International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

The Study of Construction of Language Informatization

Hulin Ren[1]✉ and Luqi Ge[2]
[12]School of Foreign Studies, University of Science and Technology Beijing, China
Corresponding Author: Hulin Ren, **E-mail**: hulinr@aliyun.com

| ABSTRACT
The paper is to explore the construction of language informatization in China. By means of introducing the concept of language informatization, the paper investigates evolutional development and application of language informatization technology in China. The paper finds that technical specifications of language informatization processing is necessary in four aspects, which are beneficial for advanced development of Chinese language informatization in the Artificial Intelligence era.

| KEYWORDS
language informatization, construction, application and technology, development and processing

| ARTICLE INFORMATION
ACCEPTED: 02 November 2024 **PUBLISHED:** 26 November 2024 **DOI:** 10.32996/ijllt.2024.7.12.1

1. Introduction

Language informatization, one primary foundation for the national strategy of informatization in China, includes constructing and stipulating linguistic and technical standards for information processing, promoting innovation and development of language informatization technology as well as building language data resources and boosting the digital development for application of language resources (Xiang, 2017). All of which should be an important mission for accomplishing better careers within the field of national languages. With the initiative to meet the requirements of the national strategy of language informatization, China has been strengthening its construction of this field since the 1980s; exerting positive influences on national social life with a series of beneficial achievements and making significant contributions to the national strategic plan of informatization construction.

Great progress was made in the national construction of Chinese language informatization in 2016. The intelligent conversion system from simplified graphemes to traditional graphemes (stage two), sponsored and developed by the National Language Committee of China, has successfully passed the project appraisal and identification, and has been granted the *Qian Weichang Award in Chinese Informatization Processing Technology* issued by Chinese Informatization Processing Society of China (CIPSC). Meanwhile, some research projects, such as *Research on the Application of Intelligent Voice and Artificial Intelligence in Language Learning*, and *On Error Analysis of English-Chinese Machine Translation* and *the Solution to the Passage-oriented Machine Translation*, have been approved to promote the technological development of language informatization, such as language intelligence, assisted learning and machine translation, etc. In addition, the first-stage construction of *Language Resource Service Platform of the National Language Committee of China* has also started to further the basic resource constructions for language informatization, such as the commonly used Chinese grapheme holographic database, and the dynamic circulation corpus of national language resources, etc.

2. Development and application of language informatization technology

Language informatization technology is the key to the construction and development of language informatization. It is an important part of artificial intelligence and can be stratified at two levels: processing graphemes or font patterns and processing words, phrases, sentences, and passages. The former is designed to input Chinese and other minority graphemes into computers

for retrieval, call, edition, display or output. The latter one, for the purpose of semantic comprehension, is to satisfy the requirements of intelligent application, such as machine translation and automatic questions and answers.

2.1. Character processing technology

The Chinese language has a large number of morpheme characters. Since the 1980s, more than seventy thousand graphemes have been successfully installed into computers. The processing technology of Chinese characters involves two aspects; the Chinese input technology, which facilitates the accurate and fast retrieval and call of different Chinese graphemes from computers, and Chinese output technology, which displays or prints Chinese graphemes onto screens or paper.

Moreover, significant improvement has been accomplished in the conversion between simplified and traditional forms of Chinese graphemes. Generally speaking, the well-developed processing technology of Chinese graphemes can meet the basic daily needs of users who wish to process Chinese graphemes using their computers and electronic devices.

2.1.1 Technology for the input of Chinese graphemes

Technology for the input of Chinese graphemes includes keyboard input technology and optical grapheme recognition technology. Keyboard input technology began in the 1970s, with the purpose to realize the scientific dream of putting Chinese graphemes into computers. Under the guidance and support of the Chinese government, many researchers exerted their utmost efforts to ensure the fast development of keyboard inputting of Chinese graphemes (Qi, 1981). According to statistics, the number of registered patent technologies on Chinese coding and inputting was close to the amount of three thousand， including almost a hundred of which had been applied in different fields at different periods of time. The input technology or system mentioned above mainly involves two technology routes， visual code and acoustic code, such as the *Wubi Method, Chinese Shuangpin Input Method, Intelligent ABC, Microsoft Pinyin,* etc. Currently, Pinyin input methods using acoustic coding technology, are widely used in China, as they have the advantage of ensuring to input exact meaning of one given sentence, while ensuring higher level of sentence intelligence. Thanks to the low cost and high efficiency, 95% of the existing market shares have been enjoyed by such methods, such as *Sougou Input Method, Baidu Input Method,* etc.

