



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

# Introduction to ACLR: Objective, Mission, and Projects

## (Beijing Advanced Innovation Center for Language Resources)



Yang Erhong

[yerhong@blcu.edu.cn](mailto:yerhong@blcu.edu.cn)





# Founding of Beijing Advanced Innovation Center for Language Resources

Beijing Advanced Innovation Center for Language Resources, officially established in 2016, is a scientific research institution accredited and funded by Beijing Municipal Education Commission, and hosted by Beijing Language and Culture University. It is centered on language resources construction, and provide language resources data for language preservation and display, and language education, technology development.

## 北京市教育委员会

### 北京市教育委员会关于认定第二批北京高等学校高精尖创新中心的公告

各有关高校：

根据市委《关于印发北京高等学校高精尖创新中心建设计划的通知》（京教研〔2015〕1号），在事前绩效评估、专家评审的基础上，经市教委主任办公会审议通过，认定北京大学未来基因诊断高精尖创新中心等4个高精尖中心为第二批“北京高等学校高精尖创新中心”（名单附后）。

请按照文件要求和高精尖创新中心建设目标，积极探索高精尖创新中心的运行体制与管理机制，广聚国际国内领军创新人才，切实做好高精尖中心的建设与管理工。

附件：第二批北京高等学校高精尖创新中心名单

北京市教育委员会  
2016年5月9日

### 第二批北京高校高精尖创新中心名单

名称	中心名称
北京航空航天大学	生物医学工程高精尖创新中心
北京建筑大学	未来城市设计高精尖创新中心
北京林业大学	林木分子设计育种高精尖创新中心
北京语言大学	语言资源高精尖创新中心
中国音乐学院	中国乐派高精尖创新中心

（注：以上中心以学校代码排序）

（本文主动公开）

Official document from Beijing Municipal Education Commission that approves the founding of the Center



# Outline

**Our Objective and Mission**



**Some Language Resources Projects**



**Future Work**





# Outline

**Our Objective and Mission**



**Some Language Resources Projects**



**Future Work**



A

B

C



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

## Stepping Stones

2016 

5 years

Rare Languages



10 years

Most of languages in Belt and Road countries



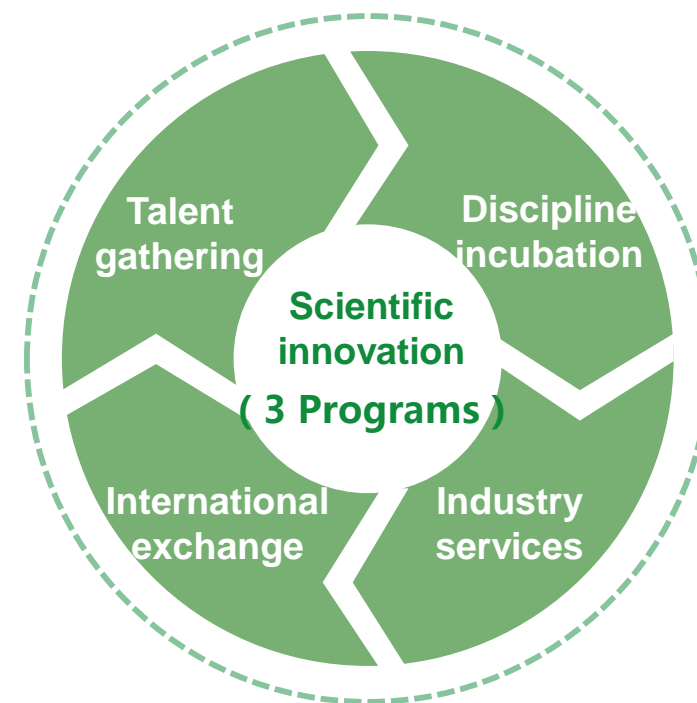
20 years

Most of the world's languages



## Objective

Objective



Focusing on scientific innovation in the Three Programs of **Language Resources Bank**, **Language and Culture Museum**, and **Language Education and Technology Service**, it conducts related work of discipline development, talent cultivation and international exchange and cooperation.



