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**

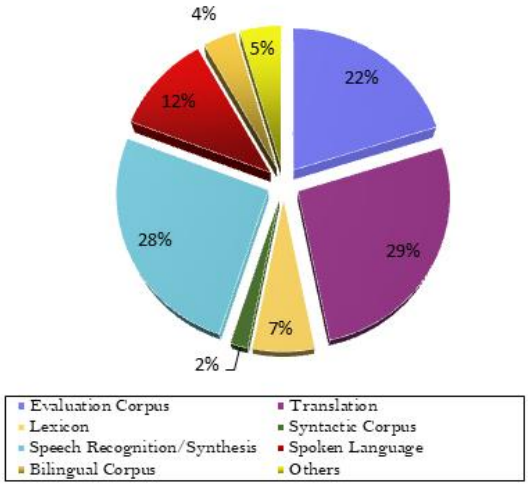
O-COCOSDA 2020， Yangon，Myanmar

Nov. 5-7, 2020

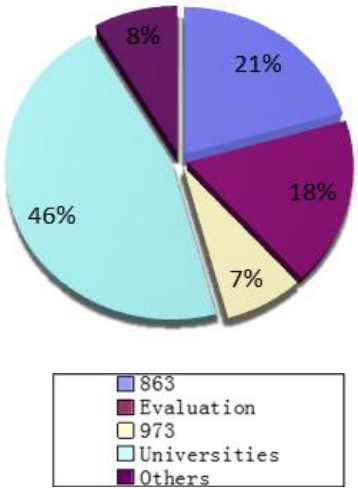
Chinese LDC

http://www.chineseldc.org/

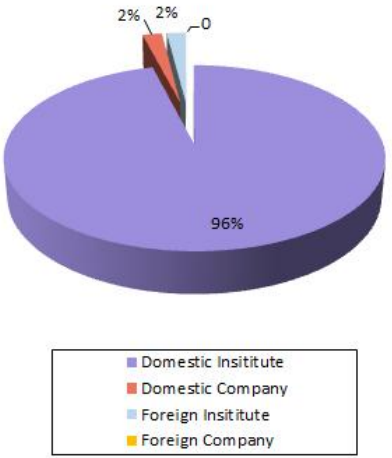
- Till now, there are 109 corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpora.
- 24 corpora (5 spoken language, 12 speech recognition/synthesis, 4 Bilingual, 1 lexicon, and 2 other corpus) have been distributed to 17 institutes and companies.



Types of the corpora



Providers of the corpora



Types of the users

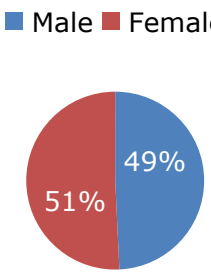
Spoken Chinese Mandarin Corpus of Cochlear-implanted Children

https://english.tongji.edu.cn/tjyy_10685/list.htm

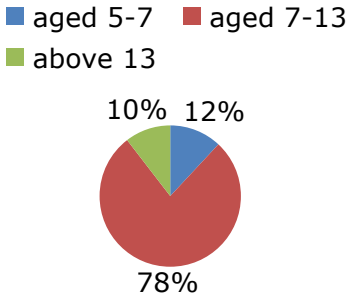
- Word list: 170 isolated Mandarin monosyllabic, disyllabic and trisyllabic words that contained all the target lexical tones and were phonemically balanced.
- Idioms: 10 Chinese popular idioms.
- Digital sequence: 10 digital sequences.
- Tongue twister: 4 popular Chinese tongue twisters.
- Read speech: 2 read narrative stories from Children's textbook.

- Children's rhymes: 7 popular Chinese Children's rhymes
- Spontaneous speech: 8 picture narrative stories from Chinese Children's book
- 99 prelingually deafened Mandarin-speaking children with unilateral multi-channel CIs were recruited and 32 NH were used as a baseline.

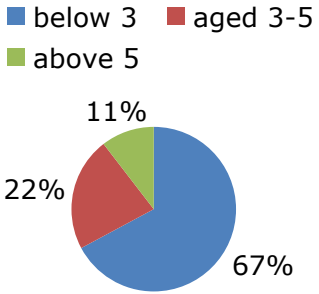
Gender



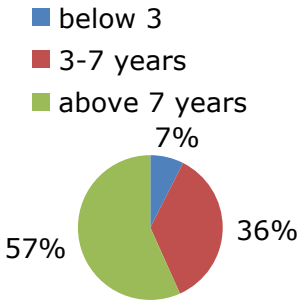
Age



Implantation age



Implantation length



语言资源高精尖创新中心
Beijing Advanced Innovation Center for Language Resources

A series of guidelines to the Prevention and Control of COVID-19 Epidemic in Foreign Languages

- Daily Precautions
- Entry Tips
- Diagnosis
- Protection Measures

In Aalbanian, Amharic, Arabic, Azerbaijani, Bulgarian, Burmese, Byelorussian, Cambodian, Catalan, Czech, English, Finnish, Filipino, French, German, Greek, Hausa, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazak, Korean, Kyrgyz, Lao, Malay, Mongolian, Persian, Portuguese, Romanian, Russian, Serbian, Spanish, Swahili, Thai, Turkish, Urdu, Ukrainian, Uzbek, Vietnamese.

