MUTUAL INTELLIGIBILITY OF CHINESE DIALECTS: PREDICTING CROSS-DIALECT WORD INTELLIGIBILITY FROM LEXICAL AND PHONOLOGICAL SIMILARITY

Chaoju Tang[1] Vincent J. van Heuven[2, 3]

[1] School of Linguistics and Literature, University of Electronic Science and Technology of China, Chengdu,
No. 2006, Xiyuan Ave, West Zone of High-tech district, 611731 Chengdu, China , P.R. China
1006946669@qq.com
[2] Phonetics Laboratory, Leiden University Centre for Linguistics,
PO Box 9515, 2300 RA Leiden, The Netherlands
[3] Dept. Applied Linguistics, University of Pannonia, Veszprém, Hungary
v.j.j.p.van.heuven@hum.leidenuniv.nl

ABSTRACT

This paper aims to predict mutual intelligibility (defined here as cross-dialectal word recognition) between 15 Chinese dialects from lexical and phonological distance measures. Distances were measured on the stimulus materials used in the experiment. Their predictive power was compared with earlier similar distance measures based on large word lists. Predictors based on just the stimulus materials used afford the better prediction. Segmental Levenshtein distance was the strongest predictor, outperforming both lexical and tonal similarity measures.

Keywords: Chinese dialects, mutual intelligibility, lexical distance, segmental distance, tone distance, word recognition, semantic categorisation task.

1. INTRODUCTION

1.1 Taxonomy of Chinese dialects

The Sino-Tibetan language family covers an enormous geographic area on the Eurasian continent. A rough estimate is that one of every four inhabitants of our planet is a native speaker of a language belonging to this family. As the name indicates the family comprises Chinese ('Sinitic') languages (or 'dialects' as Chinese linguists tend to call these languages) on the one hand, and Tibetan (Himalayan) languages on the other. In this paper we will only deal with the Sinitic branch of the language family. Chinese dialectologists agree that there is a primary split in the Sinitic dialects into a Mandarin branch and a Southern (non-Mandarin) branch, each comprising a number of (sub)groups. In this paper we will adopt the taxonomy proposed by Li [5] in his maps A1 and A2. We target 15 dialects, which are related as indicated in Table 1. A map of China showing the approximate locations where these 15 dialects are spoken is given in Figure 1.

Table 1: Taxonomy of the 15 Chinese dialects targeted.

Mandarin branch		Southern branch	
Group	**Dialect**	**Group**	**Dialect**
Zhongyuan	Xi'an	Wu	Suzhou
South-west	Chengdu		Wenzhou
	Hankou	Gan	Nanchang
Beijing	Beijing	Xiang	Changsha
Jilu	Ji'nan	Min	Fuzhou
Jin	Taiyuan		Xiamen
			Chaozhou
		Hakka	Meixian
		Yue	Guangzhou

The map shows that the Mandarin dialects are spoken in the Northern part of China, extending from North-Eastern China into the Central and Western parts of China. The Southern dialects are typically found along the South-Eastern coast.

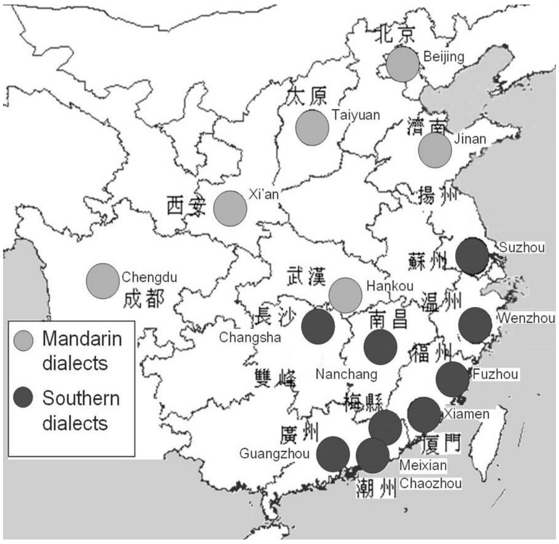

Figure 1. Approximate geographic locations of the 15 Chinese dialects targeted.

