University of Pennsylvania Working Papers in Linguistics

Volume 23
Issue 1 *Proceedings of the 40th Annual Penn Linguistics Conference*

Article 4

1-1-2017

Worldlikeness: A Web-based Tool for Typological Psycholinguistic Research

Tsung-Ying Chen
National Chung Cheng University

James Myers
National Chung Cheng University

Worldlikeness: A Web-based Tool for Typological Psycholinguistic Research

Abstract
In this paper, we introduce Worldlikeness, a web-based tool for collecting and sharing cross-linguistic wordlikeness judgments (nonce word acceptability judgments) to facilitate typological psycholinguistic research. Typological psycholinguistic research is essential since crucial factors affecting language processing vary across languages, but these factors often too confounded to tease apart by comparing just two languages at a time. This type of research is nevertheless difficult since it requires testing many speakers from each language, using materials designed with the help of expert native speakers. Worldlikeness aims to make typological psycholinguistics more feasible, by providing tools for separate groups of experimenters to design experiments with text, audio, images, and video for individual languages, collect judgments and reaction times online, and crucially, share their data with each other for typological analysis. We show that Worldlikeness successfully replicated Mandarin wordlikeness judgments collected using traditional lab-based software, and report its first use in a cross-linguistic study collecting wordlikeness judgments from bilingual speakers of Mandarin and Southern Min.

Worldlikeness: A Web-based Tool for Typological Psycholinguistic Research

Tsung-Ying Chen and James Myers*

1 Introduction

Psycholinguistic research has played an important role in providing empirical evidence for factors substantially affecting the processing of languages. However, since language processing is largely shaped by experience (e.g., tonal processing, Xu et al. 2006) and languages differ remarkably in every aspect, the crucial processing factors also vary across languages. Most psycholinguistic studies focus on just one or two major languages, such as English and Mandarin, making it difficult to extend generalizations to other languages. As processing factors are frequently confounded within a single language and even across language pairs (see Section 2), the theoretical implications of psycholinguistic studies are ambiguous unless they can tease these factors apart to discover the truly crucial ones. Not surprisingly, perhaps, very few linguists have ever carried out typological psycholinguistic research due to the tremendous effort required to coordinate different research teams and collect data using the same experimental design. Lemhöfer et al. (2008), for example, involved six authors in their word recognition study of just three languages, and Bates et al. (2003) required 22 authors to study picture naming in just seven languages. We thus set out to develop a web-based application, called Worldlikeness, intended not only to allow linguists to design their own psycholinguistic experiments but also share their experimental data, particularly wordlikeness judgments (i.e., acceptability judgments for nonwords). The key concept is that individual groups of linguists can simply study individual languages experimentally, and then share their experimental results via Worldlikeness, allowing researchers interested in typological generalizations to download these data for cross-linguistic analyses.

Web-based tools for running psychological and psycholinguistic experiments have become more prevalent over the last decade, since researchers can easily crowdsource data from a large number of speakers (e.g., tatool: von Bastian et al. 2013, turktools: Erlewin and Kotek 2016, YourMor-als.org: Graham et al. 2011, WebExp: Keller et al. 2009; Amazon Mechanical Turk: Paolacci et al. 2010). Crucially, researchers have demonstrated that responses elicited using web-based tools are as reliable as those collected in lab settings (e.g., Goslin et al. 2004). Worldlikeness is another addition to this growing body with a specific focus on phonological productivity, and differs from the existing applications by highlighting the data-sharing function and reducing the effort in setting up simple linguistic judgment experiments. To further persuade researchers to study typological psycholinguistics and help them achieve this goal more easily with Worldlikeness, we first guide readers through the main system of Worldlikeness and illustrate its unique features. Next we seek to justify the reliability of data collected using Worldlikeness by replicating results from previous monolingual phonological processing research. Finally, we report a new small-scale typological phonological judgment study made possible with Worldlikeness.

This paper is organized as follows. Section 2 gives a brief summary of recent discoveries in the field of typological psycholinguistics. Section 3 provides a simplified version of the Worldlikeness user manual and reports a wordlikeness judgment experiment that replicates findings in Myers (2015). The first cross-linguistic study of wordlikeness judgments in Worldlikeness and its results are discussed in Section 4.

2 Typological Psycholinguistics

Language users adapt themselves to their individual linguistic systems by developing diverse strategies with different trade-offs that can be observed in their language processing performance (e.g.,

*This work was funded by the Ministry of Science and Technology (Taiwan) research grant MOST-103-2410-H-194-119-MY3. We are grateful to Yi-Hsin Wu, Pei-Shan Chen, Kuei-Yeh Chen, Mei-Chun Liu, and Yu-Chu Chang of the Language Processing Lab at National Chung Cheng University for their assistance in preparing and running the experiments in this paper. We also thank Hsinhsien Li for sharing his Taiwan Sign Language materials with us.

