The Hong Kong Cantonese Corpus: Design and Uses

K. K. Luke and May L.Y. Wong

1. Introduction

At a time when language corpora are becoming something of a commonplace in Linguistics, it is perhaps worth reminding ourselves that the use of corpus data has not always been held in high regard. In fact, the idea that language can be studied by reference to what people actually say to one another, as opposed to what they think they can or cannot say, once a truism in the Structuralist Linguistics of the first half of the twentieth century, went out of fashion in the 1960s. For a time, grammar and linguistic structure were studied with little regard for everyday speech and conversation.

One of the pioneers of data-driven grammars, Charles Fries, published *The Structure of English* in 1952 on the basis of an in-depth analysis of a corpus of speech data. About this grammar, the famous scholar, discoverer of the conversational turn-taking system, Harvey Sacks had this to say (in one of his lectures delivered in 1967), "[Fries's] grammar is terribly important because I guess it's the only grammar of English that was constructed by reference to an attempt to handle actual conversation itself". (Sacks 1992: vol.1, p.189) Incidentally, Y. R. Chao's famous book, *A Grammar of Spoken Chinese* (Chao, 1968), was equally firmly based on speech data, and so we would like to suggest that Chao's grammar be recognized for what it is -- as the first grammar of Chinese "that was constructed by reference to an attempt to handle actual conversation itself".

Nowadays, it is a common expectation, if not a requirement, for grammars to be written with close reference to corpus data (Mukherjee, 2006). In studying English grammar, almost everyone would now consult Quirk et al's comprehensive grammar (Qurik et al. 1985), or Biber's Longman grammar (Biber et al (1999), or Huddleston and Pullum's Cambridge grammar (Huddleston and Pullum (2002), which are all based on sizeable, even very large, corpora, and which are built from large collections of written texts and (a smaller collection of) speech recordings.

While there is no shortage of corpora that are entirely or largely based on written texts, such as the well-known British National Corpus (BNC), the Corpus of Contemporary American English (COCA), or the Peking University Corpus of Modern Chinese, speech-based corpora are in comparison much less common. This is no doubt due to the enormous time and effort

involved in building speech corpora, and the many technical challenges associated with the task -- for instance, what to make of overlapping speech and dysfluencies, or how to transcribe features of speech and conversation. And yet the fact remains that, as the primary mode of language use, speech and conversation are in every way at least as important as, if not more important than, writing and published texts, which must be viewed, in one way or another, as secondary or derived, in that they are typically products of various degrees of editing, afterthoughts, or modifications, and tend to conform to rules and conventions that apply exclusively to writing.

In recent years, for English at least, a number of speech corpora have become available, offering researchers much better access to samples of spoken English. The more well known ones include the Santa Barbara Corpus of Spoken American English (2000-2005), the Michigan Corpus of Academic Spoken English (2002), and the British Academic Spoken English Corpus (2005). These corpora, compiled in the early years of the 2000s with large amounts of good quality data, are not without their precursors, from which some basic principles and practices are inherited. Many of these earlier spoken corpora were constructed the decade before, such as the Corpus of English Conversation (Svartvik and Quirk, 1980), the London-Lund Corpus of Spoken English (1990), the Bergen Corpus of Teenage Talk (1998), the Lancaster/IBM Spoken English Corpus (1988), and the Wellington Corpus of Spoken New Zealand English (1998).

However, two spoken corpora, both compiled much earlier, in the 1970s, must be singled out for special mention, for the simple reason that they are, as far as we know, the first spoken corpora ever constructed. The first is Carterette and Jones's Informal Speech Corpus (Carterette and Jones, 1974), and the other is Crystal and Davy's collection of British English conversations (Crystal and Davy, 1974). Cartertte and Jones's corpus was designed to capture as accurately as possible samples of speech as they occur naturally in a range of conversational settings. The rationale behind Carterette and Jones's idea is explained by the authors in the following terms:

> "In spite of a vast amount of interest in language, there is very little information available about the informal spoken language. Yet psychologists, linguists, psycholinguists, and educators need such information for both research and curricular purposes. Among other things, the native language, the spoken dialect learned mainly at mother's knee, is the most overlearned behavior in an individual's entire repertoire". (Carterette and Jones, 1974: 3)

In processing the speech recordings, the authors took pains to preserve as much as possible

the ‘original flavour’ of the conversations:

> “[An] important purpose of this monograph is to present the ‘verbatim’ transcriptions of the corpus, in both alphabetic and phonemic form, so that others, with other purposes, will be spared the arduous, expensive and time-consuming task of collecting and editing such material.” (Carterette and Jones, 1974: 4)

Crystal and Davy’s corpus was similarly conceived as a collection of spontaneous and unscripted conversations between friends and neighbours in informal settings, to be made available to linguists and students for research and language learning. Unlike Carterette and Jones’s corpus, Crystal and Davy’s was made in Britain, and contains samples of British varieties of English. The topics covered by the participants in their conversations are mostly everyday affairs, such as farm visits, Christmas get-togethers, and football (i.e., English soccer), but these were made highly interesting because of the very natural and authentic circumstances under which the conversations were conducted. Crystal and Davy’s work is also very useful for the many insightful comments made on the challenges posed by naturalistic speech data for grammatical analysis, including sentence identification, ellipsis, interlacing of clauses, and several others.

A third source of inspiration that must be acknowledged is Harvey Sacks’s work on speech and conversation, and the tradition of recording, transcribing and meticulously analyzing naturally occurring conversations that Sacks created in the 1960s , which is now widely known as Conversation Analysis (Sacks, 1992; Sacks, Schegloff and Jefferson, 1976). Sacks showed in his work, for the first time, that everyday conversation is amenable to serious and systematic study. A door was thus opened, offering a glimpse of a linguistics that is built upon the investigation of naturally occurring speech in conversational interactions.

The idea of a Hong Kong Cantonese Corpus was conceived as a response to the need for good empirical data, and guided by the pioneering work of the above scholars. The overarching objective of the Hong Kong corpus was to capture as accurately as possible a set of snapshots of the language by collecting samples of its use in the context of naturally occurring and free-flowing everyday conversational interactions. However, the reason for doing this, one must hasten to add, is not only to make it possible for scholars to study discourse and conversation, although they can certainly do that with the help of the corpus. Rather, the underlying motivation was to gain access to the vernacular, the most natural form of speech, and to ensure that the data that we collected would reflect as faithfully as possible the structure of the language. A good quality record of the vernacular is essential for the study of phonology, morphology or syntax, just as it is essential for the study of

pragmatics and discourse. As Labov puts it, “The vernacular, in which the least attention is paid to speech, provides the most systematic data for linguistic analysis” (Labov 1984: 29).

2. The Hong Kong Cantonese Corpus

As explained above, The Hong Kong Cantonese Corpus was conceptualized very much as an electronic repository of naturally occurring conversations among people in Hong Kong at the turn of the 21st century. The rationale for this decision should be plain enough: Cantonese is first and foremost a spoken language (Thompson, 2005). It is the home language of the majority of people in Hong Kong and the most commonly used language in everyday social interactions. It is also the most popular language in the media – Hong Kong newspapers are full of news articles, advertisements and commentary with identifiable spoken Cantonese features in them. It is true that Cantonese does also have written forms distinctly different from written Chinese, which is based on Mandarin (Snow 2004. However, a corpus of written Cantonese would require a rather different design. It should probably include a historical dimension, with samples of early Cantonese songs and operas, for example. For now, we will focus our discussion on our spoken Cantonese corpus.

2.1 Existing Cantonese corpora

Over the past decade, a handful of Cantonese corpora have been built in Hong Kong, offering both child language (Lee and Wong 1998; Fletcher et al. 2000) and adult language data (So 1992; Xu and Lee 1998; Leung and Law 2001). The two child language corpora are available as part of the Child Language Data Exchange System (CHILDES), which aims to support language acquisition research. The three adult language corpora vary in their purposes. So’s 1992 corpus, made up as it does of readings of 1,800 single syllables by speakers of Cantonese, offers an important resource for studying Cantonese phonetics as well as for speech synthesis and speech recognition research. Xu and Lee’s 1998 corpus comprises recordings and printed documents in Shanghainese, Cantonese and Mandarin Chinese, and allows for comparative investigations of regional language variation. In this connection, mention should also be made of the National Cheng Chi University Corpus of Spoken Chinese, which contains speech samples from Mandarin, Hakka and Southern Min, though not Cantonese (Chui and Lai 2008). Leung and Law’s 2001 corpus contains speech data taken from radio phone-in and discussion programs, which represent two of the six speech situation types mentioned in the London-Lund Corpus of Spoken English (Svartvik and Quirk

1980; Johansson 1982; Svartvik 1990).[1] As radio broadcasts represent only certain types of speech events, some of which are pre-scripted or semi-scripted, a set of radio recordings will clearly not meet the requirements of a project where the main objective is to create a representative sample of contemporary Hong Kong speech in naturally occurring situations. Thus, we believe there should still be room for a corpus based primarily on face-to-face, two- or three- party conversations in a variety of everyday settings.

