Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Lingua 119 (2009) 709–732
Lingua
www.elsevier.com/locate/lingua

Mutual intelligibility of Chinese dialects experimentally tested

Tang Chaoju[a,b], Vincent J. van Heuven[a,*]

[a] *Phonetics Laboratory, Leiden University Centre for Linguistics, PO Box 9515, 2300 RA Leiden, The Netherlands*
[b] *Foreign Language School, Chongqing Jiaotong University, Xuefu Road 66, Nan'an District, Nanping, 400074 Chongqing, PR China*

Received 3 April 2008; received in revised form 8 October 2008; accepted 8 October 2008
Available online 21 November 2008

Abstract

We argue that mutual intelligibility testing is an adequate way to determine how different two languages or language varieties are. We tested the mutual intelligibility of 15 Chinese dialects functionally at the level of isolated words (word-intelligibility) and the level of sentences (sentence intelligibility). We collected data for each dialect by playing isolated words and sentences spoken in 15 Chinese dialects to 15 listeners. Word-intelligibility was determined by having listeners perform a semantic categorization task whereby words had to be classified as one of ten different categories such as body part, plant, animal, etc. Sentence intelligibility was estimated by having the listeners translate a target word in each sentence into their own dialect. We obtained 47,250 data (15 × 150 × 15 for the word part and 15 × 60 × 15 for the sentence part). We also had at our disposal structural similarity measures (lexical similarity, phonological correspondence) for each pair of the 15 Chinese dialects published by Cheng (Computational Linguistics & Chinese Language Processing 1997, 2.1, pp. 41–72). Our general conclusion is that the degree of mutual intelligibility can be determined by both opinion and functional tests. These two subjective measures are significantly correlated with one another and can be predicted from objective measures (lexical similarity and phonological correspondence) equally well. However, functional intelligibility measures, especially at the sentence level, better reflect Chinese dialect classifications than opinion scores.

Keywords: SPIN sentences; Functional test; Word-intelligibility; Sentence intelligibility; Mutual intelligibility; Chinese dialects

1. Introduction

1.1. Linguistic distance

A very basic question that has traditionally been asked in linguistics is: 'How much do two languages differ?' When two language varieties differ only a little, linguists are inclined to consider them dialects of one language; if the differences are relatively large, the varieties will be classified as manifestations of different languages. When two language varieties differ only by a small amount, the linguistic distance between them is small; linguistic distance increases as languages differ more radically. On a higher level, the same criterion of linguistic distance can be used to set up family trees (cladistic trees) for groups of (related) languages. Languages that are relatively similar, such as Dutch and German, are considered sibling languages within the group of West-Germanic languages, whilst members

* Corresponding author. Tel.: +31 71 5272319; fax: +31 71 5272615.
E-mail addresses: chaoju.tang@gmail.com (C. Tang), v.j.j.p.van.heuven@hum.leidenuniv.nl (V.J. van Heuven).

0024-3841/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2008.10.001

of other pairs, such as Dutch and Danish, were assigned to different branches of Germanic, i.e. West-Germanic and Scandinavian, respectively. Yet, West-Germanic and Scandinavian languages, all of which are members of the Germanic branch of the Indo-European language family, are closer to each other than any of them is to, for instance, French or Spanish, which–being members of the Romance branch of Indo-European–are closer to each other than to any of the Germanic languages.

Languages resemble each other when they are related. The formidable task of establishing the family relationships among the languages of Europe has been the main issue that linguistics was concerned with in the nineteenth century. In the twentieth century the analytic tools developed in the preceding era have been successfully applied to other, non-European, languages and language families.

In spite of its apparent success and conceptual simplicity, the notion of linguistic distance, i.e. the inverse of similarity shared between languages, has persistently eluded quantification. The problem is that languages do not differ along just one dimension. Languages may differ formally in their lexicon, phonetics and phonology, morphology, and in their syntax. And again, at each of these linguistic levels, the ways in which languages may vary are further subdivided along many different parameters. Linguists have argued about family relationships among languages, and thereby implicitly about linguistic distance, largely on an intuitive basis. One difference (or a small set of differences, a 'bundle of isoglosses') between languages is held to be more important than all other differences combined, and thus sets a group of languages sharing that particular property apart from languages that do not share the property. Which isogloss or bundle is chosen as the distinguishing characteristic, seems an intuitive if not arbitrary choice. The problem is stated by Chambers and Trudgill (1980:112) as follows:

> It is undeniable that some isoglosses are of greater significance than others (. . .). It is equally obvious that some bundles are more significant than others (. . .). Yet, in the entire history of dialectology, no one has succeeded in successfully devising a satisfactory procedure or a set of principles to determine which isoglosses or which bundles would outrank some others. The lack of a theory or even a heuristic that would make this possible constitutes a notable weakness in dialect geography.

One approach to the problem is to simply ask listeners how much the speech in language B differs from their own language A. This is called 'the perception of degrees of difference between a local variety and surrounding varieties' by Preston (1987:4). Subjects listen to a recorded speech sample of a variety B and are asked to judge how different the variety is from their own variety A on some continuous rating scale. This is 'perceived linguistic distance'. Alternatively, subjects are asked to rate the distance between A and B without auditory samples but relying purely on preconceived ideas triggered by geographic names. This is called 'estimated linguistic distance' (Gooskens, submitted). In our study we prefer the first method, i.e. perceived linguistic distance. The first study using this methodology was done in the Netherlands by van Hout and Münsterman (1981), who asked listeners to rate the distance between recorded samples of nine different regional varieties of Dutch from the standard language on a 7-point scale. More recently, the same approach was used by Gooskens and Heeringa (2004), who played speech samples in 15 Norwegian dialects to groups of listeners from the same 15 dialect areas and asked the listeners to judge how much the samples differed from their own dialect. Listeners appear to have reliable (i.e. reproducible) ideas about how much language B differs from their own, even if they know the stimulus language from past exposure, and even if the recording quality of the speech samples may differ substantially. This approach is typically limited to the judgment of intelligibility among related languages.

It would seem likely that the linguistic distance judgments are essentially based on how difficult the listener thinks it would be for him to understand speakers of the other language. Therefore, we advance the concept of mutual intelligibility (sometimes also called mutual comprehensibility) as an auxiliary or alternative criterion to argue about linguistic distance. Intelligibility is best viewed as a scalar variable that expresses how well listener A understands speaker B, for instance on a scale from 0 (no understanding at all) to 100 (perfect intelligibility). Intelligibility is not necessarily a symmetrical property. There are persistent claims, for example, that Danes understand Swedish much better than Swedes understand Danish. Mutual intelligibility, therefore, is best defined as the average (mean) of the intelligibility of speaker A for listener B and vice versa (Cheng, 1997). If a procedure could be developed by which mutual intelligibility between any two languages could be established, we would have a powerful instrument, a communicatively meaningful way of arguing about linguistic distance. Obviously, if two language varieties have a high degree of mutual intelligibility, the linguistic differences between them cannot be major. As the degree of mutual intelligibility between two languages decreases, their structural differences must be more radical. Theoretically, by comparing a large number of languages differing along many dimensions we may establish the relative importance of the various dimensions using mutual intelligibility as the overall criterion variable. This, in turn, will allow us to provide a more solid, experimentally grounded, foundation for the traditional claims about genealogical relatedness among languages as proposed by linguists. The present article is an early step towards this goal.

Early attempts at establishing mutual intelligibility among related languages were made by American structuralists around 1950, trying to establish mutual intelligibility among related Amerindian languages (Voegelin and Harris, 1951; Hickerson et al., 1952; Pierce, 1952). The method was standardized and is still often used in the context of literacy programs, where a single orthography has to be developed that serves multiple closely related language varieties (Casad, 1974; Brye and Brye, 2000; Anderson, 2005). In these methods listeners either summarize, or answer questions about, the contents of a speech sample they just heard. A major problem with this approach is that it is very difficult, if not impossible, to come up with speech samples and questions of equal difficulty in each of a set of language varieties, so that reproducibility of the results is compromised. The method might work as long as the number of language varieties targeted is small (e.g. Delsing and Lundin-Åkesson, 2005 who determined mutual intelligibility among Scandinavian languages Danish, Norwegian and Swedish using a comprehension test with just five open questions). It should be realized that the amount of work involved in establishing mutual intelligibility among a set of languages is staggering. The number of comparisons to be made grows polynomially with the number of languages in the set. In our work, we target just 15 languages, which unfolds to 15 (speaker languages) x 15 (listener languages) = 225 pairs of languages that have to be compared. As a consequence, other means of establishing mutual intelligibility have to be explored.

Intelligibility testing has been a topic of considerable importance in the areas of audiology, speech technology and in foreign language testing. In the field of audiology, intelligibility tests were developed that measure intelligibility as a function of the patient's hearing loss at the level of individual sounds, of words and of sentences (see, for instance, Kalikow et al., 1977). More recently, similar techniques were adopted and extended in order to test the intelligibility of, and diagnose problems with, talking computers (see, for example, van Bezooijen and van Heuven, 1997). The same techniques were also fruitfully applied to the testing of intelligibility of foreign-accented speech (e.g. Wang, 2007 and references therein). In the present paper we will employ some of these tests to establish the degree of mutual intelligibility among Sinitic languages. We believe that Sinitic languages offer a promising testing ground for mutual intelligibility studies as the dimensionality of the comparison is somewhat reduced. Sinitic languages are characterized by the absence of morphology, and they differ relatively little in terms of their syntax. As a result, differences in mutual intelligibility are primarily related to lexicon and phonology. It is also a fortunate circumstance that Chinese linguists have established an impressive body of digital resources that can be used to study objective structural similarities and differences among the many languages spoken in China.

Basically, there are two kinds of experimental means to determine mutual intelligibility between languages (or language varieties). One is functional testing, also called the 'test the informant' method by American structuralists (Voegelin and Harris, 1951). It tests how well listener A ***actually*** understands speaker B (and vice versa). The typical measure is to count the percentage of correctly translated words from Dialect A to Dialect B (and vice versa). The other approach is judgment or opinion testing ('ask the informant' in the American structuralist method) which asks how well a listener A ***thinks*** s/he understands speaker B (and vice versa). The typical measure is a judgment along a rating scale (e.g. unintelligible . . . intelligible).

One problem that inevitably comes up in the context of functional intelligibility testing is that the same listener cannot be asked to recognize the same word twice, not even if the second instance is a translation of the target word into some other, related language. It is well known from the literature that prior recognition of a word (or stem morpheme) greatly facilitates subsequent recognition of the same word or morpheme (e.g. Morton, 1969; Murrell and Morton, 1974; Nooteboom, 1981; Cutler and van Donselaar, 2001). This so-called priming effect forces the investigator to block materials over groups of listeners, such that one individual listener never hears the same word or sentence twice. Therefore, researchers often use judgment testing as a short-cut to functional intelligibility testing. In judgment testing listeners are asked to judge how well they think a fellow native listener would understand the speaker or the language variety they hear in the test. Generally, listeners have quite clear and reproducible opinions on differences in intelligibility of materials they are exposed to, even if the same contents are heard repeatedly. We also know from studies done in the area of speech technology and foreign-language testing, that such opinions correlate very well with the results of functional intelligibility studies (van Bezooijen and van Heuven, 1997, and references therein).

