O-COCOSD Activities in China

Aijun LI
Linguistic Institute of Chinese Academy of Social Sciences, Beijing, China 100732
liaj@cass.org.cn

Fang ZHENG
Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University, Beijing, 100084, China
fzheng@tsinghua.edu.cn

There are many organizations which are dedicated to the development of CSLP corpora. Chinese Linguistic Data Consortium (CLDC) and Chinese Corpus Consortium (CCC) are two of the well known ones in China.

The goal of the establishment of CLDC is to set up a general linguistic database of Chinese Linguistic Data, which embodies the Chinese linguistic database currently in the lead internationally. After more than two year's construction, CLDC has accumulated a variety of precious linguistic data with more than 100 types.

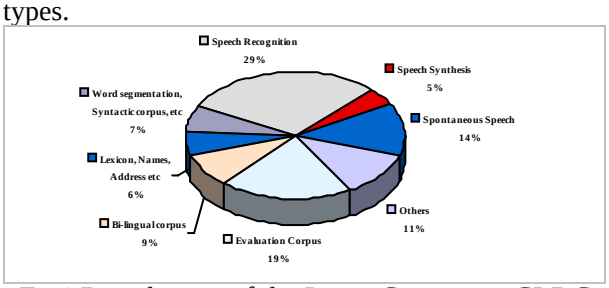

Fig1.Distribution of the Lang. Corpora in CLDC

In the past one year, much 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). Three working groups, responsible for linguistic resources, phonetic resources and evaluation respectively have been established. These groups have done much work, especially the evaluation of machine translation system and the evaluation of Chinese word segmentation, named entity recognition and part-of-speech tagging by the Evaluation Working Group in the name of CIPSC. Thanks to all these efforts, more and more data is available for research and application.(see Fig. 1)

In 2006, CCC (http://www.CCCForum.org) 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 China.

Some Newly Developed Language Resources

Many research institutions continued to collect various linguistic data supported by national funds as well as non-government organizations.

(A) Institute of Linguistics, CASS

An Expressive Speech Corpus of Standard Chinese, ESCSC (see in this conf.), phonetic annotations. (2) Accented Mandarin Chinese Speech Corpus for three Tibetan dialectal Areas,100 speakers, 400hours. phontetic annotations. (3) Multi-lingual and multi-accented speech corpus for Min,Yue and SC regions, 60 speakers, 120hours. phontetic annotations (4) Articulatory databases: EMA (AG500 system) and the EPG database. 20 speakers are being recorded with the speaker's tongue (three points), jaw, lower lip and upper lip kinematics. Four 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.

(B) Chinese Department 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 vedio 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

They are undertaking a NSF project called 'Acoustic and phonetic parameter database for Chinese Minority languages '. Several sub-corpora have been finished including Tibetan, Kazakstan, and Mongolian.

(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.