Vannest et al. 2002). Nevertheless, just as the similarities and differences across grammars can be understood with the help of typological linguistics, it should also be possible to make sense of cross-linguistic processing differences with the help of typological psycholinguistics. For practical reasons, however, typological psycholinguistics studies have been restricted to very few languages, far too few to apply the quantitative methods commonly used in typological linguistics (see Cysouw 2005) to discover universal tendencies and systematic correlations between typological features.

For example, Lemhöfer et al. (2008) studied the effect of first-language (L1) French, German, and Dutch on word recognition of second-language (L2) English. One concern of this study was the effect of orthographic transparency on how words in these languages are encoded and activated lexically (Katz and Frost 1992). They found that the semantic abstractness of L2 target words played a more important role for L1 French speakers than for German and Dutch speakers, consistent with the fact that the orthographic systems of German and Dutch are more phonologically transparent than that of French. Yet since only three L1s were tested, it is impossible to confirm that there was truly a systematic, causal relationship between orthography and the semantic effect; perhaps the French speakers differed from the German and Dutch speakers in some other way.

Another example is Bates et al. (2003), where the same pictures were named by separate speakers of seven languages (Bulgarian, Chinese, English, German, Hungarian, Italian, Spanish). The languages tested in Bates et al. (2003) were more qualitatively and quantitatively different from those in Lemhöfer et al. (2008), but the most robust findings of this study were that word frequency and goodness of depiction dominated word naming times across all languages, relating more to general cognitive processing than to language per se. There were processing differences across the languages that the authors argued were related to structural linguistic differences (e.g., the special role played by compound words in Chinese), but as typological linguists know, a set of merely seven languages (some of which are closely related) is not large or diverse enough to determine which language processing strategies truly derive from experience with particular linguistic features.

The Worldlikeness web app attempts to deal with these limitations in a particular subdomain of psycholinguistic and grammatical research: phonological wordlikeness (acceptability) judgments of nonwords. While large-scale comparisons of wordlikeness judgments across languages have yet to be made, previous research on individual languages has revealed crucial differences. For instance, in Arabic (Frisch and Zawaydeh 2001) and English (Bailey and Hahn 2001), the acceptability of nonwords is more strongly affected by their phonotactic probability (the chance of sound sequences in real words) than neighborhood density (the number of minimally different lexical words). In contrast, neighborhood density is a more decisive factor for Cantonese (Kirby & Yu 2007) and Mandarin (Myers & Tsay 2012) judges. One possible explanation would be the quantitative difference in the number of syllable types. Compared with English and Arabic, Cantonese and Mandarin have fewer syllable types (e.g., only slightly more than 1,300 unique lexical syllable-tone combinations in Mandarin), and their speakers may simply memorize whole syllables as processing units, as opposed to decomposing syllable into individual segments. As a result, Cantonese and Mandarin speakers could be less sensitive to the probabilistic distribution of specific sound sequences. This variable is nevertheless confounded with a number of other differences, including orthography, where Cantonese and Mandarin have syllable-based characters and English and Arabic have phoneme-based alphabetic writing systems. This again demonstrates the necessity of large-scale and systematic cross-linguistic study to clarify the role of individual and universal variables in wordlikeness judgments.

3 Worldlikeness

Worldlikeness is a free, open-source web application developed with Meteor® (http://www.meteor.com), a Javascript-based programming language integrated with the server-side database package MongoDB® (http://www.mongodb.com). People can thus use Worldlikeness simply with modern web browsers via the internet without the need to install any additional software. The web application is currently open for tests and hosted at http://www.worldlikeness.org (Figure 1). Worldlikeness is optimized for (but not limited to) simple phonological acceptability judgment tasks and aims at creating a typological research community. This section outlines the various user roles defined by the system, explains how experimenters can use it design experiments, and provides evidence regarding the reliability of its data collection.

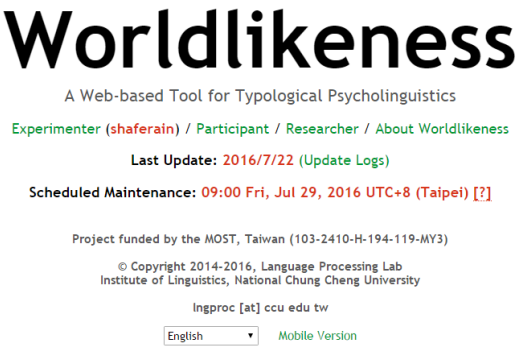

Figure 1. The home page of Worldlikeness.

