Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Does Tone Impact Mandarin Non-Word Acceptability Judgements?

Jamie Adams[1], *Sam Hellmuth*[2], *Leah Roberts*[3]

[1, 2, 3]University of York

jamie.adams@york.ac.uk, sam.hellmuth@york.ac.uk, leah.roberts@york.ac.uk

Abstract

Using non-words in psycholinguistic research allows for a high level of control over experimental stimuli. However, this relies on the assumption that they reflect natural language. Eliciting acceptability judgements from L1 speakers of the target language is one approach to ensuring the relative authenticity of non-words. For tonal languages, it is as yet unclear whether tone interacts with the perceived acceptability of non-words.

In this between-participant Mandarin non-word norming study, 72 L1 Mandarin listeners judged 750 syllables across five tones: tones 1-4 and the neutral tone (NT). Syllables were analysed as systematic gaps, which do not appear in the lexicon because they violate phonotactic constraints, and accidental gaps, which are phonotactically sound but are absent from the lexicon. Real words and malformed syllables acted as maximally and minimally acceptable controls, respectively.

Linear mixed effects models indicate that tones 1-4 do not modulate acceptability judgements. NT had a significant negative effect, but this likely arises from exposure to excised neutral tone syllables out of context rather than ungrammaticality. We suggest that Mandarin non-words can be associated with any lexical tone without concern for its effect on acceptability but that neutral tone stimuli should be presented in context to preserve authenticity.

Index Terms: Mandarin, non-word, acceptability judgement, tone

1. Background

The acceptability of non-words across languages is sensitive to numerous measures, including segment probability [2], phonotactic knowledge [1], and neighbourhood density [3]. Of these, phonotactic knowledge and neighbourhood density are consistently shown to be robust predictors of acceptability. For example, [1] demonstrate that gaps in the Mandarin lexicon are judged on a gradient as a function of their phonotactic form: real words are the most acceptable, followed by real syllables which do not co-occur with given tones (tonal gaps); non-words which are phonotactically possible but unattested in the lexicon (accidental gaps) are less acceptable, while phonotactically ill-formed non-words (systematic gaps) are the least acceptable. [1] therefore conclude that L1 listeners access and apply their phonotactic knowledge in non-word judgement.

[1] and [3] also demonstrate that neighbourhood density – the relative similarity of words to all other items in a lexicon – is also important in determining the acceptability of non-words. In both studies, higher neighbourhood density scores predicted higher acceptability judgements in Mandarin and Cantonese, respectively.

Lexical decision is also understood to be sensitive to suprasegmental factors such as lexical stress. In a lexical decision task, L1 Italian listeners were better able to distinguish non-words from real words when the non-words followed the most frequent stress pattern. This suggests that non-words are made less word-like when their assigned stress pattern is less frequent.

However, few studies have investigated the extent to which lexical tone impacts acceptability judgements. [1], for example, only tested items co-occurring with the high tone (T1). However, the finding that tonal gaps were the most acceptable gap type indicates that tone may have little effect on non-word acceptability. Moreover, [4] did incorporate tone probabilities into the modelling of Cantonese word-likeness judgements. Results indicate that neither tone probabilities nor neighbourhood density influenced judgements. Instead, phonotactic probability was the best predictor of word-likeness. We therefore do not expect tone to have a significant impact on the acceptability of non-words in Mandarin, but this is yet to be tested empirically.

2. Methods

1. Materials

The stimuli tested in this study were intended for use in research in L2 Mandarin tone acquisition, so they were selected on the basis of comparative segmental simplicity; only syllables containing segments common to both English and Mandarin were used. Similarly, while pinyin transcriptions were not used in this norming study, it was also important to select pseudo-words whose grapheme-phoneme correspondences (GPCs) matched in English and Mandarin. Thus, only words containing letters that represent the same sounds in the Romanised spelling systems for both languages were used e.g., <m> represents /m/ in Mandarin pinyin and in English, <p> represents /p/, and so on. Minor differences in vowel quality were tolerated. Finally, any Mandarin pseudo-word whose pinyin transcription resembled existing English words (e.g., king, tin, fun) was excluded.

