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 2018， Miyazaki, Japan

Commercial Activities

www.speechocean.com

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**

640 speakers, 1500 Hours

2800 speakers, 1600 Hours

- **Mandarin Corpus**

2600 speakers, 1200 Hours

1000 speakers, 180 Hours

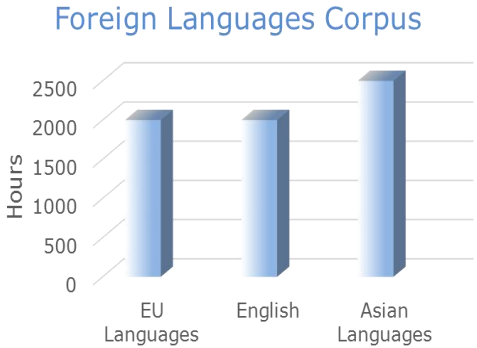
AISHELL-1:400 speakers, 170 Hours

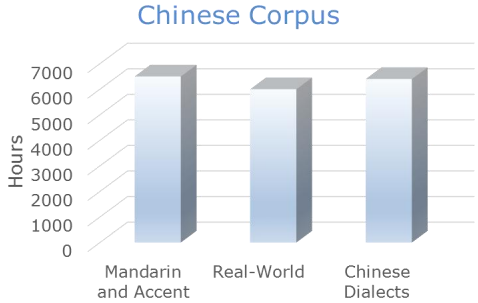
AISHELL-2:1991 speakers, 1000 Hours

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

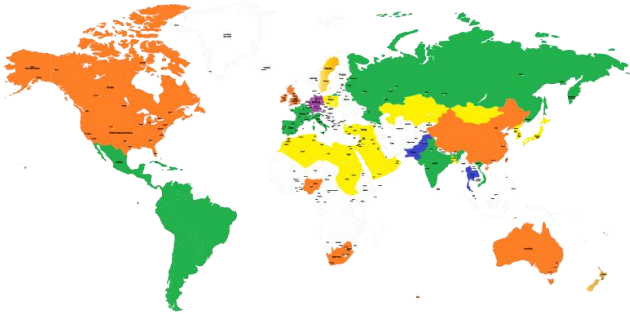
http://www.huitingtech.com/

China Mandarin Accented Mandarin Children Mandarin Elderly Mandarin Uygur / Tibetan Cantonese Mandarin English Mixed Cantonese English Mixed	18000 Speakers 11500 Hours
English UK English US English Other Accented English	1000 Speakers 700 Hours
Other Languages French, German, Italian, Spanish, Mexican Spanish, Brazilian Portuguese, Japanese	3100 Speakers 670 Hours

iFlytek Activities

- **Commercial activities**
  - Huge data collection for internal multi-lingual project (Speech-to-speech translation, NLP, ….);
  - Language coverage:

Language family	More than 30 languages
Indo-European	French, Spanish, Portugese, Italian, Russian, German, Hindi, Urdo, Turkish,…….
Asian	Vietnamese, Thai, Japanese, Korean, Malaysian,…..
Chinese Ethnic/dialects	Uyghur, Tibetan, Kazak, Mongolian, Xibe, Cantonese,Teochew, Minnan, Shanghainese,…..

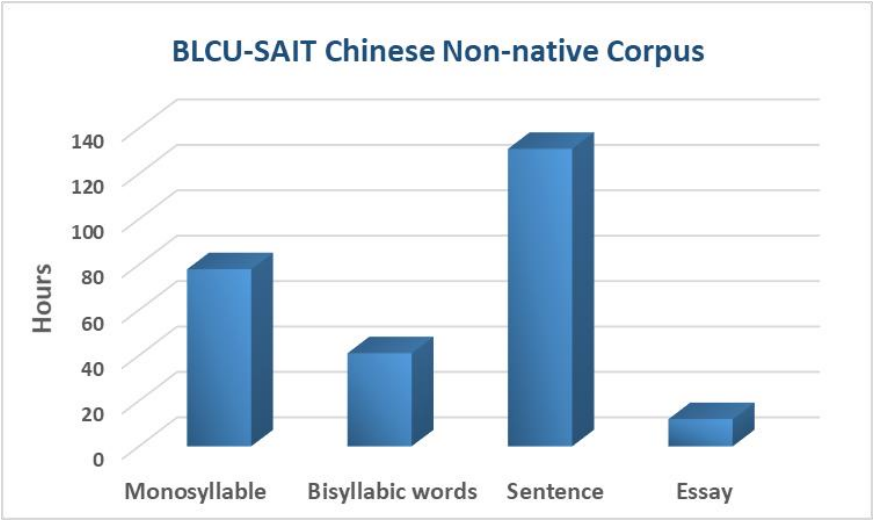


- **Academic activities**
  - Global approach for systems development: Global semantic unit, Global POS, Global phone & tone;
  - Speech replication – a speech documentary technology based on TTS for natural speech of endangered language/dialect;
    - ✓ Revealing phonetic structures, sound changes, syntax structure;
    - ✓ Translation between main languages and endangered language;
  - Data collection for language behavior and cognition study
    - ✓ Age distribution: children, teen-age, young, middle age, Senior citizen …;
    - ✓ Language distribution: minority nationality, dialects;

Academic Activities

Speech Acquisition and Intelligent Technology Lab

北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY



BLCU-SAIT Chinese Non-native Corpus

695 Speakers 243 Hours

Non-native Chinese Data 618 Speakers 243 Hours	Learners' L1 background : Indo-European : 199 Speakers, including English, Russian, Tajik and etc. Altai : 154 Speakers, including Kazakh, Kyrgyz, Turkmen and etc. Sino-Tibetan : 69 Speakers, including Thai, Burmese. Austro-Asiatic : 64 Speakers, including Vietnamese, Cambodian. Japanese : 60 Speakers. Korean : 33 Speakers. Austronesian : 32 Speakers, including Malay, Indonesian. Others : 7 Speakers, including Arabic, Swahili, Rwandan and etc.
Native Chinese Data	77 Speakers 12 Hours

Academic Activities

北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

Project for the Protection of Language Resources of China

- Huge language-culture project on the national level
- Government financed & directed
- Main contents
  - (1)investigation of Chinese dialects and minority languages
  - (2)concentration of the existing language resources
  - (3)development of the language collecting and recording platform
- 1500 sites (including hundreds of endangered languages and dialects) according to a set of unified rules between 2015 and 2019
- China Language Resources Database
- Centre for the Protection and Research of Language Resources of China in BLCU