Mutual intelligibility of Chinese dialects tested functionally

Tang, C.; Heuven, V.J. van; Koppen M. van, Botma B.

Mutual intelligibility of Chinese dialects tested functionally

Chaoju Tang[1] and Vincent J. van Heuven
Chongqing Jiaotong University / Leiden University Centre for Linguistics

1. Introduction

Chinese dialectologists agree that there is a primary split in 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 the authoritative *Language Atlas of China*. We target 15 dialects, which are related as indicated in the overview below.[2]

	group	subgroup	dialect		group	subgroup	dialect
Mandarin branch	Zhongyuan		Xi'an	Southern branch	Wu		Suzhou
	South-western		Chengdu				Wenzhou
			Hankou		Gan		Nanchang
	Beijing		Beijing		Xiang		Changsha
	Jilu		Ji'nan		Min	East Min	Fuzhou
	Jin		Taiyuan			South Min	Xiamen
							Chaozhou
					Hakka		Meixian
					Yue		Guangzhou

Mutual intelligibility has often been advanced as a criterion for linguistic distance between language varieties (e.g. Cheng 1997, Gooskens & Heeringa 2004). Dialects within the Mandarin branch (also dialects not included in our sample) are often claimed to be intelligible to each other to some extent, but are not mutually intelligible to Southern (non-Mandarin) dialects. Most dialects in the Southern (non-Mandarin) branch are mutually unintelligible to each other and, even more so, to the Mandarin dialects. Generally, mutual intelligibility is held to be much poorer for dialects within the non-Mandarin (Southern) branch than between those within the Mandarin branch (Yan 2006:2). The first goal of our study is to experimentally test these impressionistic claims on differences in mutual intelligibility among Chinese dialects.

Linguistics in the Netherlands 2008, **145–156**. DOI 10.1075/avt.25.16tan
ISSN 0929–7332 / E-ISSN 1569–9919

Basically, there are two kinds of experimental means to determine mutual intelligibility between language varieties. One is judgment or opinion testing, which asks how well listener A *thinks* he understands speaker B (and vice versa). The typical measure is a judgment along a rating scale (e.g. unintelligible … intelligible). The other approach is functional testing. It tests how well listener A *actually* understands speaker B (and vice versa). The typical measure is to count the percentage correctly translated words from dialect A to dialect B (and vice versa). Determining mutual intelligibility through functional testing is a daunting task, especially when the number of language varieties involved in the comparison is large. Therefore, opinion testing is often used as a short-cut. To our knowledge, no-one has systematically compared the results of opinion and functional tests of mutual intelligibility in the context of language variation. Therefore, as a second goal, we want to compare the results of opinion tests with those of functional tests, in order to determine (i) to what extent the two methods converge, and (ii) to the extent that they do not, which method is more credible. Therefore, we measured mutual intelligibility between pairs of Chinese dialects through experiments using opinion tests (Tang & van Heuven 2007) and functional tests (present study). We will compare our experimental data against the traditional dialect taxonomy and determine which type of test concurs best with the traditional classification of Chinese dialects.

We designed two tests: one at the level of isolated words, the other at the sentence level. The word-intelligibility test was developed from scratch. It affords fast and economical testing of the recognition of a large number of isolated words. We tested recognition through semantic categorisation. Listeners indicate to which of ten pre-given semantic categories a spoken word belongs. For instance, if the listener hears the word for 'apple', s/he should categorize it as a member of the category 'fruit'. Here, the assumption is that correct categorisation can only be achieved if the listener correctly recognizes the target word. Word recognition in sentence context was tested by a Chinese version of the SPIN ('Speech Perception in Noise') test, which was originally developed for English by Kalikow, Stevens and Elliott (1977).

2. Methods

For the word part, we prepared a list of 150 Standard Mandarin core words. These words form such categories as body parts, plants, fruits, vegetables, animals, etc. The words all denote simple concepts commonly used in everyday life in each of our 15 target Chinese dialects. For the sentence part, we selected 60 sentences based on the high-predictability section in the SPIN test sentence lists. In the SPIN test, listeners have to write down the final word (target) of each sentence they hear. Getting the final word is easier as the listener also correctly recognizes the earlier words in the sentence, as in *He wore his broken arm in a sling* (target underlined). The sentences were selected on the basis of their applicability to the Chinese situation, and translated from English into Standard Mandarin.

