# The Wenzhou Spoken Corpus

**John Newman, Jingxia Lin, Terry Butler and Eric Zhang**
**University of Alberta**

## 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, and in the following sections we describe this corpus and the associated online search tool.[1] 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.

## 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 emphasize this aspect in the case of languages where there is a strong tradition of studying earlier, historical texts. This is clearly the case in the Chinese cultural sphere, 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 usage is not unduly neglected. For the study of contemporary Mandarin, we have valuable online searchable corpora such as the Academia Sinica Balanced Corpus of Modern Chinese and the Lancaster Corpus of Mandarin Chinese.[2] In the case of Chinese, it is also important to document contemporary usage of the socially less prestigious Chinese dialects. There is a strong tendency to focus on Mandarin (Putonghua) as an object of study (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

[1] The WSC is http://corpora.tapor.ualberta.ca/wenzhou/.
[2] The Academia Sinica Balanced Corpus of Modern Chinese is http://www.sinica.edu.tw/SinicaCorpus/. The Lancaster Corpus of Mandarin Chinese is http://bowland-files.lancs.ac.uk/corplang/lcmc/.

170,000-character corpus based on phone-in programs and forums aired on Hong Kong radio (Leung and Law 2002).[3]

The initial interest in an online searchable Wenzhou corpus (similar in spirit to the HLCAC) came from Lin, at the time a graduate student in the Department of Linguistics, and Newman, her supervisor. As linguists, Lin and Newman had linguistic goals in mind in creating a corpus of Wenzhou but lacked the necessary technical skills required for its implementation – a common enough impasse in the world of digital humanities. At the University of Alberta, the solution came in the form of a joint project involving the Text Analysis Portal for Research (TAPoR) unit. TAPoR is a pan Canadian, multi-institutional assembly of computing infrastructure and expertise, supported by a Canada Foundation for Innovation grant.[4] Assistance with the more technical aspects of creating the corpus was therefore provided by the other two co-authors, Butler and Zhang, from the University of Alberta TAPoR node. The computing infrastructure for the project (server, software etc.) was also provided by TAPoR.

### 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) Interview, (5) News Commentary, and (6) Songs. 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 compares favourably in size with the HKCAC. As a point of interest, the WSC provides ample evidence for all the syntactic traits claimed for Wenzhou in Pan (1991:269-277), 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.

[3] The HKCAC is http://shs.hku.hk/corpus/main.htm.
[4] The University of Alberta TAPoR node is http://tapor.ualberta.ca/.

| | | Word Count | Unicode Character Count |
|---|---|---|---|
| 1 | Face to Face Conversation | 13009 (8.22%) | 23582 |
| 2 | Phone Call | 20885 (13.2%) | 36257 |
| 3 | Internet Chat | 7005 (4.42%) | 13132 |
| 4 | Interview | 1046 (0.66%) | 2470 |
| 5 | News Commentary | 115293 (72.9%) | 179708 |
| 6 | Songs | 894 (0.56%) | 1395 |
| | **Total** | **158132** | **256544** |

Table 1. Size of the WSC by genre

For genres (1)-(4), the method of obtaining the data was to rely upon the social networks of Lin which included family members, friends, friends of friends, etc. For the Interview, the speaker was given a short story in written form to read silently and then asked to tell the story in Wenzhou without looking at the story. News Commentary was collected from TV and includes relatively informal news commentaries as well as opinions offered by anonymous interviewees. The Songs category is composed of traditional Wenzhou children's songs. The method was clearly "opportunistic" with a resulting over-representation of some demographic categories, as shown in Table 2. More than 72% of the total word count comes from recorded News Commentary (where age and education level is unknown), reflecting the relatively easy access to this category of data. Just over 25% of the total words derives from the more spontaneous, conversational contexts (Face to Face Conversation, Phone Call, Internet Chat) and in these categories there is a clear predominance of speakers under 34 years of age and at least high school level of education. Overall, male speakers outnumber female speakers about 3:1. The uneven demographics underlying the corpus 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.

| | No. of words | Percentage of total no. of words (158132) |
|---|---|---|
| *Gender* | | |
| Male | 105880 | 66.9% |
| Female | 35175 | 22.2% |
| Unknown | 17077 | 10.8% |
| | | |
| *Age* | | |
| 0-14 | 268 | 0.2% |
| 15-24 | 27009 | 17.1% |
| 25-34 | 44 | 0% |
| 35-44 | 0 | 0% |
| 45-59 | 6792 | 4.3% |
| 60+ | 7799 | 4.9% |
| Unknown | 116220 | 73.5% |
| | | |
| *Education level* | | |
| College+ | 26513 | 16.8% |
| High school | 6914 | 4.4% |
| Elementary school | 2998 | 1.9% |
| Illiterate | 5487 | 3.5% |
| Unknown | 116220 | 73.5% |

Table 2. Distribution of demographic categories in the WSC

As with any collection of linguistic data from humans, 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 prior to the formal beginning of the project (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 cooperation in allowing that data to now be used as part of the formal project.

