A Chinese Corpus for Linguistic Research

Chu-Ren Huang
hschuren@twnas886.bitnet
Institute of History and Philology
Academia Sinica

Keh-jiann Chen
kchen@iis.sinica.edu.tw
Institute of Information Science
Academia Sinica

This is a project note on the first stage of the construction of a comprehensive corpus of both Modern and Classical Chinese. The corpus is built with the dual aim of serving as the central database for Chinese language processing and for supporting in-depth linguistic research in Mandarin Chinese.

I Background

The project being reported on is a sub-project of the on-going research of the CKIP (Chinese Knowledge Information Processing) Group. This group was founded by Hsieh Ching-chun in 1986 and is currently directed by Keh- jiann Chen and Chu-Ren Huang (Chang et al. 1989, Hsieh et al. 1989, Chen et al. 1991). The CKIP research is divided into three sub-projects according to their goals: 1) An On-line Lexicon for NLP, 2) A Corpus, and 3) A Parser. The sub-projects are designed to create a self-sufficient and mutual-supporting environment for Chinese NLP. The corpus will be the database supporting the electronic lexicon, while the lexicon will be the basic reference for automatically tagging the corpus. Moreover, both the corpus and the lexicon will support the parser. Our parser adopts the unification-based formalism of ICG (Information-based Case Grammar, Chen and Huang 1990), which encodes all grammatical information on each lexical entry. At this point in time, the lexicon consists of a fully automated earlier version with limited grammatical information and an updated version with complete grammatical version for parsing. There are more than 40 thousand entries in the completed electronic dictionary, which is available on-line in Taiwan and allows basic pattern-matching searches. There is also a PC version with reduced search capacity available from the Industrial Technology Research Institute, the primary funding agency of this pilot dictionary project. The updated version now contains roughly 30 thousand entries with complete grammatical information and another 60 thousand with basic grammatical categories. Manipulation of lexical information such as addition of entries and specification of detailed grammatical information with respect to each attribute is maintained on-line (Jian and Chen 1991). The completed 90 thousand word lexicon will be our core lexicon for pars- ing. The hierarchical arrangement will enable us to efficiently add new entries and create special lexicons for sub-domains.

Many modules of the parser are now under construction and some of them have been completed, such as a analyzer-generator for quantifier-measure compounds (Mo et al. 1991) and a look-up segmentation program (Chen and Liu 1992). Both the compound analyzer-generator and the segmentation program perform well. The recognition rate for the segmentation program, excluding proper names and derived words, is 99.77%. Since neither proper names nor word breaks are marked in the writing system, they will have to be dealt with separate modules using both morphological and heuristic information. The analyzer-generator has the perfect recognition rate of 100% while allowing over-generation and ambiguity, which also often involve proper names and derived words.

The corpus portion of the project is mainly funded through a grant from the Chiang Ching-kuo Foundation for International Scholarly Exchanges to Academia Sinica and the University of London, and is supported by matching funds from Academia Sinica. Paul Thompson is the Co-Principle Investigator at the University of London. The COBUILD project of the University of Birmingham also offers technical consultation. Of the 12 linguists working full time in the CKIP group, five are assigned to the corpus project in addition to three programmers.

II. Sources and Size of the Corpus

A strategical decision was made early in our project to build separate corpora for both classical and modern Chinese. This is not only because the same lexical computing techniques can be used for both modern and classical Chinese (data are represented as Chinese characters), but also because we think it will be interesting to compare the result of an open and artificially balanced corpus (the modern one) with a completed but randomly balanced corpus (the classical one). In addition, the impact on linguistic research will probably be more immediate and obvious in diachronical studies than in synchronical studies. Hence, the classical corpus is defined in terms of time (roughly all existing text written before 0 B.C.) and the modern corpus is defined by size (7–10 million words tagged text in two years). This also set the direction of future development: the classical corpora will develop chronologically through the time while the modern corpora will expand both in size and in domain–specific corpora.

The following texts are acquired in the first stage:

1) Modern Chinese

A. 10 million characters of texts (word breaks are not used in Chinese writing systems, but the average word-length is a little more than two characters) from three months of the Liberal Times, a daily newspaper. B. 10 million characters from the China Times group. Agreement was reached in October, 1991 with the China Times group, one of the two newspaper giants in Taiwan, to provide daily on–line text to our project. Thus we will have a dependable and unlimited source of data (up to one million characters a day). The group publishes three newspapers, several magazines, and has a separate book–publishing subsidiary. C. About one million words of data previously input by the CKIP project. This includes 10 articles from a magazine and explanation text from a dictionary. D. 30 thousand words of transcribed spoken text. This section will be input in 1992.

