A Country Report – OCOCOSDA Activities in China

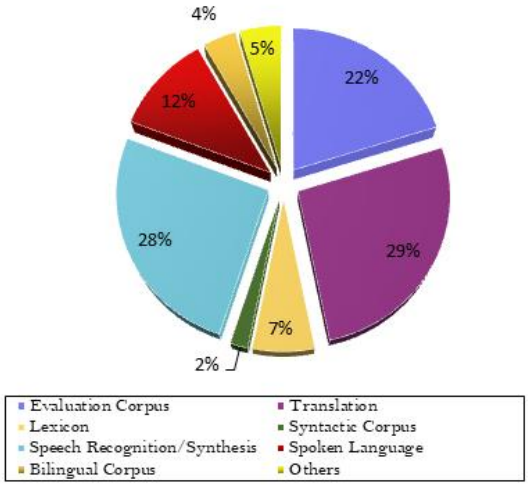
Aijun LI , *Dong WANG
Institute of Linguistics, Chinese Academy of Social Sciences
***Research Institute of Information Technology, Tsinghua University**

O-COCOSDA 2019， Cebu City, Philippines

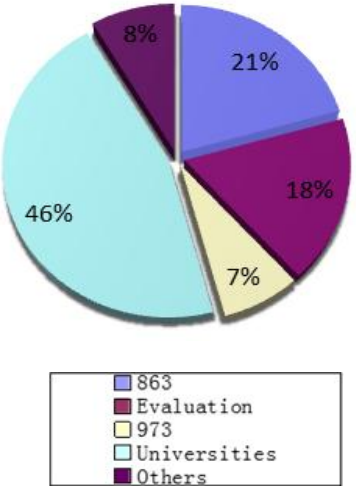
October 25-27, 2019

Chinese LDC

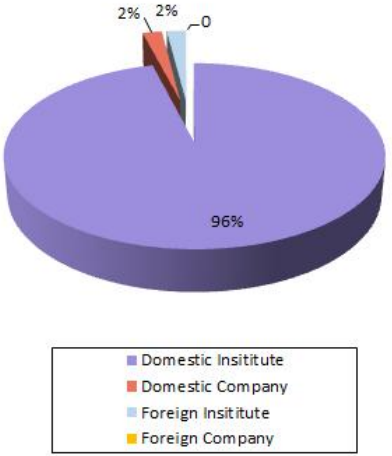
- Till now, there are 109 corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpora.
- In 2018, there are 2 new corpora added (1 Bilingual corpus and 1 other);45 corpora (3 spoken language, 4 speech recognition/synthesis, 5 Bilingual, 2 lexicon, 2 Syntactic,18 evaluation and 11 other corpus) have been distributed to 18 institutes and companies.



Types of the corpora



Providers of the corpora



Types of the users

Academic Activities

北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

Project for the Protection of Language Resources of China

- Current Data
  - Chinese Dialects: 33 provinces and cities, 1122 sites, 7383 speakers
  - Minority Languages: 8 language families, 329 sites, 861 speakers
  - Audio resource items: > 5.66 million
  - Video resource items: > 3.88 million
  - Original Data Volume: > 45TB
- Platform Provider: DCST of Tsinghua University
- URL: https://zhongguoyuyan.cn

