

PREDICTING MUTUAL INTELLIGIBILITY IN CHINESE DIALECTS FROM SUBJECTIVE AND OBJECTIVE LINGUISTIC SIMILARITY

CHAOJU TANG Y VINCENT J. VAN HEUVEN
LEIDEN UNIVERSITY CENTRE FOR LINGUISTICS (THE NETHERLANDS)
CHONGQING JIAOTONG UNIVERSITY (P. R. CHINA)

Abstract: *We determined subjective mutual intelligibility and linguistic similarity by presenting recordings of the same spoken passage 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 subjective ratings against objective structural similarity measures (lexical, phonological) for the same set of dialects as published in the literature. Our results show that subjective distance 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 (excluding the Beijing dialect, which is identical to the standard language) with $R^2 = .64$.*

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

Abstract: *We determined subjective mutual intelligibility and linguistic similarity by presenting recordings of the same spoken passage 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 subjective ratings against objective structural similarity measures (lexical, phonological) for the same set of dialects as published in the literature. Our results show that subjective distance 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 (excluding the Beijing dialect, which is identical to the standard language) with $R^2 = .64$.*

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

1. Introduction

1.1. Why study mutual intelligibility?

The more two languages are like each other, the more closely they are related. Language varieties that are very close are often called dialects of the same language. In order to determine the difference between language varieties, we need to measure the ‘distance’ between them. 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).

Useful work on structural measures of difference between non-tonal languages has been done, for instance, at Stanford University (for Gaelic Irish dialects, B. KESSLER 1995) and at the University of Groningen (for Dutch and Norwegian dialects, C. GOOSKENS and W. HEERINGA 2004), using the Levenshtein distance (a similarity metric that computes the mean number of

string operations needed to convert a word in one language to its [cognate] counterpart in the other language). This objective measure was then used to build a tree structure (through hierarchical cluster analysis) which matched the language family tree as constructed by linguists.

It is unclear how various dimensions of language difference should be weighed against each other. That is, we do not know which structural correspondences are more or less important when constructing a difference/similarity measure. Obviously, the problem gets even more complex when we apply such distance measures to tonal languages. Ideally, we want to express the difference/ similarity in a single number on a one-dimensional scale rather than as a distance between points in some multi-dimensional hyperspace. Therefore, we select a single criterion — mutual intelligibility. Mutual intelligibility exists between two languages A and B when speakers of language (variety) A can readily understand speakers of language (variety) B (and vice versa) without prior exposure, intentional study or extraordinary effort. By definition, mutual intelligibility is an overall criterion that may tell us in a psychologically relevant way whether two languages are similar/close. When two language varieties are mutually intelligible, they are probably dialects of the same language; when intelligibility drops below some threshold measure, the varieties belong to different languages.

1.2. How to determine (mutual) intelligibility experimentally?

Although methods for determining intelligibility are well-established, for instance in the fields of speech technology and audiology (e.g. R. VAN BEZOOIJEN and V. VAN HEUVEN 1997), 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 a mutual intelligibility score is available, the relative importance of structural dimensions can be found through some form of statistical optimization (multiple regression techniques). Such work has recently been done for 15 Norwegian dialects by C. GOOSKENS and W. HEERINGA (2004). Their results show that subjectively judged distance between sample dialects and the listener's own dialect correlated substantially with the objective Levenshtein distance. In this manner subjective intelligibility judgments were used to validate an objective linguistic distance measure, i.e. the Levenshtein distance.

It is implied by Gooskens and Heeringa that judged distance between a stimulus dialect and the listener's own dialect is equivalent to judged intelligibility. Indeed, we know from earlier work that language listeners are able to make accurate estimations of the intelligibility of language varieties. It should be realized, however, that perceived distance between some dialect and one's own is not necessarily the same as an intelligibility judgment. One of the aims of the present paper is to test to what extent judged distance and judged intelligibility actually measure the same property.