In this paper, we will give an account of the design principles and special features of the Hong Kong Cantonese Corpus. The broad outline of our description will follow Leech, Meyers and Thomas's very useful framework for the discussion of issues involved in developing spoken language corpora (Leech, Meyers and Thomas, 1995). In their framework, the design and construction of spoken corpora are discussed in terms of five stages – (i) Records of data collection procedures; (ii) Orthographic transcription of spoken language data; (iii) Explanation of the encoding scheme; (iv) Ascription of words to grammatical classes; (v) Corpus availability and public access to the corpus. In the final section, we will comment briefly on possible future developments of the corpus project.

2.2 Data collection

In carrying out our data collection, the guiding principle was to collect speech recordings that are naturally occurring and representative of Cantonese as it is spoken and used in Hong Kong. Our first aim was therefore to ensure a good number of speakers and a representative spread in terms of age and gender. This is necessary in order that valid generalizations can be made about the structure and use of Cantonese on the basis of samples taken from the corpus (Meyer 2002: 40). As Biber has pointed out, a corpus is representative of a language if it accurately reflects its structures and tendencies (Biber 1990; 1993a; 1993b).

At the same time, care was taken to ensure that all the participants recruited for the project were *bona fide* Hong Kong Cantonese speakers. This was done by confirming with each participant that he or she was either born in Hong Kong or had been residing in Hong Kong for a continuous duration of nine years or more, preferably since childhood (before 6 years of age). Also, they needed to confirm that Cantonese was their most prominent and frequently used home language.

[1] The London-Lund corpus comprises *six* major speech situations: face-to-face conversations, public conversations including debates and interviews, telephone conversations, radio broadcasts, spontaneous speeches and prepared speeches.

Our next decision was to give priority to face-to-face conversations over other forms of speech. Thus, while lectures and formal speeches are in a sense also a kind of spoken language, such monologues will clearly be very different in nature from face-to-face conversations which are spontaneous and unplanned and are interactive through and through.

Sinclair (2005: 6) suggests that as far as possible, audio recordings should be included in their entirety. The reason for this is to preserve as much of the context of an interaction as possible. We are in full agreement with this, and in our data collection we did record entire conversations whenever possible. However, in the transcription stage, it was not always possible to transcribe every conversation in full. Some conversations were simply too long, and to transcribe those fully would take away too much transcription time from other recordings, which would in effect mean cutting down on the number of speakers and speech samples that could be included in the corpus.

We were also conscious of the need to ensure the broad representativeness of the corpus by not imposing any pre-conception that one might have regarding particular structural or usage patterns, e.g., vocabulary choices or discourse focusing (Clear 1992). This is important because the corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it (Sinclair 2005: 8). Following this principle, our participants were asked to speak freely, in an easy and comfortable environment, with family or friends, but without any pre-determined agenda or assigned topics.

Another important consideration was the possible effects of the well-known 'Observer's Paradox' (Labov 1972: 209). What this means for the recording of everyday conversations is that the naturalness of a recording may well be compromised by the presence of the researcher and/or the recorder, which may cause participants to be self-conscious about their speech. Short of making the recordings surreptitiously, which would not be morally defensible, nor would it be deemed to have met the high standards of research integrity by the university's institutional review board, there is no real solution to this dilemma. However, we do think that there are reasons to believe that the full effects of the Observer's Paradox may have been attenuated, at least to some extent, by three factors. First, the researcher or assistant who was making the recording made it a point to leave the room as soon as the recorder was switched on. Second, in many if not all of the sessions, in spite of some initial awkwardness at the beginning, participants appeared to be able to get used to the presence of the recorder, or appeared at least to be less affected by it, some minutes into the conversation. Finally, and most importantly, care was taken to ensure that participants having conversations during the recording session were family members, close friends, or

neighbours or colleagues who are in regular contact with one another. In this way, it is hoped that participants' self-consciousness may have been reduced by the interest of their conversational topics. There were certainly moments when participants appeared to have all but forgotten about the fact that they were being recorded as they were getting into the spirit of their banter -- or so we would like to believe!