Newspapers are the mainstay of our data source for the obvious reason that the newspaper texts are online. But it is also a convenient text containing many varieties of writing, including spoken conversation (interviews), commentaries, translations (foreign dispatches), and all genres of literary styles (Chinese newspapers are different from the Western ones in having daily literature supplementaries, and are the most important venue for literary publications). However, because of difficulties in converting different character–coding and mark–up systems, we can only incorporate news from China Times to our corpus so far. Other types of texts should be available by fall 1992.

2) The Classical Corpus

Roughly three million characters of Chinese written prose have survived from the years before Christ was born (i.e. from all the periods up to the Western Han Dynasty). A corpus of roughly 1.5 million characters of the text is now available in machine readable forms from a previous project in Academia Sinica, the remaining 1.5 million characters are now being keyed in and will be on–line by mid–1992.

III. Functions and Applications

It is our long–term goals for the Corpus to have the automatic dictionary compiling abilities, fashioned after COBUILD (Sinclair 1987, 1991), adapted for both parsing and hard–copy publishing. In the first stage, however, we concentrate on developing tools for linguistic research. We will describe the search functions that we have developed so far. Most of the functions are character–based now and can be upgraded to word–based once we incorporate our tagger and word segmentation module.

The search utility in our corpus is basically a KWIC (key–word–in–context) search based on Chinese characters. Our search program allows the linguists to specify both the size of right–and left–hand side context shown as the search result. There is also a randomizer to choose a more manageable size of data if the search result is too big (more than a thousand citations). The ordering of data is done in the traditional way in terms of numbers of strokes in the character in the immediate context. This is not the ideal method but may be the best available before we can develop a system which is both linguistically and heuristically more sophisticated. A romanization–based ordering system would be even less desirable for the lack of uniform (and familiar) romanization system in Taiwan and for the failure to disambiguiate homonyms.

In addition to the basic search procedures, the following customized search commands are added to serve the need of linguistic research.
1) #<kw>#: a context where a key word is both preceded and followed by the same word.
2) AA: reduplication (one character)
3) ABAB: reduplication (two characters)
4) <kw1>*<kw2>: cooccurrences of two key words in a (possibly) discontinuous context.
5) <kw1>/l<kw2>: context where a key word (kw2) is 'left–disassociated' from another key word (kw1), i.e. where the second key word does not occur in the left–context of the first key word.
6) <kw1>/r<kw2>: context where a key word (kw2) is 'right–disassociated' from another key word (kw1), i.e. where the second key word does not occur in the right–context of the first key word.

Commands 1)–3) are helpful tools in studying morphological rules and identifying morphological constructions for Chinese. Since Chinese writing systems do not include word–breaks, and since no lexicon can ever offer a complete list of words, word segmentation is non–trivial in Chinese Language Processing (Chen and Liu 1992). Identifying and utilizing morphological information is therefore essential both in lexical computing and in natural language processing.

Command 4) is a handy tool to discover cooccurrence restrictions and their semantic consequences.

Commands 5) and 6) are used to eliminate ambiguity and to cut down the size of search results.

It can be noted that since our tagger is not running yet and since the Chinese running text seldom defines a sentence by a period (a whole paragraph often contains only one period and many commas), the above commands use number of characters rather than sentence markers to delimit search domains.

Concordance programs are also being developed for our project. The current version runs with our classical corpus. It is able to show both the text source of each concordance item as well as page numbers from the printed version for easy reference. This is originally developed on the HP workstation, the machine we now use, but a version that runs on IBM PC 486 with either SUN Unix or CCL Unix is also available now.

Another research tool that we developed deal with frequency counts of characters and words and statistical packages to compare linguistics and other textual features of the corpora. Like many other modules of our project, this module has been developed and completed independently. It has been tested on the untagged on-line classical Chinese database of the Twenty-five Dynastic Histories (Hsieh 1991). This module is ready to be incorporated into the system.

