A Country Report – COCOSDA Activities in China

O-COCOSDA 2014， Phuket, Thailand

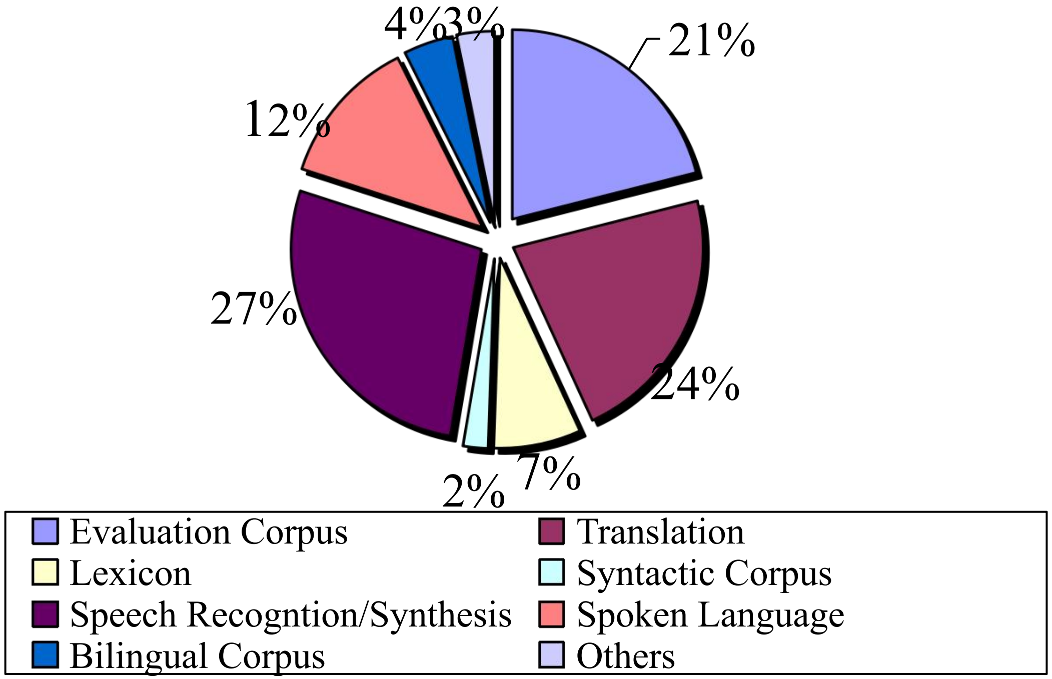
Aijun LI , *Dong WANG

Institute of Linguistics, Chinese Academy of Social Sciences

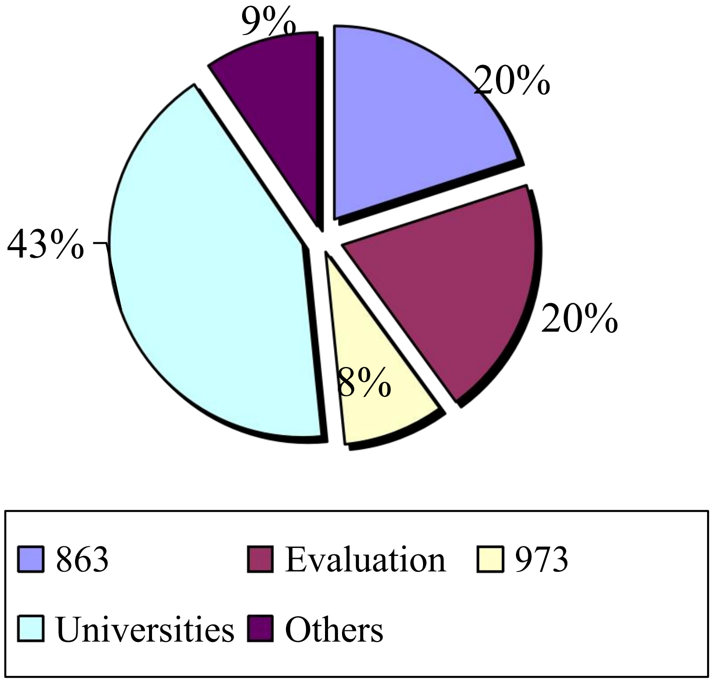
*Research Institute of Information Technology, Tsinghua University

1 Activities in Chinese LDC

- Till now, there are 97 corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpora.
- In 2014, there are 2 new corpora added;14 corpora have been distributed to 8 institutes and companies.



