CHAPTER 23

CSLP CORPORA AND LANGUAGE RESOURCES

Hsiao-Chuan Wang[†], Thomas Fang Zheng[‡], and Jianhua Tao[§]

[†]*National Tsing Hua University, Hsinchu*
[‡]*Research Institute of Information Technology, Tsinghua University, Beijing*
[§]*Institute of Automation, Chinese Academy of Sciences, Beijing*
E-mail:{hcwang@ee.nthu.edu.tw, fzheng@tsinghua.edu.cn, jhtao@nlpr.ia.ac.cn}

This chapter discusses the fundamental issues related to the development of language resources for Chinese spoken language processing (CSLP). Chinese dialects, transcription systems, and Chinese character sets are described. The general procedure for speech corpus production is introduced, along with the dialect-specific problems related to CSLP corpora. Some activities in the development of CSLP corpora are also presented here. Finally, available language resources for CSLP as well as their related websites are listed.

1. Introduction

Language resources usually refer to large sets of language data or descriptions which are in machine-readable form. They can be used for developing and evaluating natural language and speech processing algorithms and systems. A language resource can be in the form of a written language corpus, spoken language corpus, a lexical database, or an electronic dictionary. It may also include software tools for the use of that resource. A written language corpus is a text corpus which can be made up of whole texts or samples of texts. The development of corpus linguistics requires the use of computational tools to process large-sized text corpora.[1] A spoken language corpus refers to the speech database which is designed for the development and evaluation of speech processing systems. For example, a speech recognition system based on statistical model techniques needs to train acoustic models of speech units by using a large amount of speech data, as well as to train its language models by the use of a large text corpus.

The building of a large language corpus involves a huge effort that is required for data collection, transcription, representation, annotation, validation, documentation, and distribution. Some organizations have been established in the past decades to create and gather language resources, promote the reuse of these resources, and develop new technologies in the process of building language resources. The two most important and internationally-recognized organizations are the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA). The LDC, founded in 1992, is an open consortium of universities, private organizations and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes.[2,3] ELRA was established as a non-profit organization in 1995. It is active in identification, distribution, collection, validation, standardization, improvement, and production of language resources. Its focus is on the issues of making language resources available to different sectors of the language engineering community.[4,5]

Chinese is one of the major languages of the world. However, research on Chinese spoken language processing is about one or two decades behind the level of research done on English, Japanese, and some European languages. Some research on Chinese speech synthesis and recognition started in the mid-1980s. Thereafter, much more research were reported in the 1990's.[6-12] Almost in the same period, some projects involving the creation of large scale Chinese language resources were initiated in mainland China, Taiwan and Hong Kong. In mainland China, the 863 Program had created several language corpora in various domains, such as the corpora for speech recognition, speech synthesis, parallel language processing (for Chinese, English, and Japanese), information indexing, and dialogue systems.[13-16] Some corpora were created in cooperation with private companies.[16] In Taiwan, the CKIP (Chinese Knowledge and Information Processing) group was formed in the Academia Sinica to establish a fundamental research environment for Chinese natural language processing. The goal of CKIP was to construct research infrastructures with reusable resources that could be shared by domestic and international research institutes. Their accomplished sets of resources include Chinese electronic dictionaries, Mandarin Chinese corpora, and processing technologies for Chinese texts.[17] MAT (Mandarin across Taiwan) was a speech data collection project for creating telephone speech databases of Mandarin Chinese spoken in Taiwan.[18] In Hong Kong, Cantonese corpora were created for speech recognition, translation, and language understanding by a few universities.[19] Besides these, some research institutions in the United States also collected Mandarin Chinese data, such as the

CALLHOME Mandarin Chinese speech data collected by the LDC[3] and the Chinese speech data collected by the Johns Hopkins University.[20]

