# **Character-based Collocation for Mandarin Chinese**

Chu-Ren Huang Institute of History and Philology Academia Sinica Nankang, Taipei, Taiwan hschuren@ccvax.sinica.edu.tw Keh-jiann Chen Institute of Information Science Academia Sinica Nankang, Taipei, Taiwan kchen@iis.sinica.edu.tw Yun-yan Yang Computing Center National Taiwan University Taipei, Taiwan yang@iis.sinica.edu.tw

This paper describes a characters-based Chinese collocation system and discusses the advantages of it over a traditional word-based system. Since wordbreaks are not conventionally marked in Chinese text corpora, a character-based collocation system has the dual advantages of avoiding pre-processing distortion and directly accessing sub-lexical information. Furthermore, word-based collocational properties can be obtained through an auxiliary module of automatic segmentation.

## PROJECT NOTE: LARGE TEXT CORPORA

#### I. Introduction

Collocation has been established as an essential tool in computational linguistics (Church and Mercer 1993). In addition, various collocational programs have been proven to be indispensable in automatic acquisition of lexical information (e.g. Sinclair 1991, and Biber 1993). Since words are the natural and undisputed units in available text corpora, virtually all the current collocational programs are word-based. However, there are languages where texts do not conventionally mark words, such as Chinese. Unless a large tagged corpus is available, a word-based collocation system in these languages faces the following inevitable difficulties. First, hand-segmentation of a large corpus is tedious and financially nearly impossible. Second, automatic segmentation program can neither identify words not listed in the lexicon nor correctly segment all which are listed. Third, estimation of lexical probability relies on word-frequency counts based on the inaccurate results of automatic segmentation thus the deviation tends to be greater than standard tolerance.

Text corpora without wordbreaks, nevertheless, also has their advantages. Take Chinese for example, the basic units of text corpora are zi4 'character', a fairly faithful representation of the morphemic level of the language. In other words, if we take Chinese text

corpora as they are, we will be able to access sub-lexical information without additional cost. To take the full advantage of the nature of texts, reliable tools can also be devised to obtain lexical collocation. In this paper, we will describe the design and implementation of a Chinese collocational system that does not require the preprocessing of automatic segmentation but is able to allow both lexical and sub-lexical information be automatically extracted.

#### II. Background: Corpus and Computational Platform

This collocation system is developed on the 20 million character modern Chinese corpus at Academia Sinica (Huang and Chen 1992, Huang In Press). This corpus is composed mostly of newspaper texts. It is estimated to have 14 million words. Following industrial standard in Taiwan, our collocation system can deal with any corpora encoded by BIG-5 code. The program is developed under a UNIX environment on HP workstation. It should, however, be portable to any UNIX machine with compatible Chinese solution. The collocation system is currently used in research by more than 10 linguists affiliated with the Chinese Knowledge Information Processing (CKIP) group at Academia Sinica. It is also open to any visiting scholar for on-site use.

#### III. Overall Design of the System

There are two major modules in the collocation system: one deals directly with unsegmented texts and the other which incorporates automatic segmentation before collocation. The two modules share the preprocess of KWIC search module, which allows userspecified linguistic patterns (Huang and Chen 1992). They also share three common routines to detect character collocation, to identify possible collocation words through N-grams, and to contextually filter texts with user-specified strings. The overall design of the system is schematically represented in diagram 1.



# IV. Collocation Without Segmentation

There are three collocational tools available in this system without segmenting the texts into words. First, character collocation allows automatic acquisition of sub-lexical information, such as the conditions on morpho-lexical rules. This is attested by the studies on the notion of word in the mental lexicon reported in Huang et al. (1993), and the generalizations of productive derivational rules in Mandarin offered in Hong et al. (1992). Take note that when applying KWIC search to the corpus, a user has the freedom to specify a key that is a single character, a multi-character string, or even a discontinuous string of characters. These character strings may or may not be words. Thus the extracted collocational relation is not simply between characters. It can also be between characters and either a simplex word, a compound, or a phrase. The collocational relation in our system is measured and represented by both Mutual Information (Church and Hanks 1990) and frequency. The user can choose to sort and rank the collocates by either criteria. S/he can also specify threshold value by either criteria. Usually, the most effective method is to use frequency threshold and Mutual Information ranking (Huang In Press). In addition to the measures of correlation, distribution of the collocates is also indicated in terms of positions relative to the key and frequency of occurrences at each position for each collocating character.