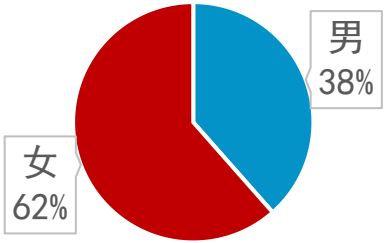
WeChat official account:
jxh-wx

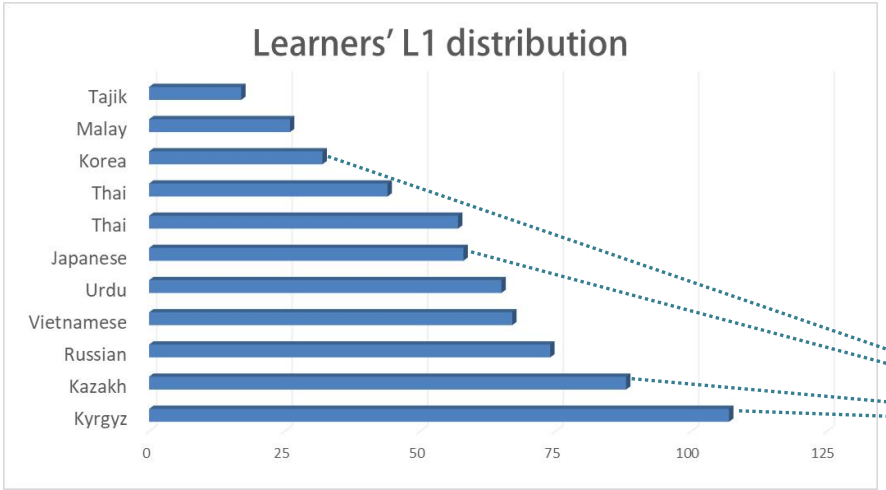
BLCU-SAIT Chinese Non-native Corpus

Totally 772 speakers： 280 Hours

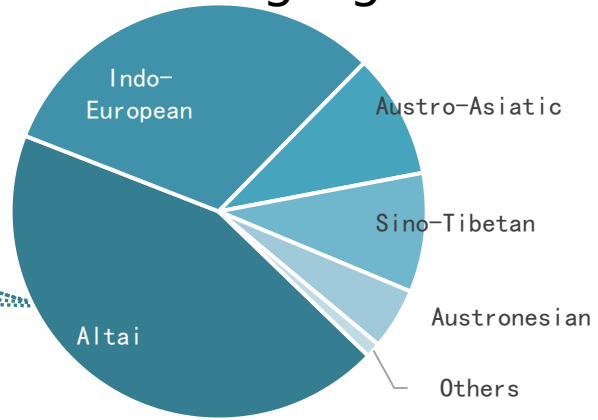
- Chinese Learners：691 speakers，from 32 L1 languages
- Chinese Native speaker：81人

Gender Distribution





Language Distribution



Commercial Activities

speechocean 海天瑞声 www.speechocean.com

希尔贝壳 AISHELL http://www.aishelltech.com

Languae in China	Speakers: 73,100 Hours: 69,000	Mandarine/Accented Mandarine, Cantonese, Hokkien, Sichuanese, Shanghainese, Taiwanese, Tibetan, Uyghur
English	Speakers: 27,500 Hours: 32,000	US English, UK English Other Accented English
Other Major Languages	Speakers: 68,300 Hours: 59,400	Covering 43 Countries & Regions, including Accented English, Spain, Russia, French, German and etc.
Low Resource Languages (Minority Languages)	Speakers: 19,800 Hours: 18,300	20 Languages, including: Urdu,Catalan,Swedish,Ukrainian,Polish,North Korean, Greek,Danish,Finnish,Filipino,Romanian,Turkish,Arabic, Hindi,Tagalog,Tamil,Gujarati,Vietnamese
TTS Speech Corpus	Hours: 600	36 Languages, including Chinese, English, Arabic, French, Spain, German and etc.
Lexicon	72 Languages,10 Million Entries. Including most of the major & minority language mentioned above.	
Text Corpus	64 Languages, including most of the major languages mentioned above.	

- **Smart Home**

  1000 speakers, 2000 Hours

- **Autonomous Driving**

  2300 speakers, 10750 Hours

- **Mandarin Corpus**

  8000 speakers, 4500 Hours

- **Chinese Children Corpus**

  1200 speakers, 180 Hours

- **US English**

  2600 speakers, 800 Hours

- **Free Corpus**

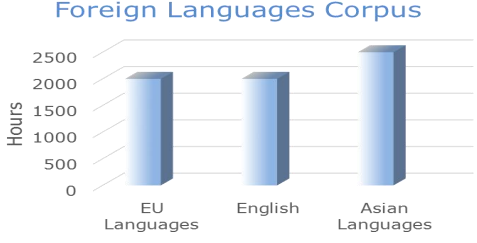
  AISHELL-1:400 speakers, 170 Hours

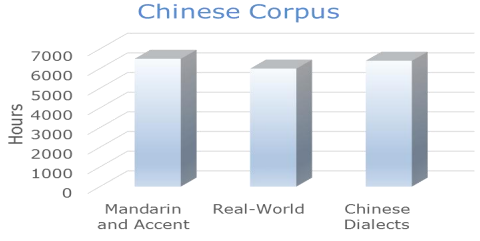
  AISHELL-2:1991 speakers, 1000 Hours

Commercial Activities

DATATANG

www.datatang.com





Others Corpus in Different Scenarios		
	Smart Home	In-Car
	Customer Services	E-Commerce
	Multi-Array Microphone	Noisy Environment
	Voice Assistant	CN-EN Mixed

慧听数据 HUITING DATA

http://www.huitingtech.com/

China **Mandarin** **Accented Mandarin** **Children Mandarin** **Elderly Mandarin** **Uygur / Tibetan** **Cantonese** **Mandarin English Mixed** **Cantonese English Mixed**	**29000 Speakers** **20000 Hours**
English **UK English** **US English** **Other Accented English**	**2000 Speakers** **1800 Hours**
Other Languages **French, German, Italian, Spanish, Mexican Spanish, Brazilian Portuguese, Japanese, Korean**	**3900 Speakers** **2100 Hours**

Free Database					
Primewords	Mandarin Speech, 1600h	**DataTang**	Mandarin Speech, 1050h	**AIShell**	Mandarin Speech, 1000h
MAGICDATA	Mandarin Speech, 755h	**Baidu**	Mandarin Speech, 50h Mandarin Text, 2.3M	**ReactiveCJ**	Madarin Text, 16G