Current practice is that functional intelligibility testing is used when the number of target languages is small. For instance, van Bezooijen and van den Berg (1999) studied the intelligibility of four Dutch and one Frisian varieties to Standard Dutch listeners; Gooskens (2007) determined mutual intelligibility among three West-Germanic languages (Frisian, Dutch, Afrikaans). When the set of languages exceeds ten, only opinion tests are employed (e.g. Gooskens and Heeringa, 2004 for 15 Norwegian varieties, Tang and van Heuven, 2007 for 15 Sinitic varieties). It is unknown at this time to what extent opinion testing may serve as a feasible alternative to functional intelligibility testing in the area of language variation studies. One important aim of the present article is to address this issue.

We now describe the mainstream view on the grouping of Sinitic languages. We will then describe the procedures we followed to establish mutual intelligibility among a set of 15 such languages. We collected functional intelligibility scores at the word and sentence level, and compare these with each other and with opinion scores obtained earlier for the same set of 15 languages (reported in Tang and van Heuven, 2007). We will then decide to what extent opinion scores may serve as an acceptable substitute for functional intelligibility testing. In order to do so we will evaluate the functional and opinion scores against traditional dialect taxonomies proposed by Chinese linguists.

1.2. The Chinese language situation

1.2.1. Taxonomy of Chinese dialects

Basically, there are six language phyla within China, viz. Sino-Tibetan, Austro-Tai, Austronesian (sub-)phylum, Altaic, Austro-Asiatic and Indo-European (Lee, 1987:A-1). The Sinitic stock (under the Sino-Tibetan phylum) comprises eight (super)groups and some unclassified language varieties: Mandarin, Min, Yue, Wu, Hakka, Gan, Jin, Xiang. The Sinitic stock is one of the few outside the languages of the Indo-European phylum that have a long tradition of linguistic scholarship of its own. In this stock, there are approximately 1500 recorded language varieties (Campbell, 2004; see http://www.glossika.com/en/dict/faq.php#1).

The classification of the Chinese varieties is tentative and still controversial. Based on the characteristics of the phonological features and the tone evolution of the Chinese dialects, various classifications were proposed by Chinese dialectologists. Basically, Chinese dialectologists agree that there is a primary split in the Sinitic varieties into a Mandarin branch and a Southern branch comprising a number of non-Mandarin (super)groups. Sources do not agree which varieties should be assigned to which of the two primary branches, nor is there agreement on the internal structure of the two main branches.

As far as our 15 target varieties are concerned, we considered two taxonomies of Chinese varieties proposed by linguists. The taxonomy in Fig. 1 is based on a map called ‘‘Chinese linguistic groups’’ (http://www.chinadata.ru/

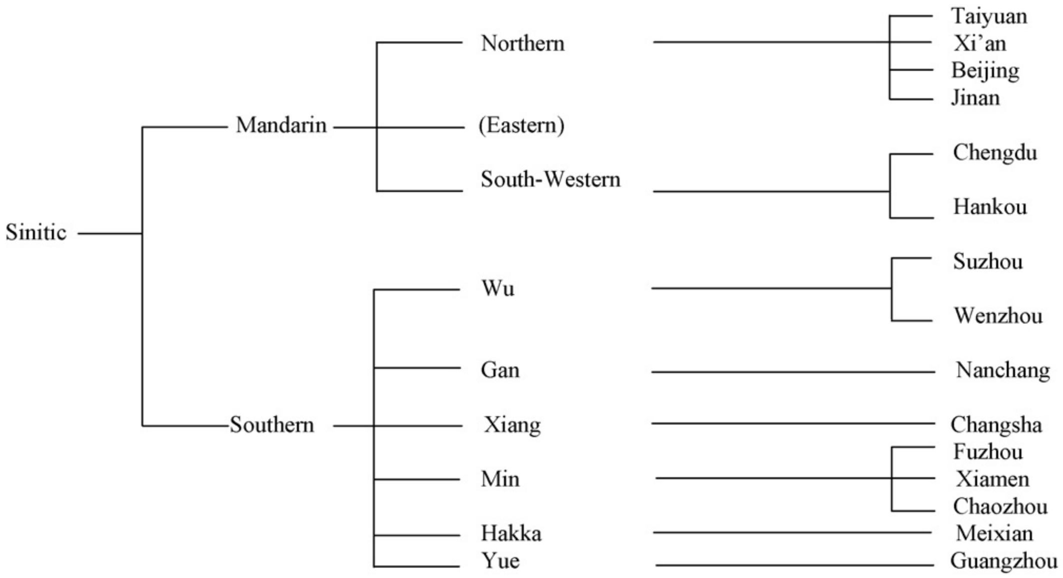

Fig. 1. Dialect taxonomy based on ‘‘Chinese Linguistic Groups’’. The subbranch in parenthesis is not represented in our 15-dialect sample.

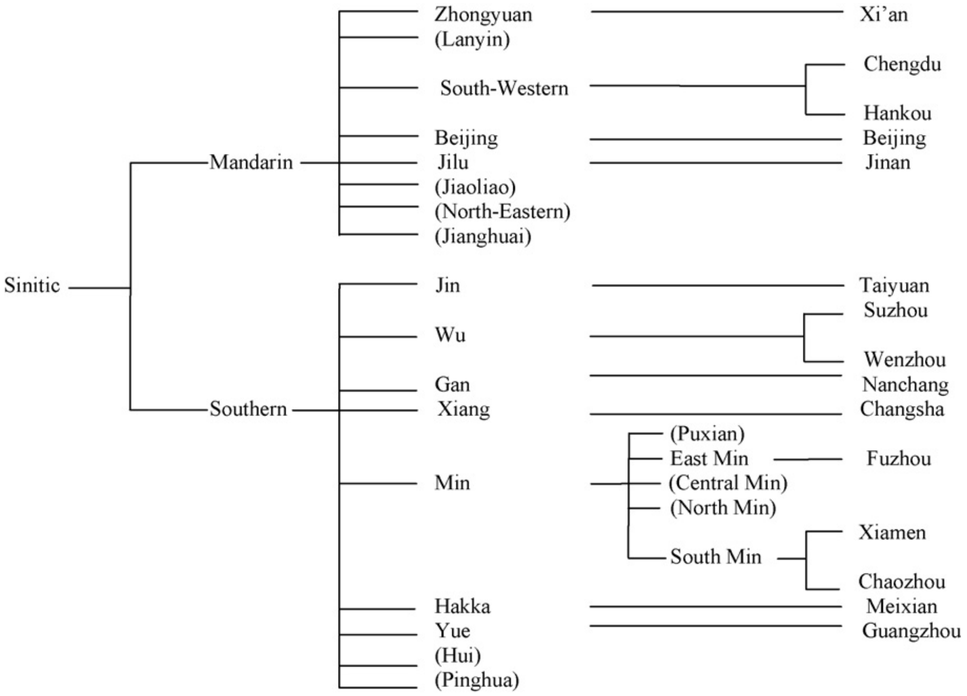

Fig. 2. Dialect classification based on *the Language Atlas of China*. (Sub)branches in parentheses are not represented in our 15-dialect sample.

linguistic_group_map.htm; for a reproduction of the map see Tang and van Heuven, 2007). The other one, in Fig. 2, follows a proposal by Li (1987:A-1, A-2) made in the *Language Atlas of China* (Wurm et al., 1987).

By and large, the dialect map published on the internet is a simplified version of the more detailed proposal by Li (1987a, 1987b). For instance, the internet map groups Xi'an, Beijing and Jinan together as Northern Mandarin varieties, whilst Li considers each of these varieties to be instances of separate branches within the Mandarin family, i.e. Zhongyuan, Beijing and Jilu, respectively. Similarly, Li sets up a number of subgroups within the Min group, grouping Xiamen and Chaozhou as varieties of South Min, and Fuzhou as an instance of East Min. Such subdivisions within the Min group are not made in the internet map.

There is, however, one major discrepancy between the two taxonomies. It concerns the status of Taiyuan. In the internet map it is grouped with the Northern Mandarin varieties, together with Xi'an, Beijing and Jinan. In Li (1987a, 1987b), Taiyuan is regarded as a variety within the Southern Sinitic branch, more specifically as an instance of a Jin variety. There are some differences that separate Jin varieties from the Mandarin group. For the moment, we will leave the status of Taiyuan undecided; later, we may have occasion to settle the issue on the basis of our own experimental data.

1.2.2. The impressionistic claims of the mutual intelligibility between Sinitic varieties

Varieties within the Mandarin branch are often claimed to be intelligible to each other to some extent, but are not mutually intelligible to varieties in the Southern branch (despite the recent influence of Standard Mandarin). Most varieties in the Southern branch are mutually unintelligible either to each other or to the Mandarin varieties. However, some cross-group intelligibility has been claimed in exceptional cases. For instance, Xiang varieties (belonging to the Southern branch) may share common terms and some degree of intelligibility with South-Western Mandarin varieties.

Generally, mutual intelligibility is held to be much poorer for varieties within the Southern branch than between those within the Mandarin branch. Many reasons may explain the mutual (un)intelligibility between Chinese varieties. For instance, the northern part of China is situated on the plains, affording easy travel, whilst the Southern part is very mountainous and difficult to travel through. Accordingly, there may have been less language contact between Southern varieties, which circumstance does not foster mutual intelligibility. It is, however, not the aim of the present paper to explain the reasons why the various varieties spoken in China grew apart to different degrees. What we are interested in, is their degree of mutual intelligibility and whether we can predict the degree of mutual intelligibility from objective, structural differences between the varieties (such as the number of cognates shared by two varieties, and the transparency of the phonological differences between two varieties).

1.3. Research aim

We are interested in determining more precisely to what extent a selection of Chinese varieties are mutually intelligible. Therefore, we aim to determine the degree of mutual intelligibility between pairs of Chinese varieties through experiments using opinion tests (done already; see Tang and van Heuven, 2007) and functional tests (present study). We will correlate the two types of subjective results with each other and with two objective structural measures (e.g. lexical similarity and phonological correspondence), and see how well we can predict mutual intelligibility from objective measures. Finally we will compare our experimental data with traditional dialect taxonomy and determine which subjective tests concur best with the traditional classification of Chinese varieties.

1.4. Previous experiments

We aim to establish the mutual intelligibility between these Chinese varieties through experimental means and then predict these results from structural measures (data obtained from Cheng, 1997) of similarity between all pairs of varieties in the set of 15, which yields 225 pairs of comparisons.

In previous experiments, we used opinion tests to obtain judged similarity and judged intelligibility of our Chinese varieties by asking naïve raters for their intuitive judgments after listening to readings of the fable *The North Wind and the Sun* spoken in these varieties. The results showed that judged similarity and judged intelligibility are highly correlated. Moreover, the two objective measures of structural similarity (lexical and phonological similarities) were always significantly correlated with the two subjective rating measures (Tang and van Heuven, 2007; see also Table 4 in the present paper).

1.5. Functional tests

In addition to the opinion tests, we also want to know to what extent mutual intelligibility of our 15 target Chinese varieties can be determined by functional tests. Furthermore, what is the correlation between functional test scores and the scores obtained from the earlier opinion tests? How much do the functional and judgment results overlap with or deviate from the results from various structural measures? Therefore, we aim at functionally testing how well a listener of language variety A actually understands a speaker of variety B (and vice versa). Specifically, we are interested in the percentage of words correctly translated from variety A to variety B and vice versa. In order to obtain experimental data, we designed two tests: one at the level of isolated words, the other at the sentence level.

The word-intelligibility test was developed by ourselves. It affords fast and economical testing of the recognition of a large number of isolated words. Target words are not translated; instead recognition is tested through semantic multiple-choice categorization. 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 categorization can only be achieved if the listener correctly recognizes the target words. Since there are as many as ten semantic categories, the role of guessing should be negligible.