Thirty speakers were recorded, i.e. one male and one female native speaker of each of the 15 target dialects. Speakers were students at Chongqing Jiao-Tong University, China. All were born and bred in the dialect region they represented. They had moved to Chongqing as young adults. They returned to their dialect area on a regular basis, for at least two months in the summer and six weeks in the winter season. In Chongqing they were part of fairly large dialect communities, and in most cases the male and female speaker representing a particular dialect were a couple who had continued to speak the dialect in their own home when in Chongqing. Also, when the recordings were made, the male and the female speaker pair spent considerable time together, speaking the dialect, in order to prepare the translations.

Prior to the recording sessions, for each dialect the designated speakers translated the target words and sentences from Standard Mandarin into their own dialects. Translations were done independently by the two speakers; the few cases where disagreement was observed, were solved by consensus.

The words and sentences were then read from paper and recorded in individual sessions. Speakers were seated in a quiet office and wore a Shure SM10A head-mounted close-talking microphone. Speakers read both the word and sentence parts in their own dialect using the translations they had prepared themselves.

In intelligibility tests it is imperative that a listener does not hear the same word (or morpheme) twice; stimulus words and sentences should be blocked over listeners, such that each listener hears each word or sentence only once, irrespective of the dialect of the speaker. Therefore, we worked out a Latin Square design (e.g. Box, Hunter & Hunter 1978), such that (i) each listener heard each of the words and sentences only once, (ii) each of the 15 listeners in one dialect group heard each version of a word or sentence in a different dialect, while at the same time (iii) every listener heard one-fifteenth of the CD materials in each of the 15 dialects.

In all, 225 CDs (15 copies of 15 different CDs) were produced. On each CD, the word part preceded the sentence part. Ten words or ten sentences formed a track, with a pause between words or sentences of 7 s and with 11-s pauses between tracks.

For each dialect in the set of 15, a local contact person was contracted. Each local contact, a native speaker of the dialect of the listener group targeted, recruited 15 native listeners of the dialect s/he represented. All local contact persons and the listener subjects were paid for their services.

Listeners took part in the experiment in individual sessions. Listeners were sampled 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.

Each listener in a dialect group listened to a different CD, one of the set of 15 CDs. All listeners were required to both read the paper instructions and to follow the spoken instructions (in Mandarin) on the CD. Stimuli were presented through twin loudspeakers in a quiet room, often in the contact person's private home.

The isolated word recognition task was presented first. Here, the listener was required to tick one of ten boxes on an answer sheet for each word representing the ten semantic categories every time a word was presented. For the subsequent sentence part, the listener had to write down the final target word in his/her own dialect after listening to each of the 60 sentences on the CD. After the last of the 60 sentences had been presented, the local contact person translated the 60 response words into Mandarin in the presence of the listener, asking the listener for clarification whenever necessary.

3. Results

We computed the mean percentage of correctly classified words for each of the 15 × 15 combinations of speaker and listener dialects, yielding 225 mean word recognition scores. For the sentence intelligibility test, the procedure was less straightforward. The first author, a native speaker of Chinese, manually checked whether the translation of target word from the local dialect back into Mandarin was semantically equivalent to the original Mandarin version. If so, the response was considered correct. If not, or if no translation had been given at all, it was considered an error. From these data we computed 15 × 15 = 225 mean sentence-intelligibility scores, i.e. one mean score for each combination of speaker and listener dialect.

Table 1 presents the mean percentage of correctly classified words for each combination of speaker dialect (rows) and listener dialect (columns).