Gooskens and Heeringa computed Levenshtein distances on the basis of only the cognate words shared between pairs of Norwegian dialects. When language varieties are less close the number of cognate words decreases, which will strongly affect mutual intelligibility. In our study we will include two predictors of mutual intelligibility, one based on the number of cognates shared between a pair of dialects and the other on the regularity phonological correspondences between the cognate words only (an alternative to the Levenshtein distance). This will allow us to estimate the relative strength of the two predictors as well as their inter-correlation.

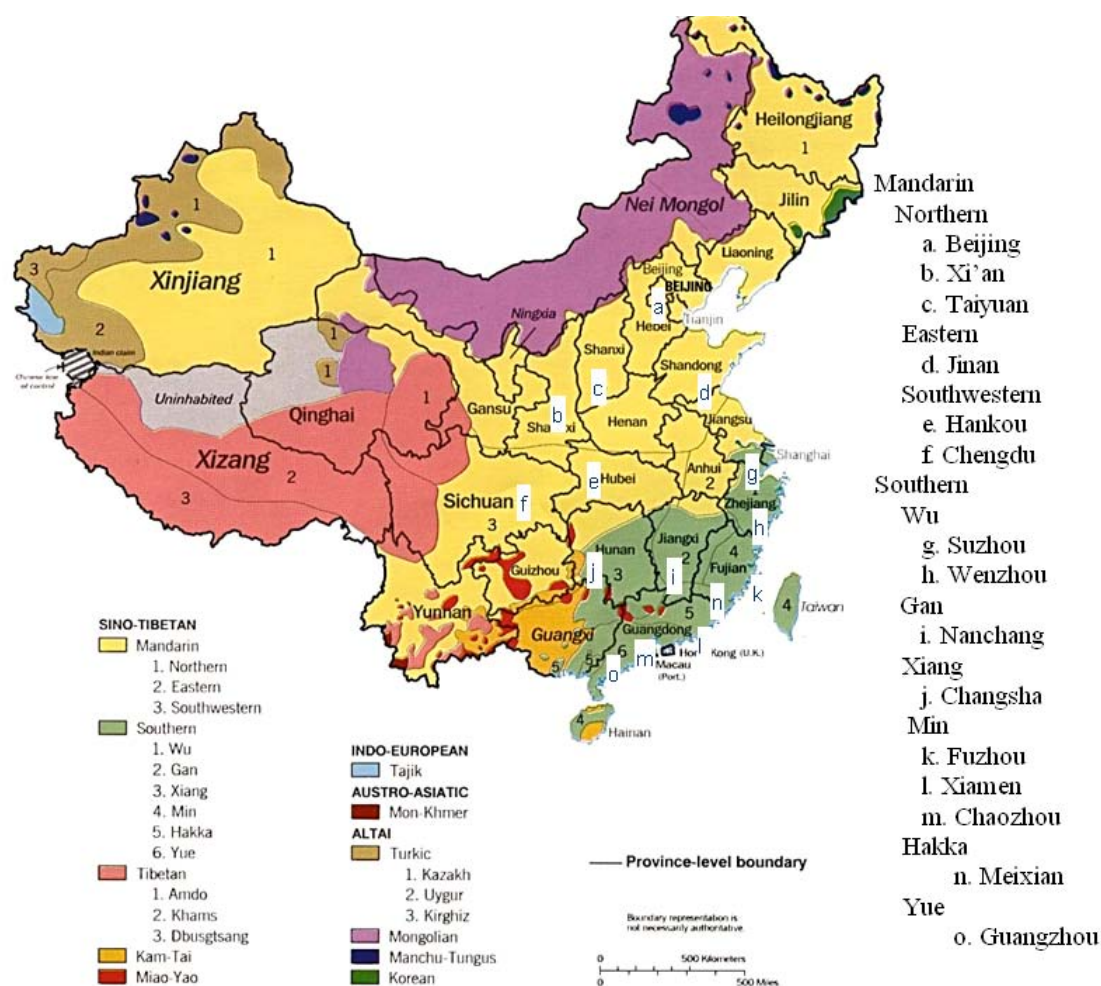


Figure 1: Language distribution in the P.R.China (downloaded from <http://www.chinatravel.com/china-travel-guides/china-maps>). The 15 target dialects (a - o) of our study are identified on the map and listed hierarchically according to dialect (sub)group.

