SELECTED PAPERS OF APPLIED LINGUISTICS

● 刘涌泉

应用语言学论文选

SELECTED PAPERS OF APPLIED LINGUISTICS

应用语言学论文选

LIU YONGQUAN

刘涌泉

The Commercial Press

1989: BEIJING

内 容 提 要

信息化社会要求语言的应用和研究逐步计算机化、工程化。本书概括地介绍了语言工程的最新进展，着重论述了以下三方面的问题：（1）作为第五代计算机主攻方向之一的机器翻译；（2）作为中国进入信息化社会钥匙的中文信息处理；（3）作为科技现代化基础的术语。作者提出了问题、分析了难点，并探讨了解决问题的途径。附录所收两篇文章是作者的早期作品，由美国人译成英文，其中一篇谈机器翻译和文字改革的关系。本书供语言学、人工智能、计算机科学以及中文信息处理等方面的专家和爱好者使用。

YĪNGYÒNG YǓYÁNXUÉ LÙNWÉNXUǍN

应 用 语 言 学 论 文 选

著者 刘涌泉

责任编辑 杨枕旦

商 务 印 书 馆 出 版

（北京王府井大街 36 号）

新华书店总店北京发行所发行

三河县二百户印刷厂印刷

ISBN 7-100-00121-8/H·52

1989 年 8 月第 1 版　　开本 850×1168 1/32

1989 年 8 月北京第 1 次印刷　　字数 149 千

印数 1.750 册　　印张 5

定价：1.75 元

Preface

Along with the development of human society and the progress of science and technology, language is constantly changing and the communicative function of language is keeping on advancing too. The study of language — linguistics, as a rule, has to progress and be modernized. Applied linguistics as an important part of linguistics is no longer limited to the study of old problems, such as language teaching. Moreover, the emergence of electronic computer has brought a series of new problems (machine translation, Chinese information processing, etc.) which are geared to machines. To solve these new problems it is imperative to adopt a set of new methods and design a number of language information systems applicable to machine processing on the basis of traditional linguistics.

Papers in this book were produced just under such a background. The author of this booklet hopes to do a bit for the promotion of linguistic modernization.

Beijing
September, 1986

Liu Yongquan

Contents 目录

1

Language Engineering in China*

1. What is Language Engineering

The term of "language engineering" was mostly used to denote language reform and the establishment of standard language. In this paper, we propose to confine the same term to the implication of language engineering in the realm of computational linguistics, or rather the computerized language engineering. The mission of this engineering is to develop various information systems easy to use, by means of softwares and hardwares of computers and on the basis of the insight of modern linguistic studies.

2. The First Item of Language Engineering — Machine Translation

2.1 *Brief Review*

The first integration of computer and linguistics was Machine

* Paper presented to the 1983 International Conference on Text Processing with a Large Character Set, held at Tokyo, October 17—19 1983. In: *Proceedings of ICTP'83*, Tokyo, 1983. Also in: *Computer Processing of Chinese and Oriental Languages*, Vol. 1, No. 3, 1984.

Translation (MT). This kind of integration is both a success in a way and a failure in another. On the one hand, like sparks, it initiated the revolution of non-numerical application of computer, thus, providing a vast experimental plant,in which quite a lot linguistic theories and methods and technical artifices turned out on the basis or, by the revelation of such an integration. On the other hand, MT is a kind of rather advanced artificial intelligence; its sophistication requires not only the full comprehension of one natural language, but also the equivalent transformation from one into another. Up to the present, MT hasn't yet been put into wide application in any proper sense, it still fails to play an exemplary role. Today, various conditions needed have greatly been improved, the prospect of the actual application of MT can be envisaged. The research of language engineering in China also started with MT. Early in 1956, MT was incorporated in the Scientific Development Programme of China; the project was entitled "Machine Translation, Translation Rules of Natural Language and Mathematical Aspects of Natural Languages." In 1957, our MT research began officially with the Russian-Chinese translation algorithm. In 1959 we succeeded in testing its routines on computer. The output was on coded form rather than characters, for no Chinese character output device was available at that time (Gao 1959).

At the end of 1958, we launched out into an English-Chinese MT algorithm. By the beginning of 1960, the algorithm was established. However, its routines were merely checked manually, without trial on computer.

From 1966 to 1975, our MT research was at a standstill. In recent years, with the increasing perfection of technical conditions and constant deepening of language research, MT has made encouraging progress. Over ten algorithms have been worked out and tested on computers. There are, for instance, MT algorithms from English, Russian and French into Chinese and that from Chinese into foreign languages. About half of these algorithms were elaborated by our post-graduates in their dissertations. Some

put out the target texts in Chinese characters, others in phonetic alphabet. There are some more systems being exploited, one of which is the Japanese-Chinese MT algorithm. Practice demonstrates that we have already found out a set of sophisticated methods, by means of which one-to-one and many-to-one MT systems could successfully be established. At present, we are seeking after the ways for the transition of MT from an experimental stage into a tentative application.