2. Chinese Spoken Languages, Transcription Systems, and Character Sets

Many dialects are spoken in China. Mandarin is a category of related Chinese dialects spoken in most of the northern, central, and western parts of China. However, *Mandarin,* as it is known to the world, refers to standard Mandarin (or modern standard Chinese) which is based on the Mandarin dialect spoken in Beijing. Standard Mandarin is the official spoken language known as *Putonghua* in mainland China and as *Guoyu* in Taiwan. In Singapore, Mandarin is one of the four official languages. Standard Mandarin is also one of the five official languages of the United Nations, and is used in many international organizations.[21] Putonghua and Guoyu are quite similar except in some areas of their vocabularies. Phonological descriptions show that the structural pattern of a Mandarin syllable is an optional initial consonant followed by the vowel, and then optionally followed by a velar or alveolar nasal ending. Another component of the Mandarin syllable is the tone which mainly specifies the syllable's pitch pattern. Technically, a syllable is presented in terms of its *initial*, *final*, and *tone*.[22] Mandarin is a tonal language because the tones, just like consonants and vowels, are used to distinguish words from each other.

Chinese linguists have proposed various transcription systems for Mandarin. But the most popular ones are *Hanyu Pinyin* and *Zhuyin Fuhao*. Hanyu Pinyin was accepted as the official transcription system for the Chinese language in 1958 by the government of China. Zhuyin Fuhao (or *Bopomofo*), a set of Chinese phonemic alphabets proposed in 1930, is used in Taiwan as an educational instrument for teaching the Chinese language. Both these transcription systems are used in the input of Chinese characters in computer systems.

Today, there are two Chinese character sets used by Chinese-language users, i.e., the traditional Chinese characters and the simplified Chinese characters. The traditional Chinese characters have been used since the 5th century. This character set is still being used in Taiwan and some overseas Chinese communities today. The simplified Chinese characters originate from the official character simplification during 1950s and 1960s. Now, this set of simplified Chinese characters is the official writing system in mainland China, and is accepted by the United Nations. In computer systems, different codes are used for these two character sets. The *Guobiao* code (GB) is a national standard character encoding in mainland China. It refers to the GB 2312-80 set issued in 1981, or the GB 18030-2000 set issued in 2000. There are 6,763 Chinese characters in the GB

3212-80 code set. *Big5* code is a character encoding method used in Taiwan for traditional Chinese characters. It contains 13,053 Chinese characters in its code set.

Mandarin Chinese is referred to as monosyllabic because the majority of words are one syllable in length. This is true for classical Chinese, but no longer true for modern Chinese. A large number of polysyllabic words are used today in daily spoken Chinese. One syllable when uttered with different tones corresponds to different characters. A word in polysyllabic form is written with two or more characters. Since Chinese texts have no spacing between words, extra effort is required to segment a sentence into word-parts. Because of these particular characteristics, the design of Chinese language corpora needs extra considerations. Most of the Chinese spoken language processing systems developed recently deal with standard Mandarin. Few of them cater for other dialects, such as Cantonese, Min-nan, Hakka, Wu, etc.[23]

3. Design of Chinese Language Resources

In general, speech corpus production involves the following procedures: (1) corpus specification, (2) preparation, (3) data collection, (4) postprocessing, (5) annotation, (6) pronunciation encoding, (7) documentation, (8) validation, and (9) distribution.[24,25] A corpus is created for the study of the language or the development of speech technology. The contents of the corpus should therefore be chosen and designed to achieve its purpose.

The specifications of a spoken language corpus includes defining the speaker profiles, the number of speakers, the spoken contents, the speaking style, the recording setup, the desired annotation, the recorded format, the corpus structure, and its validation procedure. Before data collection, preparations must be done. The instructions and prompting should be provided before the recording starts. Usually, an automatic process is used to control the actions in a recording session. A pre-test is necessary to ensure that the recording setup and equipment are functioning well. The final concern before going into the data collection phase is the recruitment of speakers. Corpus builders should obtain a sufficient number of speakers for a given data collection task.