The work done by Gooskens and Heeringa 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, Gooskens and Heeringa 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 Gooskens and Heeringa’s method on a full-fledged tone language, viz. Chinese, a language (family) with much richer tone inventories varying from four (Mandarin, e.g. Beijing, Chengdu) to as many as ten (Yue/Cantonese, e.g. Guangzhou).

The Norwegian language situation is rather unique in that Norwegian arguably has no standard language: Norwegians only use local dialects. This is a felicitous condition when trying to predict mutual intelligibility from structural differences between dialects. In the Chinese language situation (as in most other countries) one dialect has the status of national language or standard language, so that it is widely used in the education and in the media. We will test the hypothesis that mutual intelligibility can be predicted from structural differences more adequately when the standard language is excluded from the set of dialects in the study.

1.3. Earlier work on predicting mutual intelligibility between Chinese dialects

The Chinese dialect situation. In spite of a long history of linguistic and dialectological research in China, the issue of Chinese dialect classification is still controversial. Nevertheless, there seems to be broad consensus on the primary dialectal relationships within the Sino-Tibetan language family as shown in Figure 1.

Generally speaking, languages spoken in China fall into four groups. For each group, there are several sub-language groups (phyla) consisting of dialect families. The Sino-Tibetan language group covers the largest part of the country. This group is subdivided into the Mandarin dialect group (Northern Mandarin family, Eastern Mandarin family and Southwestern Mandarin family) and the Southern dialect group (Wu, Gan, Xiang, Min, Hakka and Yue families). Each dialect family comprises many language varieties or dialects. All Sino-Tibetan language varieties share the same character-based orthographic system (the same written form and essentially the same grammar). In addition, the sound and tones of one dialect can be related to those of another through systematic rules. For instance, the diphthong [ai] in Chengdu corresponds with [e] in Shanghai dialect; likewise, Beijing and Chengdu dialects typically share the same phonemes but differ in the tones only. For example, the low and falling tones in Beijing show up as falling and low tones in Chengdu, i.e., the tones have been switched between the two dialects (DUANMU 2000: 2). However, the dialects in the Mandarin group and in the Southern group are commonly thought to be mutually unintelligible across the border that separates these two dialect families. That is, speakers from different dialect families often cannot understand each other. However, we feel that such claims of mutual (un)intelligibility have to be substantiated experimentally – which is what we set out to do in the present study.

Structural similarity indices for Chinese dialects. We are in the fortunate circumstance that Chinese dialectometrists (most notably CHENG 1997) have developed structural similarity measures for pairs of Chinese dialects in three linguistic domains: (i) historical similarity, (ii) lexical similarity, and (iii) synchronic phonological correspondence (in onsets, glides, vocalic nuclei, codas and tones). In our study we will not be concerned with the historical relationships in (i). However, the quantitative measures derived under (ii), which we will call the *Lexical Similarity Index* (LSI) and (iii), which we will henceforth refer to as the *Phonological Correspondence Index* (PCI), will be used as predictors of mutual intelligibility between pairs of Chinese dialects in the present study.

2. Method

In this section we will first describe how we collected subjective estimations of intelligibility and similarity for all 225 pairs of 15 Chinese dialects (in section 2.1), and then describe the structural distance measures LSI and PCI (in section 2.2) as published by CHENG (1997). In section 3 we will regress Cheng's LSI and PCI measures against our experimental intelligibility (and similarity) judgments.