The final set of stimuli comprised 150 syllables, all co-occurring with the four Mandarin tones and the neutral tone (NT) (150 syllables * 5 tones = 750 total items). Following the criteria set out by [1], 40 of the syllables were classified as systematic gaps (SG), defined as syllables which do not exist because they violate known phonotactic constraints. For example, /bua/ is phonotactically forbidden as a [+labial] consonant cannot be followed by a [+round] vowel [1]. Crucially, however, systematic gaps maintain the GCVX structure outlined by [5]. The next 60 stimuli were so-called accidental gaps (AG): syllables which are phonotactically sound but do not appear in the lexicon. The remaining 50 syllables comprised 30 real Mandarin syllables (RW) and 20 malformed syllables (MS). The latter differ from SG in that they violate the CGVX syllable structure by introducing complex onset clusters e.g., [blai] and [snao]. RW and MS syllables served as maximally and minimally natural stimuli

10.21437/SpeechProsody.2024-85

respectively, against which to calibrate judgements on systematic and accidental gaps.

As well as these categorical groupings, a neighbourhood density (ND) score was calculated for each item (syllable + tone combination). ND was calculated within each stimulus block (see 2.2) using a Levenshtein distance of 1 as the threshold: a word *x* was considered the neighbour of word *y* if it could be formed by the deletion, addition or substitution of one segment or tone from *y* [3]. Thus, [fao1] would have [fao2], [fao3], [fao4] and [bao1] as its neighbours and therefore a neighbourhood density score of 4.

Stimuli were recorded by a 26-year-old female native speaker of Mandarin from Sichuan, China on a Marantz professional solid-state recorder PMD661 MKII at 44.1kHz 16 bit. The speaker wore a Shure SM10 professional unidirectional head-worn dynamic microphone to maintain consistent recording measurements. The speaker was presented with each syllable written in pinyin with standard tone diacritics. She produced each syllable twice, cycling through tones (T) 1-4. This ensured that each syllable was produced uniformly across all four tones. Intensity was scaled to 70dB for all sound files using Praat's *Scale Intensity* function [6].

2.1.1. The Neutral Tone

Mandarin syllables which are prosodically weak are said to carry the neutral tone (NT) [7]. In natural speech, the NT surfaces in a number of syntactic and semantic environments, usually following T1-4 syllables. In the absence of semantic information, the most effective way to elicit maximally natural NT syllables for a large number of non-words was through reduplication. Reduplication is commonplace in Mandarin and has various functions across most lexical classes, e.g., [ge1ge5] 'older brother'; [shi4shi5] 'to try'. Crucially, the tone of the reduplicated syllable is neutralised in this environment, making it an ideal candidate for NT elicitation.

In reality, the NT is not one homogenous tone; its realisation varies considerably depending on the preceding tone and neutralisation has various effects on the segments with which it co-occurs [5], [7]. As such, an exhaustive investigation would have included neutralised forms of each of the four tones, but reduplicating all target syllables was beyond the scope of this study. As such, it was decided that all NT syllables would be elicited from the reduplication of one selected tone. T3 was eliminated as a candidate as it is the only tone to neutralise to a high target [7]; the pitch of the NT after T1, T2 and T4 is, to varying degrees, low. Based on the assumption that the most frequently encountered neutralised tone would yield higher acceptability ratings, a small corpus study was conducted to determine which of these remaining tones is most frequently reduplicated in the lexicon. The MagicData-RAMC dataset [8] comprises 351 dialogues by L1 Mandarin speakers, totalling 180 hours of natural speech and approximately 850,000 syllables. The dialogues were collated and automatically converted from written characters to numbered pinyin (e.g., ni3hao3). Reduplicated syllables were extracted and the modal tone in these syllables was identified. Results from this small corpus study suggest that the most frequently reduplicated tone in Mandarin is T4, the falling tone (T1 = 27%, T2 = 14%, T3 = 21%, T4 = 38%). A caveat is that this method includes in the count instances of speaker restarts or applications of tone sandhi. Automatic Romanisation converts characters based on their canonical readings. As such, an instance of /hao3 hao3/ will be included in the count, despite the fact that tone sandhi application blocks the neutralisation of the tone in the second syllable, resulting in a surface realisation of [hao3hao1] rather than the expected *[hao3hao5]. However, we are confident that the analysis outlined above provides a reasonable metric for determining the optimal context in which to elicit productions of target syllables realised with NT.