In the data collection phase, all processes of data collection must be documented. This logging can be done on paper or online. It is desirable to perform the pre-validation after a small amount of data is collected. Besides, an ongoing quality control is necessary to detect systematic errors. The recorded data must be safely stored. Then, the recorded raw signal data are post-processed typically by the steps of file transfer, file name assignment, editing, and error

detection. Sometimes, resampling of the signals may be required to get desired sampling rate or to make format conversions.

The annotation of a speech corpus includes the following processes: segmentation and labeling, transcription and tagging, and internal validation. Segmentation is a process to get a combination of time information and categorical content. Segmental units can be phones, syllables, morphemes, words, prosodic categories, or dialogue acts. Manual segmentation is believed to be more accurate, but is extremely expensive, time- and effort-wise. Automatic or semi-automatic segmentation done by using a software tool is desirable to process a large database. Transcription is a process to represent speech in terms of its semantic contents. Typically, a recorded item is transcribed into a chain of words. Tagging refers to the markup of categorical classes on words. Some software tools can be used for transcription. These annotations should undergo a verification step to ensure good quality. The internal validation process is ideally performed by a single person or a well-trained group to ensure consistency. Documentation must be made to summarize all relevant information regarding the production and usage of the corpus. Finally, a validation process is further conducted to validate the documentation, signal data, annotation data, readability, and quality.

The production of language corpora for Chinese spoken language processing (CSLP) involves almost the same procedures as described above. Since there are many dialects spoken in different regions across China, accent differences do affect pronunciation when speaking in Mandarin. For example, the Mandarin spoken in Taiwan is somewhat different from the Mandarin spoken in Beijing, not only in terms of accent, but also in terms of the vocabulary used. For this reason, a Mandarin speech corpus should be a dialect-specified corpus.

In recent years, much language corpora have been designed by collecting data from radio and television broadcasts.[26] Some text data are collected from the internet. The option of obtaining spoken and textual data available in public channels and networks provides a quick way to gather large amounts of data. However, the collection of data alone does not imply the creation of a corpus, unless the contents are well defined and organized to meet a specific purpose. Further, the major effort of corpus building lies in the annotation, documentation and validation of the collected data. In addition to standard Mandarin, certain widely-spoken dialects need language corpora for developing their specific application systems. These also face annotation problems. Proper phonetic alphabet systems need to be developed as well.[27,28]

4. Activities in Developing CSLP Corpora

There are many organizations dedicated to the development of CSLP corpora. Among them, the ChineseLDC (Chinese Linguistic Data Consortium) is a nationwide, voluntary entity, legally-registered by researchers engaged in the creation and development of Chinese linguistic data. It is an academic and non-profit public association, with the aim of uniting various researchers in the CSLP area and producing Chinese linguistic databases to promote speech and language technology.[23] The ChineseLDC started with a project (Image, Speech, Natural Language Understanding and Knowledge Exploration Project) supported by the 973 Program (Program of National Key Foundation Research and Development) and the Chinese Hi-tech Research and Development Program (General Technical Research and Basic Database for the Establishment of the Chinese Platform). As a subordinate body to the Chinese Information Association, the ChineseLDC receives professional guidance and supervision from the association. Its office is located within the Institute of Automation, Chinese Academy of Sciences.

The goal of the establishment of the ChineseLDC is to set up a general linguistic database that is made up of the best quality Chinese databases that are currently available internationally. To achieve this goal, the ChineseLDC is creating and collecting open Chinese linguistic data that are highly integral, authorized, and systematic. It also targets data that cater to the requirements of various areas, such as lexicons, language corpora, data and instrumental references. This is to set a uniform series of standards and criteria for the users of these resources. While creating and collecting linguistic data, the ChineseLDC distributes existing data to various departments for educational, scientific research and governmental purposes, as well as for the development of industrial technology. The ChineseLDC also offers support to the fundamental research and application development in CSLP.

The Chinese Corpus Consortium (CCC), founded in 2004, is another organization for the distribution of Chinese language corpora.[20] The CCC has been sponsored by a group of universities, institutes, and private companies. Current sponsors include:

- Beijing d-Ear Technologies Co., Ltd. (d-Ear), China.
- Center for Speech Technology, Tsinghua University (CST), China.
- Institute of Linguistics, Chinese Academy of Social Sciences (CASS), China.
- Human Computer Interaction and Multimedia Lab, Tsinghua University, China.
- Chinese & Oriental Language Information Processing Society (COLIPS), Singapore.

