Computational Linguistics and Chinese Language Processing
vol.2, no.1, February 1997, pp. 41-72

Measuring Relationship among Dialects: DOC and Related Resources

Chin-Chuan Cheng*

Abstract

This paper is a synthesis of the past studies in measurements of dialect relationships. The phonological data of 17 Chinese dialects that were computerized in the late 1960s have been utilized for measurements of dialect distance. In addition, a file of over 6,400 lexical variants in 18 dialects was also used to quantify dialect affinity. This writing first explains the nature, the organization, and the coding of these files. A series of steps illustrate how the phonological file was processed to derive the needed information for calculation of correlation coefficients. The coefficients are considered as indices of dialect affinity. The dialects are then grouped by the average linking method of cluster analysis of the coefficients. The appropriateness of the correlation method to the data is then discussed. Recent work on calculation of dialect mutual intelligibility is presented to indicate the future direction of research.

Keywords: Chinese dialects, measurements of affinity, measurements of mutual intelligibility, comparative dialectology

1. Introduction

In 1966, under William Wang's direction, we came to see that linguistic theories up to that time were predominantly based on European languages and that Chinese had much to contribute to generalizations about human language. Three characteristics of Chinese made up a wealth of information for us. First, the language has an unsurpassed depth of recorded history. Second, its dialects geographically cover a major part of Asia. Third, the use of tone is distinct from that of other languages [Wang 1970]. At that time, the Hanyu Fangyin Zihui [Beijing University 1962, henceforth Zihui] was just arriving in our libraries from Beijing. It gives the sounds of over 2,700 words in 17 dialects. For each word the Middle Chinese phonological categories are also provided. So historical depth, geographical breadth, and tone are all included in this volume. We decided to use the data in the Zihui for our exploration.

*University of Illinois, Urbana-Champaign. E-mail: c-cheng1@uxl.cso.uiuc.edu

The immediate step to prepare the data for use was to put them on a computer for fast and thorough search for patterns and for testing of ideas. The contemplated computer file could be used as a dictionary for immediate reference. Thus the name "DOC" (Dictionary On Computer) was created.

In the ensuing years, we added Shanghai, the reconstructed Zhongyuan Yinyun of the 14th century, Kan-on and Go-on Sino-Japanese, and Sino-Korean to the data pool of the 2,700 words. During the developmental stage the following colleagues were involved in coding, data collection, format design, computer programming, and theory testing: Vernon Ambrose, Betty Shefts Chang, Corey Chow, Matthew Chen, Chin-Chuan Cheng, Hsin-I Hsieh, Kyungnyun Kim, Johana Kovitz, Anatole Lyovin, Thomas McGuire, Gilbert Roy, Masayoshi Shibatani, Katherine I-ping Shih, Mary Streeter, Benjamin T'sou, and William S-Y. Wang [Cheng 1994a].

When the first DOC file was implemented at Berkeley in 1969, we started to work on several issues in Chinese linguistics. The issues included phonological change of Middle Chinese initials, the time dimension of sound change, the distribution of initial consonants of the dialects, diffusion overlapping, diachronic phonology and lexical diffusion, and profile of implementation and time variable in phonological change. They all had one common theme: lexical diffusion.

We used DOC to tabulate occurrences of sounds in synchronic distribution and in historical evolution. From the patterns we saw general rules and exceptions. Exceptions were of particular interest in capturing the profile of historical change. They might occur at the onset of a change. They might be created by an incomplete change. Furthermore, competing change could be the cause of residue [Wang 1969]. So since its birth DOC has had a very distinct theoretical orientation. The papers that grew out of DOC as well as those that dealt with lexical diffusion based on other languages were collected in the 1977 anthology The Lexicon in Phonological Change [Wang 1977]. Thus lexical diffusion was the theoretical orientation of the initial phase of research using DOC.

In the search for evidence to support lexical diffusion, we focused on the history of individual words. In the late 1970s, we felt that it was time to examine the data for an overall understanding of the dialects. The computerized data allowed us to compare the dialects in various ways. As we compared them and tried to form a synthesis, not a listing of their differences, we naturally came to deal with quantitative measurements. In the early 1980s we worked on dialect similarity, and in the 1990s we ventured to calculate dialect mutual intelligibility. In this writing we will discuss such measurements of dialectal relationships.

2. Occurrence Patterns

As said earlier, the DOC file contains the pronunciations of over 2,700 syllable-words in 18 modern Chinese dialects. They are Beijing (北京), Jinan (濟南), Xi'an (西安), Taiyuan (太原), Hankou (漢口), Chengdu (成都), Yangzhou (揚州), Suzhou (蘇州), Wenzhou (溫州), Changsha (長沙), Shuangfeng (雙峰), Nanchang (南昌), Meixian (梅縣), Guangzhou (廣州), Xiamen (廈門), Chaozhou (潮州), Fuzhou (福州), and Shanghai (上海). We collected the Shanghai data ourselves, and the others were taken from the Zihui. As the Shanghai data needed rechecking and verification, we made little use of them. At one point we were interested in finding out how the pronunciations of the words differ among the dialects. For example, we wanted to know if all the words with the voiceless bilabial unaspirated stop (幫母) in Middle Chinese were pronounced as /p/. We found modern reflexes of /p, ph, b, m/ with varying number of words. Since DOC includes Middle Chinese phonological categories of initials, rimes, vowel grades, etc., it was easy to make such a tabulation. The voiceless bilabial aspirated stop (滂母) and the voiced bilabial stop (並母) were also examined to tabulate their modern reflexes and the occurrence of words in those initials. We obtained the following numbers [Cheng 1991]:

We can see in (1) that for both the voiceless stops the modern reflexes are not very different. However, the distributions of the voiced stop in Middle Chinese show distinct occurrence patterns. Suzhou, Wenzhou, and Shuangfeng have a large number of words with the voiced bilabial initial while the other dialects all have none. Thus the occurrence patterns could help us classify the dialects. But classification alone would not be of much interest to us. We wished to quantify the similarity or difference among the dialects. With a verifiable quantity index of similarity we hoped to achieve something beyond mere descriptions or listings of dialectal differences.

The numbers in (1) then looked much like a table prepared for statistical processing. Indeed, we considered the initials as cases and the dialects as variables and calculated the Pearson correlation coefficients for each pair of the dialects. The cases shown in (1) were for the bilabial initials only. Other initials, finals, and tones were actually included in the calculation. The bivariate data for each pair showing the presence or absence of certain items were considered as nominal-dichotomous. The derived coefficients were considered as indices of similarity. The correlation coefficients adapted from Cheng [1982, 1988, 1991] are given in Appendix 1. We thus established a method for measuring dialect similarity in terms of genetic relations. We will return to discuss the measurement later. For now let us examine the DOC file that allowed us to do such work.