3.1 User Roles

'Experimenters' are users who would like to design and run their experiments using Worldlikeness, or share the results of their previous psycholinguistic studies via Worldlikeness. The experimenter role is the only user role in Worldlikeness that requires setting up an account with an e-mail address to be assigned a quota for creating experiments and uploading experimental results. Worldlikeness allows experimenters to fully control whether to share their experimental data publicly on Worldlikeness, but in order to encourage data sharing for large-scale typological psycholinguistic studies, the experiment quota increases whenever experimenters set to share their experimental results. Worldlikeness also allows experimenters to add other accounts for collaborators on any specific experiment for easier management and private data sharing within a research team. Experimenters own their data in Worldlikeness (as stated in the user agreement), and they can remove their data or even their experimenter account from Worldlikeness at any time.

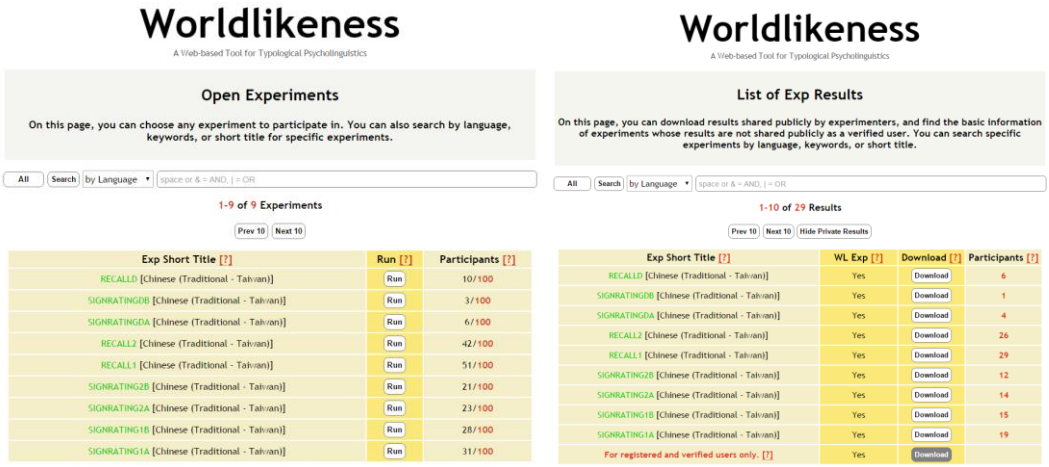

Figure 2. Public experiment page for participants (left) and experimental result download page for researchers (right).

'Participants' are users who participate in public experiments on Worldlikeness (Figure 2), are invited to private experiments via a direct link to these experiments, or participate in traditional lab-based experiments run using Worldlikeness. The consent form set up by experimenters appears before experimental sessions for participants to acknowledge their rights and give their informed consent. In addition, participants can decide whether to authorize access to their experimental results to this experimenter only, to all registered experimenters, or to all visitors to the web app even if not registered. Participants will remain fully anonymous in any context, and only non-identifying personal information is collected during an experimental session. Participants do not need to register or even provide their real names; Worldlikeness records participant IP addresses (automatically sent

by most browsers) to help prevent participants from re-running the same experiment, but this information is not made public or available to experimenters. Before an experimental session ends, participants are able to interrupt the process without leaving any experimental data on Worldlikeness by simply closing the web browser window or tab. At the end of an experiment, participants are rewarded with a result report including their own performance data and a comparison with the rest of the participants in the experiment (see 3.2).

'Researchers' are users who take advantage of the experimental data shared through Worldlikeness to conduct their own typological analyses. Anonymous researchers, without registering, can freely download a subset of the data that experimenters have agreed to share publicly (Figure 2), but these data sets will not contain the results from participants who have not authorized access to non-registered users. Registered researchers have access to fuller data sets, and are also allowed to see the background information for experiments that have not been shared publicly, making it possible for them to contact their experimenters for the data.

3.2 Experimental Design

Experimental design in Worldlikeness incorporates common features of judgment task paradigms with additional flexibility accommodating the web-crowdsourcing environment. Experimenters can choose between binary good/bad or seven-point Likert-scale judgment scales, and customize eye fixation/stimuli/trial length and choice of response keys/touchscreen/mouse clicks. Reaction times in milliseconds (measured using the clock in the participant's device) and response choices are recorded automatically for each experimental trial. Worldlikeness encourages regression-based experiment designs as opposed to factorial designs since lexical variables that affects word-likeness judgment are usually gradient and too quasi-correlated to cross effectively (Baayen 2010). Therefore, experimenters are encouraged to use random samples of nonwords as stimuli, though they may also tag stimuli with lexical variables. An ideal typological study of wordlikeness judgments would be to have speakers of different languages judge the acceptability of the same set of items that are nonwords in all of the languages. Worldlikeness will thus eventually incorporate an algorithm that can generate a universal set of nonwords from dictionary files provided by experimenters, though this has yet to be implemented.

