The Wenzhou Spoken Corpus

John Newman,[1] Jingxia Lin[2]
Terry Butler[1] and Eric Zhang[1]

Abstract

The creation of the Wenzhou Spoken Corpus, an online searchable corpus of a modern Chinese dialect, presents a number of challenges that are of interest to the corpus linguistic community. We review issues involved with collection of spoken data, its transcription and markup, as well as the functionality of the search tools. The transcription makes use of Chinese characters as well as IPA symbols for Wenzhou colloquial forms not conventionally represented by characters. XML was adopted as the standard for the basic format of files, with file searches expressed in XPath form. The search tools provide the usual options of restricting searches by age, gender, *etc*., and yield concordances and tables of collocates. Though the collection of data for the corpus was 'opportunistic' in some ways, and so not ideally balanced or representative, it is nevertheless proving to be a valuable tool for corpus-based research on Wenzhou.

1. Introduction

The Wenzhou Spoken Corpus (WSC) is an online searchable corpus of spoken Wenzhou, a southern dialect of Chinese, spoken in and around the city of Wenzhou.[3] In the following sections we describe this corpus and the associated online search tool.[4] The overall concept of a corpus like the WSC is not, in itself, novel, but we believe that the particular challenges of this project, and how we responded to them, involve a number of interesting issues of relevance to the field of corpus linguistics. The issues are wide-ranging: Chinese dialect study, spoken versus written genres, transcription, markup, exploitation of the XML encoding, search tools and collocation results.

[1] Department of Linguistics, Faculty of Arts, University of Alberta, 4–32 Assiniboia Hall, Edmonton, Alberta, T6G 2E7, Canada
Correspondence to: John Newman, *e-mail*: john.newman@ualberta.ca
[2] Department of Asian Languages, 450 Serra Mall Building 50, Main Quad, Stanford University, Stanford, CA 94305–2034, USA

[3] We are grateful to two anonymous reviewers of this paper for their very helpful suggestions.

[4] The WSC is available at: http://corpora.tapor.ualberta.ca/wenzhou/

2. Rationale and project team

While the study of contemporary usage is an important aspect of the study of any language, it is particularly important to emphasise this aspect in the case of languages where there is a strong tradition of studying historical texts. This is clearly the case in the Chinese culture, where an ancient and venerable literary tradition demands respect. Without denying the value of studying ancient texts in this tradition, one must make a special effort in such cases to ensure that the study of contemporary language use is not unduly neglected. For the study of contemporary Mandarin, there exist valuable online searchable corpora including the Academia Sinica Balanced Corpus of Modern Chinese (Huang and Chen, 1992; Huang *et al.*, 1995) and the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004).[5] In the case of Chinese, it is also important to document contemporary usage of the less prestigious Chinese dialects. There is a strong tendency to focus on Mandarin (Putonghua) as an object of study, which is understandable in light of its educational significance and its symbolic importance as a unifying factor within China; but of course the Chinese dialects are no less interesting than Mandarin from a linguistic perspective. Indeed, one could argue that they are all the more interesting on account of a relative neglect. It is against this background of ensuring that proper attention is given to the study of the Chinese dialects that the WSC was initially conceived. A corpus contributing to this kind of academic goal and which provided some inspiration for the WSC is the *Hong Kong Cantonese Adult language Corpus* (HKCAC), a 170,000-character corpus based on phone-in programs and forums aired on Hong Kong radio (Leung and Law, 2002).[6]

The initial interest in an online searchable Wenzhou corpus (similar in spirit to the HKCAC) came from two of the co-authors: Lin, at the time a graduate student in the Department of Linguistics, and Newman, her supervisor. Assistance with the more technical aspects of creating the corpus was provided by the other two co-authors, Butler and Zhang, from the University of Alberta TAPoR node. [7]

3. The corpus

It was decided that the corpus would consist of six spoken genres: (1) Face to Face Conversation, (2) Phone Call, (3) Internet Chat (audio), (4) Story,

[5] The Academia Sinica Balanced Corpus of Modern Chinese is available at: http://www.sinica.edu.tw/SinicaCorpus/. The Lancaster Corpus of Mandarin Chinese is available at http://bowland-files.lancs.ac.uk/corplang/lcmc/