Word recognition in sentence context was tested by a Chinese version of the SPIN (‘*S*peech *P*erception *i*n *N*oise’) test, which was originally developed for English by Kalikow et al. (1977). In the SPIN test the listener writes down only the last word in a number of short spoken sentences. In the materials we used, the identity of the final word was largely predictable from the earlier words in the sentence, so that this test addresses the efficient interaction of bottom-up (information from the speech signal) and top-down (expectations derived from earlier context) processes in continuous speech recognition. Earlier work has shown that this type of test is highly sensitive to differences in intelligibility due to different language backgrounds of speakers and listeners (Wang, 2007).

One additional question that we hope to answer on the basis of the present study, is to what extent the recognition of isolated words (bottom-up information only) and of words in context (interaction of bottom-up and top-down information) are predictable from each other. If recognition of words in context is largely predictable from isolated-word recognition scores, the latter type of test will suffice for future work on functional mutual intelligibility testing.

2. Methods

2.1. Preparing the recordings

2.1.1. Preparing the materials for recording; word and sentence selection

For the word part, we prepared a list of 288 standard Mandarin core words. These words are frequently used in daily life forming such categories as body part, family member, plant, fruit, house furnishing, article of clothing, word for orientation in time and space, animal, etc. The words all denote simple concepts commonly used in everyday life and thus they are assumed to be used in each of our 15 target Chinese varieties. We tried to avoid words with the same morphemes (Standard Mandarin-oriented only) in order to obviate priming effects (see introduction).

For the sentence part, we selected 70 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.[1] Recognizing the final word is easier if the listener also correctly recognizes the earlier words in the sentence, as in *He wore his broken arm in a sling* (target underlined). The 70 sentences were selected on the basis of their applicability to the Chinese linguistic/cultural situation, and translated into Standard Mandarin. We selected sentences that maintained the structure of the SPIN sentences such that each Mandarin sentence ended with in final noun as it does in English.

2.1.2. Sound recordings

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 Jiaotong 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 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. Speakers were selected for a good-quality speaking voice. We have no reason to believe that some speakers articulated more poorly than others, so that the effect of dialect on intelligibility should be much larger than the effect of speaker-individual characteristics.[2]

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.

Using Adobe Audition running on a notebook computer, the words and sentences were then read from paper and recorded by the 30 speakers in individual sessions. Speakers were seated in a quiet office and wore a Shure SM10A head-mounted close-talking microphone. Each speaker read both the word part and sentence part in their own variety (instead of Standard Mandarin) using the translations they had prepared themselves.

2.2. Listening test

2.2.1. Data segmentation and processing

For the word part, we extracted 150 words (from the original set of 288) in ten lexical categories (eight main categories, two of which were subdivided):

[1] There are two types of materials in the SPIN test. We only used the part that presents target words that are highly predictable from the earlier context (H sentences). We did not use the part with words that are not predictable from the context (L sentences)), as in *We could have discussed the dust* (target underlined). Wang (2007) showed that the H part of the SPIN test was more sensitive to differences between speaker and listener groups with different degrees of listening comprehension in English.

[2] To support this claim we determined speech rate for the set of 60 sentences read by each individual speaker and correlated this with the intelligibility score obtained for the speaker by listeners of the very same dialect. Normally, a speaker should be less intelligible as the speaking rate goes up, so that we would predict a negative correlation. We obtained a correlation coefficient of $r = 0.192$ ($N = 30$, $p = 0.310$, two-tailed), indicating that the correlation is insignificant, and that even if it were significant, the relationship between speaking rate and intelligibility would run counter to the hypothesis. We also computed the correlation between the intelligibility scores of the male-female speaker pairs, and found $r = .755$ ($N = 15$, $p < .001$, one-tailed). This fairly high correlation indicates that although there is individual speaker variation, the language effect by itself explains more than 50 percent of the variance.

1. Body parts
2. Plants
   a. Sweet fruits/nuts
   b. Vegetables
3. Animals
   a. Four-legged
   b. Other (animals)
4. Textiles/fabrics/articles of clothing, apparel
5. Orientation in time/space
6. Natural phenomena
7. Perishables (food/drinks other than fruits and vegetables)
8. Verbs of action/things people do.

Appendix 1 presents the list of 150 target words (in Mandarin only), in characters and in Pinyin (Romanized Mandarin phonological spelling plus tones), glossed and subdivided into the ten semantic categories. For the sentence part, we finally selected 60 sentences from 70. A full list of sentences in (Standard Mandarin only), in Chinese characters and in Pinyin (including tones), and English glosses is given in Appendix 2.

2.2.2. *Creating CDs*

The intelligibility tests basically require word recognition. In word recognition tests it is imperative that a listener does not hear the same word (or morpheme) twice. A word (or morpheme) which is heard for the second (or third, fourth) time within an interval of up to a day) is recognized more successfully than the first time (e.g. Morton, 1969). In order to prevent such priming effects, the stimulus words and sentences have to be blocked over listeners, such that each listener hears each word only once, irrespective of the variety of the speaker. Therefore, we worked out a completely balanced word and sentence stimulus order using a Latin Square design (e.g. Box et al., 1978). On the first CD (CD1) the 150 words were placed in a fixed random order (from nr. 1 to nr. 150). Every following word was spoken in a different variety, so that every variety was represented by 10 words. On the second CD (CD2) the words were presented in the same order with the exception that the presentation began with word nr. 150 which was then followed by words nr. 1 to nr. 149. As a result of this shift, every word on CD2 was spoken in a different variety than on CD1. On the third CD (CD3) the first item was word nr. 149, the second was nr. 150, followed by words nr. 1 to nr. 148, and so on for CDs 3 to 15. Again, every word on CD3 was spoken in a different variety than on the earlier CDs. CD15 started with word nr. 137, followed by words nr. 138 to nr. 150, and then followed by words nr. 1 to nr. 136. Through this rotation scheme we ensured that (i) each listener heard each of the words and sentences only once, (ii) each of the 15 listeners in one variety group heard each version of a word in a different variety, while (iii) at the same time every listener heard one-fifteenth of the materials in each of the 15 varieties (stimuli were blocked over listeners in a Latin square design).

Note, finally, that it was not possible to divide the materials evenly between male and female speakers in each variety, since 15 is an odd number. In order to solve this small imbalance, half of the varieties were represented by 8 male and 7 female speakers, whilst the other half of the varieties were represented by 7 male versus 8 female speakers.

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 seconds and with 11-second pauses between tracks. As a consequence, each CD contained 28 tracks including spoken instructions at the beginning, in the middle and at the end plus practice tracks containing 10 words and 10 sentences, respectively. Practice items were sampled from additional materials that were not selected as proper stimuli.

2.2.3. *Answer sheets*

For each CD, we prepared an answer sheet to match the corresponding stimulus tracks.

There were 15 blocks of word stimuli and six blocks of sentence stimuli. For each block of words, ten stimulus words were required to be categorized into one of the designated ten semantic categories. The categories were listed across the page. Listeners were asked to tick the appropriate box for each successive stimulus. The categories were repeated after every ten lines. For each block of sentences, the final words for each of ten stimulus sentences had to be written down in the listener's own dialect.

Table 1
Summary of listener characteristics broken down by dialect group.

Dialect	Age		*N* males	Education	Dialects	Standard Mandarin	
	Mean	S.D.				Understanding	Speaking
Suzhou	44.20	3.59	7	2.27	1.07	2.67	0.87
Wenzhou	45.67	3.83	8	1.47	1.13	1.93	0.73
Guangzhou	46.67	3.77	8	2.20	1.13	2.67	0.93
Xiamen	45.47	13.81	10	1.20	1.00	0.73	0.40
Fuzhou	47.53	5.58	8	1.60	1.00	1.93	0.53
Chaozhou	49.33	6.95	8	0.73	1.00	0.87	0.13
Meixian	47.93	6.97	9	2.10	1.00	2.44	0.44
Nanchang	36.33	7.68	8	2.07	1.00	2.73	0.87
Changsha	48.33	4.94	7	1.73	1.00	2.27	0.20
Taiyuan	44.07	5.71	5	2.33	1.00	3.00	0.80
Beijing	42.20	4.36	9	2.87	1.00	3.00	1.00
Jinan	51.20	4.11	7	2.40	1.13	2.73	0.33
Hankou	46.80	4.96	8	0.67	1.00	2.27	0.33
Chengdu	42.67	14.88	6	3.80	1.00	2.80	1.00
Xian	48.53	4.10	7	2.93	1.00	3.00	0.87

Mean and Standard deviation of age in years. *N* males = number of male listeners (out of 15). Education (highest level attained): 0 = none at all, 1 = primary school, 2 = junior middle school, 3 = senior middle school, 4 = vocational college, 5 = university undergraduate, 6 = university graduate. Dialects = number of dialects spoken. Understanding of Standard Mandarin: 0 = not at all, 1 = poor, 2 = moderate, 3 = good. Speaking Mandarin: 0 = no, 1 = yes. A horizontal double line divides the table into nine Southern and six Mandarin dialects.

2.2.4. Procedure

For each variety in the set of 15, a local contact person was contracted. In ten cases the local contact had also served as one of the two speakers of the dialect materials we used as stimuli. In the case of five other varieties neither the male nor the female speaker could make a trip to their dialect area, in which case we asked another contact person, one whom we had used in our earlier study (the opinion experiment).[3] Each local contact, a native speaker of the dialect of the listener group targeted, recruited 15 native listeners of the dialect s/he represented. Ideally, the listeners should be selected from the larger groups of 24 subjects who participated in our first experiment (Tang and van Heuven, 2007). For the present experiment, however, subjects had to be literate–so that some substitutions had to be made. All local contact persons and the listeners were compensated for their services.

The contact persons were instructed to enlist listeners who were monolingual rural dialect speakers in the age bracket between 40 and 65 and who had not traveled much and had never lived outside their own province. As a result most listeners belonged to the lower working class with fairly low level of education and professions of low status. Listeners filled in a questionnaire asking them about their language background, familiarity with other Chinese varieties, and some demographic details. A summary of the responses to the questionnaire is given in Table 1. There was a roughly equal split between male ($N = 115$) and female ($N = 110$) listeners. The mean age was well above 40 for most dialects; the Nanchang listener group, however, had a mean age of 36. With very few exceptions (seven listeners out of 225, and never more than two in one dialect group) all listeners declared to be monodialectal. Nevertheless, a majority of the listeners claimed to be able to speak Standard Mandarin (63%, including the 15 Beijing listeners), and most listeners claimed to be able to understand Standard Mandarin to a greater or lesser degree. This may have implications for the interpretation of the results of this study. We will return to this issue in later sections.

Listeners took part in the experiment in individual sessions. Each listener in a dialect group listened to a different CD, one of the set of 15. All listeners were required to both read the paper instructions and to follow the instructions

[3] In the case of these five contact persons, there may have been a difference between the exact town or village of the speaker of the dialect sample and that of the listeners recruited by the contact person. Due to this circumstance five listeners possibly may have listened not to their very own dialect but to a neighboring dialect within the same group. These five dialects are Nanchang (Gan family), Fuzhou (Min family), Xi'an (Northern Mandarin), Taiyuan (Northern Mandarin), and Chengdu (Southwestern Mandarin). Results show that, indeed, these five listener groups got poorer scores when responding to their ''own'' dialect than the other ten groups did. The mean word scores were 63 versus 55% correct, whilst the sentence scores were 89 versus 75%. The former difference is not significant by a paired t-test, $t(13) = 0.9$ ($p = 0.173$, one-tailed) but the latter is, $t(13) = 1.8$ ($p < .050$, one-tailed).

spoken on the CD (in Mandarin). Stimuli were presented through twin loudspeakers in a quiet room, often in the contact person's private home, using either a computer or a stereo set.