3. DOC Data Organization

DOC derived most of its data from the Zihui. Each page of the first edition of the Zihui lists nine characters across and their ronunciations in phonetic alphabet in 17 dialects. The second edition [Beijing University 1989] includes 20 dialects, three more than the first edition. The DOC file was based on the first edition with various modifications over the years. In the 1960s Chinese characters were not readily available on computers, and so they were coded in the four-digit telegraphic code. The Zihui also gives each word the traditional phonological categories of initials, rimes, and tones. We designed the coding conventions to encode these categories and the phonetic symbols.

The file consists of lines of coded categories and symbols. Each line represents either the Middle Chinese phonological categories or a dialect pronunciation. As we worked on the file, typographical errors were corrected. We also detected some errors in the Zihui and corrected them. The first stable version was established in 1971, and so we call that version "DOC 71" [Streeter 1972, 1977]. DOC 71 mainly resided on mainframe computers during the 1970s. Minor changes continued to be made. For example, originally the yang (陽) tones were coded with "B", and the yin (陰) tones were given as space. Space is also used for the tones that did not have the yin-yang differentiation. So

we changed the space that designated a yin tone to "A" later. We also migrated the file from mainframes to personal computers. These changes were made in 1984, and we call this version "DOC 84". The contents of DOC continued to change. For example, in the early 1990s we incorporated some new entries from the second edition of the Zihui. "DOC 84" identifies the format and coding but not the static contents of that year. In 1988 William Wang, Charles Wooters, Zhongwei Shen, and Jonathan Yaruss converted DOC 71 to "dBase" file format and added a user interface for querying the database [Yaruss 1990]. A function was added to DOS to display Chinese characters and IPA symbols. We may call this version "DOC 88". Then the Chinese Windows operating system emerged. In 1993 we converted most of the codes of DOC 84 to Windows format to display Chinese characters and phonetic symbols. This is version DOC 93. All these revisions of DOC are discussed in detail in Cheng [1994a].

Here we will use the data organization and coding conventions of DOC 84 to illustrate the nature of the DOC file. There are about 63,000 records. Listed in (2) are all the records for the character " 一 " and four lines for the character " 丁 " as examples of the DOC records.

(2)

```
0001   0707NNK3444Q
0001 A01A      I
0001 B01A      I
0001 C01A      I
0001 D04A    I E3   Q
0001 E01B      I
0001 F01B      I
0001 G04     I E3   Q
0001 H04A    I 01   Q
0001 I04A    I A I
0001 J04       I
0001 K01B      I
0001 L04A      I    T
0001 M04AJ     I    T
0001 N04AJ     A    T
0001 O14A      I    TL
0001 O24ATS    I    T
0001 P04A      I    K
0001 Q04A      E I  Q
0001 R04     I I1   Q
```

```
0001 W043      I
0001 X0XX      I    TU
0001 Y0XX      I    TI
0001 Z0XX      I    L
0002   2531XGK4141T
0002 A01AT     I    V
0002 B01AT     I    V
0002 C01AT     I    V
```

The telegraphic codes 0001 and 0002 encode " 一 " and " 二 " respectively. When there exist homographic words, column 5 will show "A", "B", etc. to differentiate them. For example, the character 數 'to count' is coded as "2422A" and 數 'number' as "2422B". Thus each line has a telegraphic code to identify the character. For each character the Middle Chinese record is given first. The blanks in column positions 6 and 7 of the line indicate that it is a Middle Chinese record. As we look at the other lines, we see that the code in column position 5 varies to identify the dialects: "A" for Beijing, "B" for Jinan, "C" for Xi'an, etc. The coding conventions appearing in Cheng [1970] have been revised as given in Appendix 7.

We see in (2) that the Middle Chinese records have a different structure from the dialect records. Following the two blanks for identification as Middle Chinese, the first line in (2) continues to provide the number of the page where the character in the Zihui is listed (page 070 in this example) and the position of the character on the page (character number 7). Then "NN" identifies the rime group 臻 . "K" is the code for 開口 , "3" for 三等 , "4" for 入聲 , "44" for 質韻 , and "Q" plus three blanks for 影母 . Again, Appendix 7 lists all the codes.

The dialect records follow the Middle Chinese record for the same character. Line 2 in (2) also starts with the telegraphic code 0001. The letter "A" in column 6 identifies the dialect as Beijing. For this character there is only one pronunciation and so a "0" appears in column 7. For words with two variant pronunciations, each one is given as a record, and this column will have "1" or "2". The next two characters "1A" identify the " 陰平 " tone. The next four characters for the initial are blanks representing the zero initial. The next two characters for medial are also blanks. Then the "I" and a space together code the high front nuclear vowel. We have covered the first 17 columns. Now we are looking at columns 18 and 19. They are the positions for off-glide. In this case there is no glide ending. Column 20 for vowel nasalization is blank. There is no consonantal ending specified in column 21. Column 22 is mostly used to code literary reading with "L". In this case it is blank. Thus, for each dialect record, all together there are also 22 columns.

The Zhongyuan Yinyun and the Sino-Xenic records use the same format as the Chinese dialects except for a couple of elements. The "43" for the Zhongyuan Yinyun tone is " 入作去 " (the Entering tone pronounced as the Departing tone). The Sino-Xenic languages have no tones; the string "XX" is used to fill the columns. The length of the entire file in the DOC 84 format is about 1.3 megabytes.

The more modern DOC 93 has two versions. One is a Microsoft Chinese Word document, and the other is a pure text file. The records corresponding to those given in (2) appear in (3):

Instead of the telegraphic code, a double-byte Big5 code for the character is given. The Middle Chinese record for " 一 " directly shows " 臻開三入質影 ". We retain the telegraphic code (0001) only in the Middle Chinese record. The Zihui page number and the position of the character are also kept (070 and 7), though at a different location. The dialect records show the character (internally a double-byte code), the first character of the dialect name (2 bytes), variant pronunciation (1 byte), and the tone (2 bytes) in that order. The first character of the tone identifies one of the four tones, and the second character indicates whether the tone is yin ("1") or yang ("2"). The coding for Zhongyuan Yinyun tones remains the same as in DOC 84. The initial occupies one character position. This is made possible with the font DOCIPA that we have created. All the diagram symbols of the IPA have been designed as a single character. The medial, vowel, nasalization, and ending are all of one-byte length. The character " 文 " (2 bytes) marks literal reading. The length of the entire file in the Word document format is over 1.5 megabytes. However, the pure text file without Word format is under 1 megabyte, small enough to be stored on a high-density floppy diskette.

DOC 93 does not require further interpretation of the records as displayed in Microsoft Chinese Word (version 5, 6, or 7) on Chinese Windows (version 3.1 or 95) in Big5 code. The user can directly examine the contents of the file and collect the records of interest for study. However, many in-depth inquiries would require the user to have some knowledge of computer programming in order to derive and arrange the data for making useful generalizations.