Scores in cells along the main diagonal (bolded) were obtained by listeners who listened to speakers of their own dialect. Scores in off-diagonal cells should be poorer, as these cell means are based on listeners exposed to speakers of a different dialect. Indeed, generally we do find the highest scores in the diagonal cells. The highest percentage correctly classified words is between Beijing speakers and listeners; Beijing listeners correctly recognised (i.e. classified) 83% of the words spoken in their own dialect. Some listener groups were less successful. For instance, Xiamen and Nanchang listeners could not understand the speakers of their

Table 1. Mean percent correctly classified words based on 150 responses broken down by 15 speaker dialects and 15 listener dialects. Each of 150 words was heard once, with 10 different words per dialect, for each of 15 listeners. Total number of responses is 225 × 150 = 33,750.

	Listener dialect														
speaker dialect	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**

respective dialects very well, given the mean scores of 39% and 37%, respectively. On two occasions, in fact, the native dialect listener groups were outperformed by one of the other groups.[3] Typically, listeners whose native dialect belongs to the Mandarin group were more successful (mean across the six Mandarin dialects is 72% correct) than listeners with Southern (non-Mandarin) native dialects (52%). These mean word recognition scores are summarised in Figure 1A. This figure also shows intelligibility of Mandarin and Southern speakers for listeners within the same dialect group (i.e. Mandarin speakers and listeners not sharing the same dialect, and Southern speakers and listeners not sharing the same dialect) and for listeners in the other dialect group (Mandarin speakers and Southern listeners, or vice versa).

Figure 1 shows that speakers and listeners within the Mandarin branch recognize many of the words in other Mandarin dialects (this is even the case when Beijing speakers are excluded, in which case the score is 59% instead of 61%, see discussion). Mandarin listeners get much poorer word recognition scores when listening to Southern dialects (36%). Southern listeners recognize a mere 22% of

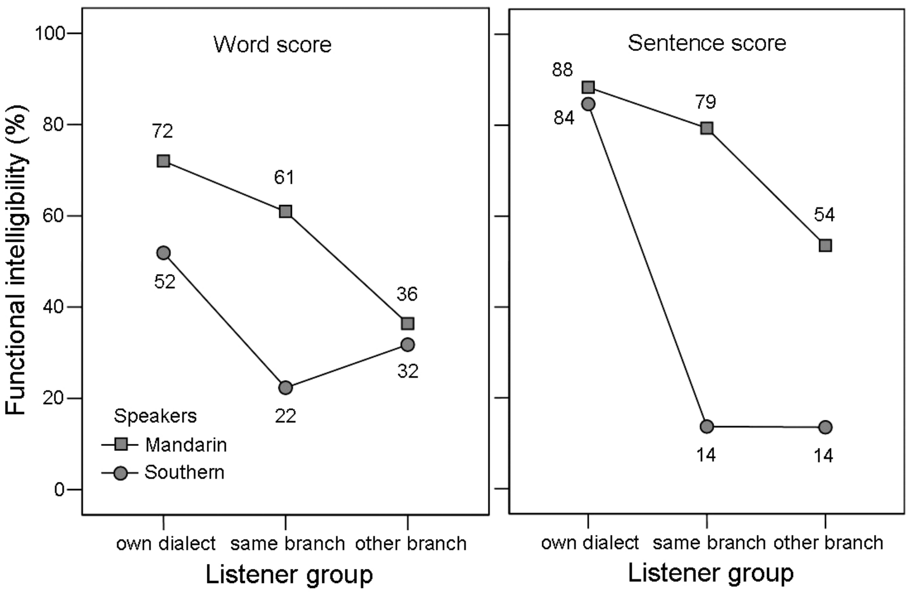

Figure 1. Intelligibility as a function of type of speaker dialect and of listener dialect at the word level (panel A) and at the sentence level (panel B). Braces enclose means that do not differ from each other by a Scheffé test ($p < .05$).

the words in other Southern dialects. They do, in fact, better on Mandarin dialects (32% correct; the same score is found even if we exclude Beijing speakers). A two-way Analysis of Variance with speaker dialect group (Mandarin, Southern)