1.2 Purpose of the study

From native listeners of 15 Chinese dialects we collected judgments of linguistic similarity and intelligibility of these dialects [7, 9]. This enterprise yielded 225 combinations of speaker and listener dialects for which we reported scores for judged linguistic similarity and for judged intelligibility. We established that judged intelligibility can be predicted rather well from judged linguistic similarity (and vice versa) with $r = 0.888$. Next, in [8, 9], we collected functional intelligibility scores for the same set of 225 combinations of speaker and listener dialects, using separate tests to target intelligibility at the isolated-word and at the sentence level. We then established, first of all, that these two functional intelligibility measures converged with $r = 0.928$; Second, we wanted to know the extent to which functional intelligibility (the 'real thing') in the more recent papers could be predicted from the 'quick and dirty' judgment tests of our earlier work. If near-perfect prediction were possible, we would not have to apply cumbersome functional tests in the future, but might rely on the more convenient judgment tests. The results revealed high correlation between the functional word and sentence intelligibility scores and the intelligibility judgment scores ($r = 0.772$ and 0.818, respectively) but not high enough to advocate the unqualified use of judgment testing as a more efficient substitute for functional testing.

In the present paper we address just one part of the functional intelligibility data, which is the cross-dialect recognition of isolated words. We will describe a functional word intelligibility test which is quick and easy to administer. We will then present the mutual intelligibility data at the word level. These results reflect the taxonomy in table 1 reasonably well, showing that words are easier to recognize in a non-native dialect as the dialect is genealogically closer to the native dialect of the listener. In the second part of the paper we will present linguistic distance measures computed on the 15 target dialects. Some of these measures were copied from existing literature, others we computed ourselves on a variety of sources. In our earlier studies, we only had at our disposal language resources (digital dictionaries, sound and tone inventories, frequency counts) based on large word lists. In the present paper we computed linguistic distance measures on the stimulus materials actually used in our word recognition experiment. The crucial question we aim to answer in this study is: which type of linguistic distance measures provides better prediction of cross-dialect word recognition: (i) overall measures collected on large corpora or specific measures computed for the stimulus materials actually used?

2. WORD RECOGNITION RESULTS

Thirty speakers (one male, one female for each of the 15 dialects) recorded 150 words (the same set of words/concepts in all 15 dialects) in their native dialects. The words were divided into ten semantic categories with 15 words in each category (eight main categories, two of which were subdivided): 1. body parts, 2A. sweet fruits/nuts, 2B. vegetables, 3A. four-legged animals, 3B. other animals, 4. textile fabrics/articles of clothing, apparel, 5. orientation in time/space, 6. natural phenomena, 7. perishables (food/drinks other than fruits and vegetables) and 8. verbs of action/things people do. Stimulus words were blocked over listeners, such that (i) each listener heard each of the 150 words only once, (ii) each of the 15 listeners in one dialect group heard each version of a word in a different dialect, so that (iii) every listener heard one-fifteenth of the materials in each of the 15 dialects.

Listeners took part in the experiment in individual sessions. They hailed from local communities, in one town or village. They filled in questionnaires indicating that they were born and raised in their local town or village and had not spent longer periods of their life outside the dialect area (for speaker-individual characteristics, see Table 4.1 in [9]). In all, 225 listeners (15 listeners for each of the 15 dialects) listened to (different) words in each of the 15 dialects and were instructed to decide to which of the ten semantic categories each word they heard belonged, and to guess when they could not recognize the word presented (for details see [10] and Ch. 4 in [9]. Correctness of the responses was established automatically. This yielded a dataset of 33,750 responses (150 words × 225 listeners). Intelligibility scores were then computed for each combination of speaker and listener dialect, yielding a 15 × 15 = 225 cell matrix, which we reproduce here as Table 2.

Mutual intelligibility was defined by Cheng [3] as the mean of the intelligibility of speaker A for listener B and of speaker B for listener A. Averaging the AB and BA intelligibility scores was applied to eliminate asymmetries. The averaging operation was performed on all pairs of contra-diagonal cells i, j and j, i in the 15 (speaker dialects) by 15 (listener dialects) = 225 cells in the score matrix we collected. We then deleted the redundant part of the matrices, keeping only the non-redundant lower triangle (without the main diagonal), and used the remaining 105 scores in the comparisons below.