One can choose to run experiments through Worldlikeness in a traditional lab setting, but Worldlikeness is particularly intended to crowdsource data from participant pools via the internet (though this has not yet been tested). In the latter case, Worldlikeness also helps experimenters deal with the additional variability arising from the use of different devices by different participants, as well as participant screening issues. Thus the size of text or image stimuli can be fixed or proportional to the actual screen resolution of individual devices for a more consistent visual experience, and experimenters can set up a forced-choice language proficiency test.

Worldlikeness is capable of displaying various types of stimuli via modern web browsers, including text, sounds, images, and videos (Figure 3), for experimenters to investigate the relation between modality and phonological processing (e.g., sign vs. oral language wordlikeness judgment). Experimenters can also design cross-modal experiments with Worldlikeness to study the role of visual-auditory interactions. Multimedia files are always preloaded to a participant's device upon their agreement to avoid issues caused by unstable internet connection during an experiment.

Figure 3. Running a binary judgment experiment with textual and signed (Li 2016) stimuli.

Participants are rewarded with an experimental result report for them to understand their own performance. This report includes the items with the highest and lowest acceptability scores from the current participant as compared with other participants, the average acceptability score and reaction time of the individual and the group, and the overall correlation (r^2) between the individual and the group's by-item judgment performance (Figure 4). Experimenters can customize the text of the report to explain the meaning of scores to participants and set correlation score criteria corresponding to customized comments on participant performance. Worldlikeness automatically blocks users from participating in same experiment again, even if the users quit partway through.

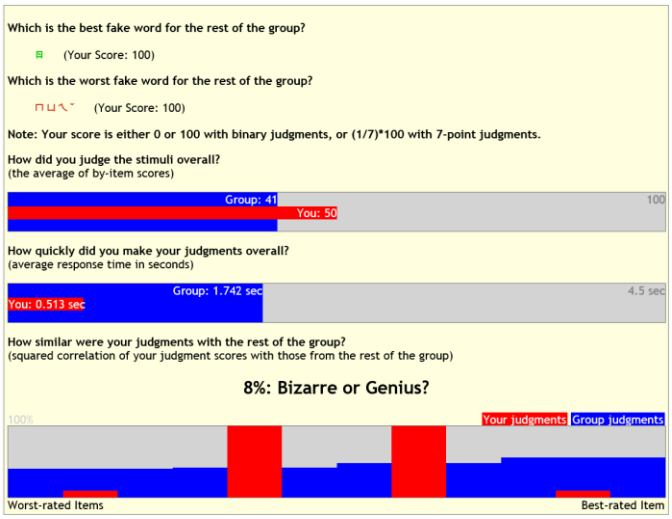

Figure 4. Sample result report given to participants at the end of an experiment.

3.3 Data Reliability: Replicating Mandarin Wordlikeness Megastudy

So far, we have tried to convince the reader that Worldlikeness is a powerful and ethical online research tool. This section attempts to justify its accuracy and precision of the judgment data and reaction times that it collects, comparable to proprietary experiment software. To this end, we seek to replicate the distribution of acceptability scores and reaction times for a large Mandarin phonological wordlikeness judgment experiment included in the Mandarin Wordlikeness Project database (http://lngproc.ccu.edu.tw/MWP/; Myers 2015). The materials, results, and other back-ground information of this replication experiment are available for download at the links labeled '中文假字實驗(一)' (Chinese Fake Word Experiment I) and '中文假字實驗(二)' (Chinese Fake Word Experiment II) in the "Researcher" section on Worldlikeness.

Myers (2015) analyzed wordlikeness judgments from more than 100 Taiwan Mandarin native speakers for more than 3,000 segmental and tonal combinations that happen to be nonlexical in Mandarin, for practical reasons divided into two subsets of roughly 1,500 items each, and written in Zhuyin Fuhao, a spelling system used in Taiwan, The experiment was run using E-Prime (Schneider et al. 2002), with each participant tested in two separate sessions with a time interval ranged from hours to days, to split up their workload. Participants were told that they would see a series of nonlexical words, and that they had to judge whether each non-lexical word is Mandarin-like or not as quickly as possible. The judgments were binary, with participants pressing either 'S' on the keyboard for 'unlike Mandarin' responses or 'L' for 'like Mandarin' responses. A major finding, consistent with the conclusion in previous Chinese wordlikeness research, was that non-words with a higher neighborhood density were more acceptable than those with fewer lexical neighbors. We aimed at replicating this basic pattern in a wordlikeness judgment task, created and run in Worldlikeness, using just a small subset of the materials.