Following common practice, informed consent was obtained from all the participants before the recordings were made, the only exception being the speakers in the radio talk shows. For one thing, it was not possible to obtain their consent before their shows. Also, radio programmes being events that are publicly broadcast are very different in nature from recordings of private conversations. We believe what we did was in line with the guidance laid down in the BAAL Recommendations on Good Practice in Applied Linguistics (http://www.baal.org.uk/dox/goodpractice_full.pdf).

Our final consideration was whether to make audio or video-recordings. In comparison with audio recordings, video recordings are evidently preferable, for the simple reason that more of the contextual features of a conversational situation can be captured on a video. However, in practice, many participants would have felt even more awkward to be placed in front of a camera, and many may even turn down our invitation to participate. It was therefore decided, with reluctance and regret, that only audio recordings would be made.

With these principles and guidelines, we proceeded to recruit participants and arrange for recording sessions to be made. The main round of recordings was made during 1997 and 1998, with more recordings added to the collection in the ensuing four years. In the event, 52 audio recordings were obtained, involving some 100 speakers (some of whom featured in more than one conversation), and roughly half male and half female in terms of gender distribution. Age-wise, in spite of our initial ambition to include an even spread of speakers across the age groups (from 15 to 60), what transpired was a greater concentration of younger speakers in their 20s and 30s, with fewer participants from the other age groups. This was mainly due to the more ready availability of participants who were friends and peers of the research assistants who helped make most of the recordings, as opposed to members of the younger and older age groups, who were relatively less accessible. Most conversations collected were two- or three-party chats, with a small number involving four participants. At a later stage, to further broaden the range of speakers and situations, a supplementary set of recordings was obtained from radio chat shows. Altogether 42 radio recordings were obtained, involving 10 different speakers. In all, we were able to acquire 30 hours of recording, each being between 3 to 40 minutes in length. On average, each sample was about 10 minutes long. When transcribed, this gave us a corpus of some 180,000 word

tokens (as opposed to Chinese characters).

2.3 Transcription

Anyone who has worked with connected speech will know that transcription is by no means a simple or straightforward affair. This is especially true of naturally occurring talk, the complexity of which (even in the absence of a record of participants' gestures, facial expressions and other visual aspects) is such that it is simply impossible to reproduce on paper all the shades, cadences and nuances that go into the making of an utterance. As Cook (1995) puts it, an orthographic transcription of a conversation is at best an incomplete representation of the original speech event. Thus, we were keenly aware, right from the start, that certain priorities had to be set, and how this should be done would depend on the uses to which the corpus was to be put. We decided early on that the corpus that we were constructing was to be used primarily for investigations at the levels of lexis, grammar, and discourse, rather than phonology. As a result, when it comes to transcription, we set ourselves the modest goal of reproducing in writing as much as possible the words and sentences that could be identified in the utterances. In that process, features of accent, prosody and intonation had to be given the backseat.

While the level of detail that we were prepared to incorporate into the transcriptions was a far cry from the usual standards of Conversation Analysis, every attempt was made to produce a transcription that is as close as possible to a verbatim record of each utterance, at least as far as the words and sentences are concerned. Thus, some common features of talk such as turn-beginning recycling, repetition and self-repair were preserved in the transcript. However, other features like stretching, pausing and overlapping were not included in the transcription, as to do so would pose insurmountable problems to the integrity and identity of words and sentences, the two basic units that we strived to preserve in our transcription.

In transcribing Cantonese speech, careful thought must be given to the use of appropriate Chinese characters that are needed for the writing of words and expressions that are unique to the language. Thousands of words and morphemes, including people's names, place names, local words and grammatical and discourse particles, do not have a ready-made written representation in the conventional Chinese character set. For a long time before the 1990s, different people, groups, companies and organizations in Hong Kong were making up new characters as they went along, with little coordination or standardization. Thanks to the Information Office of the Government of Hong Kong, a unified set of special characters was adopted in 1999 for word representation and information exchange, which was subsequently approved by ISO and included in the ISO-10646 character set. This standard, which is now

widely known as the Hong Kong Supplementary Character Set (HKSCS), was adopted and used throughout the transcription and processing of our corpus data.