Table 2. Percent correctly classified words broken down by 15 speaker dialects and 15 listener dialects. Double lines separate Mandarin from Southern dialects.

Speaker Dialect (down)	Listener dialect (across)														
	Suzhou	Wenzhou	Guangzhou	Xiamen	Fuzhou	Chaozhou	Meixian	Nanchang	Changsha	Taiyuan	Beijing	Jinan	Hankou	Chengdu	Xi'an
Suzhou	**65**	20	25	17	21	15	23	22	23	29	26	29	39	28	29
Wenzhou	23	**41**	17	19	17	17	18	21	15	24	25	25	28	18	19
Guangzhou	23	18	**55**	25	25	29	40	21	19	33	34	33	38	25	29
Xiamen	20	14	23	**39**	19	25	19	19	12	18	19	25	26	17	16
Fuzhou	17	18	17	18	**47**	14	17	16	15	22	20	23	24	20	16
Chaozhou	18	12	23	22	23	**68**	15	10	15	23	27	29	24	24	23
Meixian	31	24	35	24	23	25	**67**	31	27	43	43	43	41	37	31
Nanchang	27	26	30	25	29	22	41	**37**	29	47	51	48	57	41	42
Changsha	31	22	31	24	31	20	34	31	**48**	47	49	47	60	38	43
Taiyuan	33	30	30	29	31	21	36	36	30	**57**	59	64	55	50	48
Beijing	64	41	63	45	53	38	61	51	54	76	**83**	74	72	65	70
Jinan	40	22	31	22	36	19	39	39	31	59	61	**80**	58	51	55
Hankou	37	29	33	28	41	22	42	33	35	63	59	67	**81**	53	47
Chengdu	28	24	30	32	35	19	49	36	38	62	59	61	70	**72**	56
Xi'an	47	36	43	27	35	23	48	43	47	63	64	67	65	55	**59**

3. LINGUISTIC DISTANCE MEASURES

Two objective distance measures were copied from [3]. We call these the Phonological Correspondence Index (PCI) and the Lexical Similarity Index (LSI).[1] The PCI is a measure that expresses the complexity of the rule system that is needed to convert phonemic transcriptions (including tones) in dialect A to their cognate form in language B. The more complex the rule system, the larger is the distance between dialects A and B. Note that this is the only measure in our study that is not symmetrical: the rule set that converts A to B may be more or less complex than the set that converts forms from B to A (for details, see [3]). We transformed the PCI distance matrix to a symmetrical version as explained above. LSI was conceptually defined by [3] as the percentage of cognates shared by two dialects. This is a symmetrical measure. Obviously, the larger the percentage of shared cognates the easier it should be for a speaker of dialect A to be understood by a listener of dialect B (and vice versa). Interestingly, the correlation between the LSI and correct word cross-dialect classification was high, with r = 0.875.

Earlier, we computed a large number of objective measures on similarity between (pairs of) our Chinese dialects [9, 11]. We computed structural similarity measures based on a simple comparison of the sound and tone inventories of the 15 dialects, with and without weighing the sound units for their lexical frequency. We also determined to what extent words in all pairs of dialects are pronounced the same, separately for segmental and tonal aspects. This work was based on lists of phonetic transcriptions of 764 words (basic characters) in each of the 15 dialects made available by the Chinese Academy of Social Sciences (CASS). We will not try to summarize the results here. Suffice it to say that the predictive power of all these measures was poor (for details see [9, 11]).

