Toward a Parallel Corpus of Spoken Cantonese and Written Chinese

John Lee
The Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
jsylee@cityu.edu.hk

Abstract

We introduce a parallel corpus of spoken Cantonese and written Chinese. This sentence-aligned corpus consists of transcriptions of Cantonese spoken in television programs in Hong Kong, and their corresponding Chinese (Mandarin) subtitles. Preliminary evaluation shows that the corpus reflects known syntactic differences between Cantonese and Mandarin, facilitates quantitative analyses on these differences, and already reveals some phenomena not yet discussed in the literature.

1 Introduction

While standard Chinese, also known as Mandarin or Putonghua, is served by an ever-expanding set of linguistic resources[1], its various dialects have received relatively little attention. The use of these Chinese dialects, however, is as widespread as many other national languages. For example, Cantonese is spoken by more than 52 million people, mostly in southern China and overseas Chinese communities.

Although considered the "most widely known and influential variety of Chinese other than Mandarin" (Matthews & Yip, 1994), Cantonese currently has rather limited linguistic resources. This paucity may be due to its unofficial status, as opposed to Mandarin, which is the official language of China. Furthermore, as a primarily spoken language, it does not traditionally have any standard written form. This paper presents the first parallel corpus of transcribed Cantonese speech and its equivalent written Mandarin. The corpus is expected to be useful for language learners, linguists and developers of natural language processing applications.

The corpus provides students with authentic, parallel examples of sentences in both languages, which are not mutually intelligible. Native speakers of Cantonese must learn Mandarin for use in writing and official communication; conversely, many Mandarin speakers living in Hong Kong also want to learn Cantonese.

The corpus also serves as a repository for linguistic research. In particular, it facilitates research in comparative grammar, by lending statistical evidence, and potentially demonstrating exceptions or other differences yet unnoticed.

Finally, it can be exploited as training material for natural language processing systems, such as cross-lingual spoken document retrieval (Meng & Hui, 2001), and especially machine translation (MT) systems. For example, MT systems may be trained to automatically generate Chinese subtitles for Cantonese television programs, as has been done for Scandinavian languages (Volk et al., 2010).

2 Previous Work

Cantonese grammar has been well studied (Matthews & Yip, 1994; Cheung, 2007), and a few monolingual corpora for Cantonese have been compiled (Lee & Wong, 1998; Leung & Law, 2001; Wong, 2006). While the present corpus may also be used simply as Cantonese data, its primary contribution is as parallel data between Cantonese and Mandarin.

The main difference between Cantonese and Mandarin is in phonology and vocabulary; indeed, various bilingual dictionaries and lexical comparisons are already available (Zhang & Yang, 2008). In terms of syntax, although the "grammatical structure is similar in most major respects", the differences are not insignificant (Ouyang, 1993). So far, there have been few

[1] For example, (Chen et al., 1996), (Xue et al., 2005), and (Tsou & Kwong, 2006), among many others

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1462–1466,
Chiang Mai, Thailand, November 8 – 13, 2011.

studies on direct comparisons between the grammars of Cantonese and Mandarin (Ouyang 1993; Liang 1996), none of which was conducted on a large-scale, empirical methodology using naturally occurring Cantonese speech. This corpus is intended to lay the foundation for this direction of research.

3 Corpus

We motivate the design principles of the corpus (section 3.1), then describe how the corpus was constructed and processed (section 3.2).

3.1 Choice of material

The material of the corpus comes from television programs, including news and dramas, broadcast on a Cantonese channel in Hong Kong (see Table 1). All have Mandarin subtitles, which we aligned to the transcription of the Cantonese that was simultaneously spoken. The corpus contains 4,135 pairs of such "sentences", with a total of 36,775 characters in Mandarin, and 39,192 in Cantonese.

The choice of these sources of material follows considerations on two main issues: register variations, and speech and translation quality. Cantonese has a wide range of registers, from formal to colloquial. The formal register closely resembles Mandarin, and diverges significantly from the colloquial; this divergence is in fact a topic of active research in its own right. For any contrastive studies between Cantonese and Mandarin, a corpus balanced between formal and colloquial registers would be desirable. Thus, the TV drama provides the colloquial register; the news program contributes mostly to the formal register with the speeches of the anchor and reporters, but also some colloquial register with those of the spontaneous interviewees.

With the exception of these spontaneous interviews, all materials consist of pre-planned speech. They are thus largely free of false starts, sentence fragments, repairs, repetitions and other errors, which would have led to a considerable amount of spurious word alignments. This is an important advantage, as the parallel corpus will be used for word-level comparative studies.