- Dept. 1, ATR Spoken Language Translation Research Labs, Japan.
- Center for Language & Speech Processing, The Johns Hopkins University, USA.
- The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China.

The CCC aims to provide language corpora for the areas of Chinese automatic speech recognition (ASR), text-to-speech (TTS) synthesis, natural language processing (NLP), perception analysis, phonetic analysis, linguistic analysis, and other related tasks. The functions of the CCC include:

- Collecting and integrating existing Chinese speech and linguistic corpus resources, and continuing the creation of such resources.
- Integrating existing tools for the creation, transcription, and analysis of Chinese speech and linguistic corpus resources, improving their usability, and creating new tools.
- Collecting, organizing and introducing the specifications and standards for Chinese speech and language research and development.
- Promoting the exchange of Chinese speech and linguistic corpus resources.

Headquartered in Beijing, China, the CCC is supported by the Chinese Language Resources branch of the High-tech Enterprises Association of the Beijing Experimental Zone for the Development of New Technology Industries (HTEA), and receives supervision, inspection and management from the Beijing Municipal Commission of Science and Technology and the Beijing Social Organization Managing Office. Under the guidance of the HTEA, the CCC works for the mutual promotion of the standardization and industrialization of Chinese language resources.

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), established in Taipei in 1988, is a non-profit organization. Its goals are to conduct research in computational linguistics, to promote the utilization and development of computational linguistics, to encourage research in the field of Chinese computational linguistics both domestically and internationally, and to maintain contact with international groups who have similar goals as well as to cultivate academic exchange. To promote resource sharing, the ACLCLP also releases a wide variety of Mandarin corpora including the Sinica Corpus, the CKIP lexicon, the Chinese News Corpus, the Sinica Treebank, the Chinese Information Retrieval Testing Corpus and the Mandarin Speech Databases.[29,30] After more than a decade of effort, the International Journal for Computational Linguistics and Chinese Language Processing

(IJCLCLP) published by the ACLCLP has become one of the most important journals specializing in Chinese computational linguistics.[31]

In Singapore, the Chinese and Oriental Languages Information Processing Society (COLIPS) is a non-profit professional organization formed to advance the science and technology of information processing in Chinese and other similar oriental languages. One of the objectives is to promote the free exchange of information relating to information processing of these languages in the best scientific and professional tradition. In the past, COLIPS has organized international conferences on Chinese and Oriental Languages, computer exhibitions and Chinese input competitions. It has also held short courses and talks for members and the public at large. Its society journal, the Journal of Chinese Language and Computing (JCLC), published four times a year, is circulated world-wide.[32]

The LDC, established in the University of Pennsylvania, USA, creates and distributes corpora of many languages. These do include Chinese speech and text corpora.[3] The typical ones are the CALLFRIEND and CALLHOME Mandarin Chinese corpora collected through telephone networks. There are also Mandarin Chinese (and multilingual) text, transcripts, lexicon, and broadcast news corpora.

Many universities have produced their own Chinese Language corpus for CSLP, such as the Cantonese spoken language corpus,[20,33] developed at Chinese University of Hong Kong, the Chinese Spontaneous Speech and Wu dialectal speech[20] developed at the Johns Hopkins University, and the multilingual speech corpus for Min-nan, Hakka, and Mandarin[34] developed at the Chang Gung University, Taiwan. CSLP corpora for various application domains have been produced in recent years. Typical applications include in-car conversations, mobile phone conversations, hotel reservations, spoken dialogues, and information retrieval.[20,23,26,35]

5. Available Language Resources for CSLP

Much CSLP corpora are distributed by the language resource associations, such as LDC, ChineseLDC, CCC, and ACLCLP. Some of them are categorized and listed as follows.