Types of the corpora



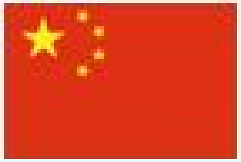
Providers of the corpora

2 Activities in Institute of Linguistics, Chinese Academy of Social Sciences-1

- **Phonetic Research on AESOP-CASS**
  - Question of learners from Northern Guan Dialect
  - Yes-no question of learners from Beijing
  - Yes-no question of learners from Wu dialect
  - Segmental errors of English learners from Shandong dialect

- **3D Articulatory speech database**
  - spoken by a female professional speaker
  - 3D articulatory data are recorded by NDI Waves
  - 2206 syllables,11204 words and 222 sentences in Standard Chinese
  - 3D IPA training database for 4 phoneticians and modeling

[redacted]

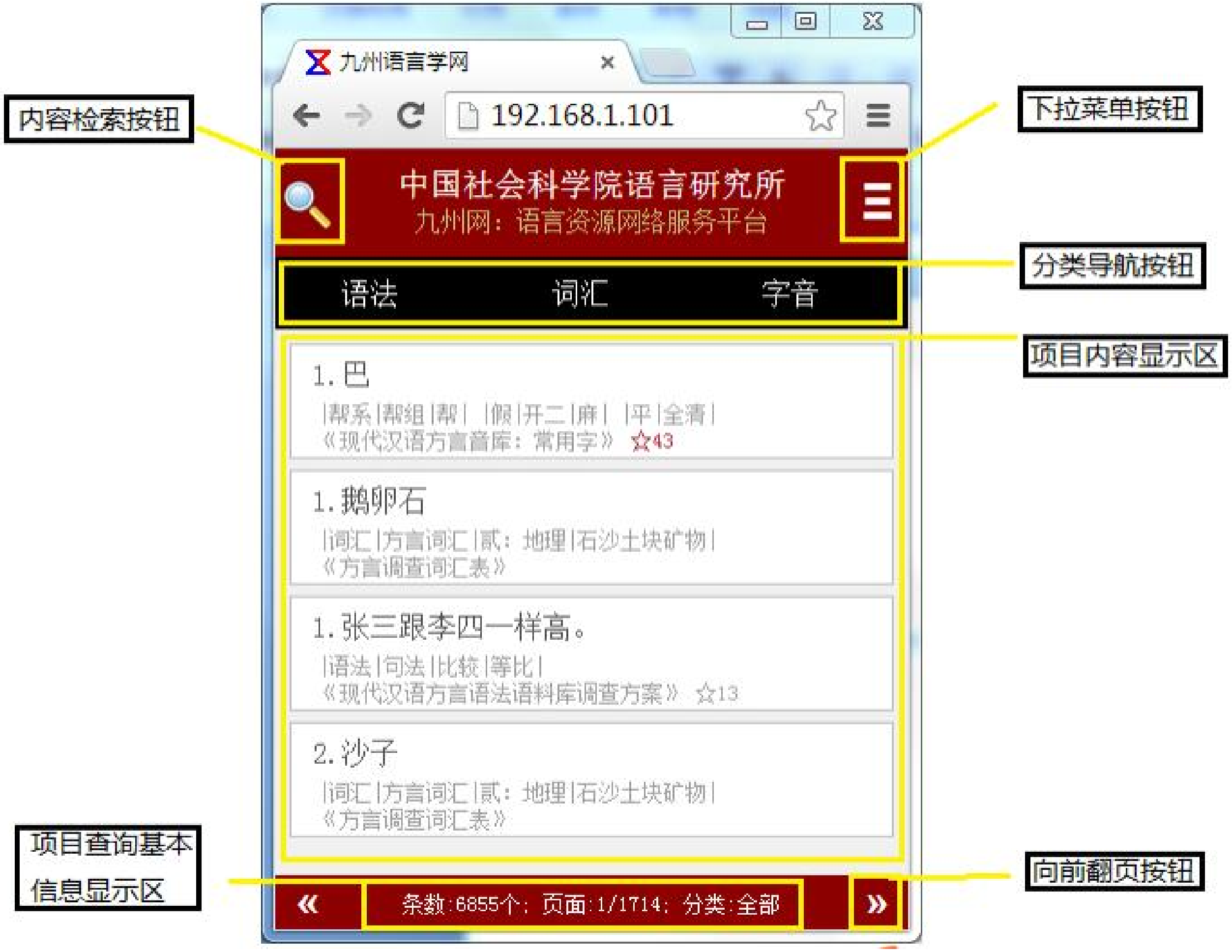


□ n of

hildren's
r each child
All speech
road phonetic

✓

✓ er aged between
: 2- or 3-month,



九州语言学网 ×

192.168.1.101

中国社会科学院语言研究所
九州网：语言资源网络服务平台

语法 词汇 字音

1. 巴
|帮系|帮组|帮| |假|开二|麻| |平|全清|
《现代汉语方言音库：常用字》 ☆43

1. 鹅卵石
|词汇|方言词汇|贰：地理|石沙土块矿物|
《方言调查词汇表》

1. 张三跟李四一样高。
|语法|句法|比较|等比|
《现代汉语方言语法语料库调查方案》 ☆13

2. 沙子
|词汇|方言词汇|贰：地理|石沙土块矿物|
《方言调查词汇表》

« 条数:6855个; 页面:1/1714; 分类:全部 »

□ **Jiuzhou linguistic resources (九州语言网）,** http://9zhou.phonetics.org.cn/.

✓ an open platform sharing resources for linguistics, phonetics and dialect research.

✓ Chinese dialect data topologically organized by pronunciation, lexicon/morphology and syntax.

3 Activities in Broadcast Media Language Research Center, Communication University of China (after Dr. HE Wei)

- **Speech corpora**
  - Broadcast Media Language Corpus
    - More than 15000h audio & video recordings of TV and Radio programs