The isolated-word recognition task was presented first. Here, the listener was required to tick one of ten boxes for each word representing the ten semantic categories/subcategories (see Section 2.2.1) every time a word was presented. For the subsequent sentence part, the listener had to write down the target word in their own dialect after listening to each of the 60 sentences on the CD. Whether the target word was in final or in pre-final position was indicated explicitly for each sentence on the listener's answer sheet.

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

In all, we collected 33,750 responses (15 × 150 × 15) for the word part and another 13,500 (15 × 60 × 15) for the sentence part. The dependent variable in the word-intelligibility test was the choice of semantic category. This choice was coded with a value from 1 to 10 and entered in a database, along with information on the dialect of the listener, dialect of the speaker and on the semantic category of the stimulus word. From this information we computed a mean percentage of correctly classified words for each of the 15 × 15 combinations of speaker and listener dialects, yielding 225 mean word recognition scores (see Table 2).

For the sentence intelligibility test, the procedure was less straightforward. The first author manually checked whether the sentence-final target word was correctly translated back into Mandarin by the local contact person. If the translation was semantically equivalent to the target specified for the item, the response was considered correct. If the translation was incorrect or if no translation was given at all, the response 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 (see Table 3).

We now first describe the analysis of the results for the word-intelligibility test (Section 3.1), and defer the presentation of the results of the sentence intelligibility test to Section 3.2.

3.1. Isolated-word-intelligibility test

Table 2 presents the mean percentage of correctly classified (recognized) words for each combination of speaker dialect (listed in the rows) and listener dialect (listed column-wise).

We expect the highest scores in the cells along the main diagonal in Table 2. These are the scores obtained by listeners who listened to speakers of their own variety. Scores in off-diagonal cells should be poorer, as these cell means are based on listeners listening to speakers of a different variety. Indeed, generally we do find the highest correct classification scores in the diagonal cells. The highest percentage correctly classified words is between Beijing speakers and listeners; Beijing listeners correctly recognized 83 percent of the words spoken in their own variety and classified them into the right categories, the listeners of Jinan and Hankou varieties recognized the speakers of their own varieties as high as 80% and 81%, respectively. Other listener groups were less successful. For instance, Xiamen and Nanchang listeners could not understand the speakers of their respective varieties 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. This is the case for the native Nanchang group, which is outperformed by no fewer than seven non-native listener groups; for the Xi'an group, which is outperformed by four non-native groups and for the groups of Changsha and Taiyuan, which are respectively outperformed by two other groups.[4] Typically, listeners whose native variety belongs to the Mandarin group were more successful in the word classification task (mean across the six Mandarin varieties is 72% correct) than the listeners with Southern native varieties (mean correct classification is 52%). The difference in mean percent correct between Mandarin and Southern listener groups is highly significant, $t(13) = 3.1$ ($p = 0.008$, two-tailed).

Table 2 also shows that, across listener groups, Beijing speakers are understood clearly better (61%) than any other type of dialect speaker. There is a gap of 13 percentage points between Beijing speakers and the next best speaker

[4] These four groups are among the set of five for which the speaker of the dialect materials did not hail from exactly the same town or village as the listeners (see further Note 3).

Table 2
Percentage correctly classified words broken down by speaker and listener dialect. Each mean is based on 150 responses. Double lines separate Mandarin from Southern dialects.

Speaker dialect	Listener dialect															
	Suzhou	Wenzhou	Guangzhou	Xiamen	Fuzhou	Chaozhou	Meixian	Nanchang	Changsha	Taiyuan	Beijing	Jinan	Hankou	Chengdu	Xi'an	Mean
Suzhou	**65**	20	25	17	21	15	23	22	23	29	26	29	39	28	29	27
Wenzhou	23	**41**	17	19	17	17	18	21	15	24	25	25	28	18	19	22
Guangzhou	23	18	**55**	25	25	29	40	21	19	33	34	33	38	25	29	30
Xiamen	20	14	23	**39**	19	25	19	19	12	18	19	25	26	17	16	21
Fuzhou	17	18	17	18	**47**	14	17	16	15	22	20	23	24	20	16	20
Chaozhou	18	12	23	22	23	**68**	15	10	15	23	27	29	24	24	23	24
Meixian	31	24	35	24	23	25	**67**	31	27	43	43	43	41	37	31	35
Nanchang	27	26	30	25	29	22	41	**37**	29	47	51	48	57	41	42	37
Changsha	31	22	31	24	31	20	34	31	**48**	47	49	47	60	38	43	37
Taiyuan	33	30	30	29	31	21	36	36	30	**57**	59	64	55	50	48	41
Beijing	64	41	63	45	53	38	61	51	54	76	**83**	74	72	65	70	61
Jinan	40	22	31	22	36	19	39	39	31	59	61	**80**	58	51	55	43
Hankou	37	29	33	28	41	22	42	33	35	63	59	67	**81**	53	47	45
Chengdu	28	24	30	32	35	19	49	36	38	62	59	61	70	**72**	56	45
Xi'an	47	36	43	27	35	23	48	43	47	63	64	67	65	55	**59**	48
Mean	34	25	32	26	31	25	37	30	29	44	45	48	49	40	39	

Table 3
Percentage correctly translated target words in sentences broken down by speaker and listener dialect. Each mean is based on 60 responses. Double lines separate Mandarin from Southern dialects.

Speaker dialect	Listener dialect															
	Suzhou	Wenzhou	Guangzhou	Xiamen	Fuzhou	Chaozhou	Meixian	Nanchang	Changsha	Taiyuan	Beijing	Jinan	Hankou	Chengdu	Xi'an	Mean
Suzhou	**77**	7	5	18	13	5	7	13	13	20	5	18	15	15	7	16
Wenzhou	5	**93**	5	12	3	2	7	10	2	7	2	10	8	7	2	10
Guangzhou	5	7	**92**	10	20	25	55	22	13	7	3	22	8	17	7	21
Xiamen	13	5	8	**97**	23	28	13	18	13	3	5	15	7	17	8	18
Fuzhou	3	3	2	17	**92**	7	3	8	5	0	0	7	2	0	3	10
Chaozhou	7	0	3	52	13	**98**	3	12	3	7	2	13	10	3	5	15
Meixian	13	2	12	28	17	20	**70**	25	18	10	3	25	15	25	8	19
Nanchang	28	13	20	25	27	17	33	**50**	32	35	18	53	43	37	23	30
Changsha	12	3	8	23	17	3	17	25	**93**	13	13	38	53	28	2	23
Taiyuan	63	35	45	63	57	25	55	68	68	**73**	77	92	92	85	73	65
Beijing	87	62	90	90	93	60	80	78	92	90	**98**	98	97	98	93	87
Jinan	52	27	32	48	48	15	40	60	70	75	77	**97**	83	82	67	58
Hankou	48	32	32	52	53	27	45	53	62	58	67	95	**100**	73	65	57
Chengdu	47	22	40	48	72	27	48	58	62	65	62	98	95	**95**	68	60
Xi'an	53	33	50	58	57	30	57	58	63	68	58	82	78	70	**67**	59
Mean	34	22	30	43	40	26	36	37	41	35	33	51	47	43	33	

dialect (Xi'an with 48%). It seems reasonable to assume that the advantage of Beijing speakers is due to the circumstance that Beijing dialect is very similar to Standard Mandarin. Listeners in every part of China have been exposed to the standard language through education and the media. Beijing listeners, however, have no advantage of their dialect being similar to the standard language. The asymmetry does not invalidate the comparison between results obtained from opinion tests and from functional tests, nor does it affect the comparison of word and sentence intelligibility, because the asymmetry affects all these results to the same degree. It could affect the structure of agglomeration trees to be drawn on the basis of matrices such as the one in Table 2. To check for this possibility we generated trees based on matrices with and without the rows and columns representing Beijing speakers and listeners. Differences were minimal, and did not affect the basic splits in the tree.

The data in Table 2 were then used to generate a dendrogram, using the average linking method that we also used in our earlier report (Tang and van Heuven, 2007). The tree structure that was generated is displayed in Fig. 3.

The tree makes a primary split between the six Mandarin varieties, and a group of nine varieties which comprises all the Southern dialects. This division concurs well with traditional taxonomies postulated by Chinese dialectologists. We will delay more detailed discussion of the internal cluster structure within the main branches until Section 6.

3.2. Sentence intelligibility test

Table 3 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.

The range of sentence scores is larger than that for semantic categorization (from 0 to 100%), and the mean scores for understanding one's own dialect are much better than that for semantic categorization (see the diagonals). It appears from the table that this sentence-level test was an easier task than the semantic categorization task with isolated words in the preceding section.

On three occasions native listener groups are outperformed by non-native groups. This occurs in the Mandarin part of the table only, where native Taiyuan listeners happen to do as well as the Xi'an group and do more poorly than all other Mandarin groups. Chengdu native listeners do more poorly than two other groups, and Xi'an native listeners are second to four other groups.

Again, we observe that the Beijing speakers were better understood across all listener groups than any other type of dialect speaker. The difference between Beijing speakers (87% correct) and the next best speaker type (Taiyuan, 65%) is 22 percentage points. However, Beijing listeners had no advantage over other listener groups. In this respect the data of the word and sentence intelligibility tests concur. We ascertained that the structure of the agglomeration tree derived from the matrix in Table 3 was the same, whether or not Beijing speakers and listener were included in the analysis.

Using the same procedure as in Section 3.1, we generated a hierarchical cluster tree for the sentence-intelligibility results. The resulting tree structure is presented in Fig. 4.

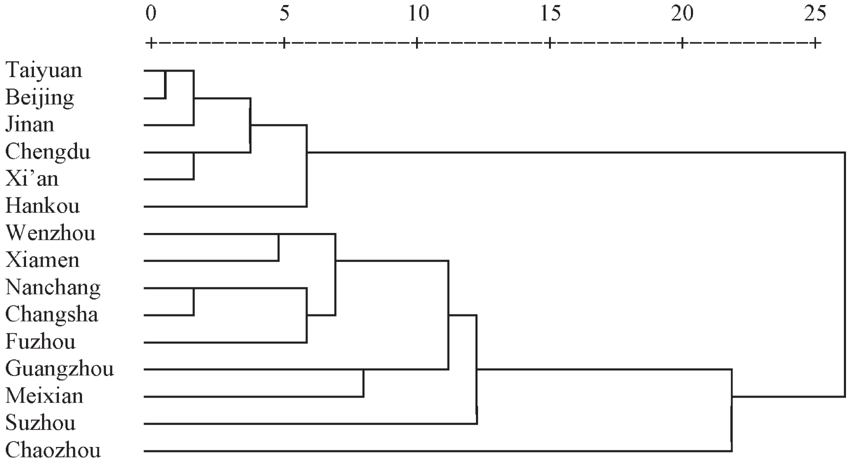

Fig. 3. Dendrogram (using average linking between groups) based on word-intelligibility scores obtained for all 225 combinations of 15 speaker and 15 listener dialects.