(1) *Telephone Speech*

- CALLFRIEND Mandarin Chinese-Mandarin Dialect (LDC)
- CALLFRIEND Mandarin Chinese-Taiwan Dialect (LDC)
- CALLHOME Mandarin Chinese Speech (LDC)
- Hub-5 Mandarin Telephone Speech (LDC)

- TSC973 – Telephone Speech Corpus (ChineseLDC)
- Telephone speech corpus for speech recognition (ChineseLDC)
- The identifiable speech database of telephone speech – the name of person, the name of place (265 people using mobile telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – the name of person, the name of place (285 speakers using stable telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – number strings (265 people using mobile telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – number strings (285 speakers using stable telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – stocks (265 people using mobile telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – stocks (285 people using stable telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – messages (64 people using mobile telephone) (ChineseLDC)
- The identifiable speech database of telephone speech – messages (86 people using mobile telephone) (ChineseLDC)
- CSTSC-Flight Corpus – Chinese Spontaneous Telephone Speech Corpus in the Flight Enquiry and Reservation Domain (CCC)
- TRSC – 500-people Telephone Read Speech Corpus (CCC)
- BIT-TeleSpeech – Telephone Read Speech Corpus (CCC)
- CHRD – Chinese Hotel Reservation Dialogue (CCC)
- MAT-160 – Mandarin spoken in Taiwan, 160 persons (ACLCLP)
- MAT-400 – Mandarin spoken in Taiwan, 400 persons (ACLCLP)
- MAT-2000 – Mandarin spoken in Taiwan, 2,000 persons (ACLCLP)
- MAT-2500Ext – Mandarin spoken in Taiwan, 2,500 persons (ACLCLP)

(2) ***Broadcast Speech***

- 1997 Mandarin Broadcast News Speech (LDC)
- TDT2 Mandarin Audio (LDC)
- TDT3 Mandarin Audio (LDC)
- Natural Broadcasting Speech Corpus (ChineseLDC)
- CASIA – The Weather Forecast Broadcasts (ChineseLDC)

(3) ***Mobile Phone Speech***

- BIT-MobileSpeech – Mobile Phone Speech Corpus for Traffic Information Query (CCC)

- BIT-MobileTalk – Mobile Phone Conversational Speech Corpus for Travel (CCC)
- BIT-TonalName – Tonally Confusing Name Speech Corpus (CCC)

(4) ***Microphone Speech***

- Tsinghua- Corpus of Speech Synthesis (ChineseLDC)
- ASCCD – Annotated Speech Corpus of Chinese Discourse (ChineseLDC, CCC)
- CADCC – Chinese Annotated Dialogue and Conversation Corpus (ChineseLDC, CCC)
- SCSC – Syllable Corpus of Standard Chinese (ChineseLDC, CCC)
- WCSC – Word Corpus of Standard Chinese (ChineseLDC, CCC)
- CASIA – Chinese Question Structures Corpus (ChineseLDC)
- CASIA – Chinese Emotion Speech Corpus (ChineseLDC)
- Chinese Part-of-Speech Tagged Corpus (ChineseLDC)
- Chinese Geographic Name Corpus (ChineseLDC)
- CASIA – Single Syllable Isolated Word Speech Corpus (ChineseLDC)
- CASIA – Northern China Accent Speech Corpus (ChineseLDC)
- CASIA – Southern China Accent Speech Corpus (ChineseLDC)
- CASIA – Mandarin Continuous Digit Speech Corpus (ChineseLDC)
- CASIA – Chinese Speech Synthesis Corpus (ChineseLDC)
- RASC863-annotated 4 regional accent speech corpus (ChineseLDC)
- Chinese and English speech corpus (ChineseLDC)
- Special Scene and special domain dialogue corpus (ChineseLDC)
- The identifiable speech database of Chinese Mandarin – wide label (ChineseLDC)
- Identifiable speech database of Chinese Mandarin – extract database (ChineseLDC)
- Identifiable speech database of tabletop speech – messages (200 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – number strings (200 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – number strings (100 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – messages (120 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – number strings (120 persons) (ChineseLDC)