Optical grapheme recognition (OGR) technology started at the end of the 1970s in China. At present, technology for the optical recognition of printed graphemes and online hand-written grapheme recognition have been put into application and implemented. The accuracy of recognizing high-quality printed text is rising to more than 99%. The enterprises, including Baidu, Hanvon, Thunis, iFLYTEK CO., LTD (iFLYTEK), Sinovoice, etc., have developed and launched proven grapheme recognition products to satisfy the requirements in the corresponding fields of secretary, education and library, etc., laying the foundation for higher-level application, such as image retrieval and translation by means of image recognition (Wang, et al., 2000).

2.1.2 Technology for the output of Chinese graphemes

The laser photo-type setting technology enabled the method of outputting Chinese graphemes to be developed in the 1980s. The library of Chinese graphemes experienced a rapid development, with the emergence of dozens of font vendors in the 1990s, such as Founder, Hanyi Font, Huawen, Huaguang, Zhongyi, Stone, the Great Wall, etc. Since the start of this new Millenium, the industry responsible for the library of Chinese graphemes has enjoyed a substantive leap, with the emergence of not only different versions of text-type fonts facilitating easy typography and reading, but also having more than 200 patterns of innovative, calligraphic fonts to satisfy the requirements of different designs. Currently, there are more than 600 types of Chinese font patterns in circulation.

Additionally, the continuous technological innovation of the library industry in charge of developing the Chinese graphemes database also benefits from the development of internet technology, such as font compression technology which solves the storage problem of mobile devices, and hint directives technology which clarifies the on-screen display of small-size graphemes as well as the font library cloud service which makes font embedding possible on web pages.

The Chinese State Press and National Publication Administration launched the project of Chinese Fonts in 2006, for the purpose of searching, organizing, coding and constructing a large scale font library system to cover ancient and modern Chinese graphemes including those in the minority (Xiang,2017). The library includes approximately 300 thousand graphemes.

2.1.3 Conversion technology between the simplified and traditional Chinese graphemes

Ever since 2012, the intelligent conversion system from simplified graphemes to traditional graphemes (S2T) has been studied and developed through joint efforts by Xiamen University of Fujian Province, China, the Institute of Language Application Research in the Ministry of Education, and Beijing Normal University, supported by the Ministry of Education and the National Language Committee of China. The system, applicable to all the Chinese graphemes in both the common standard Chinese grapheme table

and the international standard Unicode 8.0, converts simplified graphemes to their corresponding traditional ones at different levels, such as graphemes, words and phrases as well as professional terms, punctuation and so on, which provides services of online conversion and whole website conversion. The system can implement two different kinds of S2T conversion, namely, from simplified graphemes to traditional ones in Taiwan and Hong Kong, as well as from modern simplified graphemes to traditional ones in ancient books. The public can enjoy such services free of charge. The uni-processor version had been downloaded around 148 thousand times from November 2014 to March 2017; and the web version dealt with 22 million conversion requests at the average of 25.7 thousand each day in the same period.

The first stage of the system development was launched in November 2014, and the second was accomplished in June 2016. In the same year, the system was granted with the *Qian Weichang Award in Chinese Informatization Processing Technology* which was issued by the CIPSC.

3. Language processing technology

In recent years, language processing technology has grown rapidly, with relatively more fruitful achievements in voice and text processing, machine translation and knowledge graphs. The fast growth of this language processing technology has substantially enriched language life of the society, and raised new requirements for linguistic related research to move forward. Voice technology in China mainly includes voice synthesis, voice recognition and speaker recognition.

Recently, rapid improvement has been made in the quality of voice output and the naturalness of voice synthesis systems, so as to meet the requirements of various types of specific situations and implement the wide-range application of the technology in different fields, such as public information broadcasting, navigation and interactive voice response, etc. Furthermore, such applications will be gradually expanded to other fields such as entertainment, phonetic instruction and rehabilitation therapy.

Symbolizing the world-leading position attained by the national voice synthesis technology, iFLYTEK has won the first prize in different challenges and competitions for consecutive years with its own independent researched-and-developed (R&D) technology at BlizzardFest, and the world famous evaluation contest of voice synthesis.

3.1.2 Voice recognition

Up until now, the application of voice recognition in China has entered a booming era. For instance, the identification of near-field low noise is close to the level of human being. At the 4th Computational Hearing in Multisource Environments (4th-CHiME) 2016, iFLYTEK won all the first prizes in the competition test of setting, the grapheme error rates of its independent R&D technology have reached the unprecedented low level, that is, reducing from 9.15% to 2.24% in different types of fields. Meanwhile, some other Chinese businesses, such as Baidu, Sinovoice, COLOUD, etc., have also developed products with their own distinctive features in the field of voice recognition.