4. Search and Data Processing

Ideally, the DOC user would get needed information without programming. That was the goal of DOC 88 [Yaruss 1990]. However, technologies change rapidly. The database system of DOC 88 had to be changed along with the operating system. The dynamic of computer technologies made system-dependent versions of DOC unusable very quickly. Over the years we have relied on DOC 84 to do the quantitative studies in dialect similarity and mutual intelligibility. We will therefore use that version to illustrate how the data is processed. At this point we return to the modern reflexes of the Middle Chinese labial initials as given in (1). Let us take the 幫母 as an example to show how to obtain the numbers. Essentially, we need to find the Middle Chinese record that has this initial. Then collect the initials of the dialects for the character and tabulate their occurrences. Before we do anything else, we need to set up a table for tabulation. We can set up the table with an entry column and a column for the frequency of occurrence of the entry.

Following is a list of steps that we need to go through to accomplish the task. Naturally, the steps are not computer language procedures, but they can be readily translated into computer algorithms. First, let us find the first Middle Chinese record that has the 幫 initial.

(4)

a. If we have not reached the end of the file then we continue with the following steps.

b. We read a record.

c. If the record does not have a space in character position 6, i.e. not Middle Chinese, then we repeat the steps starting with step a to find another Middle Chinese record. If a Middle Chinese record is found, then we continue with the following step.

d. Now we have a Middle Chinese record. If character position 19 is not a "P" followed by three blanks (the code for 幫), then we repeat the steps starting with step a to find another Middle Chinese record. Otherwise we continue with the following step.

e. Now we have a Middle Chinese record that has the needed initial.

The next steps will collect the dialect identifications and the initials of the dialects:

(5)

a. If we have not reached the end of the file, then we continue with the following steps.

b. We read a record.

c. We examine the record to see if it has a blank in position 6. If it does, then this is a Middle Chinese record; we go back to step (4d) to see if its initial is 幫 , and continue with the subsequent steps in (4). Otherwise we have a dialect record and we continue with the following step.

d. Now we collect the 6th character, which is a dialect ID, and the characters in positions 10 through 13, which are letters coding the initial of the dialect.

e. We can put the dialect ID and the initial in a string. We may also add a ":" between the two elements for ease of reading later.

f. Now we try to find this string in the table that was set up earlier. If the string is

found in the table entry, then add 1 to its frequency of occurrence. If it is not found, then add this entry to the table and make the count 1. So now we have the dialect, its initial, and the frequency of occurrence.

g. We then go back to step a to collect all the dialects and their initials.

When we have reached the end of the file, we complete the processing of the records and the tabulation for the reflexes of the Middle Chinese initial 幫 in modern dialects. We can go through the same step to tabulate the modern reflexes of 滂 and other categories. The steps given above are for illustration of how DOC can be processed. In reality, we did not do one initial at a time. The table can be expanded, and the reflexes of all the initials can be obtained in one round of processing. Then the table can be arranged in a way to conform to the requirements of a statistical package, such as SPSS, for calculation of correlation coefficients.

In reality, we often started with an idea and tested the data in various ways. Sometimes we started out knowing very little about what would be the right thing to look for. As we worked more, the pieces gradually fell in some meaningful patterns. Then a table like that given in (1) would be formed. What could we do with the table? There were no correct answers to such a question. We had stared at the table and other similar ones for a long time trying to formulate qualitative statements of dialect relations before the idea of correlation hit us. As we were able to retrieve and arrange the data in many ways, DOC actually helped us crystallize many ideas that otherwise would be vague and fleeting.

The correlation coefficients based on initials, finals, and tones as given in Appendix 1 provide the indices of dialect closeness. They allow us to state the degrees of relationships. In addition, the coefficients can be used to make dialect grouping. Appendix 2 is a cluster analysis based on the coefficients. The grouping is compatible with our understanding of Chinese dialect classification. Thus we gained some confidence in applying this methodology to Chinese. Subsequently we used it to measure dialect affinity in terms of lexical items. We will now discuss how the lexical relationships are calculated. The evaluation of the methodology will be given later.

5. Lexical Data

The Department of Chinese at Beijing University under the leadership of Wang Futang compiled the Hanyu Fangyan Cihui (henceforth Cihui) [Beijing University 1964] in addition to the Zihui. The Cihui lists the variants of 905 Putonghua words in 18 dialect localities. For example, the item 太陽 'sun' has the variants of 太陽，日頭，爺，熱

頭，太陽佛，日，and 日頭公；月亮 'moon' has the variants of 月亮，亮月子，月光，月，and 月娘．Not all the variants are used in a single dialect. If we use 1 for occurrence and 0 for non-occurrence in the dialect localities, then we can present the occurrences of these variants as (6). The Cihui lists two Putonghua lexical items on one page. We use a string of code after the entry in Chinese characters to indicate the Cihui page number and the variant. The code 001A01 for 太陽 means that it is the first variant (01) on the left half (A) of page 1 (001).

This table shows that Beijing and Jinan have the same occurrences of 1's and 0's. That is, they share the same lexical items. This association is the basis for us to make dialect grouping. But such visual inspection will loose sight when we look at over 6,400 variants given in the Cihui. A better way to approach the data is to make a contingency table of the 1 and 0 numbers. So we established a computer file of the lexical variants in the early 1980's for the study of lexicon-based affinity [Cheng 1982]

The lexicon file based on the Cihui has the organization like that given in (6) except that the Chinese characters are not part of the file. It is a text file without any word processor format. The numbers are given one after another for the dialects list in the order in (6). The only separation mark is the space between the code for the variant and its occurrences in the dialect. The first few lines of the file are given in (7) to show the data organization.

(7) 001A01 110011111110000000
001A02 001100001011000111
001A03 000100000000000000
001A04 000000100000111000
001A05 000000000100000000
001A06 000000000000000100
001A07 000000000000000010
001B01 111111111010001000
001B02 000000010000000000
001B03 000000000101110000
001B04 000000000000000111
001B05 000000000000000110

The dialects that share the same lexical items should be closely related. Those that share few of them are remote from each other. Thus, similar to the phonological data, the lexicon file can be used to derive the correlation coefficients, and the coefficients can be used as degrees of closeness. We will explain the calculation below.

6. Measurements of Phonological and Lexical Similarity

Correlation is a matter of how entries occur in the dialects. For example, when we look at the numbers under Beijing and Meixian, in (6) or (7), we find that when a word appears in Beijing, it does not appear in Meixian. If a word exists in Meixian, then it does not occur in Beijing. Thus we can make a contingency table as follows:

(8)

		Beijing	
		"1"	"0"
Meixian	"1"	0 (a)	2 (b)
	"0"	2 (c)	8 (d)

The (a) cell is the place to hold the number of words that occur in both Beijing and Meixian. Cell (b) is for the number of words that appear in Meixian but do not exist in Beijing. Cell (c) holds the number of words that exist in Beijing but not in Meixian. And cell (d) is for entries that do not exit in either dialect. This contingency table can be used for the calculation of the *phi* coefficients. The *phi* correlation is the Pearson correlation calculated on nominal-dichotomous data. The formula is given in (9).