The Mandarin subtitles, professionally translated, are in general of high quality. However, they are sometimes condensed, likely due to constraints posed by speech timing and screen size (Prokopidis, 2008).

3.2 Corpus construction

The Mandarin side of the corpus comes from subtitles, which consist of characters only; in contrast, the Cantonese side mixes orthographic transcriptions (characters) with a small number of phonetic transcriptions and English. Phonetic transcriptions, conforming to the Jyutping standard, are used when the Cantonese morpheme does not traditionally correspond to any standard Chinese characters. Code-mixing between English and Cantonese is not infrequent, and the English words are preserved in these cases.

Sentence-final particles in Cantonese, such as 啦 *la*, present a challenge for orthographic transcription. "Many of the particles differ only in tone and in nuance of meaning. Given that there is little uniformity of representation in relation to these particles", they are written as the same form in (Leung & Law, 2001). We also follow this practice.

The metadata records both the name and the category of the speaker. Speakers in the drama are always assigned as the "Character" category; those in the news are assigned one of four, namely "Anchor", "Reporter", "Live Reporter", or "Interviewee". "Anchor" and "Reporter" are considered to belong to the formal register, and all others, to the colloquial. Overall, about 60% of the corpus belongs to the colloquial.

Automatic word segmentation was performed on the Mandarin sentences (Chang et al., 2008). A subset of these words was then manually aligned to their Cantonese counterparts to facilitate a preliminary investigation, which will be reported in the next section.

TV Program	Size
六點半新聞報道 "TVB News at Six-Thirty" (2011)	*Time*: 5 episodes x 20 min *Length (chars)*: 19,069 Mandarin; 20,900 Cantonese
溏心風暴之家好月圓 "Moonlight Resonance" (2008)	*Time*: 2 episodes x 45 min *Length (chars)*: 17,706 Mandarin; 18,292 Cantonese

Table 1. The source material of the corpus comes from two TV programs, news (top) and drama (bottom).

4 Evaluation

The usefulness of the corpus may be gauged in two ways. First, it should reflect known differences between Mandarin and Cantonese (section

4.1), and those between formal and colloquial registers in Cantonese (section 4.2). Secondly, it should not only corroborate, but also contribute new information to previous studies. For this second goal, we give examples in three specific areas, namely the plural marker for personal nouns (section 4.3), agentless passives (section 4.4), and possessive constructions (section 4.5). In what follows, 'CL' refers to "classifier" and "PL" to "plural".

4.1 Coverage of Grammatical Differences

One of the most detailed comparative study between Mandarin and Cantonese to-date is (Ouyang, 1993), which lists 18 major differences. Table 2 lists some of these[2].

To investigate the degree to which the corpus exhibits known grammatical differences the two languages, we search for examples for each of the 18 differences in the corpus. Out of these 18 differences, 15 are found. The three differences for which no examples exist are the following. The first is concerned with word order involving the gender marker of animals. For example, The marker *gong* precedes the animal in Mandarin 公鷄 *gong ji* 'rooster', but its Cantonese equivalent *gung* follows the animal, as in 鷄公 *gai gung* 'rooster'. The second deals with word order in a negated resultative verb when the direct object is a personal pronoun. In Mandarin, the pronoun is always placed after the two-character verb, but in Cantonese it may be placed between as an infix. For example, 我 *ngo* 'I' is placed between the resultative verb 打贏 *daa-jeng* 'beat' in the sentence 你打我唔贏 'you did not beat me'. Finally, the third is the use of 過 *gwo* as a dative marker in Cantonese verbs of giving, e.g., 話過你知 'tell *gwo* you know' "tell you". This marker is normally omitted in contemporary Cantonese spoken in Hong Kong.

In summary, the corpus reflects well the known grammatical differences between the two languages as set out in (Ouyang, 1993). Two of the missing differences deal with rather specific constructions, and the third is no longer valid for the Hong Kong variety of the language.

