LANGUAGES
Chinese

Language Family and Dialects

Chinese is a member of the Sino-Tibetan language family with an estimated 1.2 billion speakers around the world. It is an official language of China, Taiwan and Singapore and one of the six official languages of the United Nations. Chinese has many dialects or regional varieties led by: Mandarin or Putonghua (across China and government, media, education), Wu (Zhejiang, Jiangsu, Shanghai, Hong Kong), Yue (Guangdong, Guangxi, Hong Kong, Macau, Singapore, Malaysia) and Min (Guangdong, Zhejiang, Jiangxi).

(Image © Wyunhe / CC-BY-SA-3.0)

LDC collects and develops digital Chinese resources of all types, spanning text (newswire, web, SMS, chat), speech (telephone, broadcast), video (broadcast, web) and lexicons. It applies a range of annotations to that source data, among them segmentation, transcription, translation, parsing, word alignment and co-reference. This work, represented in language resources distributed through LDC's catalog, supports ongoing research and human language technology development including automatic speech recognition, machine translation and content extraction.

Language Resource Development Challenges and Solutions

Chinese presents various challenges to the human language technology researcher:

- Competing encoding and writing systems
- Lack of word boundaries in written form
- Sparse morphology
- Complex phonology, including tone

LDC addresses these challenges in various ways as shown by the examples below.

Encoding. LDC has developed expertise in processing and normalizing the various encoding schemes for Chinese text, principally Big 5, GB, GBK and UTF-8. Traditional Mandarin text is typically encoded in the Big 5 scheme. For data sets with both traditional and simplified text, LDC converts Big 5 encoding to GB or GBK to facilitate processing. Since that conversion carries some loss, the original Big 5 source data is often included for reference. With the first release of its Chinese Gigaword series in 2003 and continuing through subsequent editions, LDC converts all Chinese source texts (traditional and simplified) to UTF-8 encoding.

Writing systems. The Chinese government adopted a simplified writing system in the mid-twentieth century. However, some data is available in traditional Chinese only. Because language resources and tools are generally developed to handle simplified writing, LDC routinely converts traditional Chinese data to its simplified form.

Word segmentation and morphology. Natural language processing requires certain inputs, starting with a word list, followed by segmented, part-of-speech tagged and parsed text. Segmentation is a challenge for Chinese since written text lacks word boundaries. Its sparse morphology adds a further complication to NLP. LDC developed a state of the art segmenter in the 1990s which it used for over ten years. LDC currently uses several segmentation methods, including one that incorporates the work of the Chinese Treebank research team (University of Pennsylvania, University of Colorado, Brandeis University) and combines segmentation, part-of-speech tagging and parsing.

Phonetic segmentation and tone. LDC has used embedded tone modeling in experiments to improve forced alignment accuracy in Mandarin Chinese.

"Chinese" written in traditional (left) and simplified (right) charaters with Pinyin romanization
(Image, Wikimedia commons)

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Selected Resources

LDC's catalog contains over 150 Chinese resources in multiple genres.

- Mandarin Chinese broadcast speech and transcripts and parallel, word-aligned and tagged text; broadcast news and conversation, newswire, journal articles, government documents, web text
- Conversational telephone speech (some with transcripts): Mandarin, Yue, Wu, Min
- Microphone speech and transcripts: Taiwanese Putonghua
- Large text: Chinese Gigaword Fifth Edition, LDC2011T13 (newswire from 1991-2010)
- Chinese Treebanks: newswire, broadcast, web data

```
<DOC id="XIN_CMN_19980201.0003" type="story">
<HEADLINE>
李岚清会见欧盟委员会主席桑特
</HEADLINE>
<DATELINE>
新华社达沃斯(瑞士)2月1日电
</DATELINE>
<TEXT>
<P>
(记者陈维斌　严
明)正在瑞士达沃斯出席世界经济论坛年会的中国国务院
副总理李岚清1日在这里会见了欧盟委员会主席桑特。
</P>
<P>
李岚清说,近年来,中国同欧盟及其成员国的关系继
续保持良好的发展势头,双方高层互访和接触频繁,不同
层次的政治磋商和对话活跃,各个领域的合作与交流不断
扩大。双方经济互补性强,对许多重大国际问题有着一致
或相似的看法。
</P>
<P>
桑特说,欧盟非常重视发展对华关系,对中国改革开
放取得的重大成就深感钦佩。
</P>
```

Chinese Gigaword Fifth Edition Sample

LDC Research and Collaboration

Speech processing. Chinese is a tone language, and speech processing is a particular area of interest for LDC research. Activities include the development of novel techniques in speech activity detection and experiments in tone classification and phone segmentation and labeling.

Language diversity. LDC is among the first recipients of the Penn China Research and Engagement Fund Awards. Under the leadership of Director Mark Liberman, LDC uses this grant to extend its initiative on linguistic diversity in China, with specific emphasis on the documentation and analysis of variation in standard, regional, and minority languages. Collaborators include Beijing Normal University, Minzu University and Beijing Language and Culture University (School of Linguistic Sciences).

Resource development. LDC works with Hong Kong University of Science and Technology for collection and annotation of Chinese conversational telephone speech and broadcast speech, and the University of Colorado and Brandeis University for the development of Chinese treebanks and proposition banks.

Sponsored Project Support

Projects supporting the creation and distribution of Chinese language resources include ACE, BOLT, DEFT, EARS, GALE, MADCAT, TDT and TIDES. Among the NIST technology evaluations using these resources are LRE, SRE, OpenMT, Spoken Term Detection and TRECVID.

Papers and Publications

LDC has presented and published numerous papers about its work in Chinese which are found on the LDC Papers page, https://www.ldc.upenn.edu/language-resources/papers/ldc-papers. A sample follows:

- Investigating Consonant Reduction in Mandarin Chinese with Improved Forced Alignment (Yuan, et al., 2015)
- Automatic Phonetic Segmentation in Mandarin Chinese: Boundary Models, Glottal Features and Tone (Yuan, et al., 2014)
- Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus (Song, et al., 2014)
- Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures (Li, et al., 2012)
- A Very Large Scale Mandarin Chinese Broadcast Collection for the GALE Program (Yi, et al., 2010)

Conclusion

Human language technology and its related fields change rapidly, requiring new data genres, faster, cost-efficient annotation processes and flexible tools. LDC's body of work in Chinese is an example of how LDC successfully meets those challenges within a given language. Keep apprised of updates, announcements of new releases and LDC Chinese projects from our website, newsletter and social media channels.