(9) *phi* = (ad -bc)/square_root ((a+b)(a+c)(b+d)(c+d))

We calculated the coefficients for each pair of the 18 Cihui dialects in the early 1980s. Appendix 3 gives the lexical coefficients so calculated. The phonological data of DOC were also processed in a similar way and the coefficients of all the pairs of the 17 dialects were obtained using such a correlation formula. As indicated before, Appendix 1 shows the phonological coefficients so calculated. The details are given in Cheng [1982, 1988, 1991].

While the formula has been well established in statistics, it may cause unwarranted results with our data. Ma [1989], Tu and Cheng [1991], and Wang and Shen [1992] all brought up this problem. The issue is the unwarranted inflation of the weighting of 0s. In our lexical case, the variants were collected from all the dialect localities. The more entries we have the more the number of 0s will increase. If we take into consideration the 0s in the calculation, then all the items missing in a pair of dialects would contribute some positive weighting to the correlation of the pair. The result of such calculation will be skewed in favor of non-occurrences and thus inflate the value of the 0s.

We will illustrate the problem below. The calculation of the correlation between Beijing and Meixian using the contingency table in (8) is given in (10):

(10) *phi* = (0 x 8 - 2 x 2) / square_root ((0+2)(0+2)(2+8)(2+8)) = -0.2

The correlation coefficient so calculated is -0.2. Correlation values vary from 1 to -1. If the two localities of a pair are identical in number, the coefficient value is 1. If they differ in every case, then the coefficient is -1. But if we examine Beijing and Meixian again in (6), we see that the lexical variants used in these dialects are completely different. When we find a 1 in Beijing, we find a 0 in Meixian and vice versa. Therefore in terms of these variants Beijing and Meixian are perfectly negatively correlated. By visual inspection we expect the coefficient to be -1. But the calculation in (10) yields -0.2.

A way to avoid this problem is to discard the 0s appearing in both dialects of a pair. Instead of correlation, we can think of dialect relations as similarity relations. A similarity relation can be measured as the ratio of the shared items to all the items that occur in either or both dialects. Using the cell identification of the contingency table in (8), the similarity measurement is given in (11):

(11) a / (a+b+c)

The value of this similarity calculation will vary from 0 to 1. The formula can be applied to the contingency table in (8) to yield 0 (0/(0+2+2)). Thus in this calculation the similarity value for Beijing and Meixian in terms of the variants given in (6) is 0. The result is compatible with our inspection that these dialects do not share any of the items

listed here.

This similarity measurement calculates the ratio of common items to all occurring items. The coefficients and similarity values allow us to quantify dialect relationships. We have also utilized these numerical indices to group the dialects and plotted the grouping with a scale to indicate degrees of closeness. However, a fundamental question concerning Chinese dialects could not be easily answered with the quantification of closeness. The question has to do with mutual intelligibility.

7. Measurement of Dialect Mutual Intelligibility

Mutual intelligibility is one of the most significant criteria for dialect classification. However, in the past it was talked about in terms of personal impressions and experience. Since the early 1990s we have utilized the DOC data to quantify Chinese dialect mutual intelligibility [Cheng 1992, 1994b, 1996]. The basic consideration in determining mutual intelligibility is that when we talk to someone with an accent, we often try to establish a pattern of sound correspondence. As long as the sounds of the foreigner correspond to ours in a systematic way, we can understand the accented speech. Thus the first step in determining mutual intelligibility is to establish correspondence patterns. In DOC the phonological information of each word is given for all the dialects, we can find out how a sound in a dialect corresponds to the sound in another.

The correspondence is based on cognate words. As we examine all the words, general patterns of sound correspondence can be established. Some patterns have a large number of words to give an impression of a general rule. Such patterns are communication enhancing signals. On the other hand, some patterns may have only a small number of cognates. Such patterns require us to specifically memorize the exceptional words and thus are considered as disturbing noise. Furthermore, the elements of a correspondence pattern may be identical or different. For example, the dental nasal in Beijing corresponds to the sounds in Jinan in the following way:

(12) a.	n	zero	3	10.6	noise	different	zero occurs elsewhere in Beijing
b.	n	l	1	10.6	noise	different	l occurs elsewhere in Beijing
c.	n	n	27	10.6	signal	same	
d.	n		21	10.6	signal	different	does not occur in Beijing
e.	n	s	1	10.6	noise	different	s occurs elsewhere in Beijing

The "zero" in (12) is the zero initial. The Beijing nasal dental corresponds to Jinan

zero initials with three words in DOC. The five patterns listed here have a total of 53 cognate words, and the mean is 10.6. We use the mean to determine whether a pattern is signal or noise. This pattern with three items is smaller than the mean and is considered as noise. Moreover, the corresponding elements are different. Since there are other non-cognate words with the zero initial in Beijing, the zero initial for these three words in Jinan will very likely cause confusion with non-cognate words in Beijing. Here we use the dental nasal of Beijing to view the correspondence. We may call Beijing the source dialect and Jinan the target dialect.

We now need to determine a weighting scale for the signal. If the corresponding sounds are the same, then there is no problem in mutual intelligibility. If they are different and can cause confusion with non-cognate words, then the value of such a pattern is the smallest. Thus we have the following relations with ">" indicating that the preceding one is "greater than" the following one in value:

(13)

Same > Different and not occurring in the source dialect > Different and occurring elsewhere in the source dialect

To quantify the relations, we use the weighting scale in (14) for calculation. The unity 1 is the base. Signal patterns enhance communication and hence are assigned positive values. Noise patterns reduce communication possibilities and are given negative weighting.

(14)

		Signal	Noise
For each item in a pattern, the target-dialect			
a.	element is the same as that of the source dialect:	1.00	-0.25
b.	element is different from that of the source dialect		
	i. and does not occur in the source dialect	0.50	-0.50
	ii. and occurs elsewhere in the source dialect	0.25	-1.00

Since DOC provides phonological information, we use the initials, medials, nuclear vowel, ending, and tones to establish correspondence patterns. Hence a syllable-word is divided into these five elements. The weight for each element is one-fifth of 1. We are aware of the problems of such equal weighting for these five elements. But before we firmly establish a scale of perceptual distance among various sounds, we have no justifiable way to assign different values to consonants, vowels, and tones.

