File formats of tagged files.

Each file has two sections – an INFO section which documents all relevant information about the recording and a SENT section with the tagged contents.

The INFO section
Every line in the INFO section starts with a line number, followed by the name and value of an attribute. The first four attributes are tape number (TN), date of recording (DR), number of speakers (NS) and list of speakers (LS). Please note that DR is in the DDMMYY format, and 000000 stands for unknown date-of-recording. LS is the list of speaker codes in the file. Speakers are referred to by these codes in the entire file. The details of each speaker are recorded in the format <Letter-code> - <Sex> - <Age> - <Origin>. For example, A-M-22-HK indicates speaker A is a 22-year-old male speaker from Hong Kong. INFOEND marks the end of the INFO section.

The SENT section
Tagged sentences are placed in the SENT section. Each <sent> tag contains one tagged utterance. <sent_head> has the speaker code. <sent_tag> are tagged words in vertical format. <sent_tran> is reserved for the English translations. Each line in <sent_tag> contains a segmented word in Chinese character, part of speech (in the format as described in the corpus specification) and the LSHK romanisation of the word. The word, the POS and the romanisation are separated by a slash (/).

檔案格式(已標注檔案)

每個檔案都有 INFO 和 SENT 兩個部份。INFO 是關於該次錄音的詳情, 而 SENT 則爲已標注的錄音內容。

INFO 部份
該部份每行均以行數起首，然後是屬性的名稱和內容。首四行依次爲錄音帶編號(TN)、錄音日期(DR)、講話人數(NS)和講話人列表(LS)。其中 DR 以 DDMMYY 格式記錄，000000 則表示錄音日期不詳，而 LS 是講話人的代號列表。從第 5 行起是每一位講話人的背景資料，以 <代碼>-<性別>-<年齡>-<出身地>的格式記錄。譬如，A-M-22-HK 指該名被稱作 A 的講話者是一名來自香港的 22 歲男子。
INFOEND 提示 INFO 部份完結。

SENT部份
已標記的句子都放在 SENT 部份裏。每一個<sent>裏面包含一句話。<sent_head>記載了講話人的代碼、<sent_tag>記載著已標記的詞語，<sent_tran>部份保留作英語翻譯之用。<sent_tag>的內容以一行一詞的方式表示。每行均記有該詞的漢字、詞性和拼音。三者之間以 / 分隔。