## 4. File markup

An early decision was made to use XML for our file structure, recognizing 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 was 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 normalized 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 edited to ensure anonymity of participants.

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, and we wanted to maintain that kind of markup 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'. It is not always easy to decide on the morphological status of some sequences, in advance of a more complete study of this topic in Wenzhou, and we were conservative in our approach to identifying compounds. So, for example, 造 'build'+ 起 'rise' together means 'build up', but each of the parts could be analyzed as verbs in their own right (depending on criteria used for identifying verbs). In this case, the two characters were left as separate words.

We used a three-step 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[5]; (iii) finally, a Perl script automatically inserted word and punctuation tags. XMLSpy was used for construction of the DTD and validation of files. The Appendix presents an overview of the tags used, along with brief descriptions of them. (1) below is the DTD used for the XML markup and (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= '3', oid= '3'); phonetic transcription; word boundaries; punctuation.

[5] XMLSpy is an XML development program provided by Altova: 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:

```
<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>勾</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>
```

**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 provide 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, aged band, and level of education.

An example of a concordance display produced by the Concordance tool is shown in Table 4, using the default setting of 5 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 to the British National Corpus.[6] 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 the complete demographic information for each speaker.

| | | | | |
|---|---|---|---|---|
| 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 [[ts'ɳ]] 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 4. 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 of a word. A sample of collocate results by position is shown in Table 5.

[6] BNCWeb is http://homepage.mac.com/bncweb/home.html.

| -5 | -4 | -3 | -2 | -1 | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| 个(22) | 个(31) | 呢(17) | 是(22) | 卬你(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) |
| 侬(9) | 人(11) | 阿(11) | 宿(12) | 是(23) | * | 个(34) | 是(11) | 有(12) | 温州(12) | 是(9) |
| 该(8) | 侬(8) | 温州(8) | 侬(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) | 侬(6) | 温州(7) |
| 里(6) | 我(5) | 侬(7) | 就(9) | 能界(9) | * | 大学(7) | 文明(7) | 温州(8) | 人(6) | 会(6) |

Table 5. Part of the results for the keyword 温州 'Wenzhou' using the 'collocates by position' option (* = keyword)

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 state, the WSC has proved invaluable to the study of Wenzhou. Nevertheless, it was envisaged that the corpus could be expanded upon 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.[7] Naturally, we hope to be able to increase the size of the corpus. A more immediate plan is to include statistical measures of word association (MI, t-score, and z-score) which take into account overall frequencies of a keyword and collocate in the whole corpus, or sub-corpus, being searched.

So far, it has not been necessary to create the kind of relational database advocated by Davies (2005) as a means of enhancing search and retrieval performance. 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 relational databases. The XML-encoded text is in fact, structured like a database, and can be searched directly as such. Increase in the size of the corpus, and the addition of further options such as n-gram calculation and part of speech tagging might lead us to store the corpus in an XML-native database. This move would provide more powerful management and indexing features; the

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

advantage would be 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: 307-334.

Leung, M-T., and Law, S-P. (2002). HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics* 6.2:305-325.

Pan, W-Y. (1991). An Introduction to the Wu dialects. In W. S-Y. Wang (ed.), *Languages and Dialects of China*, pp. 237-293. *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 on-line at: http://www.unicode.org/standard/standard.html. Last accessed: 2006 Feb 27.

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

**Appendix:** Tag types used in the WSC

| Tag Type | Description |
|---|---|
| turn<br><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 in the search engine. |
| utterance<br><utterance> | Utterances are sentence-like segments within a turn and are also marked with ordered numerals. |
| word<br><w> | Words are one or more syllables/characters which have independent lexical status. |
| punctuation<br><c> | Punctuation marks used are: ， (comma), 。(period), “”(quotation), 《》(book title), — (sudden stop or sudden switch of topic). |
| overlapping speech<br><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<br><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<br><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<br><unclear> | Elements that are not heard clearly enough to be transcribed are so marked. |