Abstract—Though Cantonese is the most influential variety of Chinese other than Mandarin, there are only a limited number of Cantonese corpora available for linguistic studies. Among the essential steps of building a corpus, word segmentation is a necessary but highly challenging task due to the lack of clear word boundary in Cantonese. This paper reports the construction and evaluation of an open-source automatic Cantonese word segmenter developed for Cantonese. The tool is a component of the multilingual SPPAS program designed to be used directly by linguists. It is a free software distributed under a GPL license. The effectiveness of the tool was evaluated by comparing the result of segmenting some samples of a spoken Cantonese corpus manually and automatically using the tool developed. High precision and recall were found in our study. Upon completion, the tool would definitely promote the development of more Cantonese corpora for language related studies.

Index Terms—segmentation, automatic, Cantonese, software, corpus

I. INTRODUCTION

Cantonese is the native language spoken by the majority population in Hong Kong, Macau, and Guangzhou. It is used as the daily means of communication in those places. As the most widely known and influential variety of Chinese other than Mandarin, according to [1], Cantonese has been a major subject matter of researches in linguistics, sociolinguistics, psycholinguistics, neurolinguistics, bilingual studies, speech pathology, child language development, and natural language processing (NPL). However, only a limited number of annotated spoken Cantonese corpora are available for these studies because building Cantonese corpus is very labor-intensive. Since Cantonese is by and large a spoken language, there is no large quantities of machine readable text that can be converted readily into corpora. For studies that have to be based on authentic data, researchers have to collect speech data and build the database/corpus on their own.

In general, three tasks are involved in turning collected language materials into linguistic representations for further studies: 1. Orthographic and/or phonetic transcription; 2. Word segmentation; and 3. Part of Speech (POS) tagging. Among these tasks, word segmentation and POS tagging are particularly challenging for Cantonese, or Chinese in general. Between them, word segmentation is a necessary first step for language processing. Chinese is written in character strings without word delimiters, such as the white space in English. The lack of delimiters in written form is, in fact, a reflection of lack of natural word boundary in Chinese. Wordhood has long been a problematic issue in Chinese linguistics since there is no proper definition of words in the language. Unlike many European languages, Chinese makes no distinction of morphology from syntax. Word formation in Chinese is mostly achieved by recursive compounding and derivation [2]. Furthermore, the grammatical relation of morphemes in a word is basically the same as that of words in a phrase. As a result, the boundary between a word and a phrase is difficult to draw. The judgement of wordhood in Chinese is mainly based on the idiomatic meaning of the morpheme strings and the separability between the constituents. Unfortunately, there are a large number of leihapci1 離合詞 ‘separable words’ in Chinese, which can be split like phrases. All these grammatical idiosyncrasies make word segmentation notoriously challenging in Mandarin Chinese. A large body of research on the wordhood problem has been conducted in Chinese language processing, such as [3], [4], [5], [6], and [7].

As a Chinese dialect, Cantonese shares many of these idiosyncrasies. Though there is no accurate statistics, [8] estimated that about 70% of Cantonese words were cognates to Mandarin words, either with the same written form or with systematic sound correspondence. The remaining 30% were colloquial words, which were used in 50% of everyday communication. As compared with Mandarin, Cantonese has more words in monosyllabic form. Furthermore, Cantonese has a group of post-verbal particles with mainly grammatical meaning but undetermined grammatical status that are not found in Mandarin, such as saai 晒, maai 埋, hoi 開, and faan 番. Judging from grammatical meanings, they should be part of the internal structure of a word. However, they are loosely attached to the verb and can have long distance separation from the verb. Take the verb, zo 阻 ‘to hinder’ as an example, zo-zyu 阻住 ‘hindered,’ zo-zyu-saai 阻住晒 ‘totally hindered,’ zo-gwai-zyu-saai 阻鬼住晒 ‘totally hindered (in a disgusting

1The LSHK Cantonese Romanization scheme is adopted throughout the paper.
Most researchers who are not particularly interested in syntax do not want to spend their effort or do not have the resources to go through all the arguments in the literature to determine the wordhood of the character strings in their transcribed materials. It would be most helpful if there is a ready-to-use tool to provide automatic word segmentation for their reference. Unfortunately, automatic text normalization is a non-trivial task, and it is hard to have a "good" segmenter. There are quite some word segmentation approaches in NLP: dictionary-based, frequency-based, rule-based, maximum-entropy based approaches, just to name a few. Text normalization is then commonly considered as language-dependent and/or task-dependent. For normalization, rule- and regular expression-based systems are the norm for romanized languages, including the tokenizers in the RASP system [9], the LT-TTT tools [10], and the FreeLing tools [11].