- Identifiable speech database of tabletop speech – people names, place names (120 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – stocks (70 persons) (ChineseLDC)
- Identifiable speech database of tabletop speech – topics (50 persons) (ChineseLDC)
- CACSC – Cantonese Accent Chinese Speech Corpus (CCC)
- CUCorpus – Cantonese Spoken Language Corpus (CCC)
- CASS – Chinese Annotated Spontaneous Speech (CCC)
- WDCS – Wu-dialectal Chinese Speech (CCC)
- BIT-MonoSyllable – Mandarin Mono-Syllable Corpus (CCC)
- CCC-VPR2C – 2-channel Corpus for Voiceprint Recognition (CCC)
- CCC-VPR3C – 3-channel Corpus for Voiceprint Recognition (CCC)
- CCC-VPR27C – 27-channel Corpus for Voiceprint Recognition (CCC)
- CCC-VPR36C – 36-channel Corpus for Voiceprint Recognition (CCC)
- TCC-300 – Mandarin Speech, 300 persons (ACLCLP)
- Sinica MCDC – Mandarin conversations (ACLCLP)

(5) ***Multiple Language Speech***

- CSLU: 22 Languages Corpus (LDC)
- TDT4 Multilingual Broadcast News Speech (Arabic-Chinese-English) (LDC)
- Chinese-English Sentence aligned Bilingual Corpus (ChineseLDC)
- Parallel Language Corpus for the Olympics (Chinese-English-Japanese) (ChineseLDC)
- Parallel Language Corpora (Chinese-English, Chinese-Japanese) (ChineseLDC)

(6) ***Chinese Text***

- HKUST Mandarin Telephone Transcript Data (LDC)
- TREC Mandarin (LDC)
- Mandarin Chinese News (LDC)
- Chinese Gigawords (LDC)
- Hub-5 Mandarin Transcripts (LDC)
- Chinese Treebank (LDC)
- Chinese Proposition Bank (LDC)
- Tsinghua Chinese Treebank (ChineseLDC)
- Academic Sinica Balanced Corpus (ACLCLP)
- Sinica Treebank (ACLCLP)

- Sinica BOW (ACLCLP)
- Word List with Accumulated Word Frequency in Sinica Corpus (ACLCLP)

(7) ***Parallel Text***

- Chinese-English News Magazine (LDC)
- Hong Kong News Parallel Text (Chinese-English) (LDC)
- Hong Kong Laws Parallel Text (Chinese-English) (LDC)
- Hong Kong Parallel Text (LDC)
- Multiple Translation Chinese (Chinese-English) (LDC)
- TDT4 Multilingual Text and Annotations (Arabic-Chinese-English) (LDC)

(8) ***Dictionary***

- Chinese-English Olympics Dictionary (ChineseLDC)
- Modern Chinese Semantic Dictionary (ChineseLDC)
- Modern Chinese Semantic Dictionary based on International Logical Model (ChineseLDC)
- Chinese Electronic Dictionary (ACLCLP)

(9) ***Lexicon***

- Chinese Lexicon (ChineseLDC)
- Reference Lexicon for Segmentation Standard Dictionary (ACLCLP)
- Standard Segmentation Corpus (ACLCLP)
- CKIP Lexicon and Chinese Grammar (ACLCLP)
- The Grammatical Knowledge-base of Contemporary Chinese (High Frequency Words) (ChineseLDC)

(10) ***Evaluation Data***

- 863 program in 2003 speech recognition evaluation data (ChineseLDC)
- 863 program in 2004 speech recognition evaluation data (ChineseLDC)
- 863 program in 2003 speech synthesis evaluation data (ChineseLDC)
- 863 program in 2004 speech synthesis evaluation data (ChineseLDC)
- 863 program in 2003 machine translation evaluation data (ChineseLDC)
- 863 program in 2004 machine translation evaluation data (ChineseLDC)
- 863 program in 2003 automatic index evaluation data 8 (ChineseLDC)
- 863 program in 2004 automatic index evaluation data (ChineseLDC)
- 863 program in 2003 Assessment and test data of text classification (ChineseLDC)