In transcribing the data, a four-line format was used. Line 1 contains the original transcription using Chinese characters (and symbols from HKSCS). Line 2 gives a romanized version of line 1. Here, the Linguistic Society of Hong Kong's Romanization scheme was adopted, as it is one of the most systematically designed and widely used Romanization systems available. Line 3 provides a word-for-word gloss in English, and line 4 a free translation of each sentence. An example is given below to illustrate this format:

J： 嗰　我　變咗　我　自己　要
gam2 ngo5 bin3zo2 ngo5 zi6gei2 jiu3
so I turn-out I self have-to
So it turned out I had to

留　喺　嗰度　囉　一　個　人。
lau4 hai2 go2dou6 lo1 jat1 go3 jan4
stay at there FP one CL person
stay there all by myself.

A： 你　驚唔驚　啊　嗰樣？
lei5 geng1-m4-geng aa3 gam2joeng2
you scared-or-not FP like-that
Were you scared then?

J： 驚　啊，其實　最　主要　係　驚　鬼。
geng1 aa3 kei4sat6 zeoi3 zyu2jiu3 hai6 geng1 gwai2
scared FP in-fact most main be scared ghosts
I was! In fact, I was mainly scared of ghosts.

Where code switching or mixing appeared in the data, then words from languages other than Cantonese (usually English words) were spelled as they were in the foreign languages (cf. Sebba, 1995).

2.4 Encoding

As McEnery and Xiao (2005:51-53) observe, a problem in Asian corpus building is the existence of multiple and often competing encodings of Asian writing systems. In the case of

Chinese, characters can be encoded in GB2312, GB18030, BIG-5, HZ or Unicode. When our data were first transcribed, they were encoded in BIG-5, which was then the most widely used encoding scheme in Hong Kong and Taiwan. Subsequently, however, when Unicode had been upgraded to include the special Hong Kong characters (HKSCS), the entire database was converted into Unicode, and can be viewed and processed on most platforms.

2.5 Segmentation and Part-of-Speech (POS) Tagging

For our Cantonese corpus to be of use to researchers and learners, the transcription must be segmented into word-size units and POS tagged (Leech 2005: 17, Leech and Smith 1999: 26). During the initial stages, a small section of the corpus was word-segmented and tagged manually. This sub-corpus was then used as a training corpus for further segmentation and tagging to be done semi-automatically. During this stage of corpus building, we developed a method whereby segmentation and tagging were done hand in hand -- in one go, as it were, the rationale being that information concerning possible segmentation and information concerning possible assignment of POS are mutually beneficial and should be used in a complementary fashion in arriving at decisions (about segmentation and tagging). (For details of this approach, please see Fu, Luke and Wong, 2005.)

The annotation scheme is the same as the one designed for use in the PFR (*Peita-Fujitsu-Renmin Ribao*) People's Daily POS Tagged Chinese Corpus (abbreviated to PFR Chinese Corpus hereafter) Release 1.0.[2] The tagset[3] comprises 26 basic word classes, including noun (*n*), time word (*t*), space word (*s*), directional locality (*f*), numeral (*m*), classifier (*q*), non-predicate adjective (*b*), pronoun (*r*), verb (*v*), adjective (*a*), descriptive[4] (*z*), adverb (*d*), preposition (*p*), conjunction (*c*), auxiliary (*u*), modal particle (*y*), interjection (*e*), onomatopoeia (*o*), idiom[5] (*i*), fixed expression (*l*), abbreviation (*j*), prefix[6] (*h*), suffix[7] (*k*),

---

[2] The PFR People's Daily Chinese Corpus contains the set of newspaper extracts assembled in January 1998, totalling some 3,000,000 characters, corresponding to about 1 million Chinese words. It is part-of-speech tagged and freely available to the research community to use. See Zhan et al. 2006 for a brief description of the corpus.

[3] The tagset that was used in the PFR Chinese Corpus was in fact extended from the one proposed in Yu et al. (1998).

[4] Descriptives are typically formed by reduplication or compounding, for example, 实实在在 *shishizaizai* "indeed, really, honestly", 绿茵茵 *lüyinyin* "green", 久远 *jiuyuan* "far back, ages ago, remote", 烂漫 *lanman* "bright-coloured; unaffected".

[5] In Chinese, idoms, or 成语 *chengyu*, are expressions with a frozen internal structure. Their constituents and structure cannot be described in terms of morphological categories. They have to be treated as single morphological units. They should be distinguished from the fixed expressions, or 习

morpheme[8] (*g*), unclassified item[9] (*x*), and punctuation (*w*). Apart from this basic set of 26 POS markers, proper nouns were divided into personal names (*nr*), place names (*ns*), organisation names (*nt*) and other proper nouns (*nz*). Another 20 markers were added to deal with words that are specific to Cantonese. Table 1 lays out all the symbols used in our tagset.