It then occurred to us that better prediction of cross-dialect word recognition scores might be afforded by distance measures that were computed on precisely the stimulus materials used in the experiment. We asked Chinese dialect experts to provide a phonetic transcription of the words (segments plus diacritics, and tones) as recorded by our speakers. On the basis of these transcriptions we computed the same type of distance measures that were computed before on the dictionaries and word lists. Segmental similarity between all 105 pairs of target dialects were expressed in terms of Levenshtein Distance (LD), a string edit distance measure that yields a score between 0 (the string of symbols that transcribes word A is identical to that of word B) and 1.0 (words A and B do not share a single symbol. We used the GABMAP software [4] to compute the mean LD for all counterpart word pairs in all pairs of the 15 target dialects. The segmental distance was computed once on just the base IPA symbols and a second time on the base symbol plus diacritics, where a difference in diacritic was given half the weight of a difference between base symbols. The two measures turned out to be very highly corrected (r = 0.985) but the diacritic-based measure afforded a slightly better prediction of word recognition – so this measure will be used later. Note that LD was computed across all 150 words, whether cognate or not; non-cognates, obviously, yield large LD values. Next, lexical similarity (LS) was established by comparing the Chinese characters with which the two equivalent words in a pair of dialects are written. When the characters are the same, the words are cognate, i.e. historically related, and share (some of) their phonology. Cognateship was set at 0 if the equivalents shared no characters, at 1 if all characters were shared, and at 0.5 in all other cases. Finally, we computed tonal distance between pairs of equivalent words. As is customary in Chinese tonology [1, 2], tones had been transcribed

[1] For the analysis of the lexical similarity index the selection had to be narrowed down further to 13 as no data on Hankou and Taiyuan were included in the *Cihui* word list.

as sequences of one, two or three digits for each syllable in a word, one digit for each mora (tonal time slot in a syllable). Each mora could assume a pitch value on a scale between 1 (lowest) to 5 (highest). Levenshtein distances were computed on the three-digit tone strings for first syllables, second syllables and third syllables separately on all word 150 pairs for each of the 105 pairs of dialects. Then, following [12], we converted the 3-digit tone representation to onset-plus-shape sequences, with three possible onset levels: H(igh) = {4, 5}, M(id) = 3, L(ow) = {1, 2}. The remainder of the 3-digit string (if present) was coded as either E(qual) = no change, R(ise), F(all), P(eaking) = rise+fall or D(ipping) = fall+rise. Levenshtein distances were then computed on the onset+shape letter pairs. Finally, we computed a geometric tone distance measure [6] on the 3-digit strings for first, second and third syllables separately, between all pairs of equivalents in the 15 target dialects. Geometric tone distance is used in musicology to compute the auditory similarity between two melodies. After length normalisation and alignment, the measure computes the mean squared difference between the pitches (on the scale from 1 to 5) in the 3-digit tone representation for each syllable in an equivalent word pair. The tone distance is then scaled between 0 (complete similarity) and 1 (no similarity at all).

4. REGRESSION ANALYSES

Our earlier study [9] has shown that cross-dialect word recognition in our target dialects can be predicted from linguistic distance measures collected on large dictionaries and other non-experiment-specific materials with considerable success. Cheng's [6] LSI (computed on a 2,770 item word list) basically does all the work, and leaves no room for other distance measures to contribute to the prediction of word recognition. Let us now consider the question if better predictions can be obtained by using the new measures based on the specific stimulus materials used in our experiment. Table 3 is a correlation matrix for the new distance measures and the criterion variable (word recognition scores given in table 2).

The best single predictor is the segmental LD. Lexical similarity is a highly significant but poorer predictor, and much poorer than Cheng's Lexical Similarity Index (see above). Tones predict cross-dialect word recognition only to a limited extent, where the onset+shape method is better than other measures of tonal distance. Tonal predictions work best when computed on first syllables.

Table 3. Correlation matrix for criterion (% words correct) and 11 predictors. LD = segmental Levenshtein distance, LS = Lexical Similarity (% cognates shared), OS = Onset+shape, TD = Levenshtein Tone distance, GE = Geometric tone distance.

	W. corr.	OS1	OS2	OS3	TD1	TD2	TD3	GE1	GE2	GE3	LD
OS1	**.367**										
OS2	.210	**.344**									
OS3	.119	**.301**	**.415**								
TD1	.041	**.360**	**.246**	.058							
TD2	.005	.116	**.453**	.193	**.568**						
TDl3	-.211	-.074	-.016	**.250**	.136	.169					
GE1	**.235**	**.494**	.087	.186	**.310**	**.236**	-.047				
GE2	.034	.001	-.033	.002	.168	**.478**	.096	**.542**			
GE3	.060	-.090	-.006	**.316**	-.033	.138	**.317**	.189	**.328**		
LD	**-.829**	**-.243**	-.092	-.038	-.034	-.037	.146	-.196	-.125	-.038	
LS	**.600**	**.285**	.170	.097	.085	.038	-.143	.073	.068	.042	**-.633**

r > .168: p < .05 (one-tailed); r > .235: p < .01 (one-tailed)

Finally, we regressed the new, experiment-specific distance measures against the 105 cross-dialect word recognition scores to determine the relative importance of the parameters and the degree of overall success.