2.1. Collecting intelligibility and similarity judgments

2.1.1. Materials.

The Chinese dialects we targeted are the following 15 (a proper subset from Cheng 1997): Beijing, Chengdu, Jinan, Xi'an, Taiyuan, Hankou (Mandarin dialects), Suzhou, Wenzhou (Wu dialects), Nanchang (Gan dialect), Meixian (Hakka dialect), Xiamen, Fuzhou, Chaozhou (Min dialects), Changsha (Xiang dialect), and Guangzhou (Yue dialect). For their geographic location see Figure 1.

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) with different speech habits, we processed the recordings (using Praat software, P. BOERSMA and D. WEENING 1996) such that all speakers sounded like males, all had roughly the same articulation rate and speech-pause

ratio, and the same mean pitch.¹ 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 four different random orders (A, B, C, D, where C and D were the reversed order of A and B). The 15 monotonized versions preceded the 15 versions with melody. At the beginning of the CD, as part of the instructions, we recorded the reading (with melody) of the fable in the dialect of the prospective listener group. This was done to make sure that the listeners would be perfectly familiar with the contents of the fable. In all, 60 different CDs were produced.

2.1.2. Listeners

In total 360 listeners participated in the experiment. For each of the 15 dialects a group of 24 native listeners was found in the middle to older generation (ages between 40 and 60), evenly divided between males and females. All listeners were born and bred in their respective dialect areas; a further requirement was that their parents should also be native speakers of the local dialect. Listeners were mono-dialectal so that they had no experience with any other Chinese dialects (although they may have had some familiarity with the Standard Mandarin language through primary education and later media exposure).

2.1.3. Procedure

Each CD was played through loudspeakers to six listeners (three females, three males) either individually or in small groups in a quiet room with little reverberation. 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 'He will not understand a word of the other speaker' whilst '10' represented 'He will understand the other speaker perfectly'. In the second judgment the listener rated the similarity between 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 between fragments listeners were given 7 seconds to fill in their scores on both scales. In all 21,600 judgments were collected and statistically analyzed.

2.2. Structural measures of lexical and phonological similarity

We used two objective measures of structural distance between pairs of Chinese dialects. Both measures were developed by CHENG (1997). For the purpose of the present paper we changed the names of his measures so that they capture the underlying concept more clearly.

The first measure, which we call the Lexical Similarity Index (LSI), is basically the association coefficient *phi*, computed on a crosstabulation of some 2,700 words listed in a computer-readable etymological database for 18 Chinese dialects (Cihui, published by BEIJING UNIVERSITY 1964). The dictionary lists all the words used to express a concept; Cheng tabulated which words occur in which languages. For instance, across all 18 languages there are seven words for the concept 'sun' and five for the concept 'moon'. The Beijing and Meixian dialects each make exclusive selections from the pool of words for sun and moon, such that when a word occurs in Beijing it is never used in Meixian and vice versa. This yields a *phi* value of -1. Across the entire range of concepts covered by the database, however, such extreme *phi* values are never found for any pair of dialects. In fact, the *phi* coefficients established for each pair among the 18 dialects by C. CHENG (1977) range between .698 and .079 (always positive).² We used Cheng's *phi* coefficients as our measure for the lexical similarity (LSI) between all pairs of 13 of our target dialects. We did this by simply copying the values

published in C. CHENG (1997, appendix 3) but leaving out the data for the three dialects not included in our study.³ The LSI measure is symmetrical. The *phi* coefficient for the co-occurrence of lexical items in two dialects is the same whether dialect A is compared with B or vice versa. The higher the LSI, the larger the number of cognate words shared between two dialects. The LSI should be strongly related to mutual intelligibility between two dialects. 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 him to understand the message.

