CORPUS INFORMATION FORM

Check (x) or fill in the blank (____) if appropriate.

1. Name of the Corpus: CCC-TCT (Tsinghua Chinese Treebank)

2. IPR Holder: Chinese Corpus Consortium (CCC)

3. Corpus Type:

   - ■ Speech Corpus ( )
   - ■ Text Corpus (x)

4. If it is a speech corpus:

- Purpose:
  - ■ ASR ( )
  - ■ TTS ( )
  - ■ Other, please specify _____
- Language:
  - ■ Putonghua (SC) ( )
  - ■ Mandarin in Taiwan (TC) ( )
  - ■ Cantonese in HK (TC) ( )
  - ■ Other, please specify _____
- Style:
  - ■ Read speech ( )
  - ■ Spontaneous speech ( )
  - ■ Conversational speech ( )
  - ■ Other, please specify _____
- Channel:
  - ■ Close-talk Microphone ( )
  - ■ Telephone ( )
  - ■ Mobile phone ( )
  - ■ Other, please specify _____
- Sampling Rate: ____Hz
- Sampling Precision:
  - ■ PCM ( ), _____bits per sample
  - ■ A-law ( )
  - ■ Miu-law ( )
  - ■ Other, please specify _____
- Corpus size: _____hours _____ speakers
- SNR level: _____dB
- Transcriptions:
  - ■ Character tier (SC) ( )

- ■ Character tier (TC) ( )
- ■ Canonical Pinyin tier ( )
- ■ Other canonical pronunciation tier, please specify _____
- ■ Surface form IF tier ( )
- ■ Surface form IPA tier ( )
- ■ Surface form SAMPA-C tier ( )
- ■ Other surface form tier, please specify _____
- ■ Other transcription, please specify _____
- ■ Other transcription, please specify _____
- ■ Other transcription, please specify _____

5. If it is a text corpus:

- Language:
  - ■ SC (x)
  - ■ TS ( )
  - ■ Other, please specify _____
- Domain:
  - ■ Culture (x)
  - ■ Economy ( )
  - ■ Military ( )
  - ■ News (x)
  - ■ Politics ( )
  - ■ Sciences (x)
  - ■ Sports ( )
  - ■ Other, please specify literatures
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____
- Corpus size: 1.5 Mega characters
- Tag Information:
  - ■ Word segmentation: (x)
  - ■ Part-of-Speech (x)
  - ■ Other, please specify Syntactic parse trees, discourse relation trees
  - ■ Other, please specify _____
  - ■ Other, please specify _____
  - ■ Other, please specify _____

6. A brief Description of the Corpus:

This corpus was designed Chinese syntactic parsing tasks. We selected a balanced collection of journalistic, literary, academic, and other documents published in 1990s and annotated sentence-split, word segmentation, POS tagging and complete parsing trees for each sentence in these articles. We designed a two-tagset annotation scheme to describe syntactic information as

detailed as possible. The total number of annotated sentences is about 44, 600, covering about 1,000,000 Chinese words. This treebank can provide support to develop different Chinese parsers, including Chinese multiword chunk parser, Chinese functional chunk parser, Chinese dependency parser, Chinese event parser, and Chinese discourse relation parser.

In CEB-1, all sentences are extracted from Tsinghua Chinese Treebank (TCT).