A Country Report - COCOSDA Activities in China

O-COCOSDA'2007 Honia

Outline

- **Activities**
  - **CCC (Chinese Corpus Consortium)**
  - **ChineseLDC**
- **Some Newly Developed Language Resources**

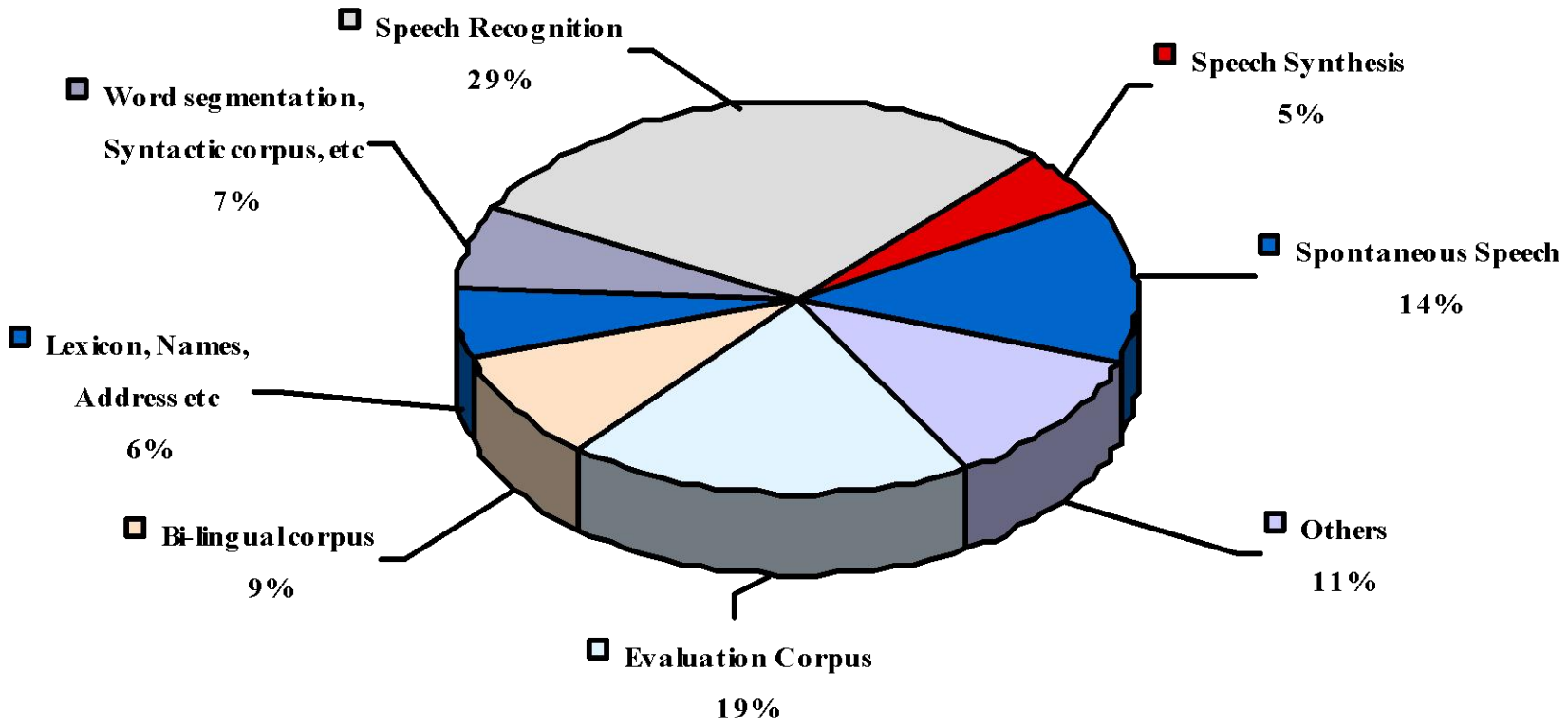
Activities - CCC

- In 2006, CCC successfully organized a special session on speaker recognition of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP 2006).
- It developed a speaker recognition evaluation (SRE) to act as a platform for developers in this field to evaluate their speaker recognition systems using two databases provided by the CCC for free.
- Eight research sites participated in the CCC 2006 SRE, which was more than expected. CCC is now preparing the 2nd SRE in ISCSLP 2008 to be held in YunNan,China.