A Guide to Understanding the Hubei Dialects for Medical Assistance Teams

in Mandarin and 9 local dialects of Wuhan, Xiangyang, Yichang, Huangshi, Jingzhou, E'zhou, Xiaogan, Huanggang, Xianning in Hubei Province.

http://yuyanziyuan.blcu.edu.cn/info/1162/2005.htm

CNCeb: Multi-Genre Speaker Recognition Corpus

- **1,000** Chinese Celebrities
- **11** copmlex genres
- **274** hours, **130k** utts.
- **Free** for research

Genre	# of Spks	# of Utters	# of Hours
Entertainment	483	22,064	33.67
Interview	780	59,317	135.77
Singing	318	12,551	28.83
Play	69	4,245	4.95
Movie	62	2,749	2.20
Vlog	41	1,894	4.15
Live Broadcast	129	8,747	16.35
Speech	122	8,401	36.22
Drama	160	7,274	6.43
Recitation	41	2,747	4.98
Advertisement	17	120	0.18
Overall	**1,000**	**130,109**	**273.73**

Commercial Activities

标贝科技 Databaker technology www.data-baker.com

Data Baker AI speech & data system

Application scenarios	• Car navigation • Self-driving Vehicles, Security • Medical Applications	• Inatelligent Education • Smart Home • Entertainment • Audio books
Speech synthesis	**Speech Synthesis** • Custom Standardized	**Libraries** •Own library: Star, animation IP, quality pronunciation and clarity Domestic and foreign languages TTS Measurement Systems
Data products	• Speech synthesis • Speech recognition	• NLP • Computer vision
Technology	**Algorithms + Models + Architecture** •Machine learning, deep neural networks, speech recognition •Speech synthesis, deep learning technology, computer vision, natural language processing	
Foundation	**Data** • Speech, image, text, etc	**Processing System** • Targeted collection, processing procedures and methods **Quality** • Quality guarantee

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

■ **Smart Home / Meeting Room**
2000 speakers, 6000 Hours

■ **Autonomous Driving**
2300 speakers, 10750 Hours

■ **Mandarin Corpus**
11600 speakers, 5800 Hours

■ **HINDI**
Speech Corpus
580 speakers, 550 Hours

■ **US English**
Speech Corpus
2600 speakers, 800 Hours

■ **Arabic**
Speech Corpus
200 speakers, 300 Hours

■ **Russian**
Speech Corpus
200 speakers, 272 Hours

■ **Free Corpora**

AISHELL-1:Mandarin
400 speakers, 170 Hours

AISHELL-2:Mandarin
1991 speakers, 1000 Hours

AISHELL-WakeUp-1:
Chinese and English
Wake-up Words Speech Data
254 speakers, 1561 Hours

Commercial Activities

DATATANG

www.datatang.com

Chinese	Mandarin/Accented Mandarin Children Mandarin Dialect(Cantonese, Sichuan, Shanghai, Min, Henan, Wuhan, Changsha and etc.) Mandarin English Mixed	24000 Hours 50000 Speakers
English	27 countries speaking English, including: US English, UK English, Other Accented English	12000 Hours 25000 Speakers
Other Languages	36 Languages, including: Japanese, Korean, Malay, Indonesian, Russia, French, German, Spanish and etc.	15000 Hours 32000 Speakers
Parallel Corpora	CH-EN, CH-RU, CH-JA, CH-FR, KO-EN, JA-EN and etc.	12 million pairs

慧听数据 HUITING DATA

http://www.huitingtech.com/

Chinese **Mandarin** **Accented Mandarin** **Children Mandarin** **Elderly Mandarin** **Dialect** **Uygur / Tibetan /Mongolian** **Cantonese** **Mandarin English Mixed** **Cantonese English Mixed**	**35000 Speakers** **25000 Hours**
English **UK English** **US English** **Other Accented English**	**2200 Speakers** **2100 Hours**
Other Languages **French, German, Italian, Spanish, Mexican Spanish,** **Brazilian Portuguese, Japanese, Korean**	**3900 Speakers** **2100 Hours**

Free Database

Prime words	Mandarin Speech, 1600h	**DataTang**	Mandarin Speech, 1050h	**AIShell**	Mandarin Speech, 2631h	**Mobvoi**	Hotwords, 36k utt
MAGIC DATA	Mandarin Speech, 755h	**Baidu**	Mandarin Speech, 50h	**ReactiveCJ**	Madarin Text, 16G	**Databaker**	TTS 10k female, 12h