieghe, & Hartsuiker, 2007). However, the effect of phonological similarity is inconsistent within and across studies (Dijkstra, Grainger, & Van Heuven, 1999; Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010; Duyck et al., 2007; Lemhöfer & Dijkstra, 2004). Several reasons may be responsible for the unstable effects. First, the phonological similarity co-varies and interacts with the orthographic similarity. Second, etymologically related words can be non-identical in different ways. For instance, some have different vowels and the others have different consonants. However, in the bilingualism considered by these studies, there are not enough cases for studying each sub-type separately.

[10] This chapter is based on Wu, J., & Chen, Y. (2014). Tonal variants in the bilingual mental lexicon. Paper presented at the The Fourth International Symposium on Tonal Aspects of Languages (TAL 2014), Nijmegen. ISBN: 978-90-9028606-8

The degree of phonological similarity becomes a more important variable when it comes to the tonal bilingualism of Jinan and standard Chinese (SC). First, unlike in the previous studies, the etymologically related words are the majority in the vocabulary of bilinguals who speak these two closely related dialects. As a result, the majority of translation equivalents are phonetically similar, providing much more cases for studying this phenomenon systematically. Second, both Jinan and standard Chinese are written with the same logographic Chinese writing system. Thus all Jinan-SC etymologically related words are orthographically identical, which controls the confusion from orthography. Third, the segmental differences between the etymologically related words are almost reduced to annihilation in the youngest generation. The tonal similarity between a Jinan word and its counterpart in standard Chinese decides the phonological similarity between them. As a result, the two etymologically related words are only different in tone. We can focus on the role of tone in the bilingual lexical representation and access.

Moreover, some Jinan words show different variants identical and non-identical to their SC counterparts in speech production. In our Jinan corpus collected in 2012, some Jinan words were produced with two or more tonal patterns, as shown in Figure 1. Here we call them multi-pattern words. Such a word usually has a variant almost identical to the word's counterpart in SC (variant_id), together with one or more variant(s) which is/are not identical to the words SC counterparts (variant_ni). Note that the segmental structure of both variant_id and variant_ni are almost always identical to the multi-pattern words' counterparts in SC and the only difference is carried by tone.

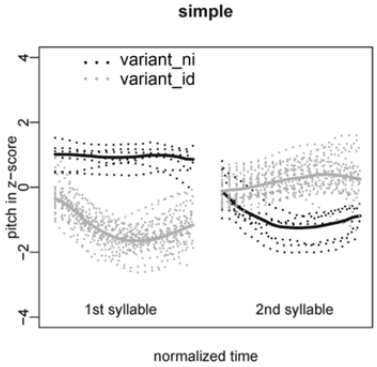

Figure 1: *The Illustration of one multi-pattern word. It is made with recordings produced by 42 Jinan speakers (1 or 2 outliers excluded). Pitch values were z-transformed semitones (the mean and SD were calculated with about 600 recordings for each speaker).*

How the two variants are stored in different speakers' memory is still an open question. It is related with the following two factors. First, for one Jinan word, a speaker can produce either a variant_id or a variant_ni in one rendition. Second, with each speaker providing one rendition of each word, the probability of variant_id can be measured for each word, by calculating the percentage of speakers who produced variant_id for the word. The former measurement represents whether

the variant_id of a lexical item is retrieved in the actual rendition of speech production (speaker's choice of variant_id, yes/ no); the latter represents the general probability of variant_id when the bilingual speakers name the shared orthographic form of the pair of etymologically related words in Jinan (word-wise probability of variant_id). For instance, the SC word 'simple' realizes with a low-plus-high tone. If we observed Speaker A produced the Jinan word 'simple' with a low-plus-high tone (variant_id) in his rendition and Speaker B produced the Jinan word 'simple' with a high-plus-rising tone (variant_ni) in his rendition, we say Speaker A made the choice of variant_id (y) and Speaker B did not make the choice of variant_id (n). With 32 out of 41 speakers producing the same word 'simple' in Jinan and, the word-wise probability of variant_id for the Jinan 'simple' is 0.78.