2. Procedure

Once stimuli had been selected, recorded and intensity-scaled, they were imported as *mp3* files into Gorilla Experiment Builder (www.gorilla.sc) where the experiment was created and hosted [9]. Although it would be ideal to present listeners with uncompressed *wav* format files, Gorilla advises against this to ensure a smooth audio experience for listeners.

Participants first provided informed consent and then answered questions about their language background, age and hearing difficulties. Next, participants were informed that they would be shown 'invented' Mandarin words and asked to judge them for how natural they sounded on a sliding (visual analogue) scale. The scale ranged from 0-100, where 0 was extremely unnatural and 100 was extremely natural. All instructions were presented on screen in simplified Mandarin characters and no time limit within which to respond was set. The audio played automatically with each item. No orthographic transcription was displayed. A large 'play' button appeared in the middle of the screen should participants wish to listen to the sound once more. The play button disappeared once it had been activated so participants could only listen to the sound up to two times. Following four practice trials, participants were automatically assigned to one of ten stimulus blocks, each containing 75 stimuli comprising 30 AG, 20 SG, 15 RW and 15 MS. Tones were distributed in a Latin-square fashion such that each block contained a unique set of syllable-tone configurations. A break screen appeared every 15 observations, encouraging participants to rest for as long as they needed.

3. Participants

72 native Mandarin speakers were recruited from universities in China and the UK. Three participants indicated that Mandarin was not their L1 and one reported hearing difficulties. Data from a further 9 participants were missing or incomplete, leaving a total of 59 eligible participants.

4. Analysis

Each of the 150 target items (syllable x tone combinations) was responded to by 3-10 participants (with the exception of blocks 4 and 5 which, after exclusions, were left with no eligible participants), yielding a total of 4425 responses for analysis.

Participants' raw responses were scaled against their individual response mean and standard deviation to produce a by-participant z-score transformation. Data from one participant were excluded due to anomalous responses that yielded an individual response mean of 95.

Results were explored in a treatment coded linear mixed effects model (lmer) using lme4 [10], predicting by-participant scaled (z-scored) acceptability scores with *gap type*, *tone* and their interaction, and neighbourhood density (ND) as fixed effects and a random intercept for *item* (where *item* = syllable + tone).

The between-participants method was adopted in an effort to avoid listener fatigue due to the large number of stimuli to be tested. However, a by-product of this design is a singular fit in the subsequent linear mixed effects model if random slopes are included, as each participant judged each item only once. Within-participant variation in responses to items cannot be captured, unfortunately, so random slopes were not included in the model random effects structure.

3. Results

Visual analysis of z-scores plotted by tone (Figure 1) suggests very little difference between acceptability judgements in each tone condition. NT syllables appear to elicit a lower median response, although there is considerable variation in all tone conditions.

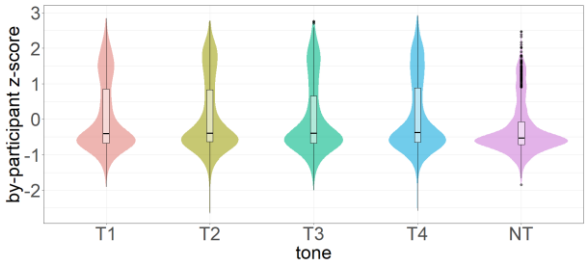

Figure 1. *By-participant z-scores in each tone condition.*

With real syllables set as the reference level for comparison, there was a main effect of *tone* such that NT was significantly less acceptable than all other tones [β = -1.155; t = -9.34; p <.001]. Significant interactions were also found between NT and all gap type conditions. The full model parameter estimates are shown in Table 1 and model predictions are visualised in Figure 2.