Two binary Mandarin wordlikeness judgment experiments were created in Worldlikeness using experimental settings similar to those reported in Myers (2015). Both experiments were run in the Language Processing Lab at National Chung Cheng University. Two subsets of 100 nonlexical monosyllables were randomly selected from each of the two complete stimulus sets used in Myers (2015).

Another four novel monosyllables were selected as stimuli for the practice session. Because Worldlikeness is not designed with factorial designs in mind, the two stimuli sets were run as separate sub-experiments (hence the separate links for data download noted above).

20 participants enrolled as undergraduate or graduate students at National Chung Cheng University in southern Taiwan were recruited, ten per stimulus set. Among the 20 participants, 15 were females and five were males, and their age ranged from 18 to 25 (mean = 20, sd = 1.7).

The experiment was run in a traditional lab setting, with Worldlikeness as the experimental control program. Each participant was randomly assigned to receive one of the two stimulus sets, and the order of stimuli was randomized for each participant (this is the obligatory setting for trial order in Worldlikeness). Participants were told to judge whether a novel Mandarin word presented in Zhuyin Fuhao on the computer screen was Mandarin-like or not as quickly as possible. As in the original study, the 'unlike' (不像) response key was set to 'S' on the left side of the keyboard and the 'like' (像) key to 'L' on the right side of the keyboard, and the trial length was set to four seconds with a one-second inter-stimulus interval showing a '+' sign at the center of the screen (all of these parameters may be adjusted in Worldlikeness). The size of each individual character of the stimuli was fixed at 76 pixels on the computer screen under the resolution of 1280 × 1024, and each stimulus was aligned to the center of the screen both vertically and horizontally (this is the only available display setting for written stimuli in Worldlikeness). Two on-screen buttons corresponding to each response options appeared right below each stimulus as reminders, but the mouse cursor was hidden to force participants to respond with key pressing (this parameter is adjustable in Worldlikeness). A thin green progress bar appeared at the bottom of the browser to show participants how far along in the experiment they were (this is an obligatory setting). Before the experimental session began, the browser's viewport was maximized and turned into the full-screen mode manually to remove potentially distracting elements. The same practice session was administered to all participants before the formal experimental session for them to become more familiar with the procedure. Each participant was paid NT$50 after the experiment.

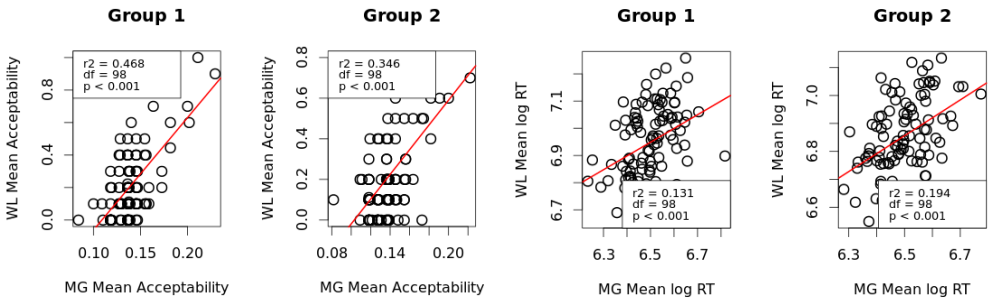

Figure 5. By-item comparisons of mean acceptability scores and reaction times of each stimulus subset in Myers's (2015) megastudy (MG) versus the Worldlikeness replication (WL).

The by-item comparison of mean acceptability scores and reaction times between Myers's (2015) megastudy and our replication is illustrated in Figure 5. Regardless of the stimulus subsets (called 'Groups' in the figure), the mean acceptability scores ($r^2 > .34$, $p < .05$) and reaction times ($r^2 > .13$, $p < .05$) in our replication results are correlated with those of the same stimulus items tested in the original study, suggesting the reliability of Worldlikeness's data. It is worth noting that the mean acceptability scores seem much lower in the original study than in our experiments (as can be seen by comparing the x- and y-axis maximums). A possible cause might be the 'satiation effect' (Snyder 2000): participants in the large-scale study had to spend hours judging more than 1,500 items in each of the two sessions, and thus their judgments had gradually become more monotonic over the long time span. Nevertheless, in linear mixed-effects logistic regression analysis using judgments as the dependent variable and neighborhood density as the sole fixed independent variable, neighborhood density effects were significant as in the megastudy for both subgroups ($\beta_{\text{Subset-1}} = 0.49$ and $\beta_{\text{Subset-2}} = 1.19$, all $p < .001$; i.e., more neighbors = higher acceptability).