No.	Tagset	POS (in Chinese)	POS (in English)
1	Ag	形语素	Adjective Morpheme[10]
2	a	形容词	Adjective
3	ad	副形词	Adjective as Adverbial[11]
4	an	名形词	Adjective with Nominal Function[12]
5	Bg	区别语素	Non-predicate Adjective Morpheme
6	b	区别词	Non-predicate Adjective
7	c	连词	Conjunction
8	Dg	副语素	Adverb Morpheme
9	d	副词	Adverb
10	e	叹词	Interjection
11	f	方位词	Directional Locality
12	g	语素	Morpheme

用语 *xiyongyu*, the internal structure of which can be broken down into meaningful morphological units.

[6] Examples include 非 *fei* "not", 超 *chao* "super", 无 *wu* "not", 过 *guo* "too", etc.

[7] Examples include 儿 *er* "little", 们 *men* "expressing plurality", 型 *xing* "model, type" , 式 *shi* "type, style", etc.

[8] Examples are 桌 *zhuo* "table", 身 *shen* "body", 鸭 *ya* "duck", etc.

[9] Unlike morphemes, unclassified items do not carry any meaning at all. They must be combined with another unclassified item to give a meaningful word. Examples are 鹌 *an* (-鹑 *-chun*) "quail", 蟑 *zhang* (-螂 *-lang*) "cockroach", 蛤 *ge* (-蚧 *-jie*) "clam", etc.

[10] The definition of morpheme was clearly stated in the institute's corpus annotation manual which states that a morpheme refers to the smallest meaningful unit which cannot be used independently. In Chinese, many characters may have their own meaning but they cannot stand alone (Norman, 1988:154-156). They have to be combined with another character or word in the word formation process. Therefore, an adjective morpheme, resembling a common adjective semantically, is a morpheme signifying an attributive meaning to the word to which it is attached.

[11] It refers to those adjectives functioning as adverbial without any modification to their morphological or phonological form.

[12] It refers to those adjectives which can fulfil nominal functions in a clause.

No.	Tagset	POS (in Chinese)	POS (in English)
13	h	前接成分	Prefix
14	i	成语	Idiom
15	j	简略语	Abbreviation
16	k	后接成分	Suffix
17	l	习用语	Fixed Expression
18	Mg	数语素	Numeric Morpheme
19	m	数词	Numeral
20	Ng	名语素	Noun Morpheme
21	n	名词	Common Noun
22	nr	人名	Personal Name
23	ns	地名	Place Name
24	nt	机构团体	Organisation Name
25	nx	外文字符	Nominal Character String
26	nz	其它专名	Other Proper Noun
27	o	拟声词	Onomatopoeia
28	p	介词	Preposition
29	Qg	量语素	Classifier Morpheme
30	q	量词	Classifier
31	Rg	代语素	Pronoun Morpheme
32	r	代词	Pronoun
33	s	处所词	Space Word
34	Tg	时间语素	Time Word Morpheme
35	t	时间词	Time Word
36	Ug	助语素	Auxiliary Morpheme
37	u	助词	Auxiliary
38	Vg	动语素	Verb Morpheme
39	v	动词	Verb
40	vd	副动词	Verb as Adverbial[13]
41	vn	名动词	Verb with Nominal Function[14]
42	w	标点符号	Punctuation
43	x	非语素字	Unclassified Item
44	Yg	语气语素	Modal Particle Morpheme

[13] It refers to those verbs functioning as adverbial without any change to their form, both morphologically and phonologically.

[14] It refers to those verbs that have acquired some nominal functions in a clause.

No.	Tagset	POS (in Chinese)	POS (in English)
45	y	语气词	Modal Particle
46	z	状态词	Descriptive

Table 1: The PFR tagset

After segmentation and tagging, the data was finally converted into an XML format to facilitate the development of concordancing and other search and processing tools. An example of this format is given below:

<info>: information about recording
<sent_head>: speaker identity
<sent_tag>: POS tag each word
<sent_tran>: English translation

```
<info>
        1-TN-026
        2-DR-280797
        3-NS-2
4-LS-AJ
        5-A-F-22-HK
        6-J-F-24-CHINA
        INFO-END
</info>
        <sent>
                <sent_head>
                        A：
                </sent_head>
                <sent_tag>
                        點/r/dim2/
                        啊/y/aa3/
                        ，/w/VQ2/
                        你/r/nei5/
                        而家/t/ji4gaa1/
                        返工/v/faan1gung1/
                        嗰度/r/go2dou6/
                        做/v/zou6/
                        成/u/seng4/
                        ？/w/VQ6/
```

```
        </sent_tag>
        <sent_tran>
            How's it going at work?
        </sent_tran>
    </sent>
```

3. Uses of the Corpus and Future Directions

As explained above, the Hong Kong Cantonese Corpus was compiled as a spoken corpus, mainly for research and learning purposes at the lexical, grammatical and discourse levels. Compared to the many large-scale corpora available for English, Chinese and other languages these days, our corpus is admittedly very small. However, the data that went into its making is, we believe, of a good quality, and its design and construction has hopefully been done with sufficient care and systematicity. Cantonese being a spoken vernacular, we believe that there is no real alternative to consulting a good quality corpus of the spoken language, however small it may be, when it comes to finding out about the structures and functions of the language.

With the help of the corpus and an accompanying toolbox, a broad range of studies and practical applications can be envisaged. The corpus will provide a solid empirical basis for such work. With a variety of examples of words and structures being used in particular contexts, the corpus can certainly be used to support investigations into the use of grammatical morphemes (e.g., aspect markers or sentence-final particles) or construction types (e.g., 'Topic-comment' or Right Dislocation constructions). For obvious reasons, the data should also be of interest to researchers working in the areas of pragmatics and discourse.

The corpus can also be used to support the development of various practical applications. For example, a Cantonese-based Chinese inputting method, 'Red Dragonfly Inputting',[15] has been constructed and made available to the public using data from the corpus. A Cantonese-English dictionary is also being compiled using real-life examples from the corpus to illustrate the meaning and usage of words and expressions in the dictionary.

Finally, reflecting upon the limitations of the corpus, one can envisage several directions towards which the corpus can be taken further. There is certainly room for more data, but

[15] The inputting method is described and available as a free download at the first author's HKU web site at http://www.linguistics.hku.hk/staff/kkl_cime.htm.

also there is room for better coverage in terms of age groups (e.g., more middle-aged speakers) and settings (e.g., from home and school to shops and markets). It would also be ideal if video recordings can be included in addition to audio recordings.

Acknowledgments

The authors wish to express, first of all, their sincere thanks to all the Hong Kong Cantonese speakers who graciously agreed to participate in the recording sessions. They would also like to thank the many colleagues and students who have helped during the many stages of the corpus's compilation. Far too many colleagues and students have contributed than can be mentioned individually by name, but a few stand out as key and indispensable helpers: Owen Nancarrow, who for several years at the initial stages of the project, was a collaborator; Fu Guohong, who converted the data into XML format and helped develop concordancing and other tools; Wong Ki Fong who served as the coordinator for much of the recording sessions and provided guidance and training to a small team of transcribers; and Lau Chaakming, who was instrumental in double-checking and confirming the accuracy of the transcriptions and in helping put up the website for the corpus during the final stages of the project.

References

BAAL: British Association for Applied Linguistics. BAAL Recommendations on Good Practice in Applied Linguistics. Available on-line at http://www.baal.org.uk/dox/goodpractice_full.pdf

Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5 (4): 257-269.

Biber, D. 1993a. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243-257.

Biber, D. 1993b. Using register-diversified corpus for general language studies. *Computational Linguistics* 19 (2): 219-241.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Carterette, Edward C. and Margaret Hubbard Jones (1974) *Informal Speech: Alphabetic and Phonemic Texts with Statistical Analyses and Tables*. Berkeley: University of California Press.

Chao, Y.R. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chui, Kawai and Huei-ling Lai 2008. The NCCU Corpus of Spoken Chinese: Mandarin, Hakka and Southern Min. *Taiwan Journal of Linguistics* 6.2: 119-144.

Clear, J. 1992. Corpus sampling. In *New Directions in English Language Corpora*, ed. G. Leitner, 21-31. Berlin: Mouton de Gruyter.

Cook, G. 1995. Theoretical issues: transcribing the untranscribable. In *Spoken English on Computer: Transcription, Mark-up and Application*, eds. G. Leech, G. Myers and J. Thomas, 35-53. Harlow: Longman.

Crystal, D. and D. Davy (1975) *Advanced Conversational English* London: Longman.

