Introduction to ACLR: Objective, Mission, and Projects

(Beijing Advanced Innovation Center for Language Resources)
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.
Outline

Our Objective and Mission

Some Language Resources Projects

Future Work
Outline

Our Objective and Mission

Some Language Resources Projects

Future Work
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.
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 solve problems in cross language communication, assist language learning, language data mining

To collect and develop language resources around the world
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 resources collected from international students

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 language investigation and language resources collection from 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 language investigation and language resources collection from target language countries, international phonetic transcription, and notes.
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
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.
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.

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

- Error detection
- Perceptual training
- Articulation feedback
- 3-D animation

Intelligent Technology for Pronunciation Teaching (ITPT)

- Inter-Chinese Speech data
- Perceptual data
- Audio-visual data
- Phonetic physiological data
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
Future: APP based collection
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
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

<table>
<thead>
<tr>
<th>序号</th>
<th>术语</th>
<th>语言</th>
<th>词性</th>
<th>术语来源文本</th>
<th>定义</th>
<th>定义来源文本</th>
<th>语境</th>
<th>语境来源文本</th>
<th>项目名称</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>curling</td>
<td>en</td>
<td>n</td>
<td><a href="http://www.beijing2022.cn/en/olympics/curling.htm">http://www.beijing2022.cn/en/olympics/curling.htm</a></td>
<td>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.</td>
<td><a href="http://www.beijing2022.cn/en/olympics/curling.htm">http://www.beijing2022.cn/en/olympics/curling.htm</a></td>
<td>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.</td>
<td>THE RULES OF CURLING and Rules of Competition</td>
<td><a href="http://www.curling.org.cn/index.php?m=news&amp;vid=460">http://www.curling.org.cn/index.php?m=news&amp;vid=460</a></td>
</tr>
</tbody>
</table>
Project undertakers 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

Some Language Resources Projects

Future Work
• 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
谢谢！
Thank you!