2.2 *Personal Experiences*

2.2.1 *Aiming High and Embarking Our Studies by Stages.*

There is an old Chinese saying: "If you aim high, you get the median; if you aim median, you get the low." The destination we endeavour to reach is the full automatic MT. However, there is still a long way to go. Translation machines today have not been equipped with automatic reading devices, therefore, their operation is merely semi-automatic. The automatic level will be even lower, if pre-editing or/and post-editing are taken into account. In our opinion, it is unlikely advisable for humans to participate frequently in MT. Nevertheless, we can not completely do away with editings, at present, at least for the input. In some cases, for instance, it is still likely necessary for the typist or operator to discriminate the decimal point from a full stop. The post-editing or the non-marked post-editing would better be done away with. A translation machine can translate several dozens or hundred sentences per minute. It would be a rather tough job for post-editors to censor and check such a large amount of output texts with a mechanical tone. At the initial stage of MT application, it seems reasonable not to expect the high quality version, but a large amount of rough readable versions for browsing. Some of the rough versions can be submitted to a human translator for further careful translation, if they are found valuable. With the constant deepening

of our research work, the quality of MT outputs can surely be perfected step by step, and it is quite possible for us to get fairly good versions, which can be compared to the human-translated versions to a certain extent. At the initial stage of application, each sentence translated may, however, be marked with three different labels, i.e. valid, tentative or fail. When a sentence is marked "fail" for translation, its original is printed. The tentative version will be directly output with the label. Consequently, only this part of version is required for post-editing.

2.2.2 *Replacement of Word Order — Central Task in Foreign-Chinese MT*

The central task of MT from foreign languages into Chinese is that of replacement of word order (Liu 1959). From the very incipient stage of our MT research, we have laid our emphasis in syntactic analysis. To fulfil this task, we have proposed a series of principles and methods in sentence analysis, particularly the verb-predicate core theory. In accordance with this theory, sentences are made up of various hierarchical levels. The predicate is the largest relating center in a sentence, and it occupies a unique level — the highest or the first level. Constituents which directly relate to the predicate are known as immediate constituents, making up the second level. Constituents which directly relate to any other constituents than predicates are called non-immediate constituents, making up the third level. The theory also proposed to determine the axes of the replacement of word order. Predicates are the major axes for the replacement of word order while the quasi-predicates are the minor ones. Most of the quasi-predicates are non-finite forms of the verbs.

2.2.3 *Relevant Analysis and Independent Synthesis — Main Feature of the Majority of Our MT Algorithms*

Most of our MT projects have been designed in accordance with the principle of relevant analysis and independent synthesis.

On the basis of this principle, we set up a special system of intermediate constituents (Liu 1982). Such constituents can express not only the functional meaning but also the distributive relationship of the constituents (i.e. the relationship between constituents and the position of constituents in a sentence), and the contrastive differences between the two languages concerned. It has been demonstrated in practice that relevant analysis and independent synthesis is a valid principle in dealing with many-to-one translation.

2.2.4 *Processing of Prepositions — Crux in Foreign-Chinese MT*

The processing of prepositions is of crucial importance for MT from foreign languages into Chinese (Liu and Jiang 1982). In addition to indicating miscellaneous kinds of relation between the full words in a sentence, prepositions can further reveal and establish the content and character of these relations. Without prepositions or their equivalents (i.e. cases in languages where morphological declension is significant, or auxiliary words in Japanese) people can merely express simple ideas. Prepositions act as important intermediaries in a sentence; many of them are used in various phrases; and their syntactic functions are not alike and their semantic values quite varied. Consequently and so to speak without any exaggeration, the validity of translation depends to a great extent on the processing of prepositions.

2.2.5 *Analysis of Conjunctions and Punctuations — Important Means for Segmenting Discourse Strings*

The analysis of punctuations and conjunctions is an important means for the segmentation of discourse strings. Without accurate segmentation, there could be no correct comprehension and translation. Sentence in written languages are segmented by conjunctions and punctuations. In colloquial speech, punctuations give way to some phonetic features such as pauses, intonations, etc. The complexity of conjunctions (mostly the word "and" in English) lies in their three different functions — coordinate, non-

coordinate and quasi-coordinate conjunctions — which are difficult to identify (Liu 1983 C). The complexity of punctuations centers on the analysis of the comma. The petty comma (,) can serve as the mark of coordinate constituents, the mark separating clauses and the mark denoting the initial or final boundary of a syntagma (phrase segment). The so-called syntagma is a discourse string between two punctuations, it can be independent or parenthetic elements, and also be subordinate or parenthetic clauses in complex sentences (Liu, Liu and Gao 1979). In dealing with the analysis of conjunctions or punctuations, pure grammatical device is unlikely feasible; in many cases, semantic device is also indispensable.