[6] The HKCAC is available from: http://shs.hku.hk/corpus/main.htm

[7] TAPoR is a pan Canadian, multi-institutional assembly of computing infrastructure and expertise, supported by a Canada Foundation for Innovation grant. The University of Alberta TAPoR node is: http://tapor.ualberta.ca/

(5) News Commentary and (6) Song. These six genres provide a mix of informal and formal contexts, as well as private and public use of language. Although there are more genres relevant to the use of Wenzhou, these particular ones were relatively easy to obtain data for. A breakdown of the size of the corpus by genre is shown in Table 1. In this table, ‘Word Count’ refers to the number of items identified as linguistic words (consisting of one or more Chinese characters) and ‘Unicode Character Count’ refers to the number of Unicode characters (Chinese character, IPA symbol or punctuation mark). While not large by comparison with some corpora, it is a respectable size for a corpus of transcribed spoken data and its size compares favourably with the HKCAC. As a point of interest, the WSC provides ample evidence for all the syntactic traits claimed for Wenzhou in Pan (1991: 269–77), including the 个 [kai] classifier functioning as a demonstrative pronoun, the 个 [kai] classifier being used in the same way as the Mandarin connective *de* 的, direct objects before indirect objects, adverbs occurring post-verbally, and agent in a passive construction obligatorily present.

	Word count	*Unicode character count*
1 Face to face conversation	13,009 (8.22%)	23,582
2 Phone call	20,885 (13.20%)	36,257
3 Internet chat	7,005 (4.42%)	13,132
4 Story	1,046 (0.66%)	2,470
5 News commentary	115,293 (72.90%)	179,708
6 Song	894 (0.56%)	1,395
Total:	158,132	256,544

Table 1: Size of the WSC by genre

For genres (1)–(4), the method of obtaining the data was to rely on the social networks of Lin which included family members, friends, friends of friends, *etc*. The Internet Chat data is based on spoken exchanges, rather than typed messages. For Story data, the speaker was given a short story written in Mandarin (characters) to read silently and then asked to tell the story in Wenzhou without looking at the text. The informal context of the

Story sessions allowed for a certain amount of free conversation, though the narration of the story dominates in each case. News Commentary was collected from the programme *Baixiao Jiang Xinwen* 'News talk by know-it-all', a netcast provided by the Economic and Science Channel, Wenzhou TV, and used with their permission (which was readily given). The programme includes relatively informal news commentaries as well as opinions offered by anonymous interviewees. More than 72 percent of the total word count comes from the recorded News Commentary (where age and education level is unknown), reflecting the relative ease of access to this category of data. The Song category is composed of the texts of traditional Wenzhou children's songs. Just over 25 percent of the total words derive from the more spontaneous, conversational contexts (Face to Face Conversation, 8 percent; Phone Call, 13 percent; Internet Chat, 4 percent) and in these categories there is a clear predominance of speakers under thirty-four years of age and educated to at least high school level. The method was clearly 'opportunistic' with a resulting over-representation of some demographic categories, as shown in Table 2. Overall, male speakers outnumber female speakers by a ratio of about 3:1.

	No. of words	*Percentage of total no. of words (158,132)*
Gender		
Male	105,880	66.9
Female	35,175	22.2
Unknown	17,077	10.8
Age		
0–14	268	0.2
15–24	27,009	17.1
25–34	44	0.0
35–44	0	0.0
45–59	6,792	4.3
60+	7,799	4.9
Unknown	116,220	73.5
Education level		
Illiterate	5,487	3.5
Elementary school	2,998	1.9
High school	6,914	4.4
College+	26,513	16.8
Unknown	116,220	73.5

Table 2: Distribution of demographic categories in the WSC

The uneven demographics underlying the corpus are far from ideal. The lack of information on age and education level for 73.5 percent of the corpus (a result of relying extensively on the News Commentary category), for example, is a severe shortcoming in our metadata. Such shortcomings can be overcome, to some extent, by restricting searches to particular domains by age band, level of education, and so on. Most of the conversational data was collected in downtown Wenzhou and Yueqing city. All the data derives from the years 2004–2005.