There are some word segmentation tools developed for Mandarin, and some of which are freely available for download. Some of the tools are reported in the following website [12]. More recent systems are also developed and are freely available on various web sites, such as GkSeg: a Chinese word segmentation package based on character-based tagging heuristics [13], and CRF algorithm or "Jieba" (Chinese for "to stutter"): a Chinese word segmentation module implemented in Python[14], just to cite two of them. However, such tools are based on a training stage that requires a large amount of manually segmented data, so that they can be adapted to a new character-based language only if such resources are available. It is just not practical to apply these tools directly to Cantonese by linguists, most of which have limited computer expertise and human resources.

Furthermore, it is highly undesirable to apply Mandarin word segmenter to Cantonese. Apart from the difference in linguistic structures, there are also differences in processing the written texts of Mandarin and Cantonese. Cantonese text processing has to face the problem of lack of standard written forms for the colloquial words in its vocabulary. Some Cantonese words have no writable characters and are customarily represented by a square, such as ngaai □ 'call for.' Some of the colloquial words are represented by homographs. For instance, the homograph meaning 'animal' kám 稈 is borrowed to represent the first syllable of kamjat 稃 'yesterday.' Some of them are represented by dialecal characters coined by the public in a non-systematic way. For instance, the negator mou 'not have' is written as either 無 or the dialectal character 行. Some of words are transliteration of loan words, mostly from English, such as baa si 巴士 'bus, dik si 的士 'taxis, tauai 喀 'tyre, tinnaaseoi 天那水 'thinner. It becomes apparent that there are more homonyms in Cantonese than in Mandarin. Consequently, greater ambiguity may arise in processing Cantonese text because it is more difficult to decide whether a character represents an independent word of just a component of a word.

Thus far, there are no tools that are open-source and readily used by linguists to perform automatic word segmentation specifically developed for Cantonese. To fill this gap, we have developed a free-to-download and ready-to-use tool to perform automatic task for linguists who have limited resources and technical knowledge in language processing. This paper reports the construction of the tool and the study we have conducted to evaluate the effectiveness of the tool developed.

II. TOOL CONSTRUCTION

A. SPPAS overview

The word segmenter we developed for Cantonese is a component of SPEch Phonetization Alignment and Syllabification (SPPAS). SPPAS is a tool to automatically produce annotations which include utterance, word, syllable and phoneme segmentations from a speech recording and its transcription [15]. It includes a succession of 4 automatic steps. The resulting alignments are a set of separate TextGrid files, the native file format of Praat software [16], for (i) utterance segmentation, (ii) word segmentation, (iii) syllable segmentation, and (iv) phoneme segmentation. SPPAS is specifically designed to be used directly by linguists in conjunction with other tools for the analysis of speech, and distributed under the terms of the GNU Public License. In addition to Praat (the most popular phonetic software among linguists), SPPAS can read and write a large set of files from other tools, such as Elan, Transcriber, HTK, SClite and CSV. It can also import data from other tools. The software fulfills the specifications listed for annotation tools in [17]; it is a linguistic tool; it is free of charge; it is ready and easy to use; it runs on any platform and it is easy to install; the maintenance is guaranteed; and it is XML-based.

SPPAS is a multilingual program which is currently implemented for French, English, Italian, Spanish, Mandarin Chinese and Taiwanese. The methodology implemented in SPPAS for all automatic annotations develops algorithms as language-and-task-independent as possible. This allows the adding of new languages with a significant reduction of time compared to the entire development of such tools. Adding a new language in SPPAS only consists of adding resources related to the annotation (such as lexicons, dictionaries, models, set of rules, etc). Consequently, the quality of the automatic annotations is largely influenced by such resources, and phoneticians can contribute to improve them. In that sense, phoneticians need automatic tools and tools need phoneticians.

B. Cantonese Word segmenter in SPPAS

The Cantonese word segmentation is a sub-task of the text normalization task in SPPAS. The input text of a corpus may contain a variety of “non-standard” token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL’s and e-mail addresses, etc. Text normalization is a task to rewrite such texts using ordinary words. It is commonly considered as language-dependent and/or task-dependent. However, the methodology implemented in SPPAS for text normalization [18] develops a method to normalize texts using a set of language-independent tools. During text
normalization, the process of word segmentation takes place by making reference to the lexicon, or word list, installed in SPPAS for a particular language. The word segmenter looks up the lexicon of the language to find out which of the characters of the input text must have to be grouped into tokens. In linguistic terms, the grouped tokens are identified as multisyllabic words. All the missing tokens are split into characters. In other words, they are considered monosyllabic words. In operation, the tool segments concatenate character strings into words based on the optimization criteria: longest matching. The character strings gwokzaifaa 國際化 serves as a good illustrating example. Since both gwokza 國際‘international’ and gwokzaifaa 國際化 ‘internationalize’ are found in the dictionary, the tool will group the whole character string 國際化, the longest matching from left to right, as one token, i.e. one word [18].