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 (such as an initial /p/ in dialect A will always be initial /f/ in dialect B). Cheng (1997) quantified the sound correspondences separately for onset consonants, for prevocalic glides, for vocalic nuclei, coda consonants, and for lexical tone, giving equal weights to these five domains. The counts were converted to a coefficient ranging between 0 (no phonological correspondence at all) to 1 (perfect sound correspondence). We call this measure (for computational details see C. CHENG 1997) the Phonological Correspondence Index (PCI).⁴ The PCI is an asymmetrical measure since the sound correspondences between dialect A to B may be different (simpler or more complex) than between B and A. For the present paper, however, we have followed Cheng's practice and used the mean of the PCI for each pair of dialects A-B and B-A. We then copied the published PCI values for all of our 15 target dialects from C. CHENG (1997, appendix 5).

3. Results

3.1. Cluster trees

Four measures were obtained for every pair out of our 15 target dialects of Chinese. Two measures derive from our experimental work (judged intelligibility and judged similarity) and two more were copied from the literature as objective measures of structural similarity. Before attempting to predict subjective measures from objective measures we will first present the results by themselves in the form of tree structures. Figure 2 presents hierarchical cluster trees for the 15 target dialects constructed from (a) objective lexical similarity (LSI, only 13 dialects), (b) objective phonological regularity (PCI), (c) subjective intelligibility judgments, and (d) subjective similarity judgments. The dendrograms were constructed from symmetrical matrices using the method of hierarchical agglomerative clustering (applying average linkage, for details see e.g. G. DIEKHOFF 1992).

Inspection of figure 2 shows a rather poor congruence between any pair of the four trees. Even the primary split between Mandarin and Southern dialects is not correctly reproduced in the trees. Typically, 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 subjective intelligibility from subjective similarity

We used the proximity (an intermediate step in the drawing of the dendrograms in figure 2) 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. The proximity matrix is symmetrical; the redundant part of the matrix was deleted before we then 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) with $r = .949$ ($p < .001$). This means that the two sets of ratings can be predicted from each other with a very high degree of accuracy (reduction of prediction error by 90%). Moreover, visual inspection of the corresponding