As shown in Figure 2, word-wise probability of variant_id ranges between 0 and 1 in our corpus. The majority of words were only produced with Jinan-only variants. The other 123 of the 400 (25.5%) recorded words were produced identical to its SC counterpart by at least one speaker. Within these words, 33 were produced identical to its SC counterpart by more than 85% of the speakers and 21 by around half (31%-76%) of the speakers. It is unlikely that all or most of the speakers coincidentally make the same code-mixing error together. This phenomenon indicates that some variant_id should be natively Jinan and tagged as a Jinan lexical item in the mental lexicon. On the other hand, the majority (66) of the multi-pattern words were produced only by a few (2%-26%) speakers as identical to its SC counterpart, which could be the real examples of code-mixing.

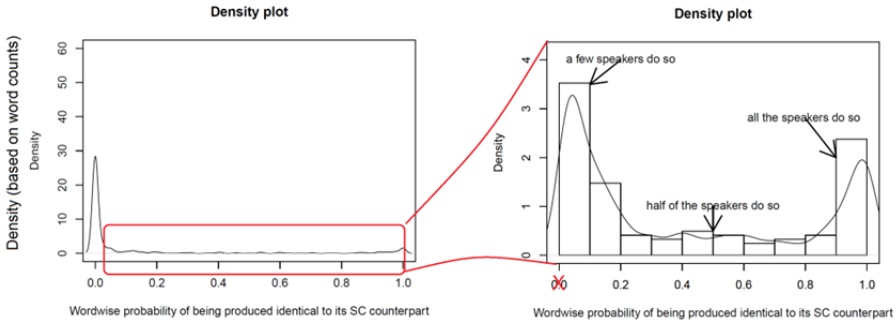

Figure 2: *Density plots of the word-wise probability of variant_id.*

Do speakers who did not produce variant_id also store it in an integrated lexicon together with variant_ni? The answer to this question decides how we should interpret the effect of word-wise probability of variant_id theoretically. The status of variant_id in the bilingual mental lexicon yields different predictions for the following experimental questions. Does the word-wise probability of variant_id affect the naming latency of the word, no matter which variant is picked in the actually rendition? Does the speaker's choice of variant_id affect the naming latency of the word, no matter how low or high the word-wise probability of variant_id is for this word? Do the two predictors interact?

Assuming the lexical representations made up by different phonemes (including tone) are also different (Dijkstra et al., 1999), variant_id and variant_ni should have

different lexical representations. As shown in Figure 3, (1) one possibility is that different variants are stored by different speakers. The speakers who produced variant_ni store variant_ni and the others who produced variant_id store variant_id as the phonological representation for the same Jinan word. The former speakers have both variant_ni and variant_id and the latter speakers only have variant_id. Under this hypothesis, the word-wise probability of variant_id only reflects the distribution of individual difference of phonological representation in the language system and should not affect the naming latency of the multi-pattern word. Instead, the speaker's choice of variant_id should show effects. Assuming an integrated bilingual lexicon (Van Heuven, Dijkstra, & Grainger, 1998), speaker who produced variant_ni should be generally slower than speakers who produced variant_id because the former speaker's variant_ni (Jinan) receives extra competition from variant_id.