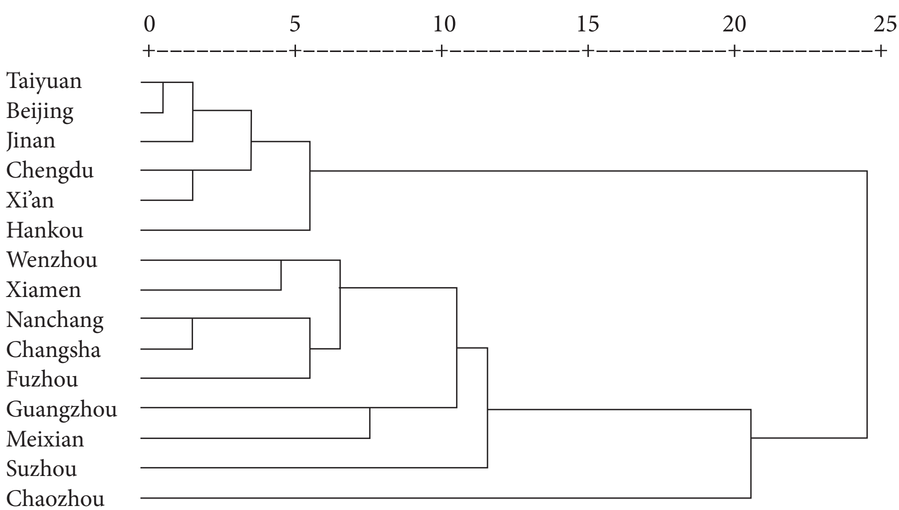

Figure 2. Dendrogram (using average linking between groups) based on word intelligibility scores obtained for all 225 combinations of 15 speaker and 15 listener dialects.

and listener group (own dialect, other dialect within same branch, dialect in other branch) as fixed factors reveals significance for both factors as well as for the interaction between the two, $F(1,219) = 120.1$, $F(2,219) = 61.8$ and $F(2,219) = 78.6$, respectively ($p < .001$ in all cases). Means in Figure 1 that are enclosed by the same brace do not differ from each other by a post-hoc Scheffé test ($p < .05$).

The data in Table 1 were then used to generate a dendrogram, using the average linking method that we also used in our earlier report (Tang & van Heuven 2007). As a first step in the procedure, the matrix was made symmetrical by averaging corresponding cells above and below the diagonal, i.e., the cell contents of every pair of contra-diagonal cells *i, j* and *j, i* were averaged. Cells on the diagonal were excluded (set to 100). The tree structure that was generated is displayed in Figure 2.

The tree makes a primary split between the six Mandarin dialects, and a group of nine dialects that comprises all the Southern (non-Mandarin) dialects. This division concurs well with traditional taxonomies postulated by Chinese dialectologists. We will discuss the internal cluster structure within the main branches later.

Table 2. Mean percent correctly translated target words is based on 60 responses in sentences broken down by 15 speaker dialects and 15 listener dialects (each of 60 sentence-final words is heard once, with 4 different words per dialect for each of 15 listeners). The total number of responses is 225 × 60 = 13,500.

	Listener dialect														
Speaker dialect	Suzhou	Wenzhou	Guangzhou	Xiamen	Fuzhou	Chaozhou	Meixian	Nanchang	Changsha	Taiyuan	Beijing	Jinan	Hankou	Chengdu	Xi'an
Suzhou	**77**	7	5	18	13	5	7	13	13	20	5	18	15	15	7
Wenzhou	5	**93**	5	12	3	2	7	10	2	7	2	10	8	7	2
Guangzhou	5	7	**92**	10	20	25	55	22	13	7	3	22	8	17	7
Xiamen	13	5	8	**97**	23	28	13	18	13	3	5	15	7	17	8
Fuzhou	3	3	2	17	**92**	7	3	8	5	0	0	7	2	0	3
Chaozhou	7	0	3	52	13	**98**	3	12	3	7	2	13	10	3	5
Meixian	13	2	12	28	17	20	**70**	25	18	10	3	25	15	25	8
Nanchang	28	13	20	25	27	17	33	**50**	32	35	18	53	43	37	23
Changsha	12	3	8	23	17	3	17	25	**93**	13	13	38	53	28	2
Taiyuan	63	35	45	63	57	25	55	68	68	**73**	77	92	92	85	73
Beijing	87	62	90	90	93	60	80	78	92	90	**98**	98	97	98	93
Jinan	52	27	32	48	48	15	40	60	70	75	77	**97**	83	82	67
Hankou	48	32	32	52	53	27	45	53	62	58	67	95	**100**	73	65
Chengdu	47	22	40	48	72	27	48	58	62	65	62	98	95	**95**	68
Xi'an	53	33	50	58	57	30	57	58	63	68	58	82	78	70	**67**