We feel that intelligibility is not necessarily symmetrical, and hence for a pair of

dialects we calculate two unidirectional intelligibility indices. Then we take the mean to be the mutual intelligibility value. Following is an example of the few lines in calculating the value using Beijing as the source dialect and Jinan as the target dialect:

(15)

		frequency	mean	weight	value	sum
zero	: zero	290	160.0	0.20	58.00	58.00
zero	:	30	160.0	-0.10	-3.00	55.00
	: zero	2	83.5	-0.20	-0.40	54.60
	:	165	83.5	0.20	33.00	87.60

We start with the zero initial in Beijing and continue to include all the other initials, medials, vowels, endings, and tones. Again, the weight is one-fifth of that given in (14). The sum is accumulated for all the elements. At the end, its value is 2004.25. There are 2,763 syllables in DOC for this pair of dialects. The intelligibility of Beijing as the source dialect and Jinan as the target dialect is 0.725 [2004.25 / 2763]. We then use Jinan as the source dialect and Beijing as the target dialect to calculate the intelligibility. The value is 0.713. We take the mean of 0.719 as the mutual intelligibility value of Beijing and Jinan. The mutual intelligibility indices for all the 272 pairs of the 17 dialects as presented in Cheng [1996] are given in Appendix 5.

Using the average linking method of cluster analysis we can establish a grouping of the dialects as given in Appendix 6. Wang [1996] presents other methods for establishing trees of relations. Here we will use the mutual intelligibility indices to illustrate the cluster analysis with average linking. To establish grouping we rank the pairs. Following are some of the highest-valued pairs:

(16)

1	.795	Hankou-Chengdu
2	.768	Jinan-Xi'an
3	.727	Beijing-Hankou
4	.726	Beijing-Chengdu

The closest pair of Chengdu and Hankou first forms a cluster. In Appendix 6 we see the lines join at .795 on the scale. Next, Jinan and Xi'an are joined at .768. Now we consider the next highest pair, Beijing and Hankou. Since Hankou was already grouped with Chengdu, we now need to join Beijing with the group of Hankou and Chengdu. The average linking method takes the average of the sum of the Beijing-Hankou (.727) and Beijing-Chengdu (.726) and link Beijing with the group of Hankou and Chengdu at the

average point .726 ((.727 + .726) / 2 = .726). Thus we see in Appendix 6 that Beijing joins Hankou and Chengdu at .726. As we complete the rank list and join all the dialects we get the grouping of Appendix 6. The grouping trees in terms of genetic phonology and the lexicon given in Appendix 2 and Appendix 4 respectively were established in the same way.

8. Future Development of DOC

In the thirty years from 1966 to 1996, DOC served as the empirical basis for lexical diffusion, for measurements of dialect similarity, and for quantification of mutual intelligibility among Chinese dialects. It has helped the formation of theories and establishment of numerical methodologies. In the near future we plan to enrich its contents and to add user interface for easy access by a larger number of users.

In 1989 the Second Edition of our source, the Zihui, was published [Beijing University 1989]. In this revision, errors in the first edition were corrected. It contains over 200 more words than the first edition. Hefei, Yangjiang, and Jian'ou localities have been added so that the new edition covers the pronunciations of 2,961 words in 20 dialects. We have been making corrections of our data and adding new words to DOC on the basis of this edition, but a large-scale update should be done soon.

Recently the Second Edition of the Hanyu Fangyan Cihui [Beijing University 1995] was published. It contains the lexical variants of 1,230 Putonghua items in 20 dialect localities. These localities are the same as those covered in the Second Edition of the Zihui. As discussed earlier, we used the first edition to measure dialect affinity in terms of the lexicon. When the second edition is implemented on the computer, we will have the same dialect localities for phonological and lexical comparisons. Dialect relations can now integrate the phonological and lexical information given in these volumes.

The user interface is not a simple function to implement. A database management system is supposed to provide retrieval of all sorts of information. And yet a query language would require some efforts on the part of the user to learn to use it . On the other hand, one could write a set of utilities for the user. But the functions of the utilities will only gather the types of data that were already used before. That is, they were for the kind of theory that required the creation of the functions. Often, they cannot be used to obtain data for new ideas. In the past we happened to be able to write our own computer programs to investigate any aspects of the DOC data. For a greater dissemination of DOC, we need to have a database management and a set of utilities to help the users who are not able to write computer programs.

REFERENCES

Beijing University, *漢語方音字匯 (Chinese Dialect Character Pronunciation List)*, 1962, Beijing: Wenzi Gaige Chubanshe.

Beijing University, *漢語方言詞匯 (Chinese Dialect Word List)*, 1964, Beijing: Wenzi Gaige Chubanshe.

Beijing University, *漢語方音字匯 第二版 (Chinese Dialect Character Pronunciation List, Second Edition)*, 1989, Beijing: Wenzi Gaige Chubanshe.

Beijing University, *漢語方言詞匯 第二版 (Chinese Dialect Word List, Second Edition)*, 1995, Beijing: Yuwen Chubanshe.

Chen, Matthew Y. and Ovid J.L. Tzeng, (eds.), *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, 1994, Taipei: Pyramid Press.

Cheng, Chin-Chuan, "DOC Handbook (Sections 3, 4, and 5)", *Monthly Internal Memorandum*, August 1970, Phonology Laboratory, Project on Linguistic Analysis, Berkeley: University of California.

Cheng, Chin-Chuan, "A Quantification of Chinese Dialect Affinity", *Studies in the Linguistic Sciences*, 12:1 1982, pp. 29-47.

Cheng Chin-Chuan, "漢語方言親疏關係的計量研究 (Quantitative Studies of Chinese Dialect Affinity)", *中國語文 (Zhongguo Yuwen)*, 203 1988, pp. 87-102.

Cheng, Chin-Chuan, "Quantifying Affinity among Chinese Dialects", in Wang (ed.) *Language and Dialects of China*, 1991, pp. 78-112.

Cheng, Chin-Chuan, "Syllable-based Dialect Classification and Mutual Intelligibility", *Chinese Languages and Linguistics 1 Chinese Dialects*, Symposium Series Number 2 1992, pp. 145-177, Taipei, Taiwan: Institute of History and Philology, Academia Sinica.

Cheng, Chin-Chuan, "DOC: Its Birth and Life", in Matthew Y. Chen and Ovid J.L. Tzeng, (eds.), *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, 1994a, pp. 71-86.

Cheng, Chin-Chuan, "漢語方言溝通度的計算 (Calculation of Chinese Dialect Mutual Intelligibility)", *中國語文 (Zhongguo Yuwen)*, 238 1994b, pp. 35-43.

Cheng, Chin-Chuan, "Quantifying Dialect Mutual Intelligibility", in Huang and Li, (eds.), *New Horizons in Chinese Linguistics*, 1996, pp. 269-292.

Huang, James and Audrey Li, (eds.), *New Horizons in Chinese Linguistics*, 1996, Boston: Kluwer Academic Publishers.

Ma, Xiwen, "比較方言學中的計量方法 (Quantitative Methods in Comparative

Dialectology)", *中國語文 (Zhongguo Yuwen)*, 212 1989, pp. 348-360.

Streeter, Mary, "DOC: 1971", *Computers and the Humanities*, 6 1972, pp. 259-270.

