ABSTRACT

The Hong Kong University Cantonese Corpus was collected from transcribed spontaneous speech in conversations and radio programs that involved two to four people. It was word-segmented, annotated with Cantonese pronunciation, and recently tagged with word classes by adopting the parts-of-speech (POS) scheme of Yu et al. (2002). This scheme, which was designed for tagging written Mandarin texts, encountered some problems in tagging spoken Cantonese. However, it is flexible for further expansion of the 26 basic word classes by customizing some subclasses for annotating other Chinese dialects (e.g., Cantonese). Its robustness was proved by the annotation of approximately 230,000 words in the HKUCC. This article will describe the format of the corpus and provide the specification that helps annotators in POS tagging and will solve problems encountered in manual annotation. Guidelines of tagging some word classes will be introduced, followed by the discussion of easily confused tags, illustrated with examples from the corpus. Further work will aim at automatic annotation by computers in order to facilitate the work of POS tagging of Cantonese and other Chinese dialects. The corpora of Hong Kong Cantonese are quite lacking. Past work focused either on a POS-tagged corpus for child language or the phonetic transcription of an adult Cantonese corpus. HKUCC fills the gap by providing a POS-tagged corpus for adult Cantonese and is believed to be of great value to the data-driven linguistic analysis and natural language processing for Cantonese.

Keywords: database; data structure; natural languages; XML

INTRODUCTION

This article addresses the issues that arise in parts-of-speech (POS) tagging of the Cantonese corpus. It aims to provide the specification that helps annotators to tag Cantonese corpus and solve the problems encountered in manual annotation.
which was tagged with Cantonese Romanization and a tagset of 33 POS tags. It provided essential resources for studying children’s language and language acquisition, but the limitedness of vocabulary and the grammar of children’s language may not be adequate to describe Cantonese grammar in general and to build a language model of Cantonese adults for NLP purposes. Two adult language corpora can be accessed online; namely, the Hong Kong Cantonese Adult Language Corpus (HKCAC) (Law, Leung & Fung, 2002) and Hong Kong Corpus of Primary School Chinese (HKCPSC) (Leung, 2002). However, they only were annotated with Cantonese pronunciation (using the International phonetic alphabet) without further processing such as word segmentation and POS tagging. Thus, this article discusses how to build a word-segmented and POS-tagged Cantonese corpus for data-driven linguistic analysis and NLP resources for Cantonese.

The Hong Kong University Cantonese Corpus (HKUCC)

The Hong Kong University Cantonese Corpus (Luke & Nancarrow, 1997) was collected from transcribed conversations that were recorded between March 1997 and August 1998. About 230,000 Chinese words were collected in the annotated corpus. It contains recordings of spontaneous speech (51 texts) and radio programs (42 texts) that involved two to four speakers, with one text of monologue. In the first phase of annotation, the texts were word-segmented and annotated with Cantonese pronunciation using the Romanization scheme of the Linguistic Society of Hong Kong (LSHK)\(^1\). The second phase that began in June 2004 was the POS tagging, following the POS scheme of Yu, et al. (2002), in which the 26 letters of the English alphabet were assigned to the 26 basic POS types in Mandarin Chinese (see Appendix I for the tag list); for instance, \(a\) stands for adjective, \(n\) for noun, \(r\) for pronoun, \(v\) for verb, \(w\) for punctuation. There were two passes in the POS tagging. The corpus was tagged by one annotator in the first pass and then checked by another annotator in the second pass. The first pass now has been completed, whereas the second pass is in process.

Organization of the Article

The article is organized as follows: It will illustrate the format of the tagged corpus in the second section and introduce some of the basic POS tags, such as categorizers and state words in the third section. The tagset is robust for expansion, and some subcategories of the basic POS tags will be described in the fourth section. Besides the close-classed words and easily confused tags in the fifth section, the article also will discuss some problematic cases involving verbs in sixth section, followed by a conclusion in the final section.

FORMAT OF TAGGED CORPUS

The HKUCC applies an XML-like format. Each element is contained in an opening tag (e.g., \(<\text{info}\>) and a closing tag (e.g., \(</\text{info}\>)). The following shows the meanings of the tags used in the corpus:

Information about the Recordings and the Interlocutors

\(<\text{info}\>) records the information about the recordings and the speakers, as shown in Figure 1. The tape number is shown on the first item, while the second item, DR, shows the date of recording — March 3, 1997. The third item, NS, stands for the