Fletcher, P., Leung, S. C.-S., Stokes, S., and Weizman, Z. 2000. Cantonese pre-school language development: A guide. Report of a project entitled Milestones in the Learning of Spoken Cantonese by Pre-school Children by the Language Fund, Hong Kong. Hong Kong: Department of Speech and Hearing Sciences, University of Hong Kong.

Fries, C.C. 1952. *The Structure of English*. London: Longman.

Fu, Guohong, K.K. Luke and Percy Ping-Wai Wong (2005) Description of the HKU Chinese word segmentation for SIGHAN Bakeoff 2005. *Proceedings of the 4th ACL SIGHAN Workshop on Chinese Language Processing*, *Jeju Island*, Korea, October 2005, pp.165-167.

Huddleston, R. and G.K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Johansson, S. ed. 1982. *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Labov, W. 1972. *Sociolingustic Patterns*. Philadelphia: University of Pennsylvania.

Labov, W. 1984. Field methods of the project on linguistic change and variation. In *Language in Use: Readings in Sociolinguistics*, edited by J. Baugh and J. Scherzer, Englewood Cliffs: Prentice Hall. 28-53.

Lee, T., and Wong, C. 1998. CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique – Asie Orientale* 27 (2): 211-228.

Leech, G. 2005. Adding linguistic annotation. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 17-29. Oxford: Oxbow Books for the Arts and Humanities Data Service. Available on-line at http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm.

Leech, G., and Smith, N. 1999. The use of tagging. In *Syntactic Wordclass Tagging*, ed. H. van Halteren, 23-36. Dordrecht, Boston and London: Kluwer.

Leech, G., Myers, G., and Thomas, J. eds. 1995. *Spoken English on Computer: Transcription, Mark-up and Application*. Harlow: Longman.

Leung, M.-T., and Law, S.-P. 2001. HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics* 6 (2): 305-325.

McEnery, A., and Xiao, R. 2005. Character encoding in corpus construction. In *Developing*

Linguistic Corpora: A Guide to Good Practice, ed. M. Wynne, 47-58. Oxford: Oxbow Books for the Arts and Humanities Data Service. Available on-line at http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter4.htm.

Meyer, C. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Mukherjee, J. (2006) Corpus linguistics and English reference grammars. In Antoinette Renouf and Andrew Kehoe (eds.), *The changing face of corpus linguistics*, 337-354. Amsterdam and New York: Rodopi.

Norman, J. 1988. *Chinese*. Cambridge: Cambridge University Press.

Quirk, R., G. Greenbaum, G. Leech and J. Svartvik 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sacks, H. 1992. *Lectures on Conversation*, 2 volumes. Oxford: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4), 696-735.

Sebba, M. 1995. Code switching: a problem for transcription and text encoding. In *Spoken English on Computer: Transcription, Mark-up and Application*, eds. G. Leech, G. Myers and J. Thomas, 144-148. Harlow: Longman.

Sinclair, J. 2005. Corpus and text: Basic principles. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 1-16. Oxford: Oxbow Books for the Arts and Humanities Data Service. Available on-line at http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm.

Snow, D. B. 2004. *Cantonese as Written Language: The Growth of a Written Chinese Vernacular.* Hong Kong: University of Hong Kong Press.

So, L. K.-H. 1992. *Hong Kong Spoken Cantonese Database*. Hong Kong: Research Grant Council.

Svartvik, J. ed. 1990. *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82. Lund, Sweden: Lund University Press.

Svartvik, J., and Quirk, R. 1980. *A Corpus of English Conversation*. Lund: Lund University Press.

Thompson, P. 2005. Spoken language corpora. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 59-70. Oxford: Oxbow Books for the Arts and Humanities Data Service. Available on-line at http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm.

Xu, L.-J., and Lee, T. 1998. *Parametric variation in Three Chinese Dialects, Cantonese, Shanghainese and Mandarin*. Hong Kong: Research Grant Council.

Yu, S.-W., Zhu, X.-F., Wang, H., and Zhang, Y.-Y. 1998. *The Grammatical Knowledge-base of Contemporary Chinese – A Complete Specification*. Beijing: Tsinghua University Press.

Zhan, W., Chang, B., Duan, H., and Zhang, H. 2006. Recent Developments in Chinese Corpus

Research. In the Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compliation and Application, Tokyo, Japan. Available on-line at: http://ccl.pku.edu.cn/doubtfire/papers/2006_Corpora_NIJL_Workshop.pdf.