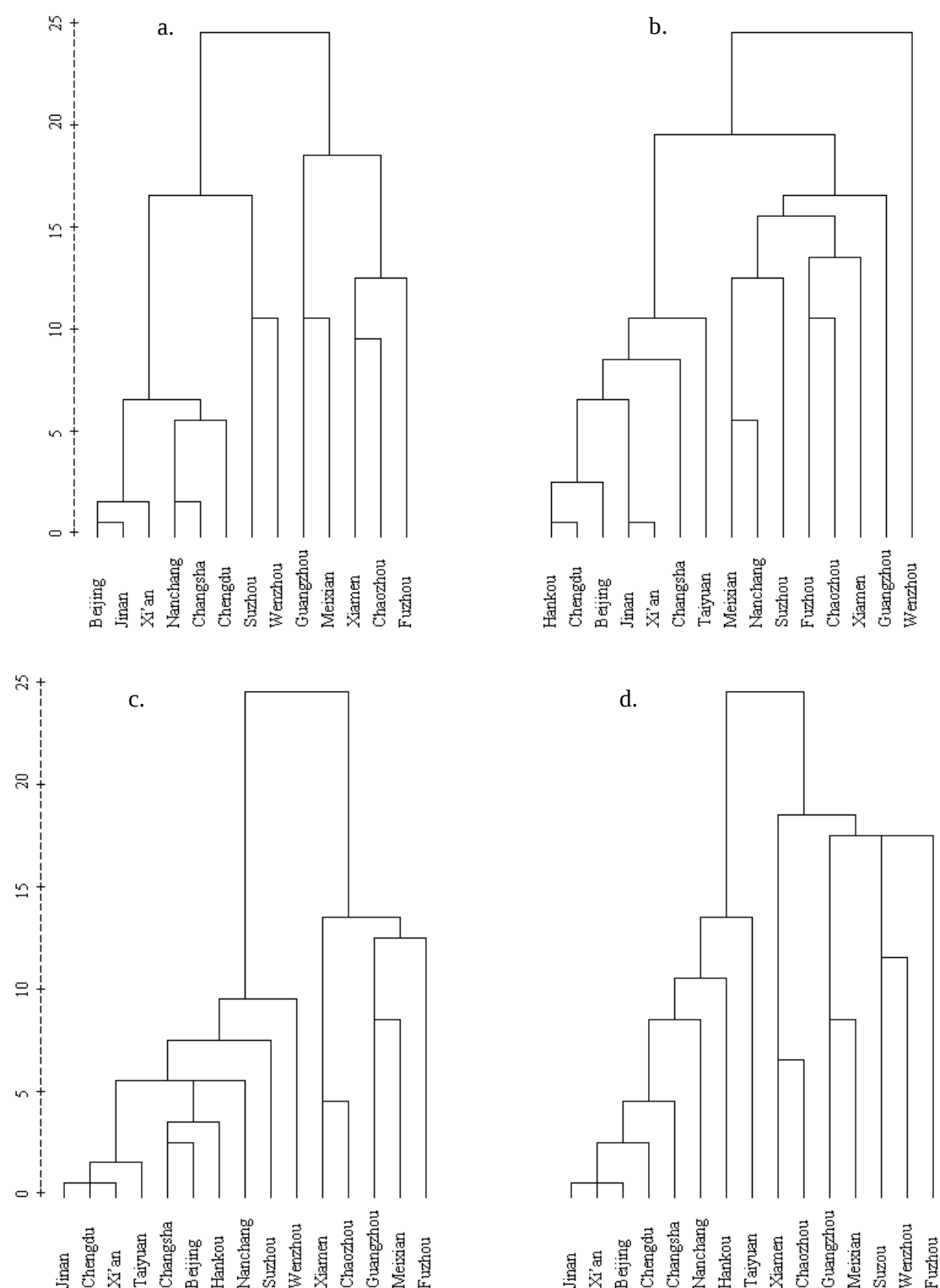


Figure 2. Hierarchical cluster trees for 15 Chinese dialects, based on (a) objective lexical similarity (b) objective phonological regularity, (c) subjective intelligibility rating, and (d) subjective similarity ratings.

scatterplot (see figure 3) reveals no specific outliers, so that the general conclusion follows that subjectively estimated similarity between pairs of languages is an exceptionally good predictor

of, or even a near-perfect alternative for, estimated intelligibility (as was assumed all along by C. GOOSKENS and W. HEERINGA 2004).

3.3 Predicting subjective ratings from objective measures

In table 1 we have specified how well judged intelligibility and judged similarity can be predicted from the objectively determined LSI and PCI measures. In the course of our computations we noticed, however, that very often the relationships between the objective and subjective measures are not linear but exponential. Therefore, 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 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 optimal combinations of conditions) in order to determine the cumulative effect of the predictors. 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.