2.2.6 *Classification of Words — the Fundamental Work*

The classification of words is the foundation for the designing of an MT algorithm. Word is the basic unit in natural languages and their processing. The information of classified words is the initial information for computer processing. The rational classification has a direct impact on the validity of computer operation. As a result, the classification of words is a fundamental task for the study of MT and other explorations of natural languages processing. In accordance with different keynotes and requirement,there can be various kinds of classification of words. Generally speaking, both grammatical and semantic classifications are indispensable to each information processing system. Considering the characteristics of MT, the transformation information must be involved in the classification of words. For a man-machine interactive system, knowledge information or extralinguistic information are indispensable. It is not difficult to provide three major classes of information and some further subclasses for MT inventory, but it is rather tough to set up a scientific system of classification which is both economical and sufficient.

2.2.7 *Make Full Use of Fixed Phrases and Fixed Structures —*

Best Way to Eliminate Polysemy and Raise Translation Quality

Our procedures seek to make full use of fixed phrases and fixed constructions. This is a viable way to eliminate polysemy and to raise the speed and efficiency of MT. New concepts and new things emerging in the development of society are designated more often by composing than by creating new words. This can be seen in the overwhelming majority of multi-word constructions in scientific and technical terms and in the constant increasing of phrasal verbs. There are, however, other kinds of combinations of words in natural languages. Such constructions are discrete and embedded. For instance, the correlative words in Chinese, some idiomatic patterns in Japanese and the relevant English constructions such as: effect of ... on..., relationship between ... and ... etc. to make full use of these facts, we can enhance the validity of MT algorithm (Wang 1982).

2.2.8 *Rely Mainly on Grammatical Analysis While Making Semantic Analysis Subsidiary*

Our consistent viewpoint for implementing MT is the grammatical analysis with the aid of semantic analysis. Homographs and polysemous words can be found throughout the whole system of MT. The major approach for disambiguation of homographs and polysemes is the grammatical analysis, which includes morphological and syntactic (or structural) analysis. The latter seems to be more important for some languages. Ignoring the status and context of a word, we can hardly disambiguate its homographs and polysemes. Sure enough, we do not mean to ignore the function of semantic analysis when we make much of grammatical analysis. Generally, semantic analysis should be increased in many algorithms but this again, does not mean to make the secondary supersede the primary.

2.3 *Problems and Prospect*

2.3.1 *Problems*

In addition to MT of spoken languages, there are still two kinds of problems to be solved. One is to contrive an optical reading device of high efficiency. The other is to strengthen language studies. Some of the above-mentioned issues have already been properly solved; others have merely just found their approaches to the possible resolution; still a large amount of work has to be done.

2.3.2 *Prospect*

In spite of the bumpy twists and turns in its development, MT has hewed out its path out of its childhood. Although there are still some problems to be solved in MT research, the day when MT can be put into actual application is drawing near. Many scholars believe, the product of MT (MTese) can be made available in 1980's.

3. The Most Crucial Item of Language Engineering — Chinese Character Information Processing

3.1 *A Brief Review*

As mentioned previously, MT like sparks, initiated the revolution in the realm of non-numerical application of computer. In order to implement MT, a series of key problems, such as adequate I/O device, large storage, speech recognition and character recognition, were posed as far back as the beginning of 1950's. In order to implement Chinese-Foreign and Foreign-Chinese MT, the encoding and I/O of Chinese characters were studied correspondingly. In 1958, Section on Chinese-Russian MT, Institute of Precision Instrument and Computing Technology, Academy of Sciences USSR proposed an input method "the new four-corner encoding

system". In 1959, the Massachusetts Institute of Technology (MIT) worked out the contrivance of Sinotype. The success of the experiment with Russian-Chinese MT algorithm in our country and that with English-Japanese MT algorithm in Japan, as well as the increasing ripening of mechanical retrieval gave an impetus to the study of Chinese character information system. It was taken into account for exploration by researchers of Institutes of Linguistics, Computing Technology and Electronics since 1960's. Large-scale research on it began in 1970's. A symposium of historic significance on Chinese character encoding was held in Qingdao in 1978, and the Association for Chinese Character Encoding was founded in the course of the Symposium (the Association was under the All-China Society of Scientific and Technological Information). The Association for Chinese Character Information Processing under the Chinese Society of Instruments and Meters, and the Section of Chinese Character Encoding under the Chinese Computer Society were founded in 1979. In order to improve the organization and coordination of the activities in the field of Chinese information processing, the Chinese Information Processing Society of China (CIPSC) was founded on the basis of the Association for Chinese Character Encoding and the Section of Chinese Character Encoding in June 1981. There are five sections in it at present:

1. Basic theories;
2. Chinese character encoding;
3. Special-purpose installations;
4. Chinese information systems, and
5. Natural language processing.