Table 1: *Parameter estimates for the linear mixed effects model: z-score ~ type*tone + tone + type + ND + (1 | item)*

	β	SE	*t*	*p*
(Int)	**1.38709**	**0.08645**	**16.045**	**< 2e-16 *****
AG	**-1.56063**	**0.10601**	**-14.722**	**< 2e-16 *****
MS	**-1.84756**	**0.13614**	**-13.571**	**< 2e-16 *****
SG	**-1.57059**	**0.11437**	**-13.733**	**< 2e-16 *****
T2	0.13344	0.12165	1.097	0.2731
T3	0.08801	0.12187	0.722	0.4705
T4	-0.07955	0.12185	-0.653	0.5141
NT	**-1.15499**	**0.12170**	**-9.491**	**< 2e-16 *****
ND	-0.04610	0.02547	-1.810	0.0709
AG:T2	-0.13907	0.14924	-0.932	0.3518
MS:T2	-0.03137	0.19238	-0.163	0.8705
SG:T2	-0.26819	0.16095	-1.666	0.0962
AG:T3	-0.09374	0.14920	-0.628	0.5301
MS:T3	-0.07726	0.19252	-0.401	0.6883
SG:T3	-0.28360	0.16113	-1.760	0.0789
AG:T4	0.12465	0.14913	0.836	0.4036
MS:T4	0.12884	0.19250	0.669	0.5036
SG:T4	0.06386	0.16108	0.396	0.6919
AG:NT	**0.97783**	**0.14934**	**6.548**	**1.30e-10 *****
MS:NT	**1.10679**	**0.19241**	**5.752**	**1.43e-08 *****
SG:NT	**0.97389**	**0.16098**	**6.050**	**2.61e-09 *****

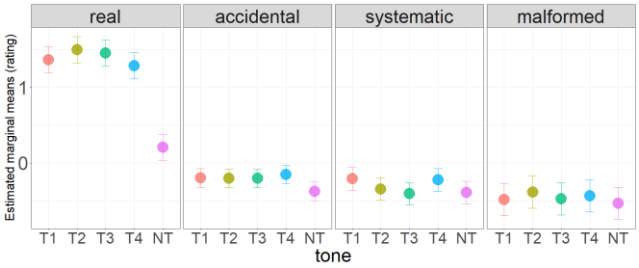

Figure 2. *Estimated marginal means of ratings, by gap type and tone.*

The model confirms that real words are significantly more acceptable than non-words of any type (β = 1.387; t = 0.086; p < 0.001). Figure 2 demonstrates that the negative effect of NT is particularly strong for real words. By relevelling the reference level of the gap type factor in a series of otherwise identical models, we find that the difference between AG and MS is marginally significant (β = -0.287; t = -2.337; p = 0.02), as is the difference between SG and MS (β = -0.277; t = -2.129; p = 0.037). No significant difference was found between AG and SG (β = -0.001; t = -0.103; p = 0.918).

Neighbourhood density appears to have no significant effect on ratings (β = -0.046; t = -1.810; p = 0.071), in line with findings by [4].

4. Discussion

1. General discussion

This study investigates the potential effects of lexical tone on the acceptability of Mandarin non-words in different gap type conditions. There was no effect of lexical tone category (T1-4) on acceptability of target syllables, for any gap type. As such, we suggest that Mandarin non-word stimuli can be paired with any of tones 1-4 without concern for their effect on acceptability.

However, items realised with the neutral tone were judged significantly less acceptable than items realised with any other tone. Given that this effect is particularly strong for real words, we posit that this is not evidence of phonological ungrammaticality, but that hearing NT syllables out of context makes them uninterpretable. As a result, NT syllables ought to be presented in context to preserve their authenticity and, thus, the ecological validity of the envisaged research.

The present study did not replicate the full set of patterns of gradient acceptability on the basis of phonotactic structure - where the acceptability of real words > tonal gaps > accidental gaps > systematic gaps - reported by [1], but real words were found here to be significantly more acceptable than any non-word type. Moreover, malformed syllables (not included in the analysis in [1]) were judged as marginally less acceptable than systematic and accidental gaps.