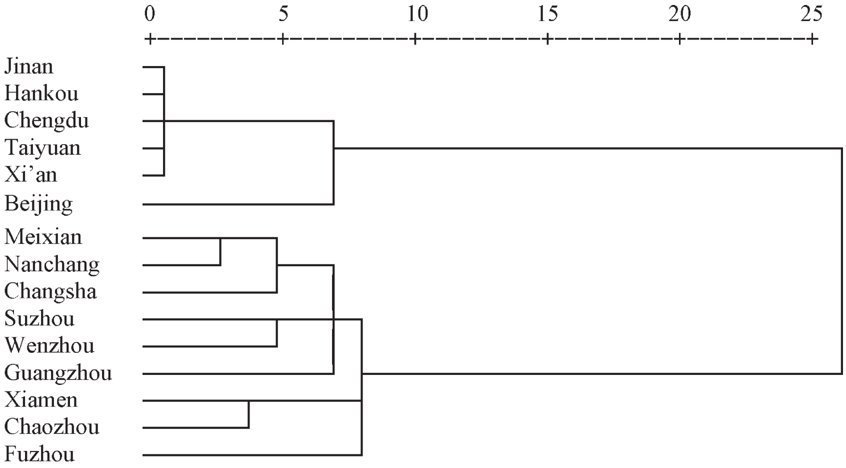

Fig. 4. Dendrogram (using average linking between groups) based on sentence-level intelligibility scores obtained for all 225 combinations of 15 speaker and 15 listener varieties.

The sentence-level tree shows, again, a clean cut between the six Mandarin and the nine Southern varieties. As before, we will not deal with the internal structure of the dialects within the main branches. This matter will be taken up in Section 6 where we will compare the clustering of the dialects in the trees above with the dialect taxonomy proposed by linguists. First, however, we will consider the question how well the functionally determined word and sentence intelligibility scores can be predicted (in Section 4) from our earlier judgment scores (on intelligibility and on linguistic distance) and (in Section 5) from objective statistical properties computed on the lexicons of the dialects.

4. Correlations between subjective measures

So far we have obtained two kinds of subjective data experimentally; one is from the opinion tests, the other is from functional tests. Within the first type (see introduction) we distinguish between judgments of (i) intelligibility and (ii) similarity between varieties. In the second type we distinguish between functional intelligibility (iii) at the word level and (iv) at the level of the sentence. In the next sections we will consider the correlation structure in this set of variables. We will first examine, in Section 4.1, the correlation between (iii) and (iv) on the basis of the data collected in the present study. In Section 4.2 we will see to what extent the opinion scores are correlated with the functional scores.

4.1. 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. In order to quantify the degree of correspondence between the two methods of functional intelligibility testing, i.e. using isolated words versus words in sentence context, we established the correlation coefficient for all cells (including those on the main diagonal) in Table 2 and the corresponding cells in Table 3.

Fig. 5 presents a scatterplot of the word (horizontal axis) and sentence-level (vertical axis) intelligibility scores. The correlation is high, viz. $r = 0.835$ (N = 225, $p < 0.01$)

In the introduction we defined mutual intelligibility between two language varieties A and B as the mean of the intelligibility of A to B and of B to A. Accordingly, we also 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 higher correlation coefficient, $r = 0.928$ ($N = 105$, $p < 0.001$). The coefficient of determination is $r^2 = .86$, which means that 14 percent of the variance is left unaccounted for.

It seems unclear, therefore, whether the word-intelligibility test (semantic categorization test) can be adequately used as a short-cut to functional intelligibility. For the moment we will assume that both the word-level and the sentence-level tests are needed. At some later stage, we compare the test results with external data (objective measures

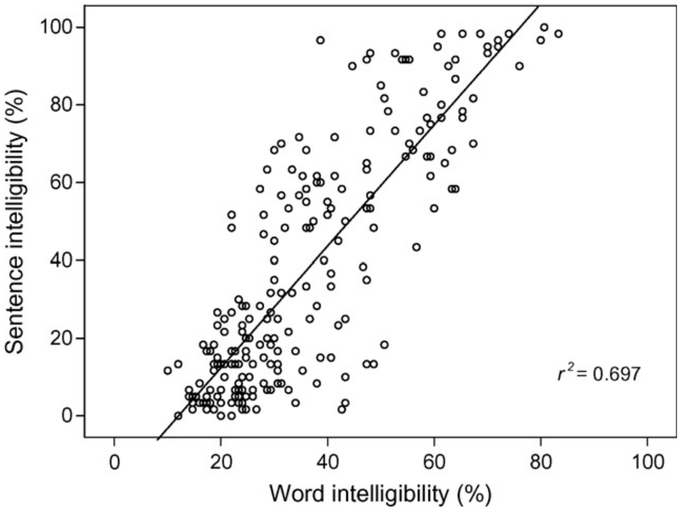

Fig. 5. Scatterplot generated from mean scores based on the isolated word-level and sentence-level.

Table 4
Correlation coefficients of the four subjective measures, using the symmetrical means (lower triangle, $N = 105$) of the matrix.

	Word-intelligibility	Sentence intelligibility	Judged intelligibility
Sent. intelligibility	0.928		
Judged intelligibility	0.772	0.818	
Judged similarity	0.738	0.779	0.888

Correlation is significant at the 0.001 level (2-tailed) for all entries.

of structural difference, traditional genealogies). We may then be able to choose between the two and consider one a more valid or representative measure of intelligibility than the other.

4.2. Functional tests versus opinion tests

In our earlier study (Tang and van Heuven, 2007), we collected opinion scores on intelligibility and on similarity between all pairs of our 15 varieties. The results revealed a very strong correlation between the two sets of opinion scores, especially when the correlation was computed on similarity matrices: $r = 0.949$ ($N = 105, p < 0.001$), which leaves only 10 percent of the variance unaccounted.[5] From this, we drew the provisional conclusion that the two opinion scores are practically interchangeable. We now determine to what extent the two sets of judgment data are correlated with intelligibility scores determined on the basis of functional test procedures. Obviously, opinion testing is much faster and easier to accomplish than functional testing. Therefore, if indeed the functional scores can be adequately predicted from the opinion scores, the latter type of testing will be preferred in the future–for reasons of economy.

Table 4 presents a correlation matrix for the four subjective measures at issue. From a range of opinion scores we selected the scores that yielded the clearest results in Tang and van Heuven (2007). These were the opinion scores for intelligibility and similarity based on readings of the *North Wind and the Sun* fable with melodic information left unaffected.[6]

We computed the correlation coefficients for all six combinations of the four functional and opinion scores. Correlation coefficients were 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 $\{j, i\}$.[7]

[5] When the complete, asymmetrical matrices are used to compute the correlation between judged intelligibility and judged similarity, we obtain $r = .854$ ($N = 225$, $p < .001$). This is still a high correlation but it leaves 27% of the variance unaccounted for.

[6] Judgment scores were generally lower, and less clearly structured, when monotonized versions of the fable were presented (using PSOLA analysis and resynthesis). The data on monotonized versions are omitted from the present paper in order to save space.

[7] This procedure is conceptually simpler than our earlier use of similarity matrices. However, it reduces the correlation between judged intelligibility and judged similarity from $r = 0.949$ to 0.888.

Table 5
Correlation coefficients between subjective measures and objective measures from the non-redundant part of symmetrical matrices.

	LSI, $N = 78$	PCI, $N = 105$	Both	
			R	R^2
PCI	0.727			
Judgment data				
Intelligibility	0.825	0.711	0.841	0.707
Similarity	0.823	0.737	0.848	0.719
Functional data				
Word Intelligibity	0.788	0.772	0.840	0.705
Sentence Intelligiblity	0.746	0.769	0.816	0.665

All entries: with simple correlations, $p < .001$ (2-tailed). All entries: with R, inclusion of second predictor significantly increases multiple R ($p < .05$).

The following observations can be made. First, the highest correlation coefficients are found between variables of the same type. That is to say, correlations between two opinion scores ($r = 0.888$) or between two functional test scores ($r = 0.928$) are better than correlations for cross-type test scores (from functional to opinion score or vice versa, all at $r < 0.820$). Within the range of r-values in the table with $N = 105$, differences between correlation coefficients larger than 0.1 are significant ($p < 0.05$, using Fisher's r to z transformation and assuming independent correlations).

Second, Table 4 shows that intelligibility judgments tend to be better predictors of the functional test scores than similarity judgments are. The difference between the coefficients, however, fails to reach statistical significance using Hotelling's t for correlated coefficients.

Third, functional intelligibility at the sentence level can be somewhat better predicted from opinion scores than at the word level but, again, the difference between the correlation coefficients is not significant (using Hotelling's t).[8]

5. Predicting subjective scores from objective measures of linguistic similarity

In Tang and van Heuven (2007) we used multiple linear regression to predict (subjective) opinion scores from two objective measures of linguistic distance, viz. a lexical similarity index (LSI, percent cognates shared between two varieties) and a phonological correspondence index (PCI) that expresses the phonological transparency between the shared cognates.[9] When two languages are related one expects to find two things. First, the languages will share a substantial number of cognate words (in a comparable lexical database), and, second, within the set of cognate word pairs one expects to find regular sound correspondences. No sound correspondences will be found between non-cognate words pairs representing the same concept. Both measures, LSI and PCI, potentially contribute to the prediction of (mutual) intelligibility scores or judged similarity between varieties.

In the present section we will present a similar regression analysis, in which we recapitulate earlier results for the opinion scores and extend the analysis to the newly collected functional intelligibility scores. We ask which of the two types of intelligibility measures are more amenable to prediction from objective measures, opinion test scores or functional test scores.

Table 5 presents correlation coefficients for combinations of objective measures of linguistic distance (LSI, PCI) and subjective, experimental measures, i.e. either opinion scores on intelligibility and similarity, or functional test scores of word and sentence intelligibility. As we did before, in Table 5 the correlation coefficients were computed on the non-redundant lower triangles of the matrix after that the matrix has been made symmetrical by averaging scores in contra-diagonal cells.

[8] Correlation coefficients were also computed for the original asymmetrical (full) matrices with and without data in the cells on the main diagonal. These correlation coefficients were lower but the relationships among them were the same as in the present data. For details see Tang (2009).

[9] The PCI (developed by Cheng, 1997, who confusingly called it ''Mutual Intelligibility''), is a number that captures the complexity of the rule system needed to transform the phonological shapes of cognates in variety A to their counterparts in B. The rule set that transforms A to B is different (and may be more or less complex) than the set that transforms B to A (i.e., the PCI is an asymmetrical distance measure). In this respect the PCI may be a more realistic measure of phonological distance between cognates than the often used Levenshtein string edit distance between phonetic transcriptions (as used by Kessler, 1995; Heeringa, 2004 and references therein). Cheng (1997) computed the LSI and PCI measures from a 2,700 word database of phonemically transcribed words in 17 Chinese dialects, including all our 15 dialects for PCI but lacking Taiyuan and Hankou for LSI.

Table 5 shows that, firstly, the two objective measures are fairly highly correlated, at $r = 0.727$. Secondly, both PCI and LSI are correlated to some extent with all subjective data (i.e. the judgment/opinion scores); interestingly, for any subjective parameter (whether judgment or functional test score), its correlation with LSI is generally better than with PCI. However, the difference is too small to reach significance in any of the four comparisons that can be made.

Within the category of functional test scores, it seems as though word-intelligibility scores can be predicted from objective measures ($r = 0.840$) better than the sentence intelligibility scores ($r = 0.816$). The difference between the R coefficients, however, is not significant.

6. Relating scores to linguistic taxonomy

In order to find out to what extent the subjective scores converge with the traditional linguistic taxonomy proposed by Chinese dialectologists, we generated hierarchical cluster schemas (trees) from the data matrices, using the average linkage method as before. The two trees based on opinion scores can be found in Tang and van Heuven (2007); the two trees based on our functional intelligibility tests are in Fig. 3 (word-intelligibility) and 4 (sentence intelligibility) of the present paper.