Modal verbs: verbs such as 能 *neng* is used in Mandarin, vs. the 得 *dak* construction in Cantonese
能 忍耐 就真的 不是 人 了 忍 **得** 嗰個都 唔係 人 嚟㗎 'can' 'tolerate' 'can' 'not' 'man' 'No man can tolerate [that]'
Plural marker of personal nouns: Suffix 們 *men* for Mandarin, prefix 啲 *di* for Cantonese
你要 照顧 **弟妹 們** 你要 睇住 **啲 細** 㗎 'you should' 'take care' PL 'young' PL 'You should take care of your younger siblings'
Double objects: different word orders
那 你 **給 我 一個地址** 咁一係 你 **俾 個地址 我** 呀 'you' 'give' 'me' 'address' 'me' 'Please give me an address'
Predicative adjectives: the adjective may be placed in front of the topic in Cantonese
她 **年紀 這麼大** 演秦香蓮? 佢 **咁大 年紀** 演秦香蓮呀? 'she' 'so big' 'age' 'so big' 'She is so old, can she still play Chin Xianglian?'
Comparison of quantities: the adjective 多 *do* is placed after the verb in Cantonese
多 補 半天假 **補 多** 半日假 'more' 'compensate' 'more' 'half-day holiday' 'compensate for another half-day holiday'
Comparison of adjectives: different word orders
保管得 **比** 你的容貌 還 **好** keep 得 **好 過** 你個樣喎 'keep' 'good' 'compare' 'your face' 'more' 'good' 'keep [one's face] in better conditions'
Use of Numerals: Certain numerals can be omitted in Cantonese in large numbers
我 出夠 **一 萬 五 千** 我 出夠 **萬 五** 'I' 'pay' 'one' '10000' '5' 'thousand' 'I pay 15000'

Table 2. Grammatical differences between Cantonese and Mandarin listed in (Ouyang, 1993). In the example sentences, Mandarin is placed on top and Cantonese at the bottom, with their words roughly aligned. A total of 18 differences are discussed in (Ouyang, 1993); please see footnote 2 for the rest.

[2] For lack of space, we describe here briefly the other differences, and refer the interested reader to (Ouyang, 1993). They include: the lack of distinction in Cantonese between inclusive and exclusive "we"; the use of the *dak* construction with 有 you 'have', and in the negated resultative verbs, both impossible in Mandarin; the use of 去 *heoi* 'go' in Cantonese without a preceding 到 *dou* 'arrive', as in Mandarin; the reduplication of verbs and adjectives in yes/no questions; and finally, the distinctive use of a number of particles in Cantonese, including the assertive particles 嚟㗎 *lai-gaa* in copular sentences; the delimitative particle 吓 *haa*, and the verbal particle 過 *gwo* for repetition.

4.2 Coverage of Register Differences

It has been observed that "almost any Mandarin grammatical pattern can be used in Cantonese and be understood, but such locutions are often not idiomatic" (Ramsey, 1987), and in general formal Cantonese is closer to Mandarin. These remarks are corroborated by our corpus. In the formal portion of the corpus, 30% of the Cantonese sentences are identical to the Mandarin; whereas in the colloquial portion, only 4% are.

Modality also highlights the differences in registers. To express modality, Mandarin typically uses modal verbs such as 可以 *ke-yi* or 能夠 *neng-gou* 'may'. While Cantonese has its equivalents *ho-ji* and *nang-gau*, in many contexts it is more idiomatic to employ syntactic constructions with 得 *dak* and 到 *dou*. The former indicates potential, and can mean possibility or permission; the latter is a verbal particle. Both can also be used in Mandarin, but much less frequently.

A comparison between the formal and colloquial registers again confirms their known differences (Matthews & Yip, 1994) and provides some quantitative evidence. In the colloquial register, there were 88 instances of *ke-yi* and *neng-gou* and their respective abbreviated forms; 27% of these instances were spoken in Cantonese with the *dak* or *dou* construction. In contrast, in the 23 instances of the same modal words in the formal register, neither *dak* nor *dou* appear.

4.3 Plural marker for personal nouns

Although not mentioned in the list of (Ouyang, 1993), it is well known that Mandarin uses the suffix 們 *men* to mark personal nouns as plural, while Cantonese has the analogous suffix 哋 *dei* for personal pronouns, and the classifier 啲 *di* for other nouns (Matthew & Yip, 1994). Our corpus shows, however, two additional details.

First, the Cantonese suffix may be omitted. For example, in the noun phrase 你兩個 *nei loeng go* 'you two CL', the suffix *dei* is expected to mark 'you' as plural but is missing. These omissions all occur in the colloquial register.

Second, besides *dei* and *di*, the classifier 班 *baan* 'group' can also serve as the plural marker. For example, 孩子們 *hai-zi-men* 'child PL' 'children' is equivalent to 班細路 *baan-sai-lou* 'group child' 'children'. These also were observed exclusively in the colloquial register. This classifier is also used for the vocative case. In Mandarin, *men* is used in vocative plural, but the Cantonese *di* itself will not do. Instead, both the plural 'you' and *baan* are prefixed before the personal noun, as in 你哋班師奶 *nei dei baan si naai* 'you PL group wife' 'O you wives'.