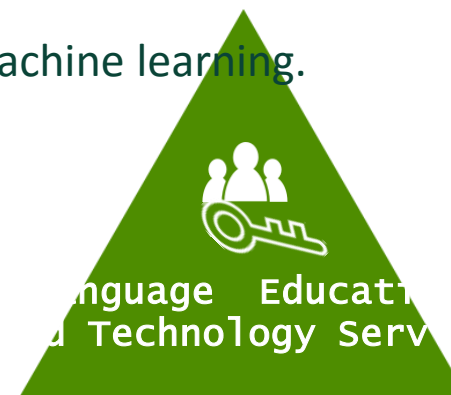
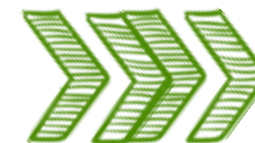
语言资源高精尖创新中心  
Beijing Advanced Innovation  
Center for Language Resources

# Mission

**Language Resources Bank** is the basic program, providing resources for Language and Culture Museum and intelligent Language Service.

Aim of Collecting language resources:

1. **Preserve languages.** The language resources contain enough information so as to **resurrect** a language.
2. **Exhibit languages.** The language resources can typically **display** a language.
3. Develop **language technology.** The Language resources can be sufficient for machine learning.



To display languages and cultures of countries around the world, to share language resources, and ensure optimal language resources utilization

To collect and develop language resources around the world

To solve problems in cross language communication, assist language learning, language data mining



## Program 1: Language Resources Bank

Basic resource database of all languages and cultures

Multi-axis multilingual parallel and comparable corpus

Multilingual and multidisciplinary thesaurus and knowledge graph

Multilingual resources database for teaching and research

Sinology (Chinese Studies) resources database

Multilingual dynamic circulating corpus

Multilingual oral, audio and text corpus

Multinational official documents database

Country-specific language learning resources database (including interlanguage corpus )

Talent information database



## Program 2: Language and Culture Museum

**Online**

Physical museum

Exhibition area

Simulation interactive experience area

Actual context area

Multi-function area





## Program 3: Language Education and Technology Service

### Machine Translation System

To solve Language communication barriers

### Computer Aided Learning System

For future language education technology

### Language Data mining

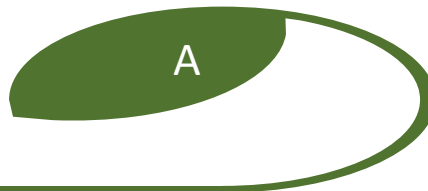
To detect and track the catchword ,hot topic, etc.,  
under complex language context



# Outline

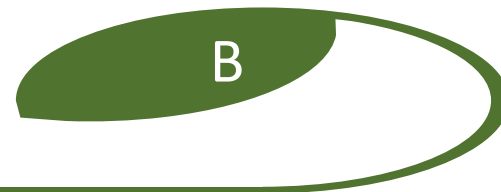
Our Objectives and Mission

---



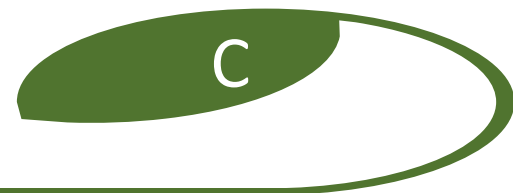
Some Language Resources Projects

---



Future Work

---





- Language resources specifically to preserve languages and display culture
- Resources for language teaching and technology research
- Field-oriented or Task-oriented language resources



## Language resources specifically to preserve languages and display culture

1

Language resources collected from international students

2

Language investigation and language data collection from target countries

3

Overseas Chinese language resources



Language and culture resources investigation, collection and shooting work is being carried out with help of the international students from over 150 countries in Beijing Language and Culture University. Recently more than 30 countries/languages resources will have been completed.