As with any collection of linguistic data for academic purposes, it was necessary to obtain approval for our data collection from the relevant university ethics board. While there was no concern raised by the board about the planned project, a question arose about the status of some data which had been collected before the formal beginning of the project (that is, the date on which ethics approval was given). It was agreed that anyone who had supplied data for the genre types (1)–(4) prior to the formal commencement of the project would be approached and asked for their co-operation in allowing that data to be used as part of the project. This would have been a daunting task had the data collection been larger, but it was in fact relatively easy to accomplish since the relevant speakers represented a small number of friends and family members.

4. File markup

An early decision was made to use XML for our file structure, recognising the widespread acceptance of the XML standard. Audio data was transcribed into Unicode Chinese characters, consistent with our decision to encode features of the corpus with XML markup. An immediate problem that presented itself concerned what to do with Wenzhou forms that do not match any Unicode Chinese characters in Unicode 4.1 (2005), e.g., [ts'ɿ] 'see' and [koŋ] 'just now'. One solution to this problem would have been to create new character encodings for these forms, using the 'Private Use Area' mechanism which is established for Unicode. However, such characters would not have appropriate glyphs associated with them, and users of our search tool would not be able to see them displayed. For these reasons, we opted to represent the Wenzhou forms currently lacking Unicode Chinese glyphs in IPA transcription. The *Wenzhou Fangyan Dictionary* (You and Yang, 1998) was used as a reference both for the Chinese characters and for the normalised phonetic transcription. Sentence-final particles are not always entered in the *Wenzhou Fangyang Dictionary* (and they are subject to considerable sociolinguistic variation), so we transcribed the actual pronunciations in these cases. Pauses were not represented. Personal names and other confidential information were removed to ensure the anonymity of participants. Spoken Wenzhou has a number of phonetically-contracted forms which required some procedural decisions concerning transcription. The negative morpheme [fu] 不, for example, combines with certain following morphemes to produce contracted forms: [fu ha] 不匄 'not give' can be reduced to [fa], [fu hə] 不好 'not good' can be reduced to [fə]. Here, too, we followed the practice of You and Yang (1998), using a single character where the dictionary lists one (孬 for [fə] 'not good') but using two characters otherwise (as in the case of [fu ha] 不匄 'not give').

Each conversation, chat, *etc*., was prepared as a separate XML file. Conversations were broken down into a series of participant turns; each participant's turn was further subdivided into one or more utterances. We had found the use of word tags in the XML files of the Lancaster Corpus of Mandarin Chinese, demarcating one or more characters as a word unit, particularly helpful; so we decided to adopt those markup conventions in our files. The Chinese characters and IPA forms were therefore grouped into word units where it was relatively clear that we were dealing with a word compound, e.g., 温州 'Wenzhou' and 学堂 'school'. In advance of a more complete study, it is not always easy to decide on the morphological status of some sequences, so we were conservative in our approach to identifying compounds. So, for example, 造 'build'+ 起 'rise' together means 'build up', but each of these parts could be analysed as verbs in their own right (depending on the criteria used for identifying verbs). In this case, the two characters were left as separate words.

We used a three-stage process to prepare the XML files: (i) transcription (Chinese characters, with words separated by spaces and IPA transcription) was carried out straightforwardly in a Microsoft Word document; (ii) some manual tagging (e.g., turns, utterances and overlapping speech) was then added to the transcription using XMLSpy,[8] (iii) finally, a Perl script automatically inserted word and punctuation tags. XMLSpy was used for construction of the DTD and validation of files. An overview of the tags used can be found in Appendix A, along with brief descriptions of them. Example (1), below, shows the DTD used for the XML markup and example (2) provides an example of this markup, illustrating a number of features: numbering of turns and provision of the 'id' number of the speaker of each turn; numbering of utterances within a turn; identification of overlapping speech by its group number within the file (gid='4') and unique identification of utterances within this group (oid='2', oid='3'); phonetic transcription (<phonetic>); word boundaries (<w>); and punctuation (<c>).