    - Over 1000h Sentence-matched speech data from TV or radio news programs
  - Mandarin Chinese Digital Sampling Corpus (MCDSC)
    - 3543h video or audio files with metadata tagging since 1933
  - An Online TV Programs Transcription Corpus (http://ling.cuc.edu.cn/RawPub/)
- **Tools**
  - The parallel corpus retrieval tool (http://ling.cuc.edu.cn/chs/download/CUC_ParaConc_V0.3.rar)
  - The dialect homophonic character generation tool TYZH_v0.4.rar (http://ling.cuc.edu.cn/chs/download/TYZH_v0.4.rar)

4 Activities in Linguistics Lab, Department of Chinese Language and Literature, Peking University （Jiangping Kong）

- **Speech Multimodal corpora**
  - X-ray vocal tract database of Mandarin
  - 2D MRI static and 3D MRI vocal tract of Mandarin
  - 2D high-speech imaging database of dynamic glottis
  - Breathing band rhythm database of Mandarin
  - electropalatography of Mandarin Tibetan and Mongolian
- **Oral culture Multimodal corpora**
  - Liao songs, Epic of Buluotuo and wizard religious rituals of Zhuang minority
  - King Pan grand song of Yao minority
  - Grand song of Dong minority
  - Bimo scriptures and Suni wizard religious rituals of Yi minority
  - Dongba scriptures and folk-song of Naxi minority
  - Tibetan Buddhism recite Buddhist
  - Humai and long tuned of Mongolian nationality

5 Activities in Institute of Ethnology & Anthropology, Chinese Academy of Social Sciences

Built Acoustic Parameter Database

- Acoustic Parameter Databases of 7 languages (multi-syllabic, reading): Mongolia, Tu, East YuGu, Uighur, Hasak, Ewenki, Xibo. The databases cover most acoustic parameters in time and frequency domains.