4 Lexical and Social Factors in Cross-linguistic Wordlikeness Judgment

However, the unique strength is not to be yet another online web experiment tool, but rather to make multi-language typological psycholinguistics feasible. As a small step in this direction, we decided to run a two-language study using test items that were nonlexical in both languages. Mandarin and Taiwan Southern Min (commonly called Taiwanese) were selected as the two target languages for a number of reasons. First, while the two languages have been constantly categorized as siblings in the same Chinese language family, they still differ remarkably and are thus not mutually intelligible. For example, there are fewer coda consonants, lexical tones, and lexical monosyllables in Mandarin than in Southern Min. Moreover, the logographic writing system has a long history in the development of modern Mandarin, whereas the education system in Taiwan did not introduce any official writing system for Southern Min until the 1990s, and it is still not widely known, let alone used, by Southern Min speakers. Such differences may shed light on the role of individual linguistic variables in phonological processing, although as reviewed earlier in this paper, two languages are far from sufficient to truly disentangle these variables. A second reason for choosing these two languages is that many speakers in Taiwan are bilinguals in them, raising cross-lexical yet speaker-internal issues of the sort also studied by Lemhöfer et al. (2008). Third, the two languages also have different social statuses in Taiwan, with Mandarin being more prestigious than Southern Min. We thus expect to see effects of social variables on language processing consistent with previous research (e.g., female speakers may tend to favor the prestige norm; Labov 2001), and possibly interactions between the social and lexical/cognitive/phonological variables as well. All experimental design information, as well as the results approved for sharing by a subset of the participants, are publicly available at the four links labeled 閩南語聽覺似詞判斷(PS) (Min wordlikeness judgment, PS), 閩南語聽覺似詞判斷(KY) (Min wordlikeness judgment, KY), 中文聽覺似詞判斷(PS) (Mandarin wordlikeness judgment, PS), and 中文聽覺似詞判斷(KY) (Mandarin wordlikeness judgment, KY) in the 'Researcher' section of Worldlikeness. For further information on this experiment, see Myers and Chen (2016).

We again adopted the binary wordlikeness judgment scale (i.e., like vs. unlike) in this experiment but used auditory rather than written stimuli to enable the same stimuli to be judged as either Mandarin-like or Southern-Min-like (as noted above, speakers of the latter are generally unfamiliar with any writing system for this language). Because the stimuli were spoken rather than simulated, we decided to tag them for potential language accent (Mandarin-accented vs. Min-accented, as defined by the talker who recorded the stimuli). There were also two different target language (Mandarin vs. Southern Min), depending on the participant instructions, making it a 2 × 2 design involving four separate Worldlikeness sub-experiments, as reflected in the links noted above.

We used all onset and rhyme types plus lexical tones in Mandarin and Southern Min to generate all logical combinations following the basic syllable template C(G)V(G)C, shared by both languages, from which lexical syllables in Mandarin and Southern Min were excluded. From these more than 5,000 non-lexical syllables, we further excluded those with a mid-level tone for its possible perceptual confusion with a high or low level/dipping tone on a lexical syllable. Syllables with an obstruent coda were also left out since they are too clear a violation of Mandarin phonotactics, and thus would be expected to have little chance to be judged as Mandarin-like. Among the remaining nonlexical syllables, we randomly selected 200 items to use. We presented the stimuli in IPA to two female lab assistants whose home language is Mandarin (speaker KY) and Southern Min (speaker PS) respectively, and asked them to read each stimulus aloud. There was no specific instruction unless the speakers had any difficulties producing the syllables naturally, in which case the first author demonstrated the pronunciation of the syllables. Recordings were made in a sound-attenuated room using Praat with a sampling rate at 44,100 Hz.

Since we used auditory stimuli, we set up a pre-test to find out if bilingual speakers of Mandarin and Southern Min misperceived any of the nonlexical syllables as real words in either language. Because the pretest was not designed for sharing or testing cross-linguistic hypotheses, it was designed and run using PsychoPy (Peirce 2007). 12 bilingual listeners were randomly assigned to one of the four subgroups with the same Accent × Target design. In each trial, one of the 200 non-lexical syllables were randomly selected and played once. The listeners had to press 'S' on the keyboard for nonlexical items and 'L' for real words, without any time pressure. If they pressed 'L', a text

box appeared for them to input the perceived word. Items that were judged as words in either Mandarin or Southern Min by more than one listener were excluded from both of the Mandarin-accented (KY) and Min-accented (PS) stimulus lists. This screening process left the same 129 non-lexical syllables from both speakers.

80 bilingual speakers of Mandarin and Southern Min (38 males and 42 females) enrolled as undergraduate or graduate students at National Chung Cheng University were recruited and divided into four groups of 20 participants for each condition. The age of the participants ranged from 18 to 32 (mean = 21.9, sd = 2.6).