[8] XMLSpy is a suite of XML development programs provided by Altova, see: http://www.altova.com

(1) DTD underlying the XML markup:

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT file (turn+)>
<!ELEMENT turn (utterance+)>
<!ELEMENT utterance (w | c | mixed | phonetic |
unclear | desc | overlap)*>
<!ELEMENT mixed (w | c | phonetic | unclear | desc
| overlap)*>
<!ELEMENT phonetic (w | c | mixed | unclear | desc
| overlap)*>
<!ELEMENT unclear (w | c | phonetic | mixed | desc
| overlap)*>
<!ELEMENT overlap (w | c | phonetic | unclear |
desc | mixed)*>
<!ELEMENT w (#PCDATA)>
<!ELEMENT c (#PCDATA)>
<!ELEMENT desc (#PCDATA)>
<!ATTLIST file type CDATA #REQUIRED>
<!ATTLIST turn tid CDATA #REQUIRED sid CDATA
#REQUIRED>
<!ATTLIST utterance uid CDATA #REQUIRED>
<!ATTLIST overlap gid CDATA #REQUIRED oid CDATA
#REQUIRED>
```

(2) Sample of XML markup:
With tags and markup

```
<turn tid="27" sid="S003">
    <utterance uid="1"><overlap gid="4" oid="2">
    <w>([娘娘</w> <w>讲</w> <w>故事</w> <w>匄</w>
    <w>你</w> <w>听</w> <c>。])</c> </overlap>
    </utterance>
</turn>
<turn tid="28" sid="S007">
    <utterance uid="1"><overlap gid="4"
    oid="3"><phonetic> <w>([[[fai]]</w> </phonetic>
    <w>吵</w> <phonetic> <w>[[fai]]</w> </phonetic>
          <w>吵</w> <c>,])</c></overlap> <w>太婆</w>
    <w>讲</w> <w>来</w> <w>先</w> <c>。</c>
    </utterance>
</turn>
```

Without tags or markup

```
S003: ([娘娘 讲 故事 匄 你 听。])
S007: ([[[fai]] 吵 [[fai]] 吵 ,]) 太婆 讲 来 先。
```

English translation (*not available through WSC*)