Table 2 presents the results of the intelligibility test at the sentence level. Percent correctly translated target words is given for each combination of speaker and listener dialects. These scores are summarised in Figure 1B.

The range of sentence scores is larger than that for semantic categorisation and the mean scores for own dialect are much better than those for the semantic categorisation (see the diagonals). Clearly, the sentence-level test was an easier task (for listeners who are familiar with the stimulus dialect) than the semantic categorisation task with isolated words. The mean scores for the native dialect listener groups (listening to their own speakers) is 84% for the nine Southern (non-Mandarin) dialects, and 88% for the Mandarin dialect groups. Again, mutual intelligibility is very good within the Mandarin dialects and very poor in the Southern branch. Southern dialects are as poorly intelligible to Mandarin listeners as they are to Southern listeners. Mandarin speakers are fairly intelligible to Southern listeners (54% intelligibility), and this effect largely remains even if we exclude Beijing speakers (48%). Main effects and the interaction in Figure 1B are significant, $F(1,219) = 165.1$, $F(2,219) = 94.8$ and $F(2,219) = 38.5$, respectively ($p < .001$).

Figure 3 shows the dendrogram based on the sentence-intelligibility results.

The sentence-level tree shows, again, a clean cut between the six Mandarin and the nine Southern dialects. As before, we will defer discussion of the internal structure of the dialects within the main branches. First we will consider the question how well the functionally determined word and sentence intelligibility scores can be predicted from our earlier judgment scores on intelligibility.

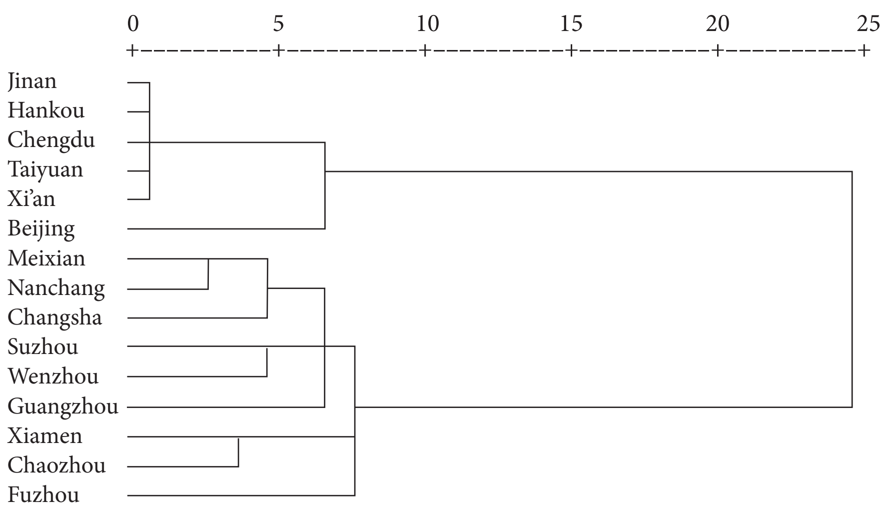

Figure 3. Dendrogram (using average linking between groups) based on sentence-level intelligibility scores (further see Figure 1).

4. Conclusions and discussion

4.1 Intelligibility within and between Mandarin versus Southern dialects

In the introduction we mentioned that Mandarin dialects are held to be more mutually intelligible amongst themselves than are Southern dialects, and that Mandarin dialects are more intelligible to Southern dialects than vice versa. Our results show that both impressions are borne out by the experimental data, both in terms of the word classification scores and of the sentence intelligibility test.