It is not difficult to see that the major task of the Society at the moment is to develop Chinese character information processing systems. Developments in recent years more and more show that Chinese character information processing is the most crucial language engineering. The automation of library information service, the modernization of editing-typesetting and the automation of office would become empty talk, if Chinese characters could

not be input into the computer.

3.2 *Chinese Character Information System — a Coherent Whole*

Several years ago some of the researchers did their encoding schemes only from the viewpoint of written language reform and indexing system for Chinese characters. To counter this tendency we have emphasized the following points at the 1978 Symposium:

1. The Chinese Character Information Processing System is usually made up of encoding, input, storage, processing, output and transmission, combined together in a proper way. Chinese encoding is the essential part and also the most difficult part;
2. We can not do encoding scheme in isolation. On the contrary, we have to study and design it in terms of the whole system;
3. To correctly appraise an encoding scheme, we must take as a whole the total system, including various indexes, i.e. ease and convenience of operation, efficiency of input and processing, reliability of transmission, economy and practicality of installation as well as the use of unambiguous codes. That is to say, we should not judge an encoding scheme simply by fewer rules and easy mastery or only by its absence of ambiguous codes (Liu 1979).

3.3 *Chinese Character Encoding*

Over 40 schemes of various types were posed for exchange at the Qingdao's Symposium in 1978. At the inauguration conference of CIPSC in 1981, encoding schemes have exceeded 200. It is estimated now that nearly 400 schemes have been designed, in which schemes tested on the computer or adopted for input

device have reached several dozens. The schemes seem to be miscellaneous, however, they can be divided into five types as follows:

1. Integral character input;
2. Input of decomposed character;
3. Input primarily through character configuration with the aid of phonetic information;
4. Input of phonetic alphabet;
5. Input primarily by Pinyin with the aid of character configuration (Liu 1983 B).

At present Chinese character encoding is in the phase of selection and optimization. A special group of our Society is in charge of the relevant research. The criteria used for selection and optimization are the same indexes, mentioned in the preceding section from the viewpoint of qualitative description. In addition, a number of institutions are sponsoring a key task team, in which researchers of configurative approach are launching out into the exploitation of a new type of encoding scheme.

3.4 *System Installation*

Research in Chinese character information processing has made gains not only in its key link — the study of Chinese character encoding, but also in some other relevant fields. A certain number of special-purpose installations for inputting and outputting Chinese characters have been produced, which involve various kinds of keyboards for inputting Chinese characters, Chinese character banks, Chinese character display terminals and Chinese character printers. Complete sets of Chinese character information processing system have been developed as well (including encoding method of Chinese characters, general-purpose keyboard for Chinese character and foreign letters, general purpose display for Chinese characters and foreign letters, installations for printing Chinese characters, Chinese

character bank and software of the system, etc.) and plans are afoot to manufacture them in batches. The updated editing and typesetting system — computer-laser editing-typesetting system for Chinese characters has also been set up. A very good base has been laid for type matrix production, thus providing favorable conditions for making various installation. Some outstanding problems at present are speed and quality of the installations and improvement in the system software.

3.5 *Internal Standard Codes*

An important publication, *Code of Chinese Graphic Character Set for Information Interchange (Primary Set) GB2312-80* which is intimately related to all aspects of work discussed above, has been accepted as the national standard. According to this, Chinese characters are divided into two grades, the first includes 3,755 characters, the second 3,008, with a total of 6,763. These are the internal codes of the computer, by means of which, the input and output devices can be designed on a unified basis, and the consistency of exchanging information among various systems can be obtained. Consequently, it will facilitate the sharing of information resources.

To meet both the need of those users who use more characters than those in the primary set and the need of Taiwan, Hongkong and other places, a supplementary set has been formulated. More than 16,000 characters are involved in the supplementary set, which are subdivided into two parts according to the frequency of the characters, and each part contains more than 8,000 characters respectively. The supplementary set is being revised at present.

3.6 *Automatic Recognition of Chinese Characters*

Automatic recognition of Chinese characters is the radical