A Country Report – COCOSDA Activities in China

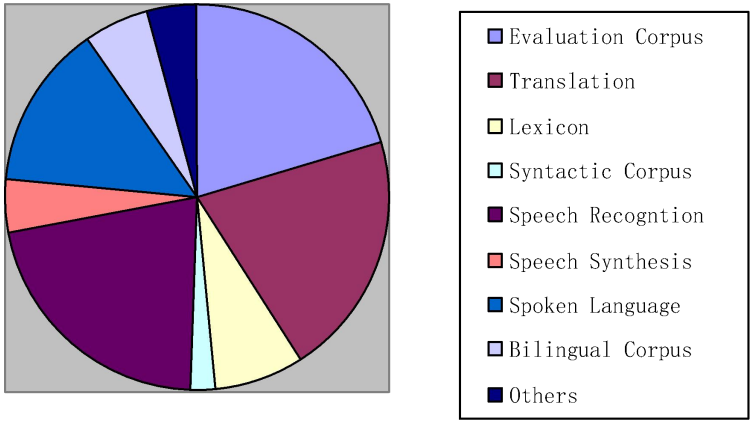
O-COCOSDA 2012

Macau, China

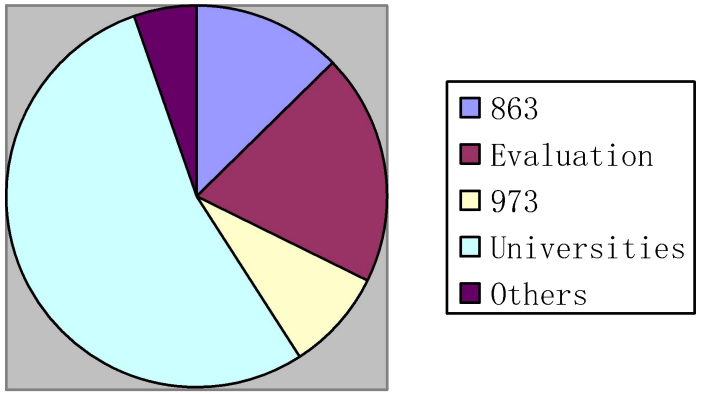
Aijun LI and Xiaojun WU

Activities in Chinese LDC

- Till now, there are 93 different corpora in total, including speech synthesis corpora, speech recognition corpora, machine translation corpora, lexicon and other natural language processing corpora; and 340 corpora are accumulated distributed among 180 institute and companies.
- 3 corpora are newly added, and 34 corpora have been distributed among 7 institutes and companies in 2012.



Types of the corpora



Resource of the corpora

Activities in Beijing Jiaotong University

- **Chinese Broadcast News Speech Corpus with Emotional Expressions**
  - Completed with Communication University of China
  - National News and Newspapers Summary (program in Central People's Radio), by 1 female professional announcer in 600 chapters
  - Annotated with hierarchal prosodies, stress, intonation and emotion

Activities in Center for Speech and Language Technologies, Tsinghua University

Time-Varying Voiceprint Corpus

60 university students (30 males and 30 females)
16 sessions of gradient time intervals from Jan 2010 to Dec 2012
100 fixed Chinese sentences per session
Speaking in standard Chinese in a reading style

Activities in the Broadcast Media Language Research Center, Communication University of China

- **Broadcast Media Language Monitoring Corpus**
  - Radio audio, TV video and the transcription texts labeled with meta-data
  - 700 million characters text data and more than 14000h audio & video data
- **Mandarin Chinese Digital Sampling Corpus**
  - 3543h speech in video or audio files generated by authoritative speakers with metadata tagging since 1933
- **Sentence-matched Broadcasting Speech Corpus**
  - Over 300h speech data from the TV or radio news programs, transcribed by text alignment and speech recognition
- **Tools**
  - The broadcasting speech retrieval system
  - The dialect homophonic character generation system, V4.0

The Corpus of Northwest University for Nationalities

Tibetan

- Tibetan Morpheme Corpus：5,000 monosyllables
- Tibetan Sound Dictionary Corpus：90, 000 entries，1 speaker
- Oral Chinese Corpus of Tibetan Students：150 speakers，half -hour for each speaker
- Tibetan Sentence Recording Corpus：5,000 Tibetan sentences ,1 speaker
- Oral Tibetan Corpus：30 speakers, half-hour for each speaker
- Text Corpus : Tibetan newspaper texts for a year (50M) ; 18 Tibetan textbooks used in primary school and middle school ； web-based text resource corpus (1G)

Mongolian

- Mongolian Vocabulary Corpus：2,000-word corpus
- Mongolian Vocabulary Recording Corpus：1,500 words， 10 speakers，about 340 minutes in total
- Oral Mongolian Corpus： 2 speakers，2 hours for each speaker