Observing the trees, we found that both trees based on functional word-intelligibility and sentence-intelligibility scores, at first sight, correspond rather well with the classification postulated by traditional dialectologists. The primary split in the trees is between a branch of the six Mandarin varieties on the one hand, and another branch of the nine Southern varieties on the other. The basic division, then, of Sinitic varieties into Mandarin versus Southern branches is correctly reflected in both the word-based and sentence-based trees.

In terms of substructure, however, the word-based tree is not very credible. The clustering of the varieties in the Southern branch differs substantially from what we would expect on the basis of dialectological taxonomy. Within the Southern branch we would like to see identifiable clusters representing the Wu dialects (Suzhou, Wenzhou) and the Min dialects (Fuzhou, Xiamen, Chaozhou). This substructure, however, is not reflected in the Southern branch of the word-based tree at all. There is not a single pair of varieties in the Southern branch that are grouped together the way they should according to the traditional linguistic taxonomy. Similarly, within the Mandarin branch, Taiyuan, Beijing, Xi'an and Jinan should form at least a ('Northern') cluster (see Fig. 1)–which is not found in the tree. To be true, Taiyuan and Beijing are grouped together, not with Xi'an but with Jinan only. Unfortunately, Xi'an, which should have grouped with the Northern subgroup is incorrectly parsed with the South-Western cluster, which should have been comprised of Chengdu and Hankou only. That is, Chengdu and Hankou do not form an independent South-Western subcluster. Instead, Hankou clusters with Xi'an, and Chengdu is added to the cluster as an isolate.

Our preliminary conclusion is that the word-intelligibility based tree only correctly reflects the primary split into Mandarin and Southern branches but otherwise shows little correspondence with traditional dialect taxonomy (see Figs. 1 and 2).

Now let us consider the tree based on sentence-intelligibility scores. This tree, too, reflects the split between the Mandarin and Southern branches, represented by six and nine varieties, respectively. This agrees with the traditional classification of Chinese dialects. Moreover, within the Southern main branch we now find clearly identifiable clusters for the Wu group (Suzhou, Wenzhou), and for the Min group. In the latter, Xiamen and Chaozhou form the South Min subcluster while Fuzhou (representing East Min) is added later. This means that even the substructure within the Min family is correctly captured by the sentence-intelligibility scores.

The other four groups within the Southern branch are represented by just one variety each of these groups. Yue (represented by Guangzhou), is relatively isolated (as it should be) but the other three groups are joined into clusters at rather low levels: first Hakka (Meixian) and Gan (Nanchang) are clustered, then followed by Xiang (Changsha).

The structure found within the Mandarin branch partly corresponds with the traditional genealogy postulated by dialectologists. Indeed Chengdu and Hankou are dialects of the Mandarin branch, but they should form a cluster of the South-Western Mandarin subgroup while they actually did not. The internal structure of the Northern and South-Western Mandarin subgroups is not reflected in the tree. Instead, Jinan, Hankou, Chengdu, Taiyuan and Xi'an seem to be rather closely related as they are joined into clusters at a very low level. The dialects of Jinan, Taiyuan, Xi'an should have clustered together with Beijing to form the Northern Mandarin sub-group. Unfortunately such a cluster is not found and Beijing assumes a special position, since it is joined to the other five varieties at a higher level on the tree.

However, the Mandarin sub-groups in the sentence-intelligibility tree are better reflections of the cladistic taxonomy postulated by Li (1987a, 1987b). In the *Language Atlas of China,* Beijing, Xi'an, Jinan are representatives of sub-groups of Beijing, Zhongyuan, Jilu, respectively (see Fig. 2). All of these three sub-groups plus the South-Western subgroup (comprising Chengdu and Hankou) form the Mandarin branch of Sinitic varieties. By and large, the traditional genealogical relationships postulated among the Southern varieties are adequately reflected in the sentence-intelligibility tree.

Acknowledging the imperfections reflected in the trees, our conclusion must nevertheless be that, within the category of functional intelligibility measures, the tree built on sentence-intelligibility scores reflects the taxonomy of Chinese dialects postulated by traditional dialectologists substantially better than the word-intelligibility tree. Also, the above comparisons have shown that, overall, tree structures based on functional intelligibility measures correspond better to traditional dialect taxonomies than the trees based on our earlier opinion-test scores, which did not even correctly reflect the primary split between Southern and Mandarin varieties (see below).

7. Discussion

7.1. Internal structure of the dendrograms

Observing the tree structures generated from the mean scores obtained from the judgment (opinion) and functional tests of mutual intelligibility, we found no perfect reflection of traditional taxonomy for Chinese varieties proposed by dialectologists.

In the trees based on opinion scores (see Tang and van Heuven, 2007), Changsha and Nanchang varieties were wrongly parsed as Mandarin members–whereas they are traditionally classified into the Southern branch. However, in the two tree structures based on the functional tests, Changsha and Nanchang are correctly classified as Southern varieties and consistently go together, i.e., they make up an identifiable sub-cluster in both trees. A survey of the traditional literature on Chinese dialectology indicates that the two dialects belong to different dialect groups, viz. Xiang, and Gan, respectively. The other subgroups in the Mandarin branch and in the Southern branch are not reflected either, especially in the word-intelligibility tree.

Within the Min group, the internal difference and uniformity are truly reflected. For instance, because their internal difference, the Min group is subdivided into several subgroups (see Fig. 2). But there is uniformity between these subgroups. According to *the Language Atlas of China*, the South Min, Puxian and the East Min share some common features and then form the Eastern cluster whilst the North Min and the Central Min share some other common features so they form the Western cluster. Our subjective trees (viz. the judged intelligibility and similarity trees as well as the functional sentence tree) reflected the internal difference and uniformity of Min group: (i) Fuzhou (East Min) did not form a cluster with Xiamen and Chaozhou (the South Min) at the same level, which shows their different degree of mutual intelligibility, viz. Xiamen and Chaozhou are more mutually intelligible than to Fuzhou but (ii) Fuzhou is added to the cluster of Xiamen and Chaozhou at a higher level, which shows that, as Min members, they are more mutually intelligible to each other than to other dialects.[10]

From the word-intelligibility and sentence-intelligibility trees, we can see that Meixian (Hakka/Kejia group) is rather close to both Guangzhou (Yue) and to Nanchang (Gan). The same relationship was seen in the trees based on Cheng's LSI and PCI (see Tang and van Heuven, 2007). Cheng's trees show that Meixian (Kejia) is lexically more similar to Guangzhou (Yue) than to other dialects but shares more phonological correspondence with Nanchang (Gan). These findings also can be explained reasonably. The Kejia dialect was formed and affected by Gan during the first immigration period, so it shares many common features with Gan, and then it was influenced by Yue during the second immigration period. Actually, Kejia becomes an interlanguage between Gan and Yue so that Kejia listeners can understand both Guangzhou (Yue) and Nanchang (Gan) to some extent. That is why some dialectologists proposed the Gan-Ke(jia) group or Yue-Gan-Ke supergroup (Li, 1937; Lau, 2002).

As for the Wu group, we do find a cluster of Suzhou and Wenzhou in the trees of judged-intelligibility, judged-similarity, and functional sentence-intelligibility, which also converge with the tree based on Cheng's objective LSI. This also correctly reflects traditional taxonomy. In other cases (viz. in the objective PCI tree and the subjective word-

[10] The internal structure of the Min group is also correctly reflected by Cheng's (1997) LSI and PCI trees (for more detail see Tang and van Heuven, 2007).

intelligibility tree), the classification of Suzhou and Wenzhou is inconsistent. We also see that Wenzhou is grouped with Xiamen (in the word-intelligibility tree). Traditionally, the Wu group comprises varieties of Northern Wu (e.g. Suzhou) and Southern Wu (Wenzhou). According to *the Language Atlas of China*, varieties in the Wu group are geographically between the Jianghuai Mandarin (to the north) and the Min groups (to the south). The Northern Wu varieties are heavily influenced by the neighboring Mandarin varieties whilst the Southern Wu varieties share some features with the Min varieties. Thus, in some cases, it might not be easy to determine their classifications.

7.2. Which intelligibility test should be preferred?

Opinion testing is generally proposed as a feasible short-cut when running full-fledged functional intelligibility tests is impractical. From the literature on intelligibility testing in speech technology we know that native listeners have very accurate intuitions (opinions) on the intelligibility of talking computers (see van Bezooijen and van Heuven, 1997), so that the use of opinion testing as a short-cut to functional testing seems warranted in that area of application. It is an open question, of course, if the same conclusion would apply to the field of dialectology. Our study would be the first that allows a direct comparison of the value of opinion testing and functional testing of intelligibility in the context of dialectology.

The two types of subjective intelligibility test used in our research, i.e. intelligibility judged through opinion tests and by functional tests, show cross-type correlations between $r = 0.7$ and 0.8. These correlations are good but not excellent, so that we may conclude that the judgment tests are sensitive to phenomena that are not picked up by the functional tests and vice versa. Moreover, the results of two types of test are about equally predictable (in terms of r^2 values) from the two objective measures (SLI and PCI, see Table 5). The only criterion left that might allow us to prefer one type of test over the other would be the correspondence with traditional taxonomy. We observed that the results of our functional tests agree clearly better with the general picture that emerges from linguistic taxonomies of the Sinitic varieties in our study. Such a clear correspondence could not be established in our earlier study in which we related the dialectological taxonomy to intelligibility measures derived from opinion tests.

We feel, then, that the correlation between the two types of tests is not good enough to recommend the indiscriminate use of opinion tests as a substitute for functional test procedures. We advocate that, whenever the resources are available, mutual intelligibility should be tested functionally.

Finally, within the category of functional tests, the sentence-intelligibility test yields more credible results than the word-intelligibility test (although word-intelligibility is highly correlated with sentence-intelligibility). The results show that sentence-intelligibility reflects traditional dialect taxonomy better than word-intelligibility does. There may be at least two reasons why this is so. First, the isolated word test is not just a word-recognition test; it also involves the additional task of semantic classification, which may introduce a source of statistical noise (error) into the data. Listeners may be quite able to recognize a word in another variety and yet fail to come up with the correct classification for the word. This problem does not arise in the sentence-intelligibility test. Second, in natural language use isolated words are the exception rather than the rule. Listeners are used to hearing words in connected speech, and to using earlier context to narrow down the range of recognition candidates. It can be argued, therefore, that the results of the sentence-intelligibility have greater ecological validity than the rather contrived semantic categorization task.

8. Conclusions

We conclude that there generally exists some degree of mutual intelligibility between pairs of Chinese dialects, and that (mutual) intelligibility can be experimentally measured both by opinion and functional tests. The experimental test results can be predicted from objective measures such as a Lexical Similarity Index (LSI) and a Phonological Correspondence Index (PCI) to some extent yet not perfectly.

Our regression analyses allow the following conclusions to be drawn:

(i) The two types of subjective measure significantly correlate with each other, either in the same type-data or cross-type data (e.g. judged similarity versus judged mutual intelligibility and functional word versus sentence intelligibility).

(ii) The results of two objective structural measures also significantly correlate with each other.

(iii) The results of the two types of subjective measures correlate reasonably with the results of the two objective structural measures.