The tokenization process for character-based languages is described in Figure 1. Romanized languages implies 2 more steps in this workflow: one to deal with apostrophes and hyphens that both can be a segmentation indicator or not, and the other one to lowerise the text. All these modules of the tokenization process are language-independent, except the number to letter conversion module. Details of this process can be found in [18]. Word segmentation is related to “Character split” and “Stick by longest matching” and is then representing only two modules of the tokenization process.

Fig. 1. The tokenization workflow for character-based languages, as implemented in SPPAS.

As explained in the above, building up the dictionary for Cantonese is vital in developing the word segmenter. The ideal dictionary has to fulfil the following 3 criteria. It should (i) contain words and not only characters (the written form of syllables); (ii) contain a reasonably large-size Cantonese word inventory; and (iii) be used under the terms of an open-source license. We have performed an examination of the existing Cantonese word lists on the Internet and found the following three word lists that matched our criteria:

1) A list contained 224,230 characters, words and phrases (http://www.ZhongWenLearner.com/);
2) The Words Commonly Used in Hong Kong (http://cpls.proj.hkedcity.net/cpls/general.jsp?headercat-teacherAreaCantons.htm). It contained over 3,000 words (4,227 characters) frequently used in Hong Kong Cantonese;
3) The Open Dictionary of Cantonese Words (http://www.kaifangcidian.com). It contained 35,037 words which included Cantonese words used in Mainland China. As compared with (2), this item contained more Mandarin cognates, literary words, and Cantonese idioms.

We finally decided to include only items (2) and (3) in our dictionary. Item (1) was not used because it contained too many entries that were too specific, such as proper names (people, streets, book titles, etc) and long phrases. With both items (2) and (3), our tool can be applied to a large geographical varieties of Cantonese of both formal and colloquial registers. As a result, a list of 46,205 entries was created and incorporated in the Cantonese dictionary of SPPAS to perform text normalization and word tokenization tasks for Cantonese.

III. Evaluation

Before we proceed to the evaluation of the SPPAS word segmenter on Cantonese materials, we have applied it to the Chinese word segmentation evaluation toolkit for written Chinese launched on the website http://projectile.sv.cmu.edu. The database of this toolkit was extracted from the Chinese Treebank, which contained written Chinese from Xinhua news, Sinorama news magazine, and Hong Kong News. We got a F-measure of 0.92, which is more or less the same as the CTSeg system of Erik Peterson, which used the same algorithm as we implemented in SPPAS. Note that this database is a written corpus, i.e. a Mandarin Chinese corpus, not a spoken Cantonese corpus. Regarding the results of the other systems, we can consider ours as a baseline.

With this encouraging result, we then evaluated the effectiveness of our tool to spoken Cantonese. Automatic word segmentation using the tool and manual word segmentation conducted by a linguist were performed and compared on some samples of a Cantonese speech corpus.

A. Cantonese Database

The corpus selected for evaluation in this study was Hong Kong Corpus for Adult Cantonese (HKCAC) [19], [20]. It contains more than 8 hours (487 minutes) of speech recordings of spontaneous spoken Cantonese comprising 20,770 utterances and about 170,000 characters. The speech data come from the recordings of phone-in programmes and forums on the radio in Hong Kong, representing the speech of about 80 native speakers. The recordings are transcribed both orthographically and phonetically. It is the only Cantonese speech
In this example, there are substitutions after precision, recall and harmonic mean $F$ are defined as:

\[
\text{Precision} = \frac{(\text{number of substitutions})}{(\text{number of separators in Hypothesis})}
\]

\[
\text{Recall} = \frac{(\text{number of substitutions})}{(\text{number of separators in Reference})}
\]

\[
F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}
\]

In the following, we illustrate insertions with the ‘|’ separator and deletions with the ‘+’ separator, for example: $C_1 C_2 S C_3 C_4 C_5 S C_6 C_7 + C_8 C_9 S C_{10}\mid C_{11} C_{12} C_{13} S C_{14}$.

C. Results and Analysis

The evaluation results of the two episodes were shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>FINANCE-21G</th>
<th>GHOST-46G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td>3123</td>
<td>1882</td>
<td>5005</td>
</tr>
<tr>
<td>Insertion</td>
<td>62</td>
<td>23</td>
<td>85</td>
</tr>
<tr>
<td>Deletion</td>
<td>219</td>
<td>112</td>
<td>331</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9345</td>
<td>0.9438</td>
<td>0.9380</td>
</tr>
<tr>
<td>Precision</td>
<td>0.9806</td>
<td>0.9879</td>
<td>0.9833</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.9569</td>
<td>0.9654</td>
<td>0.9601</td>
</tr>
</tbody>
</table>

As shown in the table, the precision and recall are relatively high. Among the two types of error made by the word tokenizer, insertion errors were limited whereas deletion errors were more common. In other words, the segmenter made more errors in ‘separating’ a word into characters, rather than ‘combining’ characters into a word. A closer examination of these two types of error are in order.