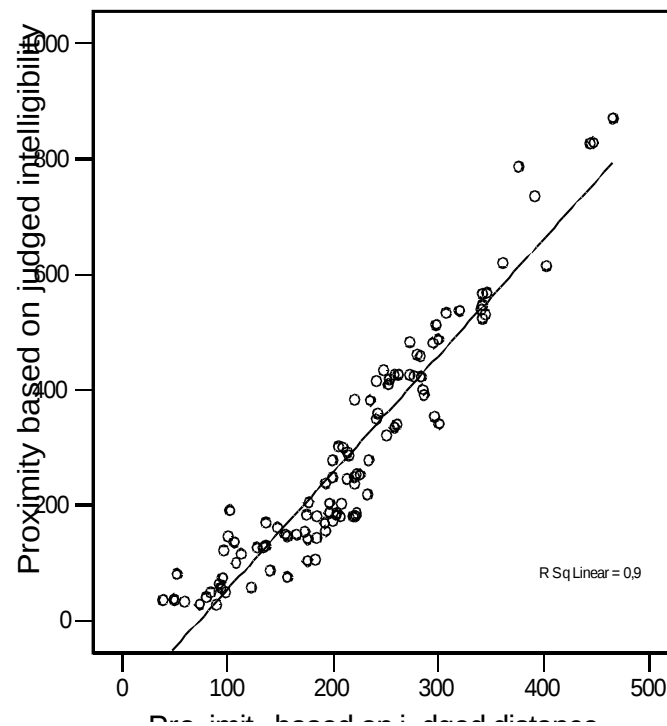


Figure 3. Relationship between judged similarity between dialects and their subjectively estimated mutual intelligibility.

4. Conclusions

A number of conclusions can be drawn from table 1. First of all, the two objective measures of structural similarity, PCI and SLI, are always significantly correlated with all of the subjective rating measures. Moreover, the two predictors are moderately intercorrelated so that there is potential room for improvement of the prediction through multiple regression, i.e., both predictors may contribute sufficient independent information. 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 from the best single predictor (57%).

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.

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 SLI		Both R
	r	N	r	N	
Cheng's SLI	.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)

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. This is in contrast to the results reported for Norwegian dialects, where monotonization merely caused confusion on the part of the listeners.

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

The overall conclusion of our study, finally, should be that Cheng's measures of lexical and phonological similarity between pairs of Chinese dialects afford reasonable to good prediction of mutual intelligibility and – even more so – of judged similarity.

Acknowledgements

The author is a guest researcher at LUCL with a grant from the China Scholarship Council, which she gratefully acknowledges. Her conference trip was subsidized by the Leiden University Fund (67%) and by the Leiden University Faculty of Arts (33%). The research was done in close collaboration with my thesis supervisor Vincent J. van Heuven. This paper is an abbreviated version of a full-length paper to be co-authored by both researchers. Both of us thank Dr. Liu Xiangbo of the Linguistics Department of the Academy of Social Sciences in Beijing for making recordings and phonetic transcriptions digitally available to us.

Notes

¹ 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 length, 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 (syllables/s) was the same for all speakers.

² In an afterthought C. CHENG (1997) regrets not having used a more straightforward measure of lexical similarity such as percent cognates shared between two dialects. He even presents a simple formula to compute such a measure but does not list his results in terms of this more adequate measure.

³ Unfortunately, CHENG (1997) does not list phi values for Taiyuan and Hankou.

⁴ Confusingly, Cheng called this measure Mutual Intelligibility. We could not adopt this term for the present paper, since our experimentally established criterion variable would then have the same name as the structural predictor.

References

- BEIJING UNIVERSITY, *Chinese Dialect Character Pronunciation List*, Beijing, Wenzhi Gaige Chubanshe, 1962.
- BEZOOIJEN, R. VAN, HEUVEN, V.J. VAN, «Assessment of speech synthesis», in D. GIBBON, R. MOORE, R. WINKSI (eds.) *Handbook of standards and resources for spoken language systems*, Berlin/New York, Mouton de Gruyter, 1997, pp. 481-653.
- BOERSMA, P., WEENINK, D., *Praat, a system for doing phonetics by computer, version 3.4*, Report 132, Institute of Phonetic Sciences, University of Amsterdam, 1996.
- CHENG, C.C., «Measuring Relationship among Dialects: DOC and Related Resources», *Computational Linguistics & Chinese Language Processing* 2.1, 1997, pp. 41-72.
- DIEKHOFF, G., *Statistics for the Social and Behavioral Sciences*, Brown, Dubuque, IA, 1992.
- DUANMU S., *The Phonology of Standard Chinese*. Oxford, OUP, 2000.
- GOOSKENS, C., HEERINGA, W., «Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data», *Language Variation and Change*, 2004, 16, pp. 189-207.
- KESSLER, B., «Computational dialectology in Irish Gaelic». In: *Proceedings of the European ACL*, Dublin, ACL, 1995, pp. 60-67.