Collecting data:

- Words: 207 Swedish core words
- Sentences: 125 commonly-used sentences
- Discourses: local story or tale, traditional song or folk song which are characteristic of students' country



Investigate the official standard languages, important languages and minority languages of the 6 countries, namely Kazakhstan, Uzbekistan, Kyrgyzstan, Laos, Myanmar, and Vietnam. Build language resources database and culture resources database of the 6 countries.

### Collecting data:

- **Words:** 3000 words from ethnic languages in Language Protection Project (without any revision) plus 5000 words complemented by the Research Group. The objective is to build a wordlist (6000-7000 words) containing common words of 8 language families and specific words of standard languages of target language countries. The wordlist structure is : Chinese words, translation of the standard languages of the target language countries, international phonetic transcription, and notes.
- **Sentences for grammar investigating:** 100 grammar items from the Language Protection Project, ,the data structure is same as the wordlist including Chinese words, translation of the standard



- 1) Basic information database of overseas Chinese people and Chinese language (Distribution information database of overseas Chinese people and Chinese language, Chinese language policies database, documents database on Chinese language studies, Chinese media usage database)
- 2) Oral history database on Chinese language and its education
- 3) Multimedia Chinese language corpus: Chinese language usage and language situation in overseas Chinese



## Resources for language teaching and technology research

01

**Global Chinese  
interlanguage corpus**

---

02

**Multi-modal Inter-Chinese Speech Corpus for  
Developing Intelligent Pronunciation Teaching  
Technology**

---

03

**BCC (BLCU Corpus Center)—  
syntax and semantic analysis**

---





# 01 Global Chinese interlanguage corpus

---

- HSK dynamic Composition Corpus is a Chinese interlanguage corpus collected from those who learn Chinese as second language. The corpus contains 11569 compositions (about 4.3 million characters) from those who take high-level HSK test (the Chinese Proficiency Test). The corpus can be used for conducting research on teaching Chinese as second language.
- The corpus begins to be built in July, 2003, and put into trial operation online in December, 2006. Now it is officially open to the outside.



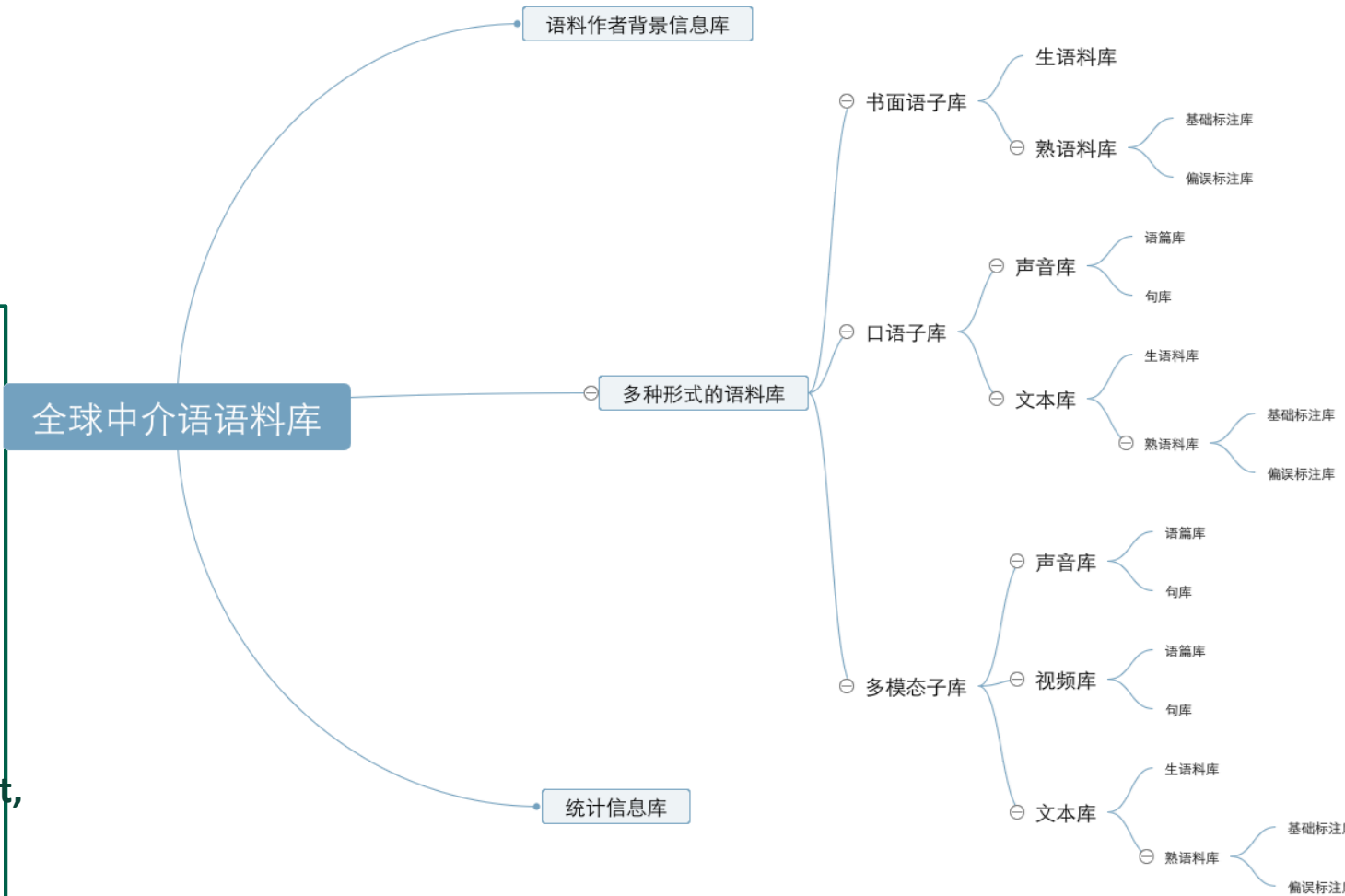
# 01 Global Chinese interlanguage corpus