(iv) All the results correspond with traditional dialect taxonomy to some extent. Opinion (judgment) tests are useful shortcuts, but they fall short of functional tests, especially at the sentence level, when it comes to the convergence with traditional dialect taxonomy. So we claim that functional sentence intelligibility is the preferred measure of mutual intelligibility between pairs of Chinese varieties in this study.

If we accept that functional intelligibility measurements, especially those determined at the sentence level, are a valid reflection of implicit decisions made by traditional dialectologists when they established family relationships among the Sinitic varieties, we may then use our results to settle disputes among linguists. One issue that was raised in our discussion of the dialectological literature on Sinitic varieties concerned the status of Taiyuan. Some sources claim that Taiyuan is a (Northern) Mandarin variety, on the grounds that it has the limited tone inventory that is typical of a Mandarin variety. However, other dialectologists classify Taiyuan as a Southern variety (see Fig. 2), because it kept the Middle Chinese Ru tone, which is typically (although not exclusively) a feature of Southern varieties. We may now turn to our intelligibility measurements to decide which characteristics should prevail in the classification of this variety. The results seem unequivocal: Both the word and the sentence intelligibility measurements yield tree structures in which Taiyuan is grouped together with only Mandarin varieties.

In this way we have come to use mutual intelligibility as a uni-dimensional, experimentally grounded criterion, based on communicative principles, that allows us to classify related languages and language varieties and establish affinity relationships among them. Ultimately, this procedure opens up the possibility of establishing the relative importance (weights) that should be attached to differences between varieties along multiple linguistic dimensions (e.g. segmental versus tonal phonology, phonology versus morphology/syntax, and so on), when arguing about dialect affinity.

Acknowledgements

The first author acknowledges the financial support from the China Scholarship Council (CSC) for tuition and subsistence in the Netherlands for a 4-year period. This article has benefited substantially from constructive criticisms made by three anonymous reviewers.

Appendix A. Stimulus words used for semantic classification task (10 categories, 15 instantiations per category).

#	English	Standard Mandarin		#	English	Standard Mandarin	
		Character	Pinyin			Character	Pinyin
(1) Body parts				(6) Textiles, articles of clothing, apparel			
1.	head	头	tou2	76.	blanket	被子	bei4zi
2.	face	脸	lian3	77.	sheet	床单	chuang2dan1
3.	eye	眼	yan3	78.	pillow	枕头	zhen3tou
4.	ear	耳	er3	79.	mosquito net	蚊帐	wen2zhang4
5.	nose	鼻	bi2	80.	thread	线	xian4
6.	mouth	嘴	zui3	81.	yarn	纱	sha1
7.	hand	手	shou3	82.	silk	丝	si1
8.	foot	脚	jiao3	83.	cloth	布	bu4
9.	neck	颈	jing3	84.	skirt	裙子	qun2zi
10.	hair	头发	tou2fa	85.	scarf	围巾	wei2jin1
11.	eyebrow	眉毛	mei2mao	86.	shirt	衬衣	chen4yi1
12.	tongue	舌	she2	87.	shoe	鞋	xie2
13.	tooth	牙	ya2	88.	sock	袜	wa4
14.	shoulder	肩	jian1	89.	earring	耳环	er3 huan2
15.	back	背	bei4	90.	sweater	毛衣	mao3yi1
(2) Plants: sweet, fruits and nuts				(7) Orientation in time and space			
16.	apple	苹果	ping2guo3	91.	above	上	shang4
17.	pear	梨	li2	92.	below	下	xia4

Appendix A (*Continued*)

#	English	Standard Mandarin		#	English	Standard Mandarin	
		Character	Pinyin			Character	Pinyin
18.	banana	香蕉	xiang1jiao1	93.	left	左	zuo3
19.	lichee	荔枝	li4zhi1	94.	right	右	you4
20.	mango	芒果	mang2guo3	95.	front	前	qian2
21.	grape	葡萄	pu2tao	96.	back	后	hou4
22.	watermelon	西瓜	xi1gua	97.	east	东	dong1
23.	peach	桃子	tao2zi	98.	west	南	nan2
24.	apricot	杏	xing4	99.	south	西	xi1
25.	pineapple	菠萝	bo1luo2	100.	north	北	bei3
26.	cherry	樱桃	ying1tao	101.	middle	中	zhong1
27.	strawberry	草莓	cao3mei2	102.	inside	里	li3
28.	date	枣	zao3	103.	outside	外	wai 4
29.	pomegranate	石榴	shi2liu	104.	tomorrow	明天	ming2tian1
30.	walnut	核桃	he2tao	105.	yesterday	昨天	zuo2tian1
(3) Plants: vegetables				(8) Natural phenomena			
31.	celery	芹菜	qin2cai4	106.	sun	太阳	tai4yang
32.	leek	韭菜	jiu3cai4	107.	moon	月亮	yue4liang
33.	eggplant	茄子	qie2zi	108.	star	星星	xing4xing
34.	pumpkin	南瓜	nan2gua	109.	rain	雨	yu3
35.	winter melon	冬瓜	dong1gua	110.	wind	风	feng1
36.	tomato	西红柿	xi1hong2shi4	111.	ice	冰	bing1
37.	potato	土豆	tu3dou4	112.	frost	霜	shuang1
38.	corn	玉米	yu4mi3	113.	snow	雪	xue3
39.	lotus root	莲藕	lian2ou3	114.	fog	雾	wu4
40.	spinach	菠菜	bo1cai4	115.	hail	冰雹	bing1bao4
41.	carrot	胡箩卜	hu2luo2bo	116.	cloud	云	yun2
42.	cucumber	黄瓜	huang2gua	117.	thunder	雷	lei2
43.	pea	豌豆	wan1dou4	118.	lightning	闪电	shan3dian4
44.	string bean	豇豆	jiang1dou4	119.	rainbow	彩虹	cai3hong2
45.	mushroom	磨菇	mo3gu	120.	flood	洪水	hong2shui3
(4) Animals: four-legged mammals				(9) Perishables (food/drinks other than fruits and vegetables			
46.	dog	狗	gou3	121.	beancurd	豆腐	dou4fu
47.	cat	猫	mao1	122.	milk	牛奶	niu2nai3
48.	pig	猪	zhu1	123.	noodle	面条	mian4tiao2
49.	ox	牛	niu2	124.	meat	肉	rou4
50.	goat	羊	yang2	125.	rice	米饭	mi3fan4
51.	tiger	老虎	lao2hu3	126.	soup	汤	tang1
52.	lion	狮子	shi1zi	127.	wine	酒	jiu3
53.	elephant	大象	da4xiang4	128.	oil	油	you2
54.	horse	马	ma3	129.	salt	盐	yan2
55.	leopard	豹	bao4	130.	soy sauce	酱油	jiang4you2
56.	giraffe	长颈鹿	chang2jing3lu4	131.	vinegar	醋	cu4
57.	bear	熊	xiong2	132.	pepper	胡椒	hu2jiao1
58.	zebra	斑马	ban1ma3	133.	egg	蛋	dan4
59.	wolf	狼	lang2	134.	sausage	香肠	xiang1chang2
60.	fox	狐狸	hu2li	135.	tea	茶	cha2
(5) Animals: other				(10) Verbs of action/things people do			
61.	cock	公鸡	gong1ji1	136.	shake hands	握手	wo4shou3
62.	hen	母鸡	mu2ji1	137.	nod	点头	dian3tou2
63.	duck	鸭	ya1	138.	shake head	摇头	yao2tou2
64.	snake	蛇	she2	139.	laugh	笑	xiao4
65.	swallow	燕子	yan4zi	140.	cry	哭	ku1
66.	magpie	喜鹊	xi2que4	141.	walk	走	zou3
67.	crab	螃蟹	pang2xie4	142.	run	跑	pao3
68.	goose	鹅	e2	143.	jump	跳	tiao4
69.	sparrow	麻雀	ma2que4	144.	stand	站	zhan4
70.	bee	蜜蜂	mi4feng1	145.	sit	坐	zuo4
71.	spider	蜘蛛	zhi1zhu1	146.	sleep	睡	shui4

Appendix A (*Continued*)

#	English	Standard Mandarin		#	English	Standard Mandarin	
		Character	Pinyin			Character	Pinyin
72.	silk worm	蚕	can2	147.	open	开	kai1
73.	ant	蚂蚁	ma2yi3	148.	close	关	guan1
74.	butterfly	蝴碟	hu2die2	149.	read	读	du2
75.	dragonfly	蜻蜓	qing1ting2	150.	write	写	xie3

Note: digits in Pinyin transcription refer to lexical tones. Tone 1 is the high level tone, Tone 2 is a mid-rising tone, Tone 3 is the low dipping tone and Tone 4 is high falling.

Appendix B. Mandarin SPIN sentences and glosses.