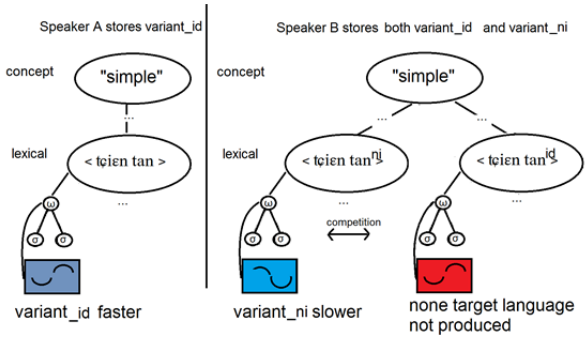

Figure 3: *Possibility (1). Different speakers store different variants.*

As shown in Figure 4 the other possibility is that all the speakers store both the variant_id and the variant_ni in an integrated lexicon. Under this hypothesis, the word-wise probability of variant_id reflects the likelihood of variant_id being selected in Jinan lexical access. In this case, the more likely the produced variant is, the shorter the naming latency should become, and no matter it is variant_id or variant_ni. We expect a higher word-wise probability of variant_id should reduce the naming latency of variant_id and increase the naming latency of variant_ni, since the condition implies a relatively lower likelihood of variant_ni. Correspondingly, a lower word-wise probability of variant_id should increase the naming latency of variant_id and reduce the naming latency of variant_ni, since the condition implies a relatively higher word-wise probability of variant_ni. Thus an interaction of speaker's choice of variant_id and word-wise probability of variant_id should be observed.

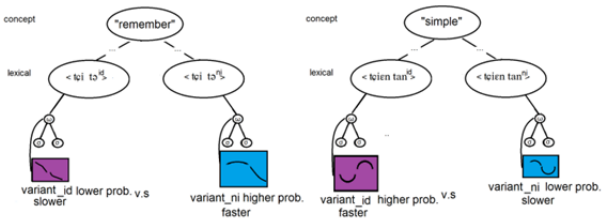

Figure 4: *Possibility (2). The variant with higher word-wise probability is produced faster.*

6.2 Experiment

6.2.1 Data preparation

Participant. The speech data used in the present study were collected from 42 Jinan native speakers in 2012.

Stimuli. Each speaker read 400 disyllabic Chinese words in Jinan. The written words were selected from a corpus of Chinese film subtitles (Cai & Brysbaert, 2010), in a way that one list of 200 high-frequency words were selected from the 10% disyllabic Chinese words with the highest word frequency. In a similar way, we selected the other list of 200 low-frequency words. In each list, each of the 20 disyllabic tonal combinations of standard Chinese contributes 10 words.

Procedure. The high and low frequency lists were presented to the speakers in two blocks with a self-paced rest break in between. The words in each list were presented in a different random order for each speaker. After the speaker finished producing a word, they pressed a key to see the next word.

A trained phonetician listened to each recording, looked at the spectrogram, and manually marked the beginning and the rhyme of the production. Also In this process, recordings with speech and recording errors were excluded from the corpus. Then naming latencies and pitch contours on the rhymes were extracted. To further remove pitch contour outliers, we calculated Local Outlier Factors (LOF) for each speaker's z-normalized pitch contours on the rhyme. Any pitch contours with an LOF greater than 1.5 (Breunig, Kriegel, Ng, & Sander, 2000) and belong to the 2.5% with the highest integral density were eliminated from the corpus. The naming latency outliers were excluded using a (method I) distributional based approach (van der Loo, 2010) on the log transformed naming latency.

Whether the Jinan word was produced almost identical to its counterpart in standard Chinese was judged by a phonetician with Putonghua Proficiency Test Certificates-Level1B. The probability of variant_id can be measured for each word, by calculating the percentage of speakers who produced variant_id for the word. The word-frequency was grouped into two levels (high-low) using the Chinese subtitle data (Cai & Brysbaert, 2010).

6.2.2 Model fitting

Only renditions of multi-pattern words (N = 3368) were taken into consideration in the following analysis. Linear mixed effects analyses were performed on the naming latency data, using R (R_Core_Team, 2013), lme4 (Bates, Maechler, Bolker, & Walker, 2013), and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2013) in the following way. A full model was built first, including the fixed effects of the speaker's choice of variant_id, the word-wise probability of variant_id, the Chinese word frequency, the tonal categories of the words' SC counterpart, and their two-way and three-way interactions, as well as the random intercept of the word and the speaker. (The random intercept model was proven to be better than alternative random slope models via model comparisons). When there were unrealized combinations of the nominal predictors, the corresponding interaction terms were removed. A backward elimination was then performed to remove non-significant effects, using p-values calculated from F test based on Sattethwaite's method (Kuznetsova et al., 2013).

