A Country Report – COCOSDA Activities in China Data

More and more companies on data resources and services suppliers are emerging in China: a new trend

O-COCOSDA 2016，Bali Indonesia

Aijun LI , *Dong WANG

Institute of Linguistics, Chinese Academy of Social Sciences

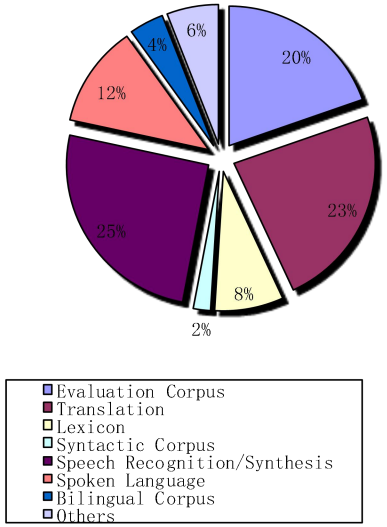
*Research Institute of Information Technology, Tsinghua University

Institute of Linguistics, CASS

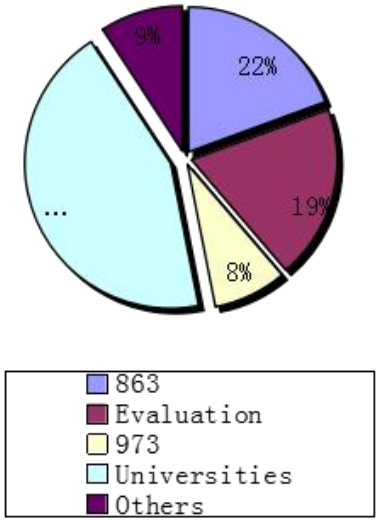
- **AESOP-CASS: 10,000h, more annotated data and research carried out.**
- **Discourse-CASS: more than 1000 dialogues and 100 discourses with rich annotation on** speech segmental and prosodic structure/Information structure/Rhetoric structure /Speech act (topics, adjacent pairs)/Referential structure/Dependency relation/expression
- **Word-Child-CASS:** 1.5-6y word database, 4000 children, all with canonical and real pronunciation annotation
- **Articulatory EMA-CASS**: English word EMA data for English L2 learners and native speakers. 10 speakers. (new in 2016)
- **Articulatory FMRI-CASS**: one Chinese speaker's syllable data.

Chinese LDC

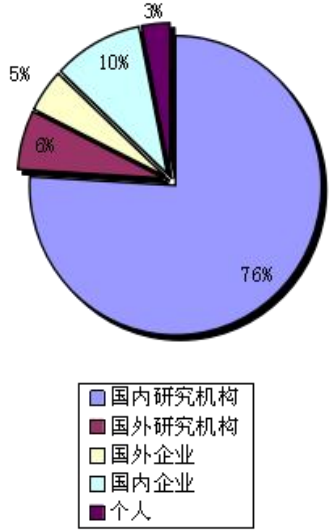
- Till now, there are 102 corpora, including speech synthesis/recognition corpora, corpora for machine translation, lexicon and other natural language processing corpora.
- In 2016, there are 4 new corpora added (1 syntactic and 3 others corpus);20 corpora (7 spoken language, 5 translation, 3 speech recognition/synthesis, 2 lexicon and 3 other corpus) have been distributed to 11 institutes and companies.



