PREDICTING MUTUAL INTELLIGIBILITY IN CHINESE DIALECTS

Tang Chaoju[*] *& Vincent J. van Heuven*[1]

Phonetics Laboratory, Leiden University (* also at Chongqing Jiaotong University)
{C.Tang, V.J.J.P.van.Heuven}@Let.LeidenUniv.NL

ABSTRACT

We determined mutual intelligibility and linguistic similarity by presenting recordings of the same fable spoken in 15 Chinese dialects to naive listeners of the same set of dialects and asking them to rate the dialects along both subjective dimensions. We then regressed the ratings against objective structural measures (lexical similarity, phonological correspondence) for the same set of dialects. Our results show that subjective similarity is better predicted than subjective mutual intelligibility and that the relationship between objective and subjective measures is logarithmic. Best predicted was log-transformed subjective similarity with $R^2 = .64$.

Keywords: Dialectology, dialectometry, linguistic distance, (mutual) intelligibility, perceptual rating.

1. INTRODUCTION

1.1. Why study mutual intelligibility?

Distance between languages is used as a criterion when arguing about genealogical relationships between languages. The more the languages resemble each other, the more likely they are derived from the same parent language, i.e., belong to the same language family. However, it is difficult to quantify the distance between languages one-dimensionally since languages differ along many structural dimensions (e.g. phonetics, phonology, morphology, syntax). It is unclear how the various dimensions should be weighed against each other. Therefore, we select a single criterion – mutual intelligibility. Mutual intelligibility is an overall criterion that may tell us whether two languages are similar/ close.

Useful work on structural measures of difference between related languages has been done, for instance, at Stanford University (for Gaelic Irish dialects, cf. [1]) and at the University of Groningen (for Dutch [2] and Norwegian dialects [3]), using the Levenshtein distance. This is a similarity metric that computes the mean number of string operations needed to convert a word in one language to its counterpart in the other language. This measure was then used to build a tree structure (through hierarchical cluster analysis) which matched the language family tree as constructed by linguists.

1.2. How to determine (mutual) intelligibility?

Although methods for determining intelligibility are well-established, for instance in the fields of speech technology and audiology, the practical problems are prohibitive when mutual intelligibility has to be established for, say, all pairs of varieties in a set of 15 dialects (yielding 225 pairs). Rather than measuring intelligibility by functional tests, opinion testing has been advanced as a shortcut. That is, the indices of the measurements of mutual intelligibility between languages are generated from listeners' judgment scores. Once mutual intelligibility scores are available, the relative predictive power of structural dimensions can be found through regression analysis. Such work has recently been done for 15 Norwegian dialects by Gooskens and Heeringa [3] (henceforth G&H). Their results show that subjectively judged distance between sample dialects and the listener's own dialect correlated substantially with the objective Levenshtein distance ($r^2 = 0.449$).

The Levenshtein distance increases rapidly when the word pairs in two languages are non-cognates. For non-cognates any sound correspondence is accidental, so that the Levenshtein distance will be close to 100. It might therefore be more informative to break the one-dimensional Levenshtein distance down into two separate parameters, i.e. (i) the percentage of cognate words shared between the vocabularies of two language varieties and (ii) the phonological distance computed for the cognate part of the vocabulary only. This is what we did in our study. We included both predictors of mutual intelligibility in order to estimate the strengths of the two predictors as well as their intercorrelation.

The work done by G&H represents a complication relative to earlier work in that their Norwegian dialects are tone languages whilst the

Gaelic Irish and Dutch dialects are not. Since it is unclear how tonal differences should be weighed in the distance measure, G&H collected distance judgments for the same reading passages resynthesized with and without pitch variations. The difference in judged distance between the pairs of versions (with and without pitch) would then be an estimate of the weight of the tonal information. Norwegian, however, is a language with a binary tone contrast. We want to test G&H's method on full-fledged tone languages, with much richer tone inventories varying from four (e.g. Beijing/Mandarin) to as many as ten (e.g. Cantonese/Yue).

Finally, it should be realized that perceived distance between some dialect and one's own is not necessarily the same as an intelligibility judgment. The third aim of our paper is to test to what extent judged distance and judged intelligibility actually measure the same property.

1.3. Earlier work

Chinese dialect classification is still controversial. Nevertheless, there is broad consensus on the primary relationships within the Sinitic languages: there is a first split between the Mandarin group (comprising the Northern, Eastern and South-western families) and the Southern group (comprising the Wu, Gan, Xiang, Min, Hakka and Yue families). Cheng [4] has computed structural similarity measures for all pairs of these Chinese dialects. We have used two of his measures (see § 2.2) as predictors of mutual intelligibility between pairs of Chinese dialects in the present study.

2. METHODS

2.1. Collecting judgments

We targeted 15 Chinese dialects (a subset from [4]), from the Mandarin group: Beijing, Chengdu, Jinan, Xi'an, Taiyuan, Hankou; from the Southern group: Suzhou, Wenzhou (Wu family), Nanchang (Gan family), Meixian (Hakka family), Xiamen, Fuzhou, Chaozhou (Min family), Changsha (Xiang family), and Guangzhou/Cantonese (Yue family).