In the final model, the main effect of Chinese word frequency was the only significant main effect, $F = 15.61$, $p < 0.05$. A higher Chinese word frequency reduces the Jinan naming latency. The main effect of the speaker's choice of variant_id, $F = 0.92$, n.s., and the word-wise probability of variant_id, $F = 0.10$, n.s., were both insignificant. However, their interaction was significant, $F = 4.00$, $p < 0.05$. As shown in Figure 5, for a word with higher word-wise probability of variant_id, the variant_id was named faster than the variant_ni; for a word with lower word-wise probability of variant_id, the variant_id was named slower than the variant_ni. All the other fixed terms were insignificant and removed in the model trimming.

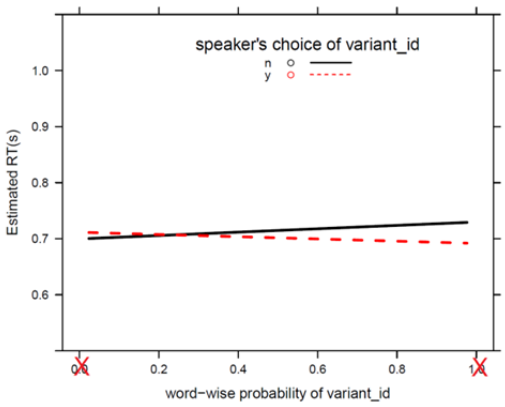

Figure 5: *Interaction of the speaker's choice of variant_id and the word-wise probability of variant_id.*

6.3 Discussion

The results support the hypothesis that all the speakers store both the variant_id and the variant_ni in an integrated lexicon (Figure 2.2), which divide the effect of Chinese word frequency. The result is also in-line with the data from a smaller corpus where each word was produced twice by each speaker in different random orders. We have observed both variant_ni and variant_id in the two renditions by the same speaker. We also have shown that the priming between such two variants spoken by the same speaker is similar (but reduced) compared with the priming between two renditions of the same variant in median-term auditory priming using lexical decision task (Wu, Chen, Schiller, & Van Heuven, accepted).

On the other hand, we have also seen in another study that individual backgrounds affect the tonal realizations of variant_ni. This indicates that individual differences have their own impact on the phonological representation (Wu et.al in preparation). However, considering this effect with the finding of the current experiment, individual differences seem to have their effect more on the shape of the stored tonal patterns than the lexical storage of variants.

Some variant_id are very likely native Jinan and stored in the Jinan lexicon. To include variant_ids in the Jinan lexicon, the theory should either allow (a) duplicated lexical nodes or (b) duplicated tagging for this variant.

6.4 Conclusions

The Jinan-SC tonal bilinguals' naming latencies of Jinan words depended on the word-wise probability of the speaker's chosen variants. No matter the chosen variant is tonally identical or non-identical to the word's SC counterpart, the higher the word-wise probability of the variant is, the shorter the naming latency is. The result supports that the speaker who produced the variant which is tonally identical to the words' SC counterpart do store the unproduced variant which is not tonally identical to the words' SC counterpart.

Acknowledgements

We would like to thank Prof. Shengli Cheng, Prof. Xiufang Du, Jia Li, Lulu Zhou, and Dianliu Neighbourhood Committee for the recruitment of participants. J.Wu's work was supported by a PhD Studentship sponsored by Talent and Training China-Netherlands Program.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 (Version R package 1.0-4.). Retrieved from http://CRAN.Rproject.org/package=lme4.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: identifying density-based local outliers.* Paper presented at the ACM Sigmod Record.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*(6), e10729.

Dijkstra, T., Grainger, J., & Van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*(4), 496-518.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*(3), 284-301.

Duyck, W., Assche, E. V., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 663.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. (2013). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition, 32*(4), 533-550.

R_Core_Team. (2013). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, version 2.15.

van der Loo, M. P. J. (2010). Distribution based outlier detection for univariate data,. *Discussion paper 10003, Statistics Netherlands, The Hague*.

Van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic Neighborhood Effects in Bilingual Word Recognition. *Journal of Memory and Language, 39*(3), 458-483.

Wu, J., Chen, Y., Schiller, N. O., & Van Heuven, V. J. (accepted). Tonal Variability in Lexical Access. *Language and Cognitive Processes*.