A Country Report – COCOSDA Activities in China

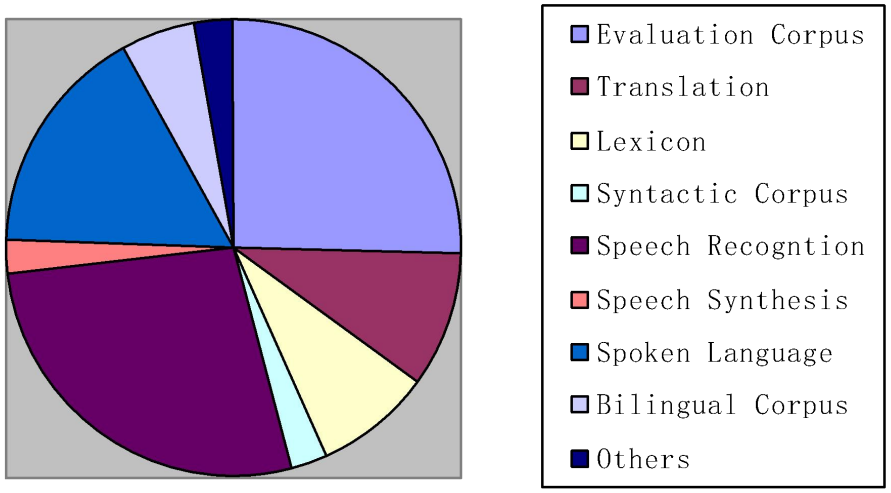
O-COCOSDA 2011

Hsinchu, Taiwan

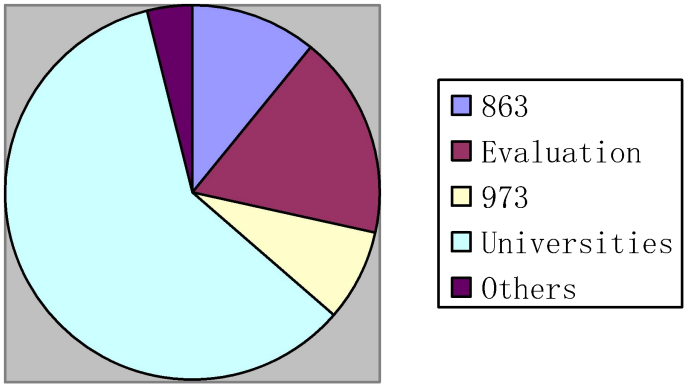
Aijun Li and Thomas Fang ZHENG

Activities in Chinese LDC

- Till now, there are 91 different corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpus.
- In 2011, there is no new corpus added, however, 32 corpora have been distributed to 26 institutes and companies within this year.

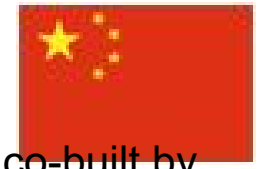


Types of the corpora



Resource of the corpora

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



- **CASIA-CASSIL Corpus(final release)**
  - 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*
- **CASS-CHILD Corpus: Child language acquisition speech corpus**
  - Longitudinal child speech database: 20 children, recorded in half ~ 1month
  - Spontaneous child speech database: 100 mother and child dyads.
  - Transcription, Phonetic annotation and POS labeling
- **AESOP-CASS Corpus**
  - English materials and Chinese PHT and dialectic materials (shared with AESOP group)
  - Regions and speakers:
    - Beijing, 10 speakers;Shandong area, 90 speakers;Zhenjiang 50 speakers;Ningbo 10 speakers;Hangzhou 10 speakers;Tianjin 5 speakers;Recording: 10-15 hours of each

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

- CASS Elemental Speech Corpus(CASS-ESS)
  - Part1: Read,60 speakers (30 F/ 30 M), 8 hours per spk.
  - Part2: incl. materials for focus, intonation and speaking styles research (each inld. hundreds of sentences)

Recording materials for each speaker(Part-I)	Number
syllable	500
Retroflexed words	40
3 syllabic word or phrase	90
4 syllabic word or phrase	100
2 syllabic word or phrase with final neutral tone	40
3 syllabic word or phrase with middle neutral tone	40
syllable (with tone info / toneless)	500
Retroflexed words	40
3-syllabic words or phrase	90
Initials/finals	21+38
Simple sentence	140
Complicated sentence	50
Phonetic balance sentence	200

Activities in the Institute of Ethnology and Anthropology , Chinese Academy of Social Sciences

- Acoustic parameter database were setup for three languages.

Summary results of corpus of three languages

Language / Item	Daur	Evenki	Oroqen
Monosyllable	414	44	34
Dissyllable	669	473	525
Multi-syllable	366	554	487
Phrase	300	110	87
Sentence	223	200	200

Activities in Chinese University of Nationality

- More than 100 languages are spoken by 55 Chinese minority ethnic groups, belonging to Sino-Tibetan language family, Altai Language Family, Austroasiatic language family, Austronesian language family, and Indo-European Language Family. Korean language's dependency is still indeterminate. In order to preserve those language data accurately, the *Recordings of Minority Languages(*《中国少数民族语言音系录像》和《中国少数民族语言词汇录音》*) in China* was started since 1992。
  - Spoken by 42 Chinese minority groups,
  - *Audio Recordings* of vocabulary of 62 languages spoken by 45 Chinese minority groups.
  - Speakers in the audio and visual records are local peoples born and brought up in their minority communities.
  - The Recordings is transcribed with the International Phonetic Alphabet (IPA)
  - Two versions: old/new versions(1992, 2007).The original version was partially sponsored by An Zijie Fund, Dai Qingxia was Leading Editor; while the new version is the result of CUN-undertaking sub-project of National "985 Project", Liu Yan & Li Dejun are leading Editors.