Developed Auto-Annotation & Auto-Extraction Software

- 8 levels annotation file format
- Auto-Annotation Software: version 1.0 – 3.3
- Auto-Extraction Software: version 1.0 – 3.9

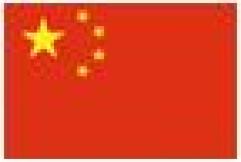
Developed Acoustic Parameter Management Software

- implement acoustic parameter inquiry & retrieval

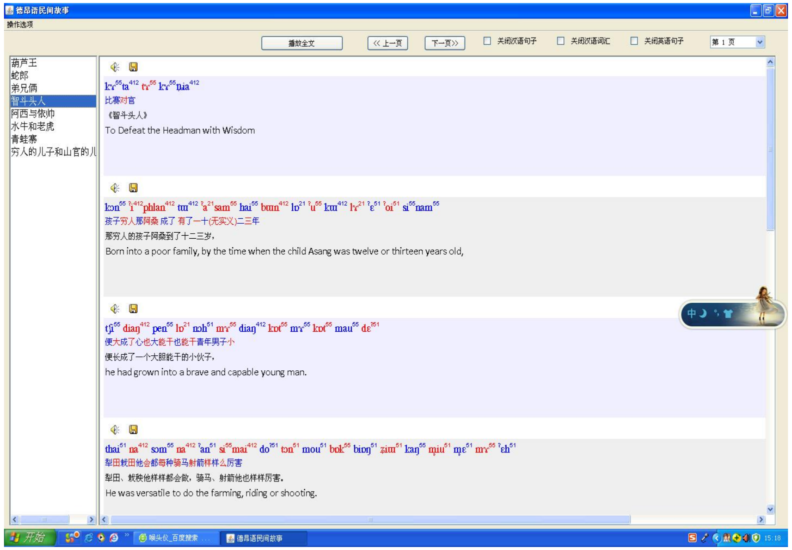
Published Related Research Papers

- 20 papers in academic journals
- 39 papers in academic conferences
- 7 thesis papers

6 Activities in Minority Languages and literature College of the Minzu University of China



- Publishing a “Series of Endangered Chinese Minority Languages Folklore Audio Book”
- This Series audio book are part of the Speech Corpus of Endangered Chinese Minority Languages, which are established by Minority Languages and literature College of the Minzu University of China, and which are financially supported by Chinese government.
- “Tujia Language Folklore” and “De’ang language Folklore” were published in 2013 and 2014.
-
- CD in the book
    - to retrieve the Folklore



葫芦王
蛇郎
弟兄俩
智斗头人
阿西与依帅
水牛和老虎
青蛙寨
穷人的儿子和山官的儿

kɤ^{55}ta^{412} tɤ^{55} kɤ^{55}ȵia^{412}
比赛对官
《智斗头人》
To Defeat the Headman with Wisdom

kɔn^{55} ^{ʔ}i^{ʔ412}phlan^{412} tɯ^{412} ^{ʔ}a^{21}sam^{55} hai^{55} bɯn^{412} lɒ^{21} ^{ʔ}u^{55} kɯ^{412} lɤ^{21} ^{ʔ}ɛ^{51} ^{ʔ}oi^{51} si^{55}nam^{55}
孩子穷人那阿桑 成了 有了一十(无实义)二三年
那穷人的孩子阿桑到了十二三岁，
Born into a poor family, by the time when the child Asang was twelve or thirteen years old,

tʃi^{55} diaŋ^{412} pen^{55} lɒ^{21} nɔh^{51} mɤ^{55} diaŋ^{412} kɒt^{55} mɤ^{55} kɒt^{55} mau^{55} dɛ^{ʔ51}
便大成了心也大能干也能干青年男子小
便长成了一个大胆能干的小伙子，
he had grown into a brave and capable young man.

thai^{51} na^{412} sɔm^{55} na^{412} ^{ʔ}an^{51} si^{55}mai^{412} do^{ʔ51} tɔn^{51} mou^{51} bɒk^{55} biɒŋ^{51} ʑiɯ^{51} kaŋ^{55} m̥iu^{51} m̥ɛ^{51} mɤ^{55} ^{ʔ}ɛh^{51}
犁田栽田他会都每种骑马射箭样样么厉害
犁田、栽秧他样样都会做，骑马、射箭他也样样厉害。
He was versatile to do the farming, riding or shooting.