```
  S003: Aunt is going to tell you stories for you to
  listen to.
  S007: Be quiet, be quiet. Great-grandma is about to
  tell some stories.
```

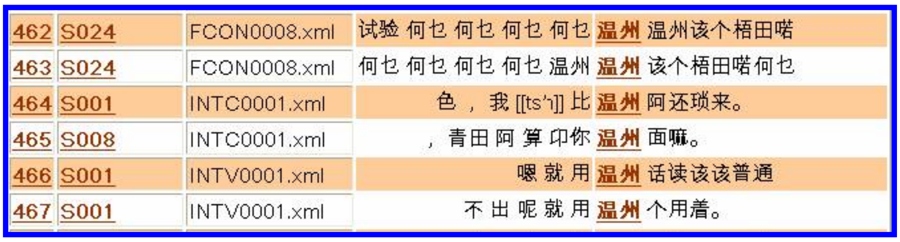
5. The search and display tools

The XML-encoded transcripts were loaded onto a web server. Here, software written in the PHP programming language was developed to create the web pages where the corpus can be searched and displayed. PHP proved to be well-suited to this task: it can generate web pages which combine a user-friendly design with the power to process the XML-encoded text. The PHP server software reads the XML files and translates the search requests into specific searches into the XML structure. These searches are expressed using the XPath standard form, which means this approach could be quickly adapted to other corpora and different XML encoding. In this way, a wide range of search requests can be executed directly on the XML data, and the results returned in a web page display.

There are two main search tools which are available as part of the interface to the WSC: Concordance and Collocates. Both tools provide the same set of options to restrict the search according to selections by gender, age group and level of education.

An example of a concordance display produced by the Concordance tool is shown in Table 3, using the default setting of five words (or punctuation marks) to the left and right of the search term, up to the beginning or end of an utterance. The display design was influenced by BNCWeb, the web-based interface of the British National Corpus (Hoffmann and Evert, 2006).[9] In particular, we incorporated the option of clicking on the speaker id (e.g., S024) to bring up the demographic details of the speaker (e.g., male, age 45–59) and clicking on the key word to bring up an expanded context of the concordance line. In the expanded context, one can toggle between displaying and hiding tags. In the case of the WSC, the expanded context amounts to the preceding and following turns. Other fields display the number of the hit and the file name. A further option allows for the display (and saving) of the concordance results, along with complete demographic information relating to each speaker.

[9] More information on BNCWeb is available from: http://homepage.mac.com/bncweb/home.html



462	S024	FCON0008.xml	试验 何乜 何乜 何乜 何乜	温州	温州该个梧田喏
463	S024	FCON0008.xml	何乜 何乜 何乜 何乜 温州	温州	该个梧田喏何乜
464	S001	INTC0001.xml	色 ， 我 [[tsʻʅ]] 比	温州	阿还琐来。
465	S008	INTC0001.xml	， 青田 阿 算 卬你	温州	面嘛。
466	S001	INTV0001.xml	嗯 就 用	温州	话读该该普通
467	S001	INTV0001.xml	不 出 呢 就 用	温州	个用着。

Table 3: Sample of a concordance display (screen image)

The Collocates tool offers two options for displaying results. An 'aggregated' display lists all the word types and the number of tokens in the selected span. A 'collocates by position' display shows the collocates by individual position, following the suggestion by Stubbs (2001: 87–96) who recommends such a display as a basis for lexical profiling. A sample of collocate results by position is shown in Table 4.

Two additional tools provide hits without any keyword searches, enabling a user to browse the overlapping speech and non-Wenzhou forms. Overlapping speech is common in conversation and is an interesting topic of study in its own right. Overlapping speech was marked in the corpus and an option has been included to allow for inspection of all the overlapping speech, restricted to any of the demographic categories. Equally, switching between languages is worthy of study and, indeed, some non-Wenzhou language, e.g., Mandarin and English, makes an appearance in the corpus. An option is provided for retrieving utterances that contain such forms.

6. Future development

Even in its first release, the WSC has proved invaluable to the study of Wenzhou. Nevertheless, it was envisaged that the corpus could be expanded on in the course of time – subject, of course, to interest and funding support. In some ways, it is surprising how much we have been able to achieve without any major research funding for the project.[10] Naturally, we hope to be able to increase the size of the corpus. Plans have already been made to incorporate an IPA transcription and dictionary meaning for each character, following You and Yang (1998), so that users may read this information by moving the cursor over a character in a concordance line. The use of 'standard' Chinese characters for Wenzhou is potentially confusing since the Wenzhou usage can be quite different from

[10] We readily acknowledge support for research assistance from the Department of Linguistics, as well as support from TAPoR at the University of Alberta.

–5	–4	–3	–2	–1	*	1	2	3	4	5
个(22)	个(31)	呢(17)	是(22)	印你(82)	*	市(48)	个(44)	个(33)	个(33)	个(34)
呢(12)	呢(15)	个(15)	个(15)	宿(28)	*	话(43)	呢(19)	里(19)	呢(14)	呢(15)
有(11)	温州(12)	人(13)	呢(15)	个(26)	*	人(35)	有(12)	一(12)	有(12)	人(10)
倈(9)	人(11)	阿(11)	宿(12)	是(23)	*	个(34)	是(11)	有(12)	温州(12)	是(9)