We used existing recordings of the fable "The North Wind and the Sun". Since each fable had been read by a different speaker (11 males and 4 females), we processed the recordings (using [5]) such that all speakers sounded like males, all had roughly the same articulation rate and speech-pause ratio, and the same mean pitch.[2] Also, each reading of the fable was produced in two melodic versions, i.e., one with the original pitch intervals kept intact, and one with all pitch movements replaced by a constant pitch (monotone), which was the same as the mean pitch of the fragment with melody (and the same as all other fragments).

The 2 × 15 readings of the fable were recorded onto audio CD in one of four different random orders. The 15 monotonized versions preceded the 15 versions with melody.

For each of the 15 dialects 24 native listeners were found in the middle to older generation (ages between 40 and 60), evenly divided between males and females. All 360 listeners were born and bred in their respective dialect areas. Listeners were mono-dialectal so that they had no experience with any other Chinese dialects (though all had some familiarity with Standard Mandarin).

Each CD was played through loudspeakers to six (three female, three male) listeners per dialect. Listeners rated the materials twice: the first time they estimated on a scale from 0 to 10 how well they believed a monolingual listener of their own dialect, confronted with a speaker of the dialect in the recording for the first time in their life, would understand the other speaker. Here '0' stood for 'S/He will not understand a word of the other speaker' whilst '10' represented 'S/he will understand the other speaker perfectly'. In the second judgment the listener rated the similarity between her/his own dialect and the dialect of the speaker in the recording, where '0' meant 'No similarity at all' against '10' meaning 'This dialect is exactly the same as my own'. In all 21,600 judgments were collected and statistically analyzed.

2.2. Structural measures

We used two objective measures of structural distance between pairs of Chinese dialects. Both measures were generated by [4].

The first measure, which we call the *Lexical Similarity Index* (LSI), can be conceived of as the percentage of cognates shared between the vocabularies of two language varieties. Obviously, the higher the number (and token frequencies) of cognate words a listener encounters in a non-native dialect, the easier it will be for her/him to understand the message. We simply copied the values published in appendix 3 of [4].[3]

Cheng's second measure basically captures the regularity of the sound correspondences in the sets of cognate words shared between two dialects. Cognates between two dialects will be easier to

recognize if they contain the same sounds in the same positions in the words, or if the sounds can be converted from one dialect to the other by a simple and general rule. In [4] the counts were converted to a coefficient ranging between 0 (no phonological correspondence at all) to 1 (perfect sound correspondence). We call this measure the *Phonological Correspondence Index* (PCI). We copied the PCI values in appendix 5 of [4]).

3. RESULTS

3.1. Objective and subjective measures

We generated 15 x 15 matrices for each of the six measures for the 15 target dialects: (a) objective lexical similarity (LSI, only 13 dialects), (b) objective phonological correspondence (PCI), (c-d) subjective intelligibility judgments for stimulus versions with and without melody, and (e-f) subjective similarity judgments for versions with and without melody. From the matrices (not presented due to lack of space) hierarchical cluster trees were derived using the method of average linking.

The trees (not presented) show a rather poor congruence. Even the primary split between Mandarin and Southern dialects is not correctly reproduced in the trees. Typically, the arguably Southern dialects Changsha and/or Nanchang are incorrectly parsed with the Mandarin dialects. Generally, the degree of congruence is better between the two subjective ratings than between the objective measures. We will now first examine the relationship between the two subjective measures, and then see how well these subjective ratings can be predicted by some combination of objective similarity measures.

3.2. Predicting intelligibility from similarity

We used the proximity between the members of every single pair (N = 105) of dialects out of the set of 15 as our measure of closeness between the members. Proximity matrices are symmetrical; the redundant part of the matrices was deleted before we correlated the proximity values obtained from the intelligibility ratings and similarity ratings. The result shows that judged intelligibility correlates with judged similarity (N = 105 pairs of values) at r = .949 ($p < .001$). This means that the two sets of ratings can be predicted from each other with a very high degree of accuracy. Moreover, visual inspection of the corresponding scatterplot (not presented) reveals no specific outliers, so that the conclusion follows that subjectively estimated similarity between pairs of languages is an exceptionally good predictor of, or even a near-perfect substitute for, estimated intelligibility.

3.3. From objective to subjective measures

In Table 1 (next page) we have specified how well judged intelligibility and judged similarity can be predicted from the objectively determined LSI and PCI measures. We also computed correlation coefficients between objective and log-transformed subjective measures; these generally yield higher r-values. A separate series of computations was done on the scores after excluding Beijing (which is almost identical to Standard Mandarin) as one of the dialects. Moreover, all the computations were done once with the judgments based on the sound stimuli with full melodic information and a second time with judgments based on the monotonized versions. Finally, we list the results of selected multiple regression analyses (with LSI and PCI entered in the analysis together for only the optimal combinations of conditions) in order to determine the cumulative effect of the predictors.