Table 4. Results of multiple regression. In the stepwise method the R^2 values are cumulative. The absolute value of the beta weight indicates the relative importance of a predictor.

Simultaneous entry			Stepwise entry		
Predictors	R^2	β	Predictors	R^2	β
LD segm		−.732	LD segm	.687	−.786
O+S tone		.164	O+S tone	.716	.176
Lex Sim		.098			
All	.721				

When all predictors are entered simultaneously, 72% of the variance in the cross-dialectal word recognition scores can be predicted. Using stepwise entry, the most powerful predictor by far is the segmental Levenshtein distance, which by itself explains 69% of the variance. The onset+shape tone similarity (in the first syllables of words) makes a significant contribution, adding another 3 percent. Other parameters do not make further contributions. In [9], seven parameters were selected that could account for 88% of the variance in simultaneous entry mode and 81% in stepwise mode. However, the contribution of the best single parameters (simple correlation) was always smaller than that found in the present attempt with predictors based on the materials used in the experiment.

5. CONCLUSION

We conclude that predicting cross-dialect word recognition is better when linguistic distance measures are based on the stimulus materials used in the experiment, but poorer when all distance measures are used in multiple regression.

6. ACKNOWLEDGMENT

The first author's travel expenses for this conference were partly-financially supported by the School of Foreign Linguistics and Literature, UESTC. The previous research reported on in this study was supported by a grant from the Chinese Scholarship Council and by the Leiden University Fund (LUF). Some Chinese dialectologists contributed to transcription of the 150 words for each of 15 dialects. The first author paid them. We also thank the anonymous reviewers.

6. REFERENCES

[1] Chao, Y.-R. 1928. *Studies in the modern Wu dialects*. Beijing: Tsinghua College Research Institute (Monograph 4).

[2] Chao, Y.-R. 1930. A system of tone letters. *Le Maître Phonétique* 45, 24-27.

[3] Cheng, C. C. 1997. Measuring relationship among dialects: DOC and Related Resources. *Computational Linguistics & Chinese Language Processing* 2, 41-72.

[4] Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. & Leinonen, T. 2011. Gabmap — A Web Application for Dialectology. *Dialectologica* Special issue II, 65-89.

[5] Li, R. 1987. Chinese dialects in China. In S. A. Wurm, B. T'sou, D. Bradley, R. Li, Z. Xiong, Z. Zhang, M. Fu, J. Wang & Dob (eds.), *Language atlas of China*. Hong Kong: Longman.

[6] Mongeau, M. & Sankoff, D. 1990. Comparison of musical sequences. *Computers and the Humanities* 24, 161-175.

[7] Tang, C. & van Heuven, V. J. 2007. Mutual intelligibility and similarity of Chinese dialects, in B. Los & M. van Koppen (eds.), *Linguistics in the Netherlands*. Amsterdam: John Benjamins, 223-234.

[8] Tang, C. & van Heuven, V. J. 2008. Mutual intelligibility of Chinese dialects tested functionally. In M. van Koppen & B. Botma (eds.), *Linguistics in the Netherlands*. Amsterdam: John Benjamins, 145-156.

[9] Tang, C. 2009. *Mutual intelligibility of Chinese dialects: An experimental approach*. LOT dissertation series nr. 228. Utrecht: LOT.

[10] Tang, C. & van Heuven, V. J. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119, 709-732.

[11] Tang, C. & van Heuven, V. J. 2015. Predicting mutual intelligibility of Chinese dialects from objective linguistic distance measures. *Linguistics* 53, 285-311.

[12] Yang, C. & Castro, A. 2008. Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing* 2, 205-219.