Another difference is that [1] found a significant, positive effect of neighbourhood density on ratings of non-words in the high tone condition, but there was no effect of neighbourhood density in the present study. The absence of a neighbourhood density effect in the presence of tone shown in the present study corroborates the findings of [4], although this may also reflect differences in the reference population against which neighbourhood density is calculated. [1] calculated ND using a database of 384 existing and 3136 unattested syllables in the Mandarin lexicon. [4], on the other hand, calculated ND for

non-words by comparing them to real words in the Chinese Character Database [11]. Due to the small scale of the present study, it was only possible to calculate ND on the basis of the other non-word stimuli. It is therefore possible that ND effects are only found when both real words and non-words are taken into account.

2. Future research

Although the present study reveals a highly significant difference between the acceptability of real words and non-words, the difference between malformed syllables and other non-word gap types is less clear. As such, we suggest that future studies focus on a subset of the present data such that the effect of phonotactic knowledge on acceptability can be more thoroughly explored.

Due to the large number of items to be tested, participants in the present study were only exposed to each item once, so it was not possible to investigate within-participant variation in the statistical modelling. We propose that future studies of this nature target a subset of items to allow collection of more trials (repetitions) from participants, to confirm the results of the present study.

It has also been suggested that the use of a sliding scale might introduce an artefact. Although the sliding scale allows for greater freedom in participants' judgements, values are much less memorable, making comparisons between scores more difficult. A Lickert scale, while more restrictive, might allow participants to more accurately recall previous judgements and thus compare judgements against each other.

Finally, while lexical tone does not appear to impact the acceptability of a given non-word syllable, excised neutral tone syllables taken out of context are less acceptable and less natural. Subsequent research of this type should therefore present any neutral tone items needed for the purposes of the study within a suitable context. For example, if neutral tone syllables are, as in the present study, elicited through reduplication, higher acceptability ratings might be achieved by presenting target neutral tone syllables in context with their non-reduced base syllable.

5. Acknowledgements

We would like to thank all those who took part in this study, as well as those who helped to recruit them. We are equally grateful to our speaker who kindly attended two long recording sessions without complaint. Finally, we thank the Department of Education at the University of York for providing funding for the project.

6. References

[1] S. Gong and J. Zhang, 'Modelling Mandarin speakers' phonotactic knowledge', *Phonology*, vol. 38, no. 2, pp. 241–275, May 2021.

[2] Large, N. R., Frisch, S., & Pisoni, D. B., 'Perception of wordlikeness: Effects of segment probability and length on subjective ratings and processing of non-word sound patterns', *Research on Language Processing: Progress Report*, vol. 22, pp. 95–125, January 1998.

[3] J. P. Kirby and A. C. L. Yu, 'Lexical and phonotactic effects on wordlikeness judgments in Cantonese', in *Proceedings of the international congress of the phonetic sciences xvi (Vol. 13891392)*, 2007, pp. 1389–1392.

[4] Y. Do and R. K. Y. Lai, 'Incorporating tone in the modelling of wordlikeness judgements', *Phonology*, vol. 37, no. 4, pp. 577–615, Nov. 2020.

[5] S. Duanmu, 'Chinese syllable structure', *The Blackwell Companion to Phonology*. John Wiley & Sons, Ltd, Oxford, UK, pp. 1–24, 28-Apr-2011.

[6] P. Boersma, *Praat: doing phonetics by computer (version 5.2. 19)*. 2011.

[7] Y. R. Chao, *A grammar of spoken Chinese*. University of Calif. Press, 1968.

[8] Z. Yang *et al.*, 'Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset', *arXiv [cs.CL]*, 31-Mar-2022.

[9] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder', *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020.

[10] D. Bates, M. Mächler, B. Bolker, and S. Walker, 'Fitting Linear Mixed-Effects Models using lme4', *arXiv [stat.CO]*, 23-Jun-2014.

[11] Chinese Character Database. http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/