The use of the established voice recognition technology in China has contributed to the innovation of technology and services in many other fields, such as language education (Mandarin training and testing in particular), public service, e-commerce, personal assistance and secretarial work as well as Chinese national security, etc.

3.1.3 Speaker recognition

As a relatively better established modern applicable technology, speaker recognition can satisfactorily meet the requirements of different situations, such as search and rescue, shipping, radio and television, public security, etc. The technologies and products developed by the Institute of Automation, Chinese Academy of Sciences (IACAS) and iFLYTEK have already matched the advanced standards around the world. The aggregate indicators of recognition systems developed by iFLYTEK were ranked first in 2008 and second in 2010 at the Speaker Recognition Evaluation held by National Institute of Standards and Technology (NIST).

3.2 Text processing technology

Text processing technology can be stratified into three levels: lexical analysis, syntactic analysis and semantic analysis. Lexical analysis contains Chinese word segmentation and parts of speech annotation. Syntactic analysis refers to the attainment of syntactic structures by means of automated analysis of sentences. Semantic analysis is to comprehend the true meaning of one given sentence.

3.2.1 Technology of lexical and syntactic analysis

Currently, the major methods of technology in analyzing words and sentences involve statistics, in-depth learning, to train automatically and construct lexical and syntactic analyzing systems in word segmentation and parts of speech annotation and tree-bank. In Chinese Word Segmentation Evaluation held by ACL SIGHAN in 2014, the F1 value of the most optimized system reached 97.3% accuracy at the setting competition test and 71.8% accuracy in the same year test at Combined Category Grammar

Setting in Chinese syntactic analysis evaluation. Basically, technologies of Chinese word segmentation, parts of speech annotation and syntactic analysis can be used to facilitate the upper and further applications, including semantic analysis, information retrieval and extraction (Xing, 1985).

3.2.2 Technology of semantic analysis

Semantic analysis is one focal and difficult point of text analysis. At its current level, Chinese technology of shallow semantic parsing plays an important part in information retrieval, extraction and interactive voice response. Nevertheless, the in-depth semantic parsing is difficult; the semantic analysis at sentence and text level cannot achieve satisfactory results. The maximum F1 value of semantic marking by the most optimized system reached 68.6% accuracy at ACL SemEval in 2016.

3.2.3 Technology of machine translation

Ever since the start of the 1950s, machine translation has gone through a long development and now enters a new practical stage to benefit society and the general public as a whole. It is supported by the Twelfth Five-year Plan and National "863" Projects in China, Institute of Automation, Chinese Academy of Science (CASIA), Zhejiang University, Harbin Institute of Technology, Institute of Computer Science, Chinese Academy of Science (CASICS) and Tsinghua University. Under the strong support of Baidu, great efforts have been implemented to ensure the industrialization project of machine translation can be based on big data statistics from the internet. Such implementation was awarded the second prize at the Chinese National Science and Technology Progress Award in 2015.

A bunch of machine translation products with their basic applicable functions were put on the market by many Chinese national enterprises in 2016, such as Baidu, Sougou, iFLYTEK, Tencent, Youdao, etc, indicating considerable progress in machine translation and an important role in many fields where foreign and minority languages need to be noticed.

3.2.4 Knowledge graph

In recent years, the knowledge graph, as the basic infrastructure for the internet intelligence service, has demonstrated its strong power in intelligent Q&A and brought dynamics to the internet semantic retrieval. In the field of intelligent Q&A, the answers are retrieved from the existing structured knowledge bank after the semantic analysis of the questions whose non-structured sentences are parsed into structured query statements. Semantic retrieval functions to improve the accuracy of the search results by means of large-scale knowledge graph to perform semantic marking to the keywords, and the text content are searched for by users (Cheng, 1992). The typical applications of knowledge graph are *ZhiXin* developed by Baidu and *Zhili Fang* by Sougou. In addition, the knowledge graphs in specific fields are also developed by many colleges, universities and science institutes within China.

3.3 Standard specifications for language informatization

Standard specifications for language informatization consist of Chinese coding standards, Chinese fonts standards and technical specifications of language informatization processing.

3.3.1 Chinese coding standards

Chinese coding, the plan and rule of Chinese storage in computers, is one crucial step to informatize Chinese inputting into a computer, and also one of the most important parts of the standardization of Chinese language. Since the 1970s, many standard specifications for solving coding problems of Chinese graphemes have been compiled and renovated by the departments interested in language informatization and standardization. The ten volumes of national standards and one industry standard can satisfactorily solve the problems challenging Chinese language storage, switch and processing within computers in practice (Qi, 1981).