## Universally-used Chinese interlanguage corpus

50 million characters (about 15 million characters in the first 3 years), including written corpus, spoken corpus and video corpus, emphasizing spoken sub-corpus and multimodal sub-corpus construction.

Develop standards and codes for building Chinese interlanguage corpus, such as standards for corpus construction, standards for corpus collection and input, standards for spoken and video corpus transcription, codes for corpus annotation.

### 全球中介语语料库



Structure of the corpus



# 02

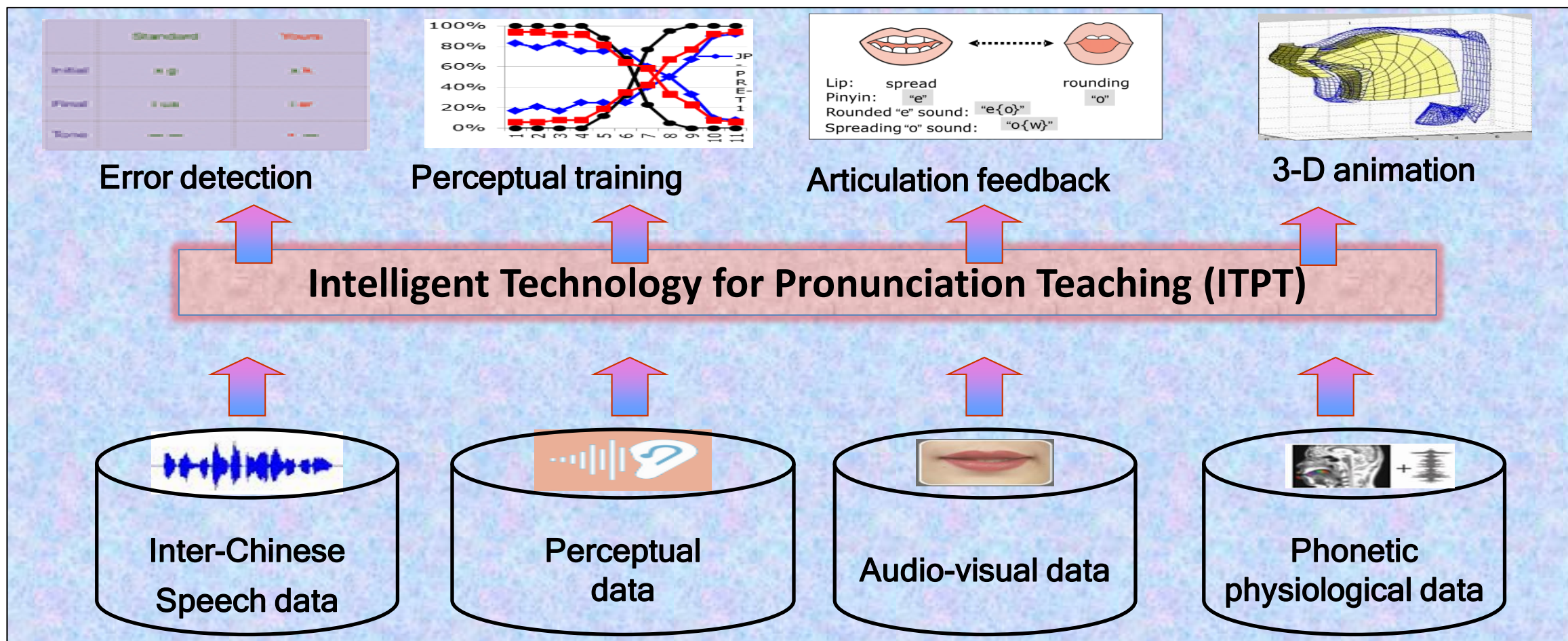
## Multi-modal Inter-Chinese Speech Corpus for Developing Intelligent Pronunciation Teaching Technology

---

Aim: Constructing a multi-modal inter-Chinese speech corpus for developing intelligent technology for pronunciation teaching (ITPT).



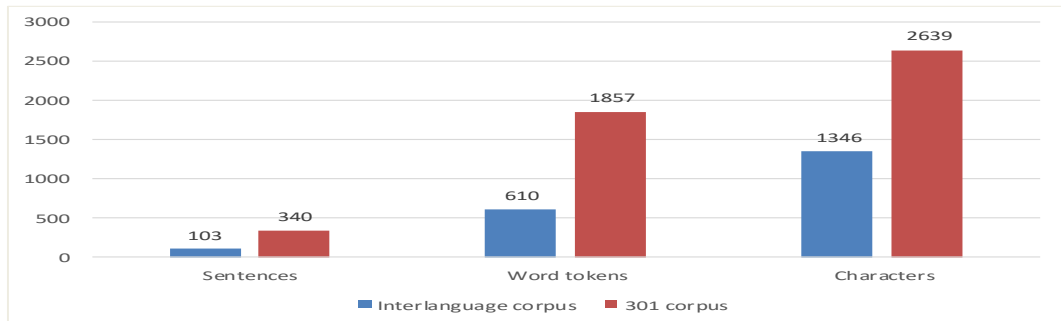
# Multi-modal Inter-Chinese Speech Corpus of L2 Learners