- 863 program in 2004 Assessment and test data of text classification (ChineseLDC)
- 863 program in 2003 Assessment and test data of chinese recognition (ChineseLDC)
- 863 program in 2004 information index evaluation data (ChineseLDC)
- 863 program in 2003 full text retrieval evaluation data (ChineseLDC)
- 863 program in 2003 name entry identification evaluation data (ChineseLDC)
- 863 program in 2003 part-of-speech evaluation data (ChineseLDC)
- 863 program in 2005 machine translation evaluation data (ChineseLDC)
- 863 program in 2005 information index evaluation data (ChineseLDC)
- 863 program in 2005 speech recognition evaluation data (ChineseLDC)
- CASIA98-99 speech testing library (ChineseLDC)

There are CSLP corpora currently being developed in various universities and research institutes. They should be available from these associations in the near future. It can be expected that more advanced technology will be applied to speed up corpus production. Language corpora will inevitably increase rapidly in size and number of types and categories due to the increasing diversity of speech processing applications.

6. Conclusion

The development of language corpora is a major part in the advancement of natural language and speech processing technologies. Building a re-usable, expandable, and consistent speech corpus for research and development purposes is a requirement for improving the research infrastructure. For CSLP, we need more good-quality language corpora to boost NLP and speech processing research and techniques. In other words, more effort should be channeled to promote the production of good language corpora in the future.

References

1. C. R. Huang, "Corpus-based studies of Chinese linguistics", *Computational Linguistics and Chinese language Processing*, vol.2, no.1. (1997)
2. M Liberman and C Cieri, "The Creation, Distribution and Use of Linguistic Data", *Proc. First International Conference on Language Resources and Evaluation*. (1998)
3. LDC – Lingustic Data Consortium. http://www.ldc.upenn.edu/
4. K. Choukri, "European language resources association: History and recent developments", *Proc. Oriental COCOSDA Workshop 1999*, (1999), pp. 15-23.
5. ELRA – European Language Resources Association. http://www.elra.info/