Streeter, Mary, "DOC, 1971: A Chinese Dialect Dictionary on Computer", in Wang (ed.), *The Lexicon in Phonological Change*, 1977, pp. 101-119.

Tu, Wen-Chiu and Chin-Chuan Cheng, "A Linguistic Classification of Rukai Formosan", Paper presented at the Sixth International Conference on Austronesian Linguistics, May 20-24, 1991, Honolulu, Hawaii.

Wang, William S-Y, "Competing Changes as a Cause of Residue", *Language*, 45 1969, pp. 9-25.

Wang, William S-Y, "Project DOC: Its Methodological Basis", *Journal of the American Oriental Society*, 90 1970, pp. 57-66.

Wang, William S-Y, (ed.), *The Lexicon in Phonological Change*, 1977, The Hague: Mouton.

Wang, William S-Y, (ed.), *Language and Dialects of China, Journal of Chinese Linguistics Monograph*, 3 1991.

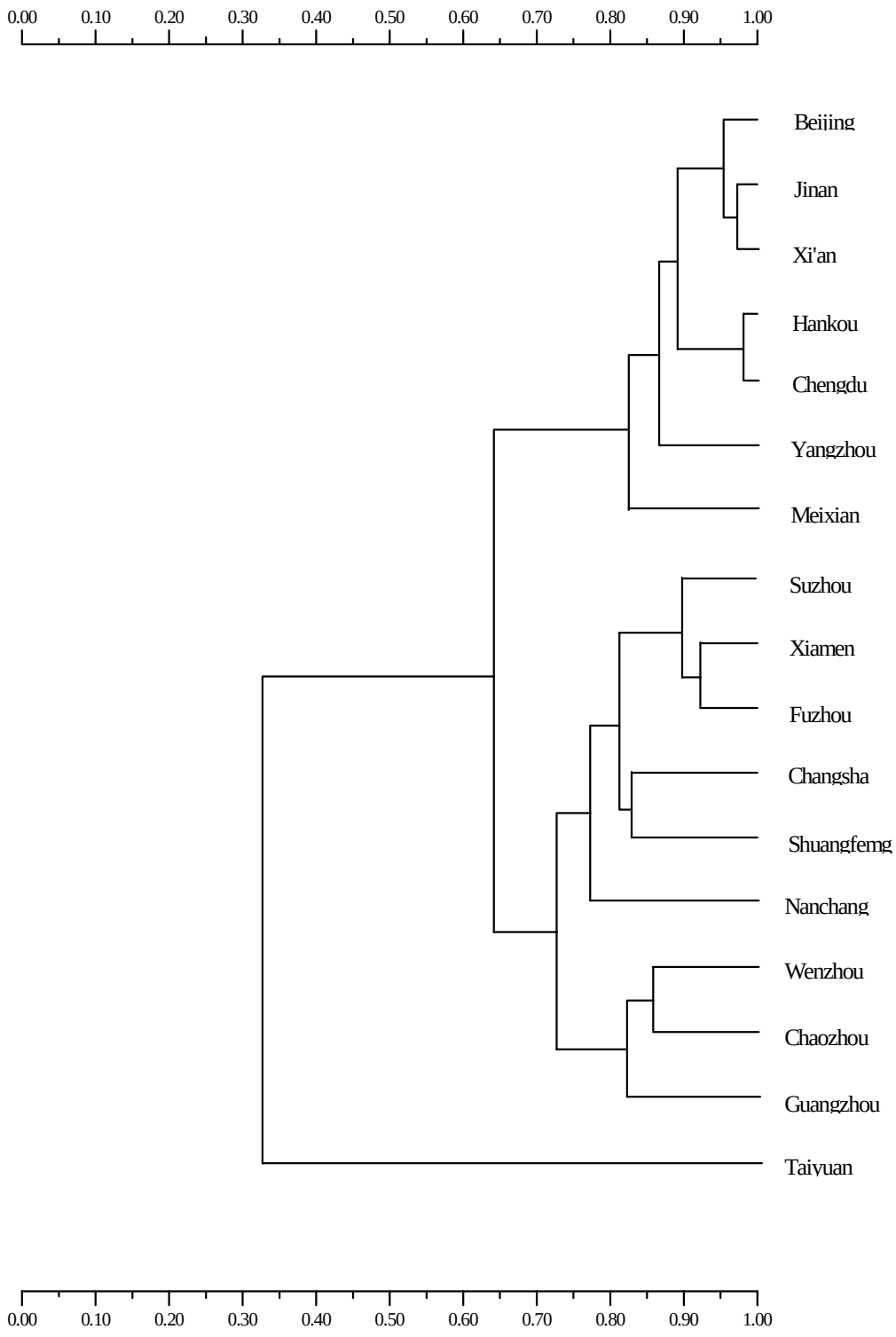
Wang, William S-Y., "Linguistic Diversity and Language Relationships", in Huang and Li, (eds.), *New Horizons in Chinese Linguistics*, 1996, pp. 235-267.

Wang, William S-Y., and Zhongwei Shen, "方言關係的計量表述 (A Quantitative Description of the Relationship Among Chinese Dialects)", *中國語文 (Zhongguo Yuwen)*, 227 1992, pp. 81-92.

Yaruss, Jonathan Scott, "DOC 1988: The Modernization of a Chinese Dialect Dictionary on Computer", *Computers and the Humanities*, 24 1990, pp. 207-219.