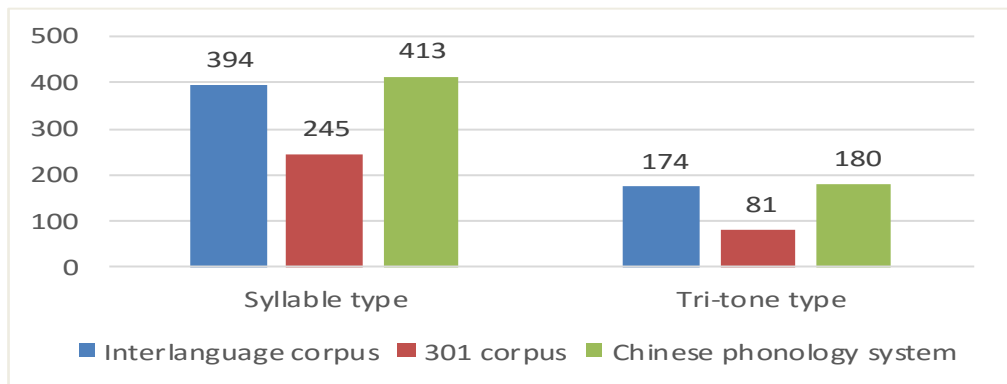


## Progress: Phonetically Rich 103 Sentences for Speech Collection

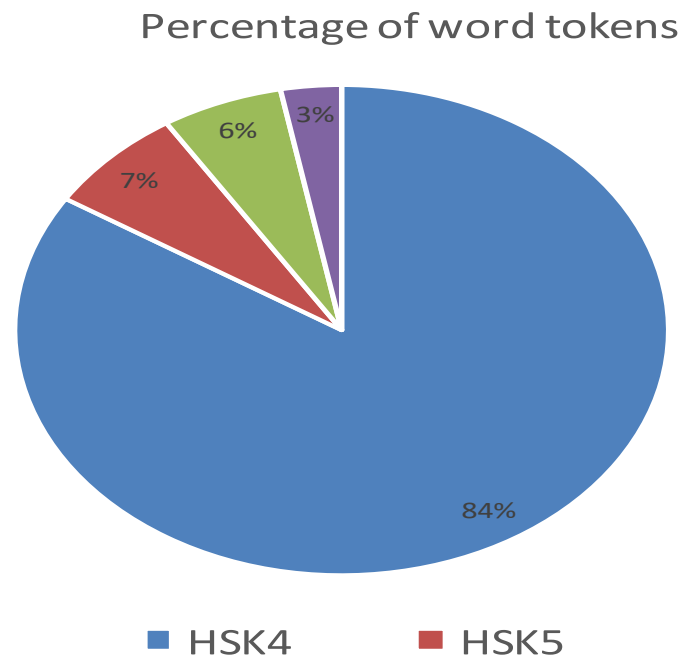
Phonetically Rich 103 Sentences for Speech Collection, , 610 word tokens, 1346 characters, average length=13.2char/sent.



(a) Basic statistics for 2 corpora.



(b) Phonetic varieties of 2 corpora.



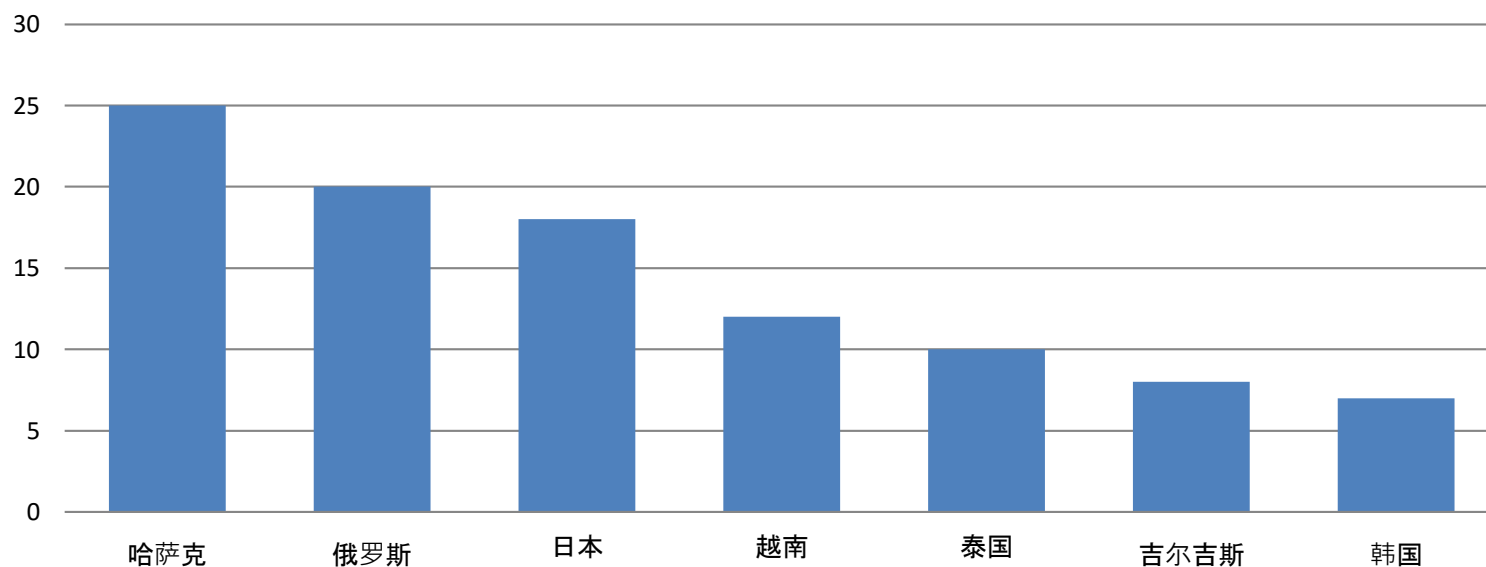
(c) Lexical difficulty level statistics.