#	Mandarin		English
	Characters	Pinyin	
1.	他捕鱼用网。	ta1 bu4 yu2 yong4 wang3	He caught the fish in his net.
2.	关上窗户，挡住风。	guang1 shang4 chuang1 hu4, dang3 zhu4 feng1	Close the window to stop the draft.
3.	我的电视是十二英寸的屏幕。	wo3 de dian4 shi4 shi4 shi2 er4 ying1 cun4 de ping2 mu4	My T.V. has a twelve-inch screen.
4.	舰长指挥舰队。	nong2 min2 shou1 ge2 zhuang1 jia.	The farmer harvested his crop.
5.	舰长指挥舰队。	jian4 zhang2 zhi3 hui1 jian4 dui4	The admiral commands the fleet.
6.	喝啤酒的人举起了酒杯。	he1 pi2 jiu3 de ren2 ju3 qi2 le jiu3 bei1	The beer drinkers raised their mugs.
7.	白蚁看起来象蚂蚁。	bai2 yi3 kan4 qi3 lai2 xiang4 ma2 yi3	A termite looks like an ant.
8.	他膝盖上的伤口结了一个疤。	ta1 xi2 gai4 shang4 de shang1 kou3 jie1 le yi2 ge4 ba1	The cut on his knee formed a scab.
9.	农民堆码干草。	nong2 min2 dui1 ma3 gan1 cao3	The farmer baled the hay.
10.	为了你的生日，我做了蛋糕。	wei4 le ni3 de sheng1 ri4 wo3 zuo4 le dan4 gao1	For your birthday I baked a cake.
11.	火车脱离了轨道。	huo3 che1 tuo1 li2 le gui3 dao4	The railroad train ran off the track.
12.	那只孤独的鸟在找它的同伴。	na4 zhi1 gu1 du2 de niao3 zai4 zhao3 ta1 de tong2 ban4	The lonely bird searched for its mate.
13.	他们喝完了一整瓶酒。	ta1 men he1 wan3 le yi4 zheng3 ping2 jiu3	They drank a whole bottle of gin.
14.	我们在沙滩上玩沙。	wo3 men zai4 sha1 tan1 shang4 wan2 sha1	On the beach we play in the sand.
15.	我们迷路了，所以要看地图。	wo3 men mi2 lu4 le,suo3 yi3 yao4 kan4 di4 tu2	We're lost so let's look at the map.
16.	飞机丢下一颗炸弹。	fei1 ji1 diu1 xia4 yi4 ke1 zha4 dan4	The airplane dropped a bomb.
17.	把香肠切成条。	ba3 xiang1 chang2 qie1 cheng2 tiao2	Cut the bacon into strips.
18.	这把钥匙不配这把锁。	zhe4 ba3 yao4 chi2 bu2 pei4 zhe4 ba3 suo3	This key won't fit in the lock.
19.	男孩在踢足球。	nan2 hai2 zai4 ti2 zu2 qiu2	The boy gave the football a kick.
20.	为了安全，警察穿了防弹衣。	wei4 le an1 quan2 jing3 cha2 chuan1 le fang2 dan4 yi1	The cop wore a bullet-proof vest.
21.	洗完澡，他穿上睡衣。	xi2 wan2 zao3,ta1 chuan1 shang4 shui4 yi1	After his bath he wore a robe.
22.	装汤用碗。	zhuang1 Tang1 yong4 wan2	The soup was served in a bowl.
23.	工人们在挖一条水沟。	gong1 ren2 men zai4 wa1 yi4 tiao2 shui3 gou1	The workers are digging a ditch.
24.	船长召集他的船员。	chuan2 zhang3 zhao1 ji2 ta1 de chuan2 yuan2	The ship's captain summoned his crew.
25.	他们在玩猫捉老鼠的游戏。	ta1 men zai4 wan3 mao1 zuo1 lao3 su3 de you2 xi4	They played a game of cat and mouse.
26.	黑腥腥是猿猴。	hei1 xing1 xing shi4 yuan2 hou2	A chimpanzee is an ape.
27.	垫子里面塞的是塑料泡沫。	dian4 zi li3 mian4 sai1 de shi4 su4 liao4 pao4 mo4	The cushion was filled with foam.
28.	他抛给那个快淹死的人一根绳子。	ta1 pao1 gei3 na4 ge4 kuai4 yan1 si2 de ren2 yi4 gen1 sheng2 zi	He tossed the drowning man a rope.
29.	扫地用扫帚。	sao3 di4 yong4 sao4 zhou3	To sweep the floor with a broom.
30.	我们听见钟的滴答声。	wo3 men ting1 jian4 zhong1 de di1 da1 sheng1	We heard the ticking of the clock.
31.	医生开了处方。	yi1 sheng1 kai1 le chu2 fang1	The doctor prescribed the drug.
32.	下棋是一种乐趣。	xia4 qi2 shi4 yi4 zhong2 le4 qu4	Playing checkers can be fun.
33.	他早餐喝了一些牛奶。	ta1 zao3 can1 he1 le yi4 xie1 niu2 nai3	At breakfast, he drank some milk.
34.	国王戴的金制的皇冠。	guo2 wang2 dai4 de jin1 zhi4 de huang2 guan1	The king wore a golden crown.
35.	沙粒堆成了山。	sha1 li4 dui1 cheng2 le shan1	The sand was heaped in a pile.
36.	狮子发出一声怒吼。	wei4 fang4 mu4 cai2,ta1 da1 le yi2 ge4 peng2	To store his wood, he built a shed.
37.	狮子发出一声怒吼。	shi1 zi fa1 chu1 yi4 sheng1 nu4 hou3	The lion gave an angry roar.
38.	高速公路有六条车道。	gao1 su4 gong1 lu4 you3 liu4 tiao2 che1 dao4	The super highway has six lanes.
39.	汽车滚下了悬崖。	qi4 che1 gun3 xia4 le xuan2 ya2	The car drove off the steep cliff.
40.	扔掉那些无用的垃圾。	reng1 diao4 na4 xie1 wu2 yong4 de la1 ji1	Throw out all this useless junk.
41.	她给他做了一吨丰盛的饭菜。	ta1 gei3 ta1 zuo4 le yi2 dun4 feng1 sheng4 de fan4 cai4	She cooked him a hearty meal.
42.	房东提高了房租。	fang2 dong1 ti2 gao1 le fang2 zu1	The landlord raised the rent.
43.	我们的座位在第二排。	wo3 men de zuo4 wei4 zai4 di4 er4 pai2	Our seats were in the second row.
44.	我们跟着狮子找到了它的窝。	wo3 men gen1 zhe shi1 zi zhao3 dao4 le ta1 de wo1	We tracked the lion to his den.

Appendix B (*Continued*)

#	Mandarin		English
	Characters	Pinyin	
45.	她给她自己倒了一杯茶。	ta1 gei2 ta1 zi4 ji2 dao4 le yi4 bei1 cha2	She poured herself a cup of tea.
46.	大雨引起了洪灾。	da4 yu3 yin3 qi3 le hong2 zai1	The heavy rains caused a flood.
47.	警察寻找线索。	jing2 cha2 xun2 zhao3 xian4 suo3	The police searched for a clue.
48.	洗地板用抹布。	xi2 di4 ban3 yong4 mo2 bu4	Wash the floor with a mop.
49.	小鸡啄玉米用嘴。	xiao3 ji1 zhuo2 yu4 mi3 yong3 zui3	The chicken pecked corn with its beak.
50.	池塘里满是呱呱叫的青蛙。	chi2 Tang2 li3 man4 shi4 gua1 gua jiao4 de qing1 wa1	The pond was full of croaking frogs.
51.	游泳的人跳进了游泳池。	you2 yong3 de ren2 diao4 jin4 le you2 yong3 chi2	The swimmer dove into the pool.
52.	牧羊人看着他的羊群。	mu4 yang2 ren2 kan1 zhe ta1 de yang2 qun2	The shepherd watched his flock of sheep.
53.	把肉切成小块。	ba3 rou4 qie1 cheng2 xiao3 kuai4	Cut the meat into small chunks.
54.	西瓜有很多籽。	xi1 gua you3 hen3 duo1 zi3	Watermelons have a lot of seeds.
55.	新娘穿着白色的婚纱。	xin1 niang3 chuan1 zhe bai3 se4 de hun1 sha1	The bride wore a white gown.
56.	生病的孩子吞下了药片。	sheng1 bing4 de hai2 zi tun1 xia4 le yao4 pian4	The sick child swallowed the pill.
57.	自行车有两个轮子。	zi4 xing2 che1 you3 liang2 ge4 lun2 zi	A bicycle has two wheels.
58.	她有一条玻璃珠的项链。	ta1 you3 yi4 tiao2 bo1 li2 zhu1 de xiang4 lian4	She had a necklace of glass beads.
59.	船驶出了港湾。	chuan3 shi2 chu1 le gang2 wan1	The boat sailed across the bay.
60.	奶牛生了牛犊。	nai3 niu2 sheng1 le niu2 du2	The cow gave birth to a calf.

References

Anderson, H., 2005. Intelligibility testing (RTT) between Mendankwe and Nkwen. SIL Electronic Survey Reports 2005-002. http://www.sil.org/silesr/abstract.asp?ref=2005-002 (accessed 23.09.2008).

Bezooijen, R., van den Berg, R., 1999. Word intelligibility of language varieties in the Netherlands and Flanders under minimal conditions. In: van Bezooijen, R., Kager, R (Eds.), Linguistics in the Netherlands 1999. John Benjamins, Amsterdam, pp. 1–12.

Bezooijen, R., van Heuven, V.J., 1997. Assessment of speech synthesis. In: Gibbon, D., Moore, R., Winksi, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin/New York, pp. 481–653.

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

Brye, E., Brye, E., 2000. Rapid Appraisal and Intelligibility Testing Surveys of the Eastern Beboid Group of Languages. Yaounde, Cameroun.

Campbell, J., 2004. Chinese language FAQ. *Glossika Language Web*. http://www.glossika.com/en/dict/faq.php#1. (Updated 25.01.2004; last accessed 23.09.2008).

Casad, E.H., 1974. *Dialect Intelligibility Testing*. Summer Institute of Linguistics Publications in Linguistics and Related Fields 38. Norman, Oklahoma: The Summer Institute of Linguistics and the University of Oklahoma.

Chambers, J.K., Trudgill, P., 1980. Dialectology. Cambridge University Press.

Cheng, C.-C., 1997. Measuring relationship among dialects: DOC and related resources. Computational Linguistics & Chinese Language Processing 2.1, 41–72.

Cutler, A., van Donselaar, W., 2001. Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. Language and Speech 44, 171–195.

Delsing, L.-O., Lundin-Åkesson, K., 2005. Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska. [Does language keep the Nordic countries together? A research report on how well young people understand Danish, Swedish and Norwegian]. Köpenhamn: Nordiska ministerrådet.

Gooskens, C., 2007. The contribution of linguistic factors to the intelligibility of closely related languages. Journal of Multilingual and Multicultural Development 28, 445–467.

Gooskens, C., submitted. Non-linguists' judgments of linguistic distances between dialects. Nordic Journal of Linguistics, submitted.

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

Hickerson, H., Turner, G.D., Hickerson, N.P., 1952. Testing procedures for estimating transfer of information amongst Iroquios languages and dialects. International Journal of American Linguistics 18, 1–8.

Heeringa, W., 2004. Measuring dialect pronunciation differences using Levenshtein distance. Doctoral diss. Rijksuniversiteit Groningen.

Kalikow, D.N., Stevens, K.N., Elliott, L.L., 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.

Kessler, B., 1995. Computational dialectology in Irish Gaelic. ACL, Dublin, Proceedings of the Association for Computational Linguistics, European Chapter. pp. 60–67.

Lau, C.-f. 刘镇发 2002. Hanyu Fangyan de fenlei biaozhun yu "Kejiahua" zai Hanyu fangyan fenlei shang de wenti 汉语方言的分类标准与"客家话"再汉语方言分类上的问题 [Criteria for Chinese dialect classification and the problem of the position of the "Hakka dialect" in Chinese dialect grouping]. Journal of Chinese Linguistics, 30, 82–96.

Lee, M.W., 1987. Languages in China. In Wurm, S.A. et al. (Eds.), The Language Atlas of China, A-1, backside.

Li, F.-k. 李芳桂 1937. Languages and Dialects of China. The Chinese Yearbook. Shanghai: Commercial Press. (Reprinted in Journal of Chinese Linguistics 1, 1–13, 1973).

Li, R., 1987a. Languages in China. In Wurm, S.A. et al. (Eds.), The Language Atlas of China, A-1.

Li, R., 1987b. Chinese Dialects in China. In Wurm, S.A. et al. (Eds.), The Language Atlas of China, A-2.

Morton, J., 1969. Interaction of information in word recognition. Psychological Review 76, 165–178.

Murrell, G.A., Morton, J., 1974. Word recognition and morphemic structure. Journal of Experimental Psychology 102, 963–968.

Nooteboom, S., 1981. Lexical retrieval from fragments of spoken words: Beginnings vs endings. Journal of Phonetics 9, 407–424.

Pierce, J.E., 1952. Dialect distance testing in Algonquian. International Journal of American Linguistics 18, 203–210.

Preston, D.R., 1989. Perceptual dialectology. Nonlinguists' views of areal linguistics. Foris, Dordrecht.

Tang, C., van Heuven, V.J., 2007. 'Mutual intelligibility and similarity of Chinese dialects'. In: Los, B., van Koppen, M. (Eds.), Linguistics in the Netherlands 2007. John Benjamins, Amsterdam, pp. 223–234.

van Hout, R., Münsterman, H., 1981. Linguistische afstand, dialect en attitude [Linguistic distance, dialect and attitude]. Gramma 5, 101–123.

Voegelin, C.F., Harris, Z.S., 1951. Methods for determining intelligibility among dialects of natural languages. Proceedings of the American Philosophical Society 95, 322–329.

Wang, H., 2007. English as a lingua franca: Mutual Intelligibility of Chinese, Dutch and American speakers of English. LOT Dissertation Series nr. 147. Utrecht: LOT.

Wurm, S.A., T'sou, B., Bradley, D., Rong, L., Zhenghui, X., Zhenxing, Z., Maoji, F., Jun, W., Dob, 1987. Language Atlas of China. Jointly compiled by the Chinese Academy of Social Sciences and the Australian Institute of Humanities. Longman, Hong Kong.