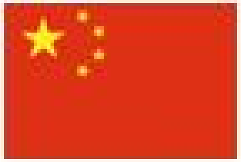


A Country Report – COCOSDA Activities in China

O-COCOSDA 2010

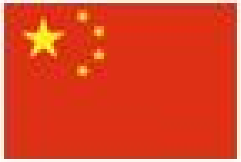
Kathmandu, Nepal

Thomas Fang ZHENG and Aijun Li

Activities in Chinese LDC

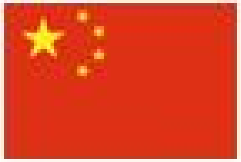
- Web Based Bilingual Parallel Database (CLDC-2010-001)
  - 1,075,162 parallel sentences from bilingual website, e.g. Government website, news website, language learning website, etc.
- Speaker Recognition Database for mobile application (CLDC-2010-002)
  - Large corpus which covers 303 speakers in four different recording channels, microphone, PDA, telephone and mobile phone
- Chinese and Mongolian Bilingual Database (CLDC-2010-003)
  - 400 Chinese sentences were selected for translation. Each sentence was translated by four Mongolian with different methods
- Mongolian Text Corpus (CLDC-2010-004)
  - It contains 100,000,000 Mongolian word entities with Latin transcription, and covers 22 different teaching books, 10 political books, 4 novels and lots of news.
- Chinese and Mongolian Bilingual Parallel Database (CLDC-2010-005)
  - It contains 60,000 bilingual parallel sentences, and can be used for the research on translation
- Summary of Chinese LDC
  - Till now Chinese LDC has collected 90 different corpora (more than 2T bytes).
    - for research on translation, word segmentation, syntactic parsing, spontaneous speech analysis, speech recognition and synthesis, speaker recognition, evaluation, etc.
  - 32 different databases have been shared via Chinese LDC since last December

Activities in Chinese Academy of Social Sciences



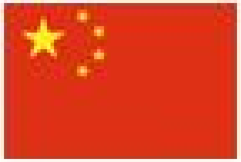
- CASS-CHILD Corpus: Child language acquisition speech corpus
  - Parental Input language database
    - words and sentences designed for checking parental input speech cues
    - 64 parents recorded while talking to babies aged 4-26 months
  - Spontaneous speech database
    - parent and child dyads every month or two weeks, 1 hour each time
- CASS-EEMA Corpus: The CASS EMA emotional corpus
  - 1 actor and 1 actress
  - 400 neutral sentences, 111 sentences in 7 kinds of emotions (*disgust*, *sad*, *angry*, *happy*, *surprise*, *fear*, and *neutral*), and emotional discourses (230 sad emotion sentences and 300 happy emotion sentences)
- CASIA-CASSIL Corpus
  - A large-scale corpus of Chinese spontaneous telephone conversations co-built by CASIA and CASSIL
  - Recorded in 5 real scenarios concerning tourism domain
    - air, hotel, restaurant, travel, and taxi
  - Annotation
    - a multi-leveled framework including *Turns*, *Speaker Gender*, *Orthographic Transcription*, *Chinese Syllable*, *Chinese Phonetic Transcription*, *Prosodic Boundary*, *The Stress of the Sentence*, *Non-Speech Sounds*, *Voice Quality*, *Topic*, *Dialogue-Act* (*DA*) *and Adjacency Pairs* (*AP*), *Ill-formedness*, and *Expressive Emotion*

Activities in Tsinghua University



- Time-Varying Voiceprint Database
  - To examine solely the time-varying impact on speaker recognition
  - Reading-style speech instead of free-style conversations
  - Fixed prompt texts
    - 100 Chinese sentences and 10 Chinese words
  - Gradient time intervals between two adjacent recording sessions
    - 16 sessions during the 3-year project
    - Initial sessions are of shorter time intervals, and following ones of longer and longer time intervals
  - Speakers
    - 60 fresh students with 30 male and 30 female aged 16-21

Activities in Other Research Institutes - 1



- Xinjiang University – Uyghur Corpus
  - Uyghur text data (1.8GB)
    - L3 labeling of 20,000 Uyghur sentences, and POS dictionary of 100,000 Uyghur words
  - Uyghur reading speech data (14GB, 125 hours)
    - 358 speakers with 190 female and 168 male
  - Telephone speech data
    - Uyghur dialog speech (800, 100 hours), and Nonnative (Uyghur people speaking Putonghua) dialog speech (300, 30 hours)
  - Nonnative reading speech data
    - 50 speakers, for mispronunciation detection
- Northwest University for Nationalities – Tibetan Corpus
  - Video corpus
    - More than 6,000 monosyllables pronounced by 3 speakers, respectively
    - Mainly focus on facial characteristics and lip position changes
  - Text corpus
    - Basic corpus of Tibetan text data, Parallel Tibetan-Chinese(-English) corpus, and POS-labeled corpus of primary school textbooks
  - Speech corpus
    - Word corpus (100 thousand), Speech synthesis corpus (7,000 news sentences), Speech recognition corpus (spontaneous, 600+ hours, 100+ speakers), and Respiratory rhythm corpus (100 poems and 100 pieces of news)

Activities in Other Research Institutes - 2

- Beijing Language & Culture University – Inter-Chinese Speech Corpus
  - For study on Chinese pronunciation acquisition as a second language and developing computer aided pronunciation training
  - Including Japanese, English, Korean, Arabic, *etc*
  - Inter-Chinese speech data by 120 Japanese speakers have been collected.
    - 526 chosen monosyllables by 100 speakers, and 301 daily utterances by 20 speakers (phonetically annotated)
- Beijing Jiaotong University – Broadcast News Speech Corpus
  - 60 discourses of broadcast news speech by a female professional speaker, 1 hour
  - Transcription in Chinese characters, pronunciation labeling in Pinyin, and segment boundary in syllables
  - Prosodic annotation: *Prosodic hierarchy*, *Stress*, *Intonation* & *Mood*
    - Three dimensions to specify *Mood:* Commendation (CO) – Criticism (CR), Lightness (LI) – Seriousness (SE), and Happiness (HA) – Sadness (SA)
- Communication University of China – Broadcast News Speech Corpus
  - 24h speech data so far (segmented, transcribed and labeled)
    - 10h labeled data focus on sentence intonation
    - 14h labeled data focus on discourse level prosodic feature of broadcasting announcing
  - From the flagship news programs i.e. *Xinwen he Baozhi Zhaiyao* on China National Radio and *Xinwen Lianbo* on China Central TV