A Country Report – COCOSDA Activities in China

O-COCOSDA 2015，Shanghai, China

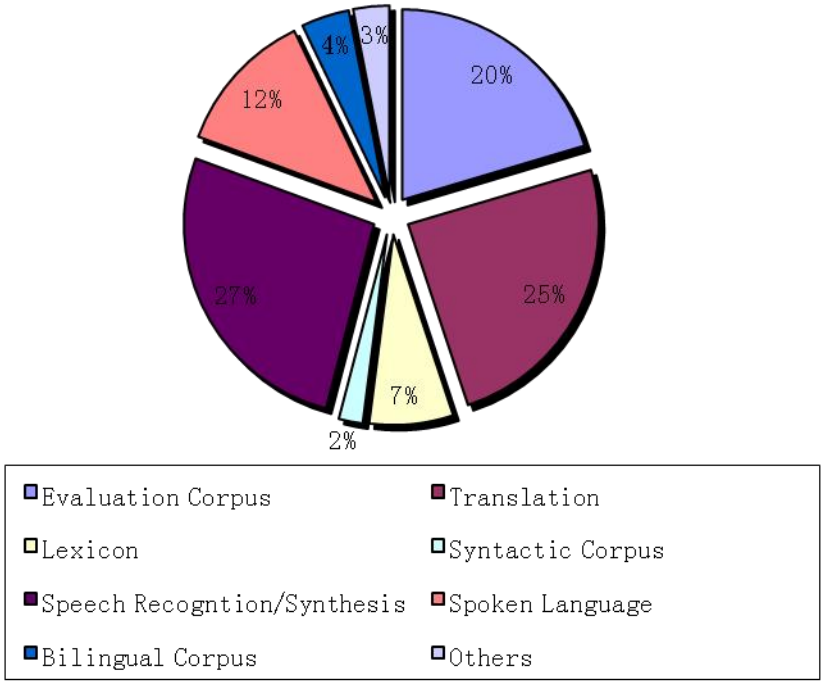
Aijun LI , *Dong WANG

Institute of Linguistics, Chinese Academy of Social Sciences

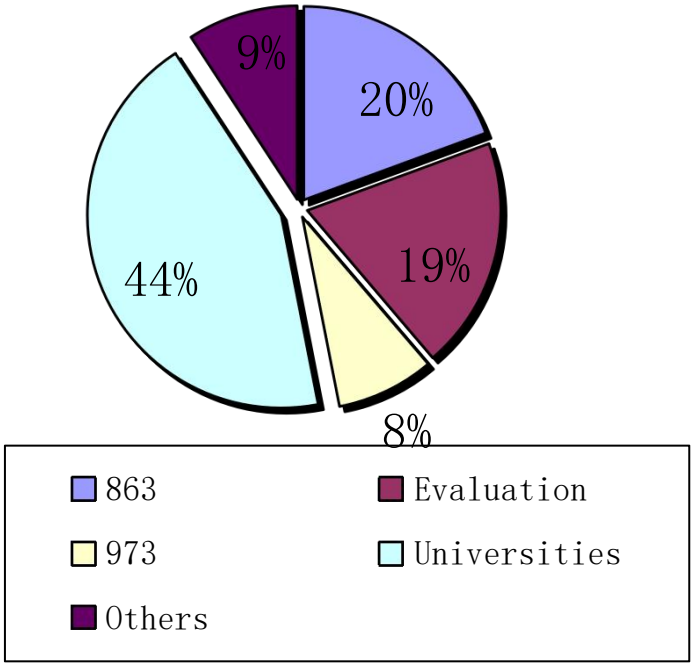
*Research Institute of Information Technology, Tsinghua University

1 Activities in Chinese LDC

- Till now, there are 98 corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpora.
- In 2015, there are 1 new corpora added;37 corpora have been distributed to 10 institutes and companies.



Types of the corpora



Providers of the corpora

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

- **Progress on AESOP-CASS**
  - Types of accents errors of English learners from Dalian and Shenyang
  - Yes-no question with focus from learners of Ningbo
  - Yes-no question with focus from learners of Dalian and Shenyang
  - Segmental errors from learners of Changsha
- **Progress on Discourse-CASS**
  - More speech data collected through online APP and CALL-center
  - Annotation on discourse level
    - Information structure/Rhetoric structure /Speech act (topics, adjacent pairs)/Referential structure/Dependency relation

- ***Jiuzhou* linguistic resources on Chinese dialects (九州语言网）**

Aiming at dialect resources collecting, sharing and analysis, we developed an online APP and a WeChat APP, which have the functions including:

- Phonetic transcription for dialect recording
- xRecorder: a corpus recording, anlysis and management tools
- Word pronunciation presented in both ancient and present times
- Grammar materials for 14 dialects
- syllable materials and their sounds of 70 speakers from Beijing dialect.
- http://9zhou.phonetics.org.cn

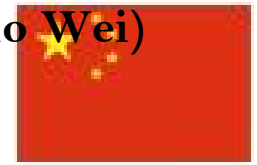
3 Activiites in CSLT@Tsinghua University

- Free Chinese speech database THCHS-30 (http://cslt.org/resources.php?Public%20data)
  - 60 spks, 30h reading speech by microphone
  - Associated lexicon and LM
  - DNN ASR baseline results and receipe
  - Totally free, public downloadable
- Free Uyghur speech database THUYG-20 (http://cslt.org/resources.php?Public%20data)
  - 360 spks, 20 h reading speech by microphone
  - Associated lexicon and LM
  - DNN ASR baseline results and receipe
  - I-vector SID baseline results and receipe
  - Totally free, public downloadable
  - Call for challenges
- Open source tools (http://cslt.org/resources.php?Public%20tools)
  - JCLTS: a tool for sequential symbol segmentation. Used for letter to sound prediction.
  - THUYG-LM: a tool for cleaning Uyghur text data and building morpheme LM.

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

- Speech corpora
  - Broadcast Media Language Corpus
    - 15000h audio & video recordings of TV and Radio programs
    - Over 1500h Sentence-matched speech data from TV or radio news programs
  - Mandarin Chinese Digital Sampling Corpus (MCDSC)
    - 3486h video or audio files with metadata tagging since 1932
- Text corpora
  - An Online TV Programs Transcription Corpus (http://ling.cuc.edu.cn/RawPub/)
  - 70 year *people's daily* corpus
  - Annual newspaper corpus

5. Activiites in CCA@Tianjin University (Jianguo Wei)



- Ultrasound based articulatory database
  - Recorded by Tarason T3000 system that Synchronization of ultrasound image, high-speed video and audio signals at 60Hz
  - 5 spks(2 females and 3 males), about 6h reading speech by microphone
  - Corpus contains 1000 sentences with 28721 monophones, 14482 and 26721 triphones

- Multi-modal speech database for Tibetan and Chinese
  - Speaker： 20 spks (10 males, 10 females)
  - Corpus 30 Tibetan letters and 4 vowels and 25 monophones
    Sentences consist of 41 Tibetan sentences and 27 Chinese sentences
  - Audio + EMA + Ultrasound + HD Camera

- Chinese MRI database with speech
  - five female speakers and two male speakers of Chinese (Dialect is not controlled.)
  - MRI scanner: 3T MRI (Siemens Verio, at BAIC, ATR-Promotions, Kyoto)