Uyghur

- Uyghur Sound Dictionary Corpus：10,000 entries，1 speaker

Activities in the Institute of Linguistics, Chinese Academy of Social Sciences

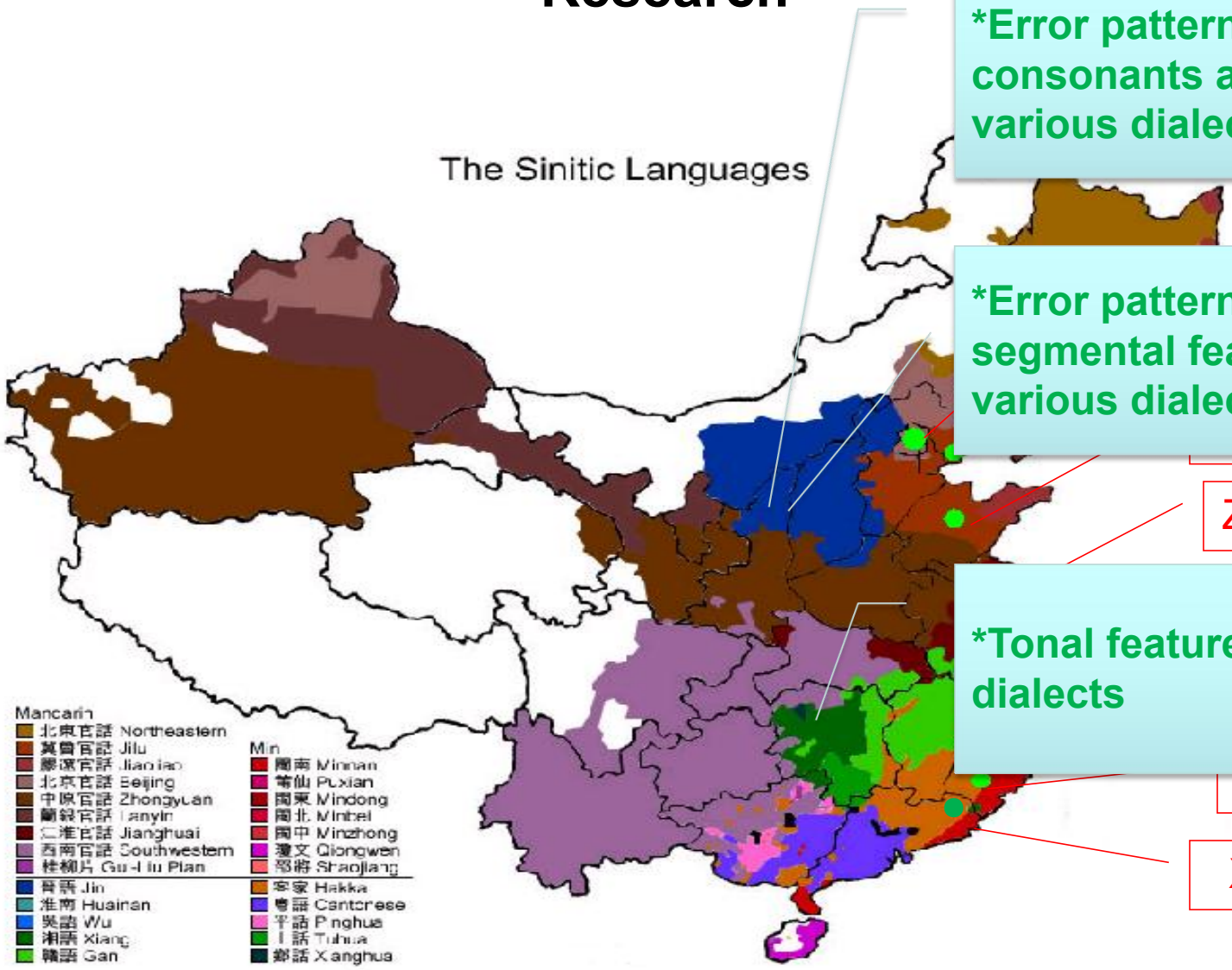
- **CHILD-CASS Corpus: Child language acquisition speech corpus**
  - Longitudinal child spontaneous speech: 23 children, recorded in half ~ 1month. The audio files are transcribed orthographically and annotated phonetically in conjunction with linguistic and paralinguistic information.
  - **Elicited recordings**

    The Question-Answer (Q-A) dialogues between parents and children, and children's production of words.
  - **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 CHILD-CASS Corpus

Progress of AESOP-CASS Corpus and Phonetic Research



*Error patterns of consonants and vowels in various dialectal regions

*Error patterns of supra-segmental features in various dialectal regions

Zhenjiang **50**

*Tonal features of various dialects

Fuzhou **20**

Xiamen **30**