该(8)	倈(8)	温州(8)	倈(11)	走(19)	*	呢(19)	站(10)	阿(8)	里(8)	有(9)
温州(7)	该(7)	讲(7)	该日(10)	讲(13)	*	有(12)	里(8)	局(8)	多(7)	一(7)
是(7)	一(7)	渠(7)	一(10)	拉(10)	*	火车(9)	哪(7)	呢(8)	倈(6)	温州(7)
里(6)	我(5)	倈(7)	就(9)	能界(9)	*	大学(7)	文明(7)	温州(8)	人(6)	会(6)

Table 4: Part of the results for the keyword 温州 ‘Wenzhou’ using the ‘collocates by position’ option (* = keyword)

that in Mandarin. For example, 賺 represents *zhuan* ‘earn’ in Mandarin, but is used to mean ‘wrong’ in Wenzhou; 近 is *jin* ‘near’ in Mandarin, but is used to mean ‘earn’ in Wenzhou. Obviously, providing the dictionary meanings will be of great assistance to users unfamiliar with Wenzhou written practice. A future version of WSC will also include statistical measures of word association (MI, *t*-score and *z*-score) which take into account overall frequencies of a key word and collocate in the whole corpus, or sub-corpus, being searched.

For WSC, it was not necessary to create a full relational database as a means of enhancing search and retrieval performance (as advocated, for example, by Davies, 2005). The modest size of the corpus and the relatively straightforward nature of the searches we have allowed for have meant that we are able to achieve fairly immediate display of results without relying on a relational database. The XML-encoded text is, in fact, structured like a database, and can be searched directly as such. Increasing the size of the corpus, and adding further options such as n-gram calculation, part of speech tagging, and dictionary definitions of characters might lead us to store the corpus in an XML-native database. This move would provide more powerful management and indexing features, with the advantage that the existing coding scheme and structure would not need to be changed at all.

References

Davies, M. 2005. ‘The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation’. International Journal of Corpus Linguistics 10 (3), pp. 307–34.

Hoffmann, S. and S. Evert. 2006. ‘BNCweb (CQP-Edition) – The marriage of two corpus tools’ in S. Braun, K. Kohn and J. Mukherjee (eds) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 177–95. Frankfurt am Main: Peter Lang. Available from http://es-sebhoff.unizh.ch/Hoffmann-Evert.pdf

Huang, C-R. and K-J. Chen. 1992. ‘A Chinese corpus for linguistic research’, Proceedings of COLING 92, pp. 1214–17.

Huang, C-R., K-J. Chen, L-P. Chang and H-L. Hsu. 1995. ‘An introduction to Academia Sinica Balanced Corpus’, [In Chinese], Proceedings of ROCLING VIII, pp. 81–99.

Leung, M-T. and S-P. Law. 2002. ‘HKCAC: The Hong Kong Cantonese Adult Language Corpus’, International Journal of Corpus Linguistics 6 (2), pp. 305–25.

McEnery, A. and Z. Xiao. 2004. ‘The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study’.

Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004, pp. 1175–78.
Available from:
http://www.lancs.ac.uk/postgrad/xiaoz/papers/231.pdf
Also available from: http://eprints.lancs.ac.uk/65/

Pan, W-Y. 1991. 'An introduction to the Wu dialects' in W.S-Y. Wang (ed.) Languages and Dialects of China, pp. 237–93. Journal of Chinese Linguistics, Monograph Series, Number 3.

Stubbs, M. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Unicode 4.1. 2005. 'The Unicode Standard for character encoding'. Available from: http://www.unicode.org/standard/standard.html. Last accessed: 27 February 2006.

You, R. and Q. Yang. 1998. 温州方言词典 [Wenzhou Fangyan Dictionary]. Nanjing: Jiangsu Jiaoyu Chubanshe [Jiangsu Education Press].

Appendix A: Tag types used in the WSC

Tag Type	*Description*
turn <turn>	A turn is a speaker's uninterrupted speech, marked with ordered numerals. A turn also includes the speaker id, which is linked to the background information of a specific speaker.
utterance <utterance>	Utterances are sentence-like segments within a turn and are also marked with ordered numerals.
word <w>	Words are one or more syllables/characters which have independent lexical status.
punctuation <c>	Punctuation marks used are: ，(comma), 。(period), “ ” (quotation), 《 》(book title), — (sudden stop or sudden switch of topic).
phonetic transcription <phonetic>	Phonetic transcription is marked with [[…]].
overlapping speech <overlap>	Overlapping speech occurs when one speaker begins to speak while another is still speaking. Overlapping speech is marked with <overlap gid=“” oid=“”></overlap>, “gid” stands for the group number of overlapping; “oid” stands for the segment number in a single overlapping. Transcriptions of overlapping speech are displayed with ([…]).
non-speech <desc>	Events other than speech include laughing, crying, shouting, and advertisements in News Commentary and so on. The non-speech descriptions are displayed in (…).
non-Wenzhou languages <mixed>	The corpus sometimes contains stretches of speech that are not Wenzhou, e.g., English, Japanese. These stretches of speech are displayed in {{}}.
unclear elements <unclear>	Elements that are not heard clearly enough to be transcribed are so marked.