One reason why Mandarin dialects are mutually more intelligible than are Southern dialects could be that the former are intrinsically more similar than the latter. It is also the case however, that most Chinese listeners are familiar, through education and the media, with Standard Chinese, which is very similar to Beijing dialect. If we recompute the word and sentence intelligibility scores after eliminating the two Beijing dialect speakers, the results of our study are hardly affected, as we observed on several occasions in the previous section. The clearest way of testing the intrinsic greater similarity of Mandarin dialects is by including the scores obtained by Beijing listeners only (60 and 68% correct word and sentence scores) and comparing these with the scores obtained for the Southern listeners exposed to other Southern dialects (22 and 14% correct recognition, cf. Figure 1AB). Clearly, intelligibility of other Mandarin dialects for Beijing listeners is much better than mutual intelligibility within the Southern dialects, $t(75) = 13.2$ and 11.0 for word and sentence scores, respectively ($p < .001$). These comparisons show that intrinsic linguistic similarity is a more important determinant here than the possible advantage of Mandarin dialects being close to the Standard language.

4.2 Correlations between measures

We have obtained two kinds of functional data experimentally; one is for isolated word-intelligibility, the other is for sentence-intelligibility. We will first examine the correlation between functional intelligibility at the word level and at the level of the sentence, and then relate these measures to our earlier opinion scores. Finally, we will relate these measures to traditional Chinese taxonomy.

Intelligibility at the word and sentence level. The results obtained from the word-intelligibility and the sentence-intelligibility tests presented above converge to a great extent. We define mutual intelligibility between two language varieties A and B as the mean of the intelligibility of A to B and of B to A. Thus, we computed the correlation coefficient for the word and sentence scores after averaging the contra-diagonal cells in the matrix (i.e., averaging the contents of every pair of cells *i*, *j* and *j*, *i*), which makes it a symmetrical matrix, of which only the non-redundant part ('lower triangle') is used in the computation of r). This procedure yields a high correlation coefficient, $r = .928$ ($N = 105$, $p < .001$). The coefficient of determination is $r^2 = .86$, which means that still 14% of the variance is left unaccounted for. It seems unclear, therefore, whether the word-intelligibility test (semantic categorisation test) is a fully adequate short-cut to functional intelligibility. For the moment we will assume that both the word-level and the sentence-level tests are needed. Later, when we compare the test results with external data (traditional genealogies), we may be able to choose between the two.

Opinion versus functional testing. In our earlier study (Tang & van Heuven 2007), we collected opinion scores on intelligibility between all pairs of our 15 dialects. Our opinion scores were based on natural readings of *the North Wind and the Sun* fable. We will now determine to what extent the judgment data are correlated with intelligibility scores obtained from functional test procedures. Table 3 was computed on the non-redundant parts (lower triangles) of the matrices after they had been made symmetrical by averaging the contents of all contra-diagonal cells *i, j* and *j, i*.

Table 3. Correlation coefficients for three mutual-intelligibility measures, using the symmetrical means (lower triangle, $N = 105$) of the intelligibility matrices.

	Word intelligibility	Sentence intelligibility
Sent. intelligibility	.928*	
Judged intelligibility	.772*	.818*

* Correlation is significant at the 0.001 level (2-tailed).

Our earlier opinion scores (judgments of intelligibility) correlate reasonably well, i.e. around $r = .8$, with both functional intelligibility measures. However, this correlation coefficient leaves some 35% of the variance in the functional intelligibility measure unaccounted for, so that it is now important to see which type of intelligibility measure is more credible. We will crossvalidate the trees derived from each of the intelligibility measures concur against linguistic taxonomy.

Relating scores to linguistic taxonomy. We compare three trees based on experimental data with the traditional linguistic taxonomy as proposed by Chinese dialectologists (see Section 1). The first tree was based on the opinion scores on intelligibility collected in Tang & van Heuven (2007). These generated a tree structure that failed to reflect the primary split between Mandarin and Southern dialects. Nanchang and Changsha dialects were incorrectly classified with the Mandarin group, whilst Suzhou and Wenzhou were wrongly grouped with the Mandarin branch.