Participants were first asked to read a paragraph in Southern Min to establish their bilingual language competence. Each qualified participant was randomly assigned to one of the four conditions (i.e., Worldlikeness sub-experiments) representing each accent and target language combination. Participants were told that they would hear a series of monosyllables on headphones, which were not real words in the target language, and they would have to judge whether they sounded like a word in the target language as quickly as possible. At the beginning of each trial, the eye fixation symbol '+' with a font size of 76 pixels appeared on the computer screen for one second before the onset of an auditory stimulus. The auditory stimulus was then played once, and participants had to make their judgment within four seconds by pressing 'S' for 'unlike' and 'L' for 'like'. Reaction times were measured from trial onset (rather than stimulus offset) to response. Prior to the formal session, a practice session including four nonwords not used in the formal session was administered to help familiarize participants with the foregoing experimental procedure. These experiments were run in a sound-attenuated room of the Language Processing Lab at National Chung Cheng University, and each participant was paid NT$50 after completing an experiment.

Our discussion of the experiment results focuses on the effect of Mandarin and Southern Min neighborhood density, accent, target language, and gender on wordlikeness judgments. Phonological neighbors of a nonwords were defined as lexical words that only differed from the stimulus item in one segment, ignoring tone. Binary wordlikeness judgments were analyzed in a linear mixed-effects logistic regression model with the above five predictors and their interactions.

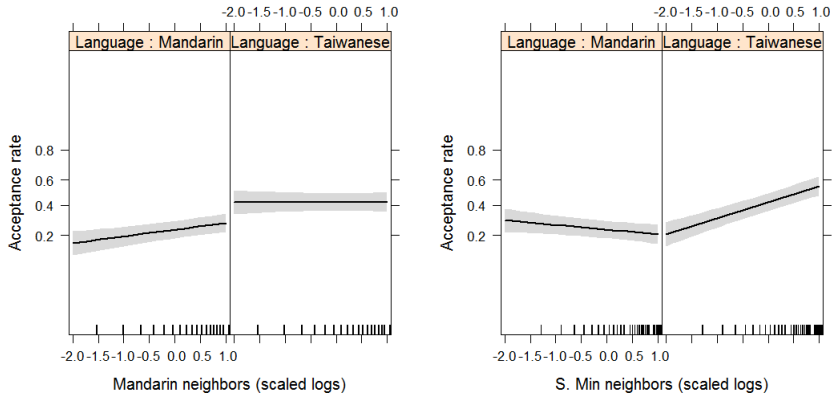

Figure 6. Target Language × Mandarin (left) or Southern Min (right) Neighborhood Density.

Figure 6 illustrates the significant interaction between the target language and the neighborhood density in either Mandarin (β = 0.11, p < .01) or Southern Min (Taiwanese; β = -0.33, p < .001). Crucially, a greater number of Mandarin neighbors predicted higher Mandarin wordlikeness (without affecting Southern Min judgments), but while a greater number of Southern Min neighbors predicted higher Southern Min wordlikeness, it lowered Mandarin wordlikeness. These results suggest not just bilingual lexical activation (i.e., Southern Min neighbors affect Mandarin wordlikeness judgments) but also an asymmetry in attitudes toward the target languages (i.e., a negative effect of neighbors from the less prestigious language on wordlikeness judgments for the more prestigious language).

The different attitudes toward the target languages were also reflected in the three-way interaction with gender (β = -0.09, p < .001) illustrated in Figure 7, which indicated that female speakers showed a significantly stronger negative effect of Southern Min neighbors on Mandarin wordlikeness, in line with the common finding that females favor prestige norms (Labov 2001).

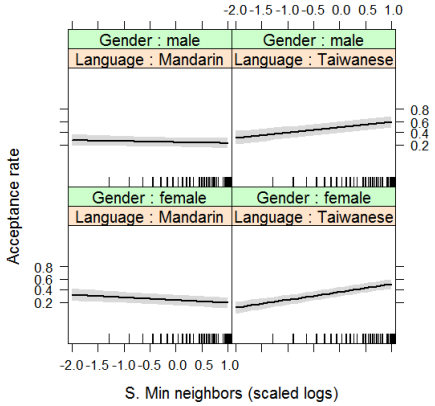

Figure 7. Southern Min Neighborhood Density × Target Language × Gender.

Finally, as shown in Figure 8, while Southern Min neighbors generally facilitated Southern Min wordlikeness judgments, the effect was stronger (β = -0.07, p < .01) if the auditory stimuli were produced by the speaker with a Southern Min accent (i.e., speaker PS, see Materials). One possible explanation would be that lexical words are encoded with fine phonetic details (e.g., accent) and that phonological neighborhood is evaluated via acoustic similarities rather than the abstract segmental or featural distance between individual tokens. If this is the case, however, we might ask why Mandarin-accented stimuli did not elicit a stronger effect of Mandarin neighbors on Mandarin wordlikeness as well. Either the Mandarin-accented speaker (KY) does not have the typical Mandarin accent required to strongly activate Mandarin neighbors, or there could be some possible qualitative between-language differences in the representation of lexical words, which awaits further investigation.