# Speech data collected

- Native Chinese: 25 males, 25 females
- L2 learners: 28 males, 72 females

Distributions of L2 learners collected







# Future: APP based collection

“汉语说”  
智能汉语语音教学系统

外国学生自主发音训练工具  
对外汉语教学辅助工具




字词  
短语  
对话



	Standard	Voice
Initial	x	x
Final	ia	ia
Tone	—	—

标识学生错误发音，针对声韵母、声调  
不同层级纠正反馈，正确率接近 95%



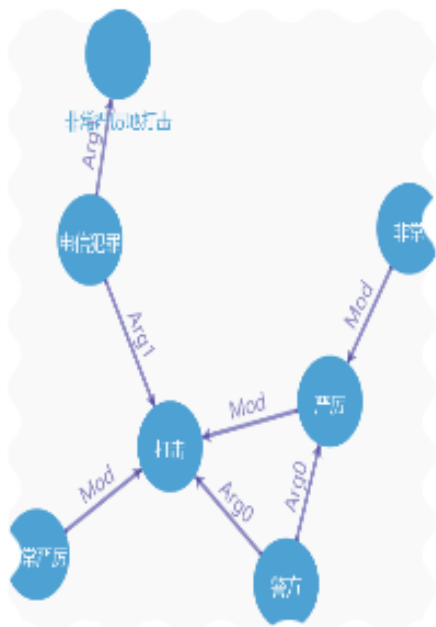
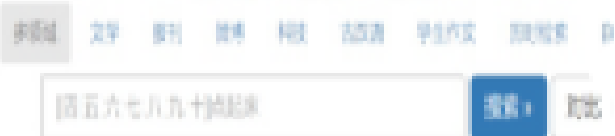
Georgia Tech  
SALT  
北京智能语言习得实验室  
语言资源  
高精尖创新中心



# 03

## BCC (BLCU Corpus Center)——syntax and semantic analysis

# BCC



### BCC (BLCU Corpus Center)——syntax and semantic analysis

BCC, totally 15 billion characters, including multi-field corpus: newspaper (2 billion), literature (3 billion), micro-blog (3 billion), science and technology (3 billion), general (1 billion), and ancient Chinese (2 billion). It is a large-scale corpus to fully reflect today's language situation.

The Project of Syntax and Semantic Analysis and the Applications has collected all kinds of Chinese language corpus of hundred Giga. Multi-level annotation is done by human and mainly computer. Chinese typical structures as bound phrases and various chunks are accessed, including preposition phrase, verb phrases, adverbial phrases, and verb-object phrases.





## Field-oriented or Task-oriented language resources

Russian Chinese bilingual resources

Olympic Winter Games Multilingual Termbase



## Russian Chinese bilingual resources

---



Quantities of Russian-Chinese dictionaries have been collected to build Russian-Chinese bilingual language knowledge database.