IV. Accomplishments and Future Developments

In this first stage of the development of a Chinese corpus for NLP and for linguistic research, we have achieved the primary goals of acquiring the core text data, and establishing a mutual-supporting environment between the lexical computing research on the corpus and the computational linguistics research on NLP. Our systems are developed on HP workstations under Unix. The system should be portable to any Unix machine with compatible Chinese solution. For instance, we are porting each of our modules to a IBM PC 486 running Unix for the use of our collaborator in London.

In this preliminary stage, the most encouraging sign is that our human linguists have established productive interaction with the corpus. Human expertise helped to design search utilities pertinent to linguistic research and corpora provided both a convenient source of linguistic facts and a solid basis for deducing useful generalizations. With the basic KWIC search utilities, the CKIP project has finished several linguistic studies with the help of the corpus. Take Mo et al. (1991) for example, the quantifier-measure rule that our analyzer-generator uses is based on generalizations extracted from the corpus, and the program itself is tested on texts randomly selected from the corpus. Other linguistic works based on the corpus include Hong, Huang and Chen (1991) on morphological rules for Chinese, and Mo, Huang and Chen on serial verb construction in Chinese (1991). This corpus is currently used by the CKIP project in their ac- counts of A-not-A questions, of resultative compounding, of nominalization, and of various reduplications.

Incorporation of an automatic tagger and extraction of grammatical information from the tagged corpora is the most important immediate future goals of this project. This, of course, depends crucially on a tagging system with theoretically well-structured attributes. A database for attributes is being developed with the INFORMIX software. And we will follow the TEI guidelines (Sperber-McQueen and Burnard 1990) whenever possible. Our word segmentation program now doubles as a category-tagger. But this can only be viewed as a research aid for the linguists to detect categorical ambiguities and unlisted words. We also expect the on-going linguistic research to identify more search functions and refine the existing utilities. Direct extraction of dictionaries and grammars should be feasible in five years.

Acknowlegements:

Research of this project was partially funded by the Chiang Ching-kuo Foundation for International Scholarly Exchanges. The first author wants to express his gratitude to the Center for the Study of Language and Information, Stanford University for the surport he received when he visited there, during which a draft of this note was completed. Responsibility of any errors remain ours alone.

Bibliography

Chang, L.L. and CKIP. 1989. The Categorical Analysis of Mandarin Chinese (Revised edition, in Chinese). Taipei: Academia Sinica.

Chen, K.-J. and CKIP. 1991. The Chinese Knowledge and Information Project and Chinese Electronic Dictionary (in Chinese). Paper presented at the Joint Chinese-Japan Symposium on Information Processing. Taipei.

Chen, K.-J. and C.-R. Huang. 1990. Information-based Case Grammar. COLING-90. Vol. II. 54-59.

Garside, R., G. Leech, and G. Sampson. Eds. 1987. The Computational Analysis of English. London: Longman.

Hsieh, C.-C. 1991. Statistics of the Text of the Twenty-Five Dynastic Histories. Paper presented at ROCLING IV.

Jian, L.-F. and K.-J. Chen. 1990. The Hierarchical Representation and Management of Lexical Information. Proceedings of the Third R.O.C. Computational Linguistics Conference (ROCLING III). pp. 295-310.

Chen, K.-J. and S.-H. Liu. 1992. Word Identification for Mandarin Chinese Sentences. To be presented at COLING-92.

Hong, W.M., C.-R. Huang, and K.-J. Chen. 1991. The Morphological Rules of Chinese Derived Words. To be presented at the 1991 International Conference on Teaching Chinese as a Second Language. Dec. 1991. Taipei.

Mo, R. J., Y.-R. Yang, K.-J. Chen, and C.-R. Huang. 1991. A Analyzer-Generator for Mandarin Chinese Quantifier-Measure Compounds. Proceedings of ROCLING IV.

Mo, R. J., C.-R. Huang, and K.-J. Chen. 1991. Serial Verb Constructions in Mandarin Chinese -- Their Definition and Control Relations. To be presented at the 1991 International Conference on Teaching Chinese as a Second Language. Dec. 1991. Taipei.

Sinclair, J. M. 1987. Ed. Looking Up --An account of the COBUILD Project in Lexical Computing. London: Collins.

Sinclair, J. M. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sperberg-McQueen, C. M. and L. Burnard. 1990. Giudelines for the Encoding and Interchange of Machine-Readable Texts. (TEI P1). ACH-ACL-ALLC.