Where do the deletion errors come from? There were altogether 331 deletions in our study. The detailed breakdown are listed in the following:

- 97 errors were proper names, that occurred only in FINANCE-21G, i.e. 44.3% of the deletion errors in this file.
- 10 errors concerned 科技+股
- 10 errors concerned 大姑+奶
- The others errors were equal or less than 3 occurrences each.

The high percentage of the proper names found in FINANCE-21G may be caused by the addressing of the personal names of the speakers repeatedly in the programme, such as Sek Gengcyun 石鏡+泉 and Sin Zunggaaui 單件+佳 in their exchanges. Furthermore, some company names and geographic names, such as Jumbo盈科 ‘PCCW Ltd.,’ Soumaagong 數碼港 ‘Cyberport,’ Hoenggong 香港 ‘Hong Kong’ were frequently mentioned in the forum. Interestingly, Hoenggongjan 香港人 ‘Hongkongers’ was in the dictionary but not Hoenggong 香 港 ‘Hong Kong.’ The term fogen 付戈+股 ‘tech stocks,’ is a relatively new word, which occurred frequently in this programme because it was one of the subject matter of the forum. The other high frequency error daaigunaai 大姑+奶 ‘husband’s bigger sister’ was a kinship term used in highly limited context in contemporary Cantonese. It occurred frequently only in GHOST-46G because she was the leading
character of the paranormal story in the programme. It is obvious that adding these tokens in the dictionary can considerably reduce the amount of deletion errors.

Where did the insertion errors come from? There were 85 insertion errors detected. The breakdown is shown in the following.

- 15 errors were related to: 一|個
- 14 errors were related to: 咪|個
- 7 errors were related to: 亦|都
- The others were equal or less than 3 occurrences each.

A total of 29 errors involve the classifier go 個. Our dictionary does not treat classifiers as independent words. However, it was treated as an independent word in manual segmentation based on the fact that classifiers in Cantonese can occur alone without a numeral or a demonstrative as in go haak zou laa 個客走嘅 'The guest is leaving.' The other frequent error concerned the adverbs jik 亦 'also' and dou 都 'also,' which can occur independently despite their co-occurrence in many occasions. By correcting these 'wrong tokens' in our dictionary, the error rate would largely reduced.

The other source of insertion error concerned the ambiguity problem caused by homographs. This phenomenon is well illustrated by the following error: baakut-mau Soumaagong 包+ 括埋| 數+ 碼+ 港 'including Cyberport as well.' The character 埋 'included' in this utterance. However, the character was also the written form of the lexical verb 'bury,' which could combine with the noun 'number' to form the word 埋數 'closing the account.' If the popular name 'Cyberport,' a newly established corporation, was added in our dictionary, the error would have been avoided. As introduced in the above, ambiguity is a major problem in Cantonese text processing. However, errors caused by ambiguity were not commonly captured in our study.

IV. CONCLUSION AND FUTURE WORK

To conclude, our tool has generated impressive results for this highly challenging spontaneous spoken corpus. Most of the errors found in our study can be traced back to the dictionary. However, no dictionary will be perfect. In fact, OOVs problem and named entities are the main challenges in Chinese word segmentation, in particular under the maximum match paradigm adopted in our tool. The advantage of using SPPAS is that any user can easily edit the dictionary, and add or remove lexical items as needed in order to get the tokenization to generate the expected result as far as possible. Furthermore, the consensus of many syntactic status of Cantonese words has yet to be reached. The dictionary provides one possible framework for the users as a quick reference. Our segmenter will be a handy tool for researchers who do not have to be involved in serious syntactic arguments.

We are going to edit our dictionary in SPPAS to better reflect the development of the Cantonese syntactic study. Then, we will generate a word frequency list based on HKCAC. At the moment, there is no information of word frequency based on spoken Cantonese. However, word frequency is an important reference for studies of language change, and a much needed data in forming stimuli for a variety of psycholinguistic and neurolinguistic experiments. With word segmentation handled, POS tagging could also be proceeded. Upon completion, the word tokenizer of Cantonese SPPAS will definitely be a useful tool for language researchers. It will also promote the development of corpus by reducing the technical challenges to many linguists.

ACKNOWLEDGMENT

The collaboration of this study has been supported by PROCORE-France/Hong Kong Joint Research Scheme.

REFERENCES