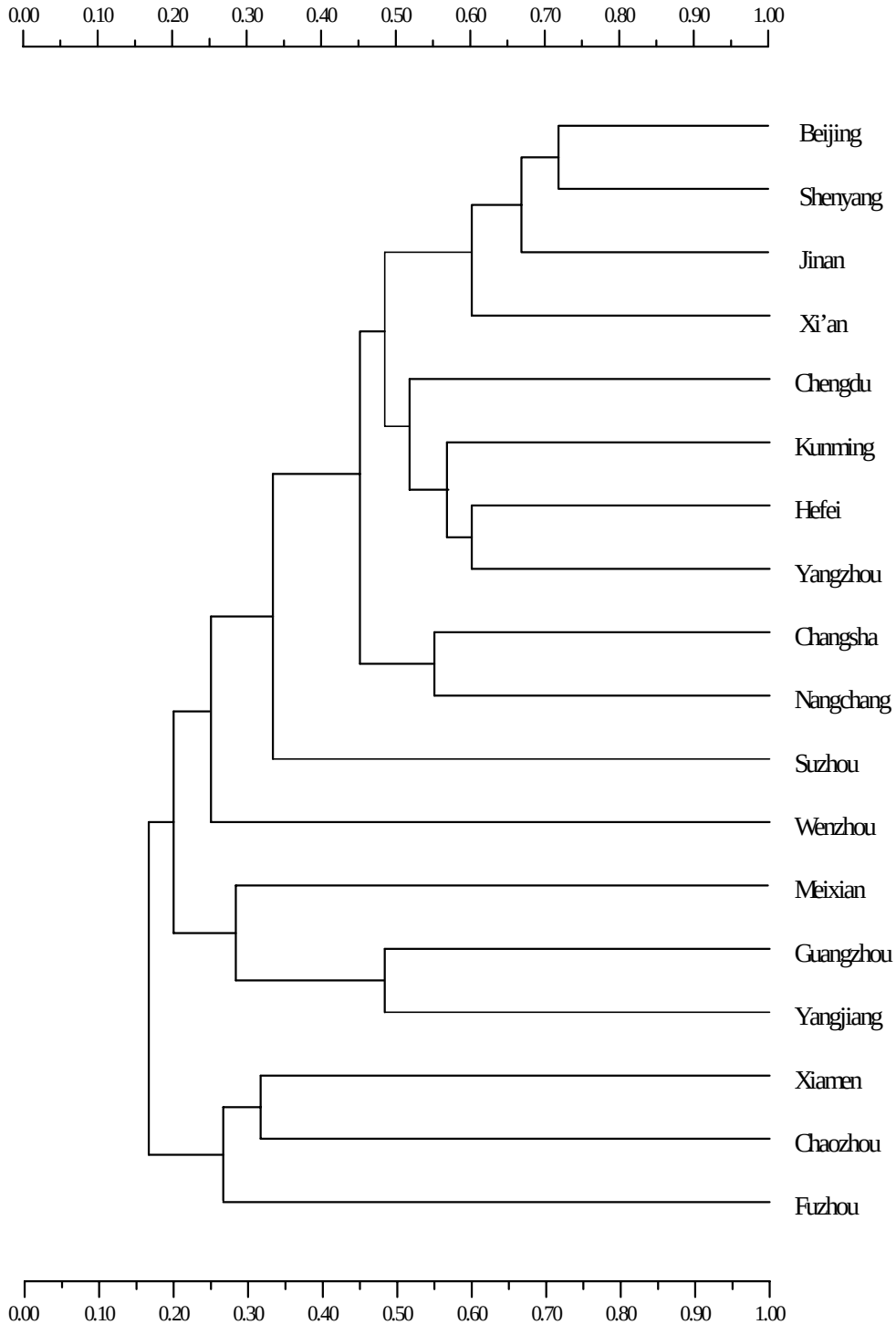
Appendix 1. Correlation Coefficients -- Initials, Finals, and Tones

Appendix 2. Dialect Affinity Based on Genetic Relations of Initials, Finals, and Tones



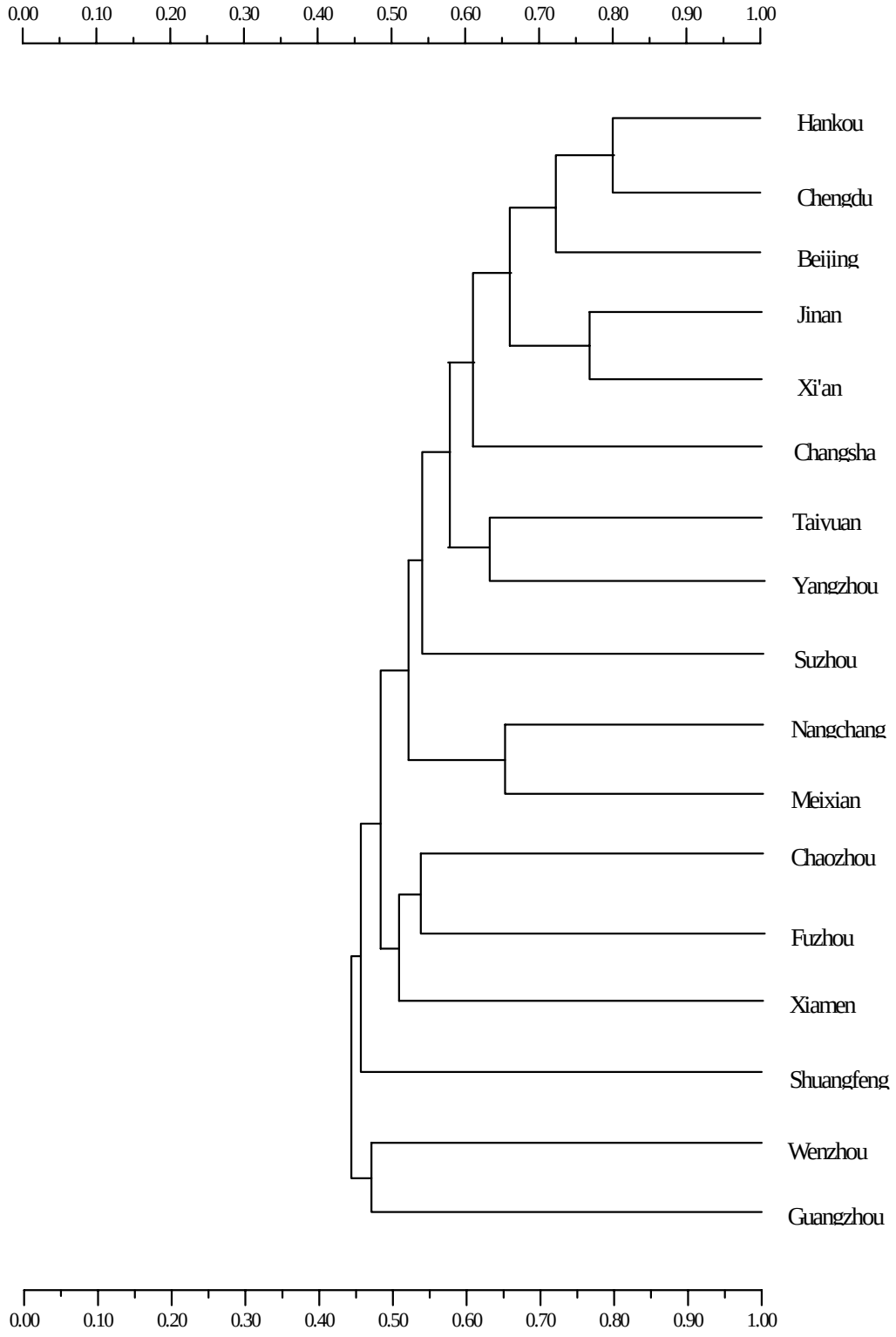
Appendix 3. Correlation Coefficients -- Lexicon

Appendix 4. Dialect Affinity Based on Lexicon


0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Beijing
Shenyang
Jinan
Xi'an
Chengdu
Kunming
Hefei
Yangzhou
Changsha
Nangchang
Suzhou
Wenzhou
Meixian
Guangzhou
Yangjiang
Xiamen
Chaozhou
Fuzhou
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