Types of the corpora



Providers of the corpora



Types of the users

speechocean 海天瑞声

http://www.speechocean.com/

ASR Corpus

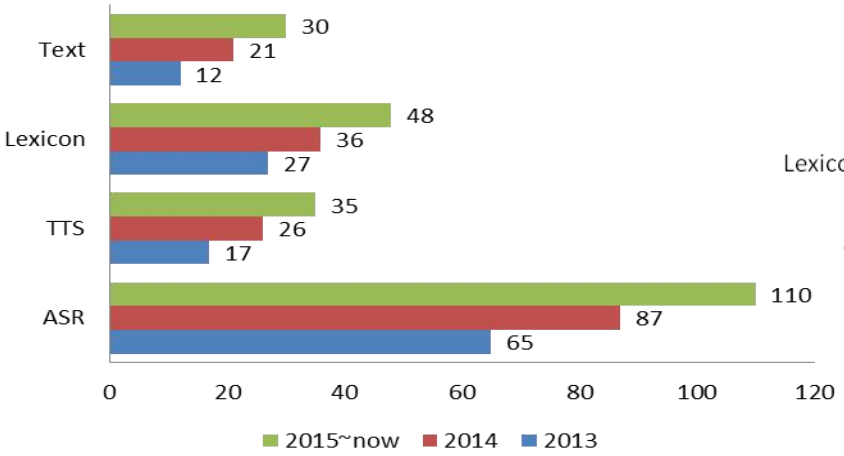
TTS Corpus

Lexicon Corpus

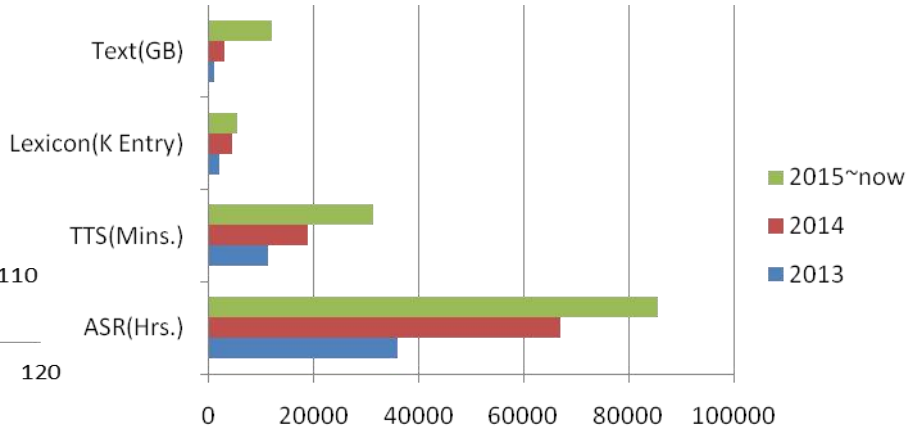
Text Corpus

A worldwide data resources & services supplier with 15 years experience in the fields of Human Computer Interaction and Human Language Technology, such as speech synthesis, speech recognition, machine translation, web search, image recognition, and natural language understanding.

Languages Increasing



Volume Increasing



慧听数据 HUITING DATA

http://www.huitingtech.com/

About Huiting Data

Founded in 2011, Huiting Data is a leading multimedia data and technology service provider. Huiting Data collaborates with international artificial intelligence technology companies providing high quality speech, image, text and other multimedia databases.

Award-Winning Corpus

- Multi-accent Mandarin Speech Recognition Database (Link)
- Awarded as one of the five creative products by Speech Industry Alliance of China (SIAC)
  - Speakers' Dialects include Cantonese, Min Dialect, Gan Dialect, Sichuan Dialect, Wu Dialect and Xiang Dialect
  - Sentence-wise Error Rate < 2% | Recorded by Smart Phone
  - 3000 Hours | 3000 Speakers | Mono PCM | 16kHz | 16bit

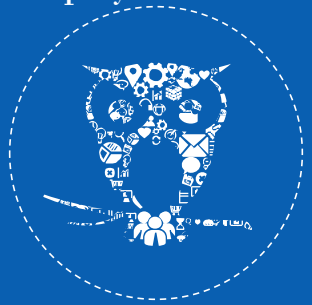
Featured Corpuses

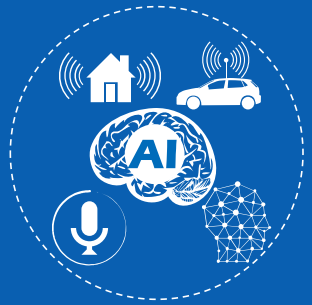
Cantonese Speech Recognition Database	1500 Speakers | 1000 Hours	Link
Cantonese-English Mixed Recognition Database	1400 Speakers | 500 Hours	Link
Mandarin-English Mixed Recognition Database	2700 Speakers | 1000 Hours	Link
Multi-language Speech Recognition Database	300 Speakers | 90 Hours for Each Language	Link

Data Mall is:

- The largest data exchange platform
- 45,000 datasets in all domains
- 1,600,000+ exchanges in 2015

- Has 500,000 certified collectors distributing globally.
- Releases 100 new collecting missions weekly.
- Collects various types of data stably including Speech, Image and Text
- Trusted partner of Fortune 500 companies
- Processed over 10 million images/100,000 hours of speech in 2015
- Employees: 1000+





DATA RESOURCE

Datatang is capable of collecting data from various industries including government, finance, health care, and traffic, etc.

DATA MARKET

Datatang

Data Service
Data Exchange
API/SDK

APPLICATION

Datatang aims to make your products even smarter. Training algorithm and building machine learning model have never been easier.

globalsales@datatang.com