4.4 Agentless passive

Both Cantonese and Mandarin mark passives with the word 被 bei, followed by the agent. If the agent unknown, it can be simply dropped in Mandarin, but in Cantonese the "generic" agent 人 *jan* 'person' must still be supplied.

Of the 16 sentences with passives, 9 are agentless in Mandarin. As for their Cantonese counterparts, 7 conform to the normal practice using *jan*, but the other two are agentless. These latter may be considered a form of "Mandarinism", i.e., usage that is not ungrammatical, but atypical of Cantonese speech. As expected, one of these occurs in the formal portion of the corpus; the other, in the colloquial, turns out to be a read speech in the drama.

4.5 Possessive constructions

Mandarin uses the possessive marker 的 *de*, whose Cantonese counterpart is 嘅 *ge*. In Cantonese, the marker may be omitted when expressing kinship or a "close" and "inalienable" link (Matthews & Yip, 1994; Pacioni 1998), as in 佢哋老豆 *keoi-dei lou-dau* 'they father' 'their father', without *ge* in between 'they' and 'father'.

The corpus shows, on the one hand, that this phenomenon extends to other nouns such as 佢心願 *sam jyun* 'wish'. On the other hand, for some expressions of kinship, the marker is not simply omitted but replaced by a classifier, such as 我個仔 *ngo-go-zai* 'I CL son' 'my son'. The number of syllables may be a determining factor.

5 Conclusion

We have presented the first large-scale parallel corpus of transcribed spoken Cantonese and written Chinese. Have shown its coverage of grammatical differences between the two languages, and its potential in corroborating and adding to known issues, we plan to further exploit it for quantitative studies in comparative grammars.

Acknowledgments

The author gratefully thanks Man Chong Mak for transcribing the TV programs and performing initial analyses. This work was partially supported by a Small-Scale Research Grant from the Department of Chinese, Translation and Linguistics at City University of Hong Kong.

References

Pi-Chuan Chang, Michel Galley, and Chris Manning, 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. Proc. ACL 3rd Workshop on Statistical Machine Translation.

K.-J. Chen, C.-R. Huang, L.-P. Chang, and H.-L. Hsu, 1996. Sinica Corpus: Design Methodology for Balanced Corpora. Proc. 11th Pacific Asia Conference on Language, Information and Computation (PACLIC). Seoul, Korea.

Samuel Hung-nin Cheung, 2007. Cantonese as Spoken in Hong Kong 香港粵語語法的研究. The Chinese University of Hong Kong Press, Hong Kong.

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

Man-Tak Leung and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese Adult Language Corpus. International Journal of Corpus Linguistics 6(2):305---325.

Stephen Matthews and Virginia Yip, 1994. *Cantonese: A Comprehensive Grammar*. Routledge, London.

Helen M. Meng and Pui Yu Hui. 2001. Spoken Document Retrieval for the Languages of Hong Kong. Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong, China.

Jueya Ouyang 歐陽覺亞, 1993. 《普通話廣州話的比較與學習》。北京：中國社會科學出版社。

Patrizia Pacioni, 1998. Possessive Constructions, Classifiers and Specificity in Cantonese. *Studies in Cantonese Linguistics*, Stephen Matthew (ed.), Linguistic Society of Hong Kong.

Prokopis Prokopidis, Vassia Karra, Aggeliki Papagianopoulou, and Stelios Piperidis, 2008. Condensing Sentences for Subtitle Generation. *Proc. Linguistic Resources and Evaluation Conference (LREC)*.

S. R. Ramsey, 1987. The Languages of China. Princeton University Press.

B. K. Tsou and O. Y. Kwong, 2006. Toward a Pan-Chinese Thesaurus. Proc. 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström, 2010. Machine Translation of TV Subtitles for Large Scale Production. *Proc. 2nd Joint EM+/CNGL Workshop on Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC)*. Denver, CO.

Ping-Wai Wong, 2006. The Specification of POS Tagging of the Hong Kong University Cantonese Corpus. International Journal of Technology and Human Interaction 2(1):21---38.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer, 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11:207-238.

Yaling Liang 梁雅玲, 1996. 《普通話與廣州話常用句型對譯》。香港：香港文化出版社。

Bennan Zhang 張本楠, Ruowei Yang 楊若薇, 2008. 《同形異義：粵普詞語對比例釋》。香港：三聯書局。