Appendix 5. Dialect Mutual Intelligibility

Appendix 6. Dialect Affinity Based on Mutual Intelligibility



Appendix 7. DOC 84 File Formats and Coding Conventions

A. Formats

A DOC record contains 22 characters. The format for Middle Chinese entries and that for dialect information are different. "Blank" and "space" are used interchangeably to refer to the space character.

a. Middle Chinese Records

Position	Contents
1-4	4-digit telegraphic code
5	A letter beginning with "A" to differentiate homographic characters; Blank if none
6-7	Blank to identify Middle Chinese record
8-10	Zihui page number
11	Character position on the Zihui page (1 through 9)
12-13	Code for She (攝) , the rime group (see coding below)
14	Kai-He lip rounding ("K" for spread and "H" for rounding)
15	Deng (等) or vowel grade (1 through 4)
16	Tone (1 through 4) for " 平上去入 "
17-18	Yun (韻) or rime (see coding below)
19-22	Initials (see coding below)

b. Dialect Records

Position	Contents
1-4	4-digit telegraphic code
5	A letter beginning with "A" to differentiate homographic characters; space if none
6	Dialect ID (see coding below)
7	0 to indicate no variant pronunciations; 1, 2, etc. to show variants
8-9	tonal category (see coding below)
10-13	Initial (see coding below)
14-15	Medial (see coding below)
16-17	Nucleus (see coding below)
18-19	Off-glide (see coding below)
20	Vowel nasalization: Z for nasalization; blank for none

21 Ending (see coding below)
22 Miscellaneous: "L" for Chinese literary form; vowel for Sino-Japanese

B. Coding Conventions

I. Middle Chinese Record Coding

a. Rime Groups and Rimes

The rime groups (攝) and rimes (韻) each taking up two character positions are coded as follows. The rime groups are traditionally classified as inner (內轉) or outer (外轉) groups. The first letter of the inner groups is "N" or "O" and that of the outer groups is "W", "X", or "Y". (The rime name with an asterisk indicates that a more common character is used for the original one which I could not find in the Big5 Chinese character set and that I did not create it in the user font.)

通 NG

東	11	董	12	送	13	屋	14
冬	21			宋	23	沃	24
鍾	31	腫	32	用	33	燭	34

江 WG

江	11	講	12	絳	13	覺	14

止 NI

支	11	紙	12	寘	13
脂	21	旨	22	至	23
之	31	止	32	志	33
微	41	尾	42	未	43

遇 NO

模	11	姥	12	暮	13
魚	21	語	22	御	23
虞	31	噳 *	32	遇	33

蟹 WI

咍	11	海	12	代	13
		泰	01		
灰	21	賄	22	隊	23
皆	31	駭	32	怪	33

佳	41	蟹	42	卦	43		
		夬	02				
		祭	03				
		廢	04				
齊	51	薺	52	霽	53		

臻 NN

痕	11	很	12	恨	13	紇	14
魂	21	混	22	慁	23	沒	24
臻	31			節	34		
真	41	軫	42	震	43	質	44
諄	51	準	52	稕	53	術	54
欣	61	隱	62	焮	63	迄	64
文	71	吻	72	問	73	物	74

山 WN

寒	11	旱	12	翰	13	曷	14
桓	21	緩	22	換	23	末	24
刪	31	潸	32	諫	33	轄 *	34
山	41	產	42	襉	43	黠	44
仙	51	獮	52	線	53	薛	54
元	61	阮	62	願	63	月	64
先	71	銑	72	霰	73	屑	74

效 WU

豪	11	皓	12	號	13
肴	21	巧	22	效	23
宵	31	小	32	笑	33
蕭	41	篠	42	嘯	43

果 XO

歌	11	哿	12	箇	13
戈	21	果	22	過	23

假 WO

麻	11	馬	12	禡	13

宕 YG

唐	11	蕩	12	宕	13	鐸	14
陽	21	養	22	漾	23	藥	24

梗 XG

庚	11	梗	12	映	13	陌	14
耕	21	耿	22	諍	23	麥	24
清	31	靜	32	勁	33	昔	34
青	41	迥	42	徑	43	錫	44

曾 OG

登	11	等	12	嶝	13	德	14
蒸	21	拯	22	證	23	職	24

流 NU

侯	11	厚	12	候	13
尤	21	有	22	宥	23
幽	31	黝	32	幼	33

咸 WM

覃	11	感	12	勘	13	合	14
談	21	敢	22	闞	23	盍	24
咸	31	豏	32	陷	33	洽	34
銜	41	檻	42	鑑	43	狎	44
鹽	51	琰	52	豔	53	葉	54
嚴	61	儼	62	釅	63	業	64
凡	71	范	72	梵	73	乏	74
添	81	忝	82	㮇 *	83	帖	84

深 NM

侵	11	寢	12	沁	13	緝	14

b. Initials

Each code of the Middle Chinese initials occupies four character positions. In order to align the code here the underscore character is used to represent the space character.

唇	幫	P___	滂	P__H	並	B___	明	M___
	非	F___	敷	F__H	奉	V___	微	MV__
舌	端	T___	透	T__H	定	D___	泥	N___

	知	T_P_	徹	T_PH	澄	D_P_	娘			
牙	見	K___	溪	K__H	群	G___	疑	NG__		
齒	精	TS__	淸	TS_H	從	DZ__	心	S___	邪	Z___
	莊	TSR_	初	TSRH	崇	DZR_	生	S_R_	俟	Z_R_
	章	TSP_	昌	TSPH	船	DZP_	書	S_P_	禪	Z_P_
喉	影	Q___	曉	X___	匣	GR__	云	J___	以	0___
舌齒	來	L___	日	N_P_						

II. Dialect Record Coding

a. Dialect Identification

A: 北京	I: 溫州	Q: 福州
B: 濟南	J: 長沙	R: 上海
C: 西安	K: 雙峰	X: Kan-on Sino-Japanese
D: 太原	L: 南昌	Y: Go-on Sino-Japanese
E: 漢口	M: 梅縣	Z: Sino-Korean
F: 成都	N: 廣州	
G: 揚州	O: 廈門	
H: 蘇州	P: 潮州	

b. Tone (2 character positions)

DOC 71 did not use "A" for Yin tones. "A" was added later to show the Yin-Yang distinction.

Modern dialects:

平： 1	陰平 : 1A	陽平 : 1B	上： 2
陰上 : 2A	陽上 : 2B	去： 3	陰去 : 3A
陽去 : 3B	入： 4	陰入 : 4A	陽入 : 4B
中入 : 4C			

Zhongyuan Yinyun:

陰平 : 1A	陽平 : 1B	上： 2	去： 3
入作陽： 41	入作上： 42	入作去： 43	

Sino-Xenic:

XX for no tone

c. Initials, Medials, and VowelsThe DOC 84 code is given in the first line, the corresponding IPA in the second line, and the key code of DOC 93 in the third line of each group.