4. CONCLUSIONS

A number of conclusions can be drawn from Table 1. First, the two objective measures of structural similarity, PCI and LSI, are always significantly correlated with all of the subjective ratings. Moreover, the two predictors are only moderately intercorrelated so that there is potential room for improvement of the prediction through multiple regression. The success of multiple regression is demonstrated most clearly in the prediction of log-transformed similarity for versions with melody and Beijing dialect excluded: here the accuracy of the prediction (coefficient of determination, i.e. r^2 or R^2) from both objective measures together (64%) is 7 percentage points better than that from the best single predictor (57%). It is even 19 percent than the single r^2 in G&H [3] (see § 1.2). The latter result shows that better prediction of judged similarity and intelligibility can be obtained when a one-dimensional objective phonological distance measure is broken down into two separate parameters, one covering the proportion of cognates shared between two vocabularies and the other targeting the phonological similarity in the shared cognates only – as was assumed all along by [4].

Second, similarity judgments can be predicted more successfully (higher r-values) than the corresponding mutual intelligibility judgments.

Third, the prediction of log-transformed judgments is better than of the corresponding linear measures. This effect has been found in many other studies on the relationship between objective counts on language use and the subjective impression of such phenomena, e.g. in the area of word token frequency.

Fourth, the ratings based on versions with full melodic information can be predicted substantially better from the objective measures than those based on monotonized versions. This indicates that melodic information should carry a rather heavy weight in the ultimate prediction of ratings in the Chinese language situation.

Fifth, leaving out the Beijing dialect yields clearly better predictions of judged similarity and of mutual intelligibility. It would make sense, in the Chinese language context, where almost every language user has had some basic exposure to the standard language (which is very close to the Beijing dialect), that the naive raters may appreciate the structural difference between dialects better than the mutual intelligibility.

Table 1. Correlation coefficients (*r*) and number of dialect pairs involved (*N*) between two measures of objective structural similarity and subjective intelligibility and similarity ratings. Multiple *R* is indicated for optimal conditions only (see text).

Variables and conditions	Cheng's PCI		Cheng's LSI		Both
	r	*N*	*r*	*N*	*R*
Cheng's LSI	.763**	77			
Judged intelligibility, melody	.527**	105	.423**	77	
Judged intelligibility, monotone	.482**	105	.378**	77	
Judged similarity, melody	.622**	105	.558**	77	
Judged similarity, monotone	.523**	105	.482**	77	
Log judged intelligibility, melody	.647**	105	.591**	77	.636**
Log judged intelligibility, monotone	.600**	105	.536**	77	
Log judged similarity, melody	.703**	105	.694**	77	.742**
Log judged similarity, monotone	.616**	105	.626**	77	
Judged intelligibility, melody, no Beijing	.591**	91	.576**	65	
Judged intelligibility, monotone, no Beijing	.548**	91	.537**	65	
Judged similarity, melody, no Beijing	.648**	91	.701**	65	
Judged similarity, monotone, no Beijing	.552**	91	.629**	65	
Log judged intelligibility, melody, no Beijing	.703**	91	.710**	65	.753**
Log judged intelligibility, monotone, no Beijing	.658**	91	.667**	65	
Log judged similarity, melody, no Beijing	.696**	91	.753**	65	.798**
Log judged similarity, monotone, no Beijing	.631**	91	.713**	65	

**: $p < .01$ (two-tailed)

NOTES

1. The first author acknowledges the Leiden University Fund / Van Walsem Fund for a (partial) travel grant in order to attend the 16th ICPhSc.
2. The mean pitch was normalized to the mean of the 11 male speakers. Relatively small shifts in pitch (in semitones) were performed (using the PSOLA pitch manipulation implemented in the Praat software) on the male speakers, larger shifts were required for the female voices. For the female speakers a gender transformation was carried out by decreasing the formants by 15%. Longer pauses were reduced to 500 ms, and the remaining speech was linearly speeded up or slowed down (in the same PSOLA manipulation that changed the pitch) such that the articulation rate (syll./s) was the same for all speakers (sound files on CD).
3. No LSI values are listed for Taiyuan and Hankou in [4].

5. REFERENCES

[1] Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proc. European ACL*, Dublin, 60–67.

[2] Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenshtein distances*. Groningen dissertations in linguistics nr. 46, Groningen University.

[3] Gooskens, C., Heeringa, W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16, 189–207.

[4] Cheng, C.C. 1997. Measuring Relationship among Dialects: DOC and Related Resources, *Computational Linguistics & Chinese Language Processing* 2.1, 41–72.

[5] Boersma, P., Weenink, D. 1996. *Praat, a system for doing phonetics by computer, version 3.4*, Report 132, Institute of Phonetic Sciences, University of Amsterdam.