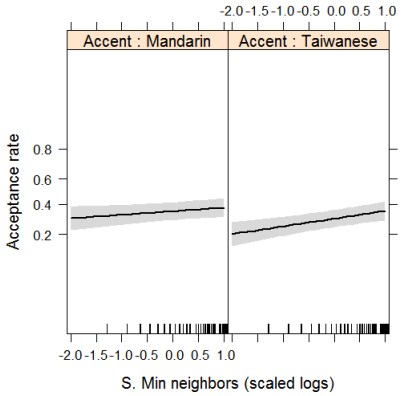

Figure 8. Southern Min Neighborhood Density × Accent in Southern Min wordlikeness.

5 Conclusion

This paper aimed to highlight the importance of typological psycholinguistic research and introduce a web app, Worldlikeness, intended to remove practical barriers to cross-linguistic experimental studies (wordlikeness judgments in particular). Hopefully, our attempt will inspire other researchers to follow the same track, whether by joining the Worldlikeness community, or adopting our open-source code for their like-minded projects.

References

Baayen, R. Harald. 2010. A real experiment is a factorial experiment? *The Mental Lexicon* 5:149–157.
Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language* 44:568–591.
Bates, Elizabeth, Simona D'Amico, Thomas Jacobsen, Anna Székely, Elena Andonova, Antonella Devescovi, Dan Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, Nicole Wicha, Kara Federmeier, Irini Gerdjikova, Gabriel Gutierrez, Daisy Hung, Jeanne Hsu, Gowri Iyer, Katherine Kohnert, Teodora Mehotcheva, Araceli Orozco-Figueroa, Angela Tzeng, and Ovid Tzeng. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin and Review* 10:344–380.
Cysouw, Michael A. 2005. Quntitative methods in typology. In *Quantitative Linguistik: Ein internationals Handbuch* [Quantitative Linguistics: An international handbook], eds. R. Kohler, G. Altmann, and R. G. Pitrowski, 554–578. Berlin: Walter de Gruyter.
Erlewine, Michael Yoshitaka, and Hadas Kotek. 2016. A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory* 34:481–495.
Frisch, Stefan A., and Bushra Adnan Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77:91–106.
Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spanssena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101:366–385.
Goslin, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver John. 2004. Should we try Web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist* 59:93–104.
Katz, Leonard, and Ram Frost. 1992. The reading process is different for different orthographies: The orthographic depth hypothesis. *Advances in Psychology* 94:67–84.
Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. 2009. Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods* 41:1–12.
Kirby, James P., and Alan C. L. Yu. 2007. Lexical and phonotactic effects on wordlikeness judgments in Cantonese. In *Proceedings of the International Congress of the Phonetic Science XVI*, eds. J. Trouvain and W. J. Barry, 1389–1392.
Labov, William. 2001. *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford, UK: Blackwell.
Lemhöfer, Kristin, Ton Dijkstra, Herbert Schriefers, R. Harald Baayen, Jonathan Grainger, and Pienie Zwitserlood. 2008. Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34:12–31.
Li, Hsinhsien. 2016. A comparative study of the phonology of Taiwan Sign Language and Signed Chinese. Doctoral Dissertation, National Chung Cheng University.
Myers, James. 2015. Markedness and lexical typicality in Mandarin acceptability judgments. *Language and Linguistics* 16:791–818.
Myers, James, and Jane Tsay. 2012. The interaction of markedness and experience in phonotactic judgments. Paper presented at SLE 2012: 45th Annual Meeting of the Societies of Linguistica Europaea, Stockholm, Sweden.
Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5:411–419.
Peirce, Jonathan W. 2007. PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods* 162:8–13.
Schneider, Walter, Amy Eschman, and Anthony Zuccolotto. *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Incorporated.
Vannest, Jennifer, Raymond Bertram, Juhani Järvikivi, and Jussi Niemi. 2002. Counter-intuitive cross-linguistic differences: More morphological computation in English than in Finnish. *Journal of Psycholinguistic Research* 31:83–106.
Von Bastian, Claudia C., André Locher, and Michael Ruflin. 2013. Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods* 45:108–115.
Xu, Yisheng, Jackson T. Gandour, and Alexander L. Francis. 2006. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of the Acoustical Society of America* 120:1063.

Graduate Institute of Linguistics
National Chung Cheng University
Chiayi, Taiwan 621
tsungyin@ualberta.ca
lngmyers@ccu.edu.tw