39 dictionaries  
over 100 classified lexica bases in professional dictionary  
Parallel corpus for **economic and trade fields**



## Olympic Winter Games Multilingual Termbase

### **Languages:**

8 languages: Chinese, English, French, Japanese, Korean, Russian, Arabic, Spanish

### **Terminology Statistics:**

Total terms to be collected: 20,000-30,000

Current work:

English: 5000+, Chinese and other languages: 3000+

Current work is to validate the term information, and expand the basic information of every term.

### **Sources:**

Over 200 digital books related to Olympic Winter Games, and Olympic Games.



## Term information example

序号	术语	语言	词性	术语来源 文本	术语来源 链接	定义	定义来源 文本	定义来源 链接	语境	语境来源 文本	语境来源 链接	项目名称
0	冰壶	cn	n		<a href="http://www.beijing2022.cn/cn/olympics/curling.htm">http://www.beijing2022.cn/cn/olympics/curling.htm</a>	冰壶，两队之间比赛，每队四人，两队轮流掷球，不仅需要使冰壶准确到达营垒的中心，同时让对方的冰壶远离圆心，最后以冰壶距离营垒圆心的远近决定胜负。冰壶1998年正式列入冬奥会比赛项目。		<a href="http://www.beijing2022.cn/cn/olympics/curling.htm">http://www.beijing2022.cn/cn/olympics/curling.htm</a>	如果冰壶在比赛中损坏，该队应按照冰壶精神，决定这个（或这些）冰壶所在的位置。如果双方未能就位置达成一致，该局比赛重新开始。	《奥林匹克冬季运动冰壶运动和竞赛规则》	<a href="http://www.curling.org.cn/index.php?m=news.view&amp;id=450">http://www.curling.org.cn/index.php?m=news.view&amp;id=450</a>	冰壶
0	curling	en	n		<a href="http://www.beijing2022.cn/en/olympics/curling.htm">http://www.beijing2022.cn/en/olympics/curling.htm</a>	Starting from the Nagano 1998 Olympic Winter Games, curling was adopted as an official sport, with two curling events - men's and women's curling.		<a href="http://www.beijing2022.cn/en/olympics/curling.htm">http://www.beijing2022.cn/en/olympics/curling.htm</a>	Curling is a game of skill and of tradition. A shot well executed is a delight to see and it is also a fine thing to observe the time-honoured traditions of curling being applied in the true spirit of the game.	<i>THE RULES OF CURLING and Rules of Competition</i>	<a href="http://www.curling.org.cn/index.php?m=news.view&amp;id=450">http://www.curling.org.cn/index.php?m=news.view&amp;id=450</a>	Curling



## Project undertakers:

Project undertaker and team member: **scholars and experts from Research institute and universities all of world.**

.

.

- Language resources specifically to preserve world languages and display culture
- 

- Resources for language teaching and technology research
- 

- Field-oriented or Task-oriented language resources
-



# Outline

Our Objectives and Mission



A

Some Language Resources Projects



B

**Future Work**



C



- Language resources need to be constructed from the aspects of resource protection, cultural inheritance and information processing.

---

- There is lack of language resources along the Belt and Road countries. Research on language technology will be conducted with low resources.

---

- Standards or specification of language resources construction.

---

- The evaluation on the language resources application

---



# Workshop on the Belt and Road Language Resources and Evaluation (B&R LRE)

- a new LREC Workshop
  - 8 May 2018, co-located with LREC
  - The Phoenix Seagaia Resort, Miyazaki, Japan
  - Submission deadline     January 15, 2018
- Topics include, but not limited to, the following areas:
  - The Belt and Road language resources collection, processing and sharing
  - Standards and codes of the Belt and Road language resources
  - Application of tools and language technologies in the Belt and Road language resources
  - Development of language technologies evaluation method
- Panel discussion on low-resource language collection and development





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

谢谢!  
**Thank you!**

