A Country Report – COCOSDA Activities in China

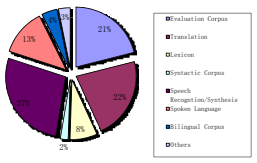
O-COCOSDA 2013

Gurgaon, India
Aijun LI and Xiaojun WU

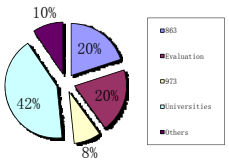
1

Activities in Chinese LDC

- Till now, there are 95 different corpora, including corpora for speech synthesis/recognition, machine translation, lexicon and other natural language processing corpus.
- In 2013, there are 2 new corpus added, and 16 corpora have been distributed to 9 institutes and companies within this year.



Types of the corpora



Resource of the corpora

2

Activities in Chinese Corpus Consortium

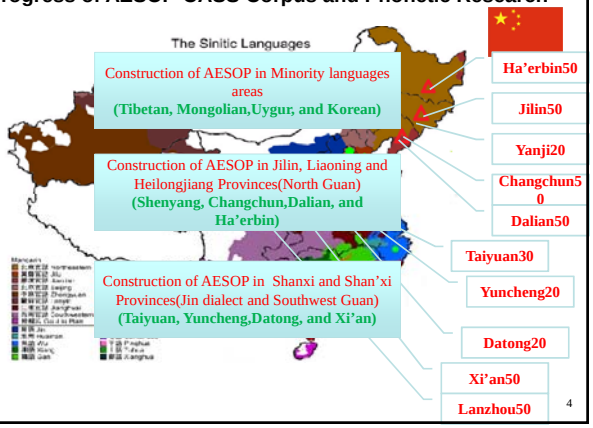
- Speech corpus
  - 18 for speech recognition
  - 3 for speech synthesis
  - 9 for speaker recognition
- 3 corpus as language resources

Activities in Center for Speech and Language Technologies, Tsinghua University

- A longitudinal voiceprint database
  - Focuses on the time-varying issue
  - 60 university students (30 males and 30 females)
  - 16 sessions of gradient time intervals in 3 years
  - 100 fixed Chinese sentences per session, reading style

3

Progress of AESOP-CASS Corpus and Phonetic Research



4

Activities in Institute of Linguistics, Chinese Academy of Social Sciences

- **CASS_CHILD Corpus: child language acquisition speech corpus**
  - Longitudinal child spontaneous speech: 23 children, recorded once for one hour per month in both audio and video format. The audio files are transcribed orthographically and annotated phonetically in conjunction with linguistic and paralinguistic information.
  - **Elicited recordings**
    The Question-Answer (Q-A) dialogues between parents and children based on prepared prompts, and children's production of words (Beijing Articulation Norms Project) , 4000 children, 1.5y~6y.
  - **Web-based recording platform (co-developed with iFlytek)**
    http://pslab.cass.cn
  - **Retrieval tool (co-developed with IOA of CAS)**
    A & V clips can be retrieved for multimodal CASS_CHILD Corpus
- Online linguistic and speech resource sharing platform: http://pslab.cass.cn/
- Online Chinese dialectal speech resources: http:// dialect.phonetics.org.cn

5

Activities in Broadcast Media Language Research Center, Communication University of China (after Dr. HE Wei)

- Speech corpora
  - Broadcast Media Language Corpus
    - 800 million characters text data
    - More than 15000h audio & video data
  - Mandarin Chinese Digital Sampling Corpus (MCDSC)
    - 3543h video or audio files with metadata tagging since 1933
  - Sentence-matched Broadcasting Speech Corpus
    - Over 1000h speech data from TV or radio news programs
- Tools
  - The sentence-matched speech retrieval system
  - The dialect homophonic character generation tool

6