In contrast to this, we now find that both trees based on functional scores correspond rather well with the classification by dialectologists. The primary split in the trees is between the six Mandarin dialects on the one hand, and the nine Southern (non-Mandarin) dialects on the other. The basic division of Chinese dialects into Mandarin versus Southern branches is correctly reflected then. In terms of substructure, however, the word-based tree is less credible as there is not a single pair of dialects within the Mandarin branch, nor in the Southern branch, that are grouped together the way they should according to the traditional linguistic taxonomy. The functional sentence-intelligibility tree reflects the substructure within the main branches considerably better. It yields identifiable clusters such as the Wu (Suzhou, Wenzhou) and Min groups (Xiamen, Chaozhou and Fuzhou).

We conclude that mutual intelligibility can be experimentally measured both by opinion and functional tests. However, although opinion tests converge to a considerable extent with the results of actual functional intelligibility tests, the correlation between the two is not good enough to unequivocally recommend the former as an adequate (and more efficient) substitute for the latter. Opinion testing seems defensible only if functional testing is unfeasible, for instance in situations where even a larger number of language varieties have to be compared than in our set of 15.

Within the category of functional tests, word intelligibility and sentence intelligibility are highly but not perfectly correlated, so that again a choice must be made. We recommend that the sentence-intelligibility test should be the preferred method, since the dendrogram derived from this test, rather than the tree based on the word-intelligibility scores, concurred substantially better with the taxonomy of Chinese dialects postulated by traditional dialectologists.

Notes

1. The first author acknowledges the financial support from the China Scholarship Council (CSC) for tuition and subsistence in The Netherlands for a four-year period.

2. The family tree we used in Tang & van Heuven (2007) is a condensed version (i.e. with less internal structure) of the taxonomy proposed in the Atlas. Dialect (sub)groups not represented in our sample have been omitted from the taxonomy. For geographic locations of the 15 dialects, see our earlier paper or consult www.taiwandocuments.org/map08.htm. We depart from the Atlas in one detail: we parse Taiyuan (Jin group) with the Mandarin branch. The main reason for Li (1987a, b), and others, to consider Taiyuan a Southern (non-Mandarin) dialect is that it kept the so-called Ru tone. However, several other Mandarin dialects (not represented in our sample) also kept the Ru tone, so that the classification of Taiyuan as a Southern (non-Mandarin) dialect does not follow.

3. In these cases the local contact person hailed from a different town or village (although within the same dialect region) as the speakers on the CD. This reduces intelligibility of the own dialect but should not affect mutual intelligibility across dialects.

References

Box, G. E. P., W. G. Hunter and S. J. Hunter, 1978. *Statistics for Experimenters*. New York, NY: John Wiley & Sons.

Cheng, Ch-Ch. 1997. 'Measuring Relationship among Dialects: DOC and Related Resources'. *Computational Linguistics & Chinese Language Processing* 2.1.41–72.

Gooskens, C. and W. Heeringa. 2004. 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data'. *Language Variation and Change*, 16.189–207.

Kalikow, D. N., K. N. Stevens and L. L. Elliott. 1977. 'Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability'. *Journal of the Acoustical Society of America*, 61.1337–1351.

Li, R. 1987a. *Languages in China. The Language Atlas of China*. Hong Kong: Longman, A-1.

Li, R. 1987b. *Chinese Dialects in China. The Language Atlas of China*. Hong Kong: Longman, A-2.

Tang, C. and V. J. van Heuven. 2007. 'Mutual Intelligibility and Similarity of Chinese Dialects'. ed. by B. Los and M. van Koppen, 223–234. *Linguistics in the Netherlands 2007*. Amsterdam: John Benjamins.

Yan, M. M. 2006. *Introduction to Chinese Dialectology*. LINCOM Studies in Asian Linguistics. München: LINCOM.