Activities- ChineseLDC

- Work on the construction, evaluation and distribution of phonetic and linguistic resources has been done by the Committee for the Construction and Management of Linguistic Resources, Chinese Information Processing Society of China (CIPSC).
- After more than two year's construction, CLDC has accumulated a variety of precious linguistic data with more than 100 types.

Activities- ChineseLDC



Distribution of the Lang. Corpora in CLDC

Activities- ChineseLDC

- Three working groups have been established
  - linguistic resources
  - phonetic resources
  - evaluation
- Evaluation work done by the Evaluation Working Group in the name of CIPSC
  - machine translation system
  - Chinese word segmentation,
  - named entity recognition
  - part-of-speech tagging
  - 43 participates, 28 returned results

Some Newly Developed Language Resources – according to affiliations

Institute of Linguistics, CASS

- (1) An Expressive Speech Corpus of Standard Chinese, ESCSC,phonetic annotations. 100 hours
- (2) Accented Mandarin Chinese Speech Corpus for three Tibetan dialectal Areas,100 speakers, 400 hours. phonetic annotations.
- (3) Multi-lingual and multi- accented speech corpus for Min,Yue and Beijing (Putonghua) regions, 60 speakers, 120hours. phontetic annotations

(A)Institute of Linguistics, CASS

- (4) Articulatory databases: EMA (AG500 system): 20 speakers are being recorded with the speaker's tongue (three points), jaw, lower lip and upper lip kinematics.
- (5) the EPG database:4 speakers were recorded wearing customized artificial palates.

  The speech material is composed of segments, tonal syllables, phonetically balanced disyllabic words, phonetically balanced sentences, paragraph and spontaneous speech.
- (6) Tel-corpus with 5 domains: hotel, restaurant, travel agency, airport service, taxi service, 2000 dialogues for each domain

(B)Chinese Dept. of Peking Univ

- (1) the Speech Therapy and Physiological Research Corpora, including the Tone Acquisition and Voice Corpus of Hearing-Impaired Children, and the Speech and Vocal Tract video of Standard Chinese.
- (2) the Speech Corpora of Chinese Ethnic Minorities, including the Ethnical Linguistic Corpus, the Corpus of Tones in Kam-Tai Languages, the Prosody Corpus of Tibetan Anduo Dialect, the Voice Corpus of Ethnic Singing Methods:Praying in Mi Zong of Tibetan Buddism, Humai (Traditional Mongolian Throat Singing), Japanese No Drama and some Ethical Musical Instruments.

(C)Institute of Automation, CAS

- (1) the CASIA Corpora of Continuous Telephone Speech (spontaneous dialogues on given topics through telephone, 514 speakers, 257 dialogues, around 120 hours, 10 minutes each speaker).
- (2) the CASIA Speech Corpora of English Learners of Various Levels (mainly for the evaluation of pronunciation quality, 500 speakers classified into 5 level groups, 1.5 minutes each speaker)

(D) TsingHua University

- Dialectal Chinese database – Wu (Shanghai), Min (Xiamen) and Chuan (Chengdu), the speakers are 100, 36 and 36 respectively. 200 long sentences, 200 phrases, 10 digits, 26 English letters per speaker, transcribed in Chinese Character, syllable and Initial-Final layers.

(E) The Institute of Ethnology and Anthropology, CASS

- ‘Acoustic and phonetic parameter database for Chinese Minority languages ‘. Several sub-corpora have been finished including Tibetan, Kazakstan, and Mongolian. Parameters:F0, F1-F3, VOT
- EPG database: 2 speakers, words and sentences

(F) CASS of Inner Mongolian

- Funded by the government of Inner Mongolian autonomous region of China, Chinese Academy of Social Sciences of Inner Mongolian is undertaking a 15 year project to produce a comprehensive and huge corpus including spoken, writing and literature of Mongolian, as well as culture and anthropology related A/V materials in almost every counties and villages.
- Total 80 regions in inner Mongolian, Mongolian, Russia and other Chinese regions.