6. S. H. Chen, S. H. Hwang and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE Trans. on Speech & Audio Processing*, vol. 6, (1998), pp. 226-239.
7. K. W. Gan, K. T. Lua and M. Palmer, "A statistically emergent approach for language processing: application to modeling context effects in ambiguous Chinese word boundary perception", *Computational Linguistics,* (1996), vol.44, pp. 531-553.
8. L. S. Lee., C. Y. Tseng and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system", *IEEE Trans. On Acoustics, Speech, & Signal Processing,* vol. 37, (1989), pp. 1309-1320.
9. L. S. Lee, "Voice dictation of Mandarin Chinese", *IEEE Signal Processing Magazine*, vol. 14, no. 4. (1997).
10. T. Lee and P. C. Ching, "Cantonese syllable recognition using neural networks", *IEEE Trans. On Speech & Audio Processing*, vol. 7, (1999), pp. 466-472.
11. T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng and B. Mak, "Tone recognition of isolated Cantonese syllables", *IEEE Trans. On Speech & Audio Processing*, vol. 3, (1995), pp. 204-209.
12. Y. R. Wang and S. H. Chen, "Tone recognition of continuous Mandarin speech assisted with prosodic information", *J. Acoustical Society of America*, vol. 96, (1994), pp. 2637-2645.
13. L. Du, "Recent activities in China: Speech corpora and assessment", *Proc. Oriental COCOSDA Workshop 2000*,. (2000)
14. A. Li, Y. Zu, Z. Li, "A national database design and prosodic labeling for speech synthesis", *Proc. Oriental COCOSDA Workshop,* (1999)
15. B. Xu, T. Y. Huang, X. Zhang and C. Huang, "A Chinese spoken dialogue database and its application for travel routine information", *Proc. Oriental COCOSDA Workshop 1999*. (1999)
16. C. Zheng, X. Liu and Z. Li, "A Chinese database for network service", *Proc. Oriental COCOSDA Workshop 1998*. (1998).
17. CKIP -- Chinese Knowledge and Information Processing Group, Academia Sinica. http://godel.iis.sinica.edu.tw/new/
18. H. C. Wang, "MAT—A project to collect Mandarin speech data through telephone networks in Taiwan", *Computational Linguistics and Chinese language Processing*, vol.2, (1997), pp. 73-90..
19. S. Li, F. Zheng, M. Xu, Z. Song and D. Fang, "A Cantonese accent Chinese speech corpus", *Proc. Oriental COCOSDA Workshop,* (1999).
20. CCC -- Chinese Corpus Consortium. http://www.CCCForum.org
21. Encyclopædia Britannica. 2006. "Chinese languages." http://www.britannica.com/eb/article?tocId=75050
22. C. N. Li and S. A. Thompson, Mandarin Chinese : A functional reference grammar, University of California Press. (1981)
23. Chinese LDC – Chinese Linguistic Data Consortium. http://www.chineseldc.org/
24. M. Wynne, Ed., *Developing Linguistic Corpora: a Guide to Good Practice*, AHDS: Arts and Humanities Data Service, (2006). http://www.ahds.ac.uk/.
25. F. Schiel and C. Draxler, The Production of Speech Corpora, Version 2.5 , *BAS- Bavarian Archive for Speech Signals*, (2004). http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/node1.html
26. H. M. Wang, B. Chen, J. W. Kuo and S. S. Cheng, "MATBN – A Mandarin Chinese broadcast news corpus", *Computational Linguistics and Chinese language Processing*, vol.10, (2005), pp. 219-236.
27. C. Y. Tseng, "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan", *Proc. Oriental COCOSDA Workshop 1998*. (1998).

28. J. Zhang, "A SAMPA system for putonghua (Standard Chinese)", *Proc. Oriental COCOSDA Workshop 1999* (1999)..
29. H. C. Wang, "Speech research infra-structure in Taiwan – From database design to performance assessment", *Proc. Oriental COCOSDA Workshop 1999.* (1999)
30. H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, "MAT-2000 — Design, collection, and validation of a Mandarin 2000-speaker telephone speech database," *Proc. ICSLP 2000* (2000).
31. ACLCLP – The Association for Computational Linguistics and Chinese Language Processing
32. COLIPS – Chinese and Oriental Language Information Processing Society. http://www.colips.org/
33. T Lee, W. K. Lo, P. C. Ching, H. Meng, "Spoken language resources for Cantonese speech processing", *Speech Communication*, vol. 36, (2002), pp. 327-342.
34. R. Y. Lyu, M. S. Liang and Y. C. Chiang, "Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka, and Mandarin Chinese", *Computational Linguistics and Chinese language Processing*, vol. 9, (2004) , pp. 1-12.
35. H. C. Wang, C. H. Yang, J. F. Wang, C. H. Wu, and J. T. Chien, "TAICAR – The collection and annotation of an in-car speech database created in Taiwan", *Computational Linguistics and Chinese language Processing*, vol.10, (2005), pp. 237-250.
36. R. H. Wang, "National performance assessment of speech recognition system for Chinese", *Proc. Oriental COCOSDA Workshop 1999*. (1999)
37. AHDS -- Arts and Humanities Data Service. http://www.ahds.ac.uk/.
38. BAS- Bavarian Archive for Speech Signals. http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/node1.html
39. COCOSDA – The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques. http://www.slt.atr.co.jp/cocosda/
40. IPA – The International Phonetic Association. http://www.arts.gla.ac.uk/IPA/ipa.html
41. NIST Speech Group. http://www.nist.gov/speech/index.htm
42. O-COCOSDA – Oriental COCOSDA. http://www.slt.atr.jp/o-cocosda/org.html
43. Wikipedia, "Mandarin," http://en.wikipedia.org/wiki/MandarinChinese