

CORPUS INFORMATION FORM

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

1. **Name of the Corpus:** WCSC (Word Corpus of Standard Chinese)

2. **IPR Holder:** Phonetics lab., the Institute of Linguistics,
Chinese Academy of Social Sciences

3. **Corpus Type:**

- Speech Corpus (X)
- Text Corpus ()

4. **If it is a speech corpus:**

- Purpose:
 - ASR ()
 - TTS ()
 - Other, please specify phonetic and acoustic analysis for disyllables
- Language:
 - Putonghua (SC) (X)
 - Mandarin in Taiwan (TC) ()
 - Cantonese in HK (TC) ()
 - Other, please specify _____
- Style:
 - Read speech (X)
 - Spontaneous speech ()
 - Conversational speech ()
 - Other, please specify _____
- Channel:
 - Close-talk Microphone (X)
 - Telephone ()
 - Mobile phone ()
 - Other, please specify _____
- Sampling Rate: 16 k Hz
- Sampling Precision:
 - PCM (X), 16 bits per sample
 - A-law ()
 - Miu-law ()
 - Other, please specify _____
- Corpus size: over 1,000 disyllables each speaker, 3 speakers
- SNR level: _____ dB

- Transcriptions:
 - Character tier (SC) (X)
 - 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 ()
 - TS ()
 - Other, please specify _____
- Domain:
 - Culture ()
 - Economy ()
 - Military ()
 - News ()
 - Politics ()
 - Sciences ()
 - Sports ()
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____
- Corpus size: _____ Mega characters
- Tag Information:
 - Word segmentation: ()
 - Part-of-Speech ()
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____
 - Other, please specify _____

6. A brief Description of the Corpus:

- More than 3,000 words or phrases are designed covering segmental and super-segmental co-articulations.
- A management software can be used to access the speech data and display

- the spectrum of the disyllables.
- Speaker contents are non-overlapping.