With the developmental experience of forty years, coding specifications of the Chinese language can successfully meet the social requirements by incorporating the minority languages, and have been applied universally in China and are now in line with the international norms.

3.3.2 Chinese font standards

Chinese font standards, dot matrix and vector font specifications in particular, are the fundamental work for outputting Chinese informatization in virtual space. Font pattern refers to the image of individual Chinese grapheme in image output from a computer. The collection of the grapheme in the same pattern constitutes a font. Font standard specifications include setting grapheme standards, grapheme library formats, font patterns and font designs (Guo, 1983). The currently practiced twenty volumes of national standards concerning font (containing font information switching and font checking specifications) and nine volumes of industry standards can meet the social and technological requirements of Chinese grapheme displaying better. Corresponding

to the font standards, the setting of Chinese coding graphemes decides the range covering the font patterns (Tang, 1981). The standard font patterns in multiple styles and sizes for Chinese coding grapheme sets (GB 2312 and GB 18030), the CJK Chinese grapheme set and universally used multi-coded grapheme set (MCGS) have been stipulated in the existing national standard specifications. Meanwhile, the Song Typeface and Boldface in the MCGS are compatible with all Unicode have standardized font patterns in different sizes in line with the national standard. Finally, the standardization of the four basic printing typefaces, i.e. Song Typeface, Imitation Song, Kai Ti, Boldface, has been reputed the most frequently used base of Chinese coding grapheme set (Xing, 1990). It can be seen that technical specifications of language informatization processing are urgent in the Artificial Intelligence era.

4. Technical specifications of language informatization processing

Standard specifications vary from technologies of different language informatization processing. Basic and well-established technologies are relatively more advanced and better-rounded, while cutting-edge and fast-developing technologies are relatively soft and technologically oriented, focusing on the evaluation.

4.1 Standard specifications for word processing technology

Word processing is the most basic language processing task aside from grapheme processing. Currently, there are two major word processing technologies: Chinese word segmentation and part-of-speech tagging in line with the national standards (Zhang, 1989). They function as the guiding force for many language informatization tasks, including voice recognition, machine translation and systematic construction of corpus. Regarding standard specifications for voice technology, voice recognition and voice synthesis are two important directions for language informatization. The functions and features are all standardized in terms of their features, and thus have two different national standard specifications.

4.2 Standard specifications for technology-oriented evaluation

The major boosting power for fast developing cutting-edge technologies is public tests being held regularly, so that the public can test their products and give opinions for improvement. Therefore, it bears the great significance that standards should be constructed for the cutting-edge technologies and the related tests. In recent years, the technology of eight-test standards has been formulated in China (Xu, 1990). 4.4 As for language informatization standards for keyboard input, handsets and user interface, the national and industry standards have also been drafted in China to perform the guiding task of solving language processing problems in voice recognition, keyboard input, handsets and the user interface of informatization system (Zhu, 2005).

5. Conclusion

The paper aims to examine the construction of language informatization in China. In this context, the concept of language informatization is introduced, and the evolutional development and application of language informatization technology in China are investigated. The paper finds that four aspects of technical specifications of language informatization processing are crucial, i.e. standard specifications for word processing technology and voice technology, standard specifications for technology-oriented evaluation, as well as Language informatization standards for keyboard input, handsets and user interface, all of which are the solid foundations for further improvement of Chinese language informatization in the Artificial Intelligence era.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Cheng, X. (1992). *Lunheng fuyinci yanjiu*. Jinan: Shandong Education Press.
Guo, M. (1983). *Buci tongzuan*. Beijing: Science Press.
Qi, S. (1981). *Zhongguoshi tanyan* 中国史探研. Beijing: Zhonghua Book Company.
Tang, L. (1981). *Yinxu wenzi ji*. Beijing: Zhonghua Book Company.
Wang, Y et al. (2000). *Zhonggu hanyuyanjiu* 中古汉语研究. Beijing: The Commercial Press.
Xiang, X. (2017). Jianming hanyushi 简明汉语史. Beijing: The Commercial Press.
Xing, F. (1990). *Xiandai hanyu* 现代汉语. Beijing: Higher Education Press.
Xiang, X. (2017). Jianming hanyushi. Beijing: The Commercial Press.
Xing, F. (1985). *Fuju he xiangguanci*. Heierbin: Heilongjiang People's Publishing House.
Xu, Z. (1990). *Jiagu wenzi dianxu*. Chengdu: Sichuan Lexicographical Press.
Zhang, S. (1989). *Lvshi chunqiu cihui yanjiu*. Jinan: Shandong Education Press.
Zhu, Q. (2005). *Zhonggu Hanyuyanjiu*. Beijing: The Commercial Press.