Second, lexical information can also be derived from this collocational system regardless of its lack of demarcation of lexical items. This is achieved through a simple Markov model. Once the KWIC search extracted the relevant contexts, a simple N-gram routine can be performed on the context(s) specified by the user. Depending on the purpose of the study and the size of relevant texts, the length of the target sequence as well as the threshold number can be specified. For instance, a linguist may want to look for all two or three character sequences that occur over 5 times after a key verb. This would likely turn out a list of possible arguments (i.e. syntactic words) for that verb. Hence lexical information such as semantic restriction of the predicates on its post-arguments can be indirectly extracted. In our system, the user is allowed to iterate the N-gram search by designating different contexts and string length (N). The following is an example of collocation without segmentation. Huang et al. (1994) argue that Mandarin light verbs select the verbs they nominalize. This is supported by the N-gram collocation results in diagram 2. The collocation is extracted from a 20 million character corpus and the collocation window is 5 characters to the right of the key word. It shows that the verb *jin4xing2* typically nominalizes a process verb.

#### Diagram 2. N-gram Collocation (By Frequency) Bi-syllabic Collocation with the verb jin4xing2

gong1zuo4 'te	o work'		437	
diao4cha2 'te	) investigate'	354		
gong1cheng2 'engineering work' 233				
tao3lun4 '1	o discuss'	223		
goultongl 'to	o communicate	;'	198	
xie2tiao2 'to	coordinate'	185		
yan2jiu4 't	to study'		185	
liao3jie3 'to	understand'	166		
gue1hua4 'to	o plan'		156	
xie2shang1 'te	negotiate'	154		

Last, the user can specify a character string in the context as a filter. The most useful application is to specify a string that forms a syntactic word. This is a technique commonly used to resolve categorical or sense ambiguities. Combining both N-gram search and string filtering, frequency-based word collocation is achieved without segmentation.

#### V. Collocation After Segmentation

When lexical or phrasal relation is the focus of the study, the above collocation module may sometimes be

inadequate. In this case, we will need to apply the automatic segmentation/tagging program such that we can acquire information involving word pairs as well as grammatical categories. The automatic segmentation procedure is an revised version of the program reported in Chen and Liu (1992). The on-line lexicon is the CKIP lexicon of more than 80 thousand entries (Chen 1994).

We did not automatically segment and tag the whole corpus for very good reasons. First, without a correctly tagged corpus, no statistically-based tagger can perform satisfactorily yet.

Second, there is no practical way to recover incorrectly identified words. That is, when the automatic tagger takes a character from a target word to form an inappropriate word with a neighboring character; that target word is lost and cannot be identified in this context. Thus, it will be linguistically more felicitous to allow KWIC to identify all matching strings and allow filtering of incorrect matched words in later steps.

Last, segmented texts restrict the available collocation information exclusively at word level. For instance, not only morpheme-morpheme collocation will not be available, neither can correlations between a morpheme and a word be extracted.

In contrast, when optional segmentation is performed on-line on the result of KWIC search, the collocational system can be applied to any electronic text corpora with

minimal pre-processing. This current approach also allows us to mix sub-lexical, lexical, and extra-lexical conditions according to our research need.

Even though the post-segmentation module shared three routines with the module without segmentation, they do differ non-trivially in their applications. First, the character collocation module is basically the same. The additional step of segmentation excludes accidental string matches. For instance, with *qu4shi4* 'to pass away' as the keyword, KWIC may extract the incorrect context '*tal qu4 shi4 jie4 ge4 di4 lu3 xing2*'. This error in identifying word boundaries can be easily avoided when the text is correctly segmented. In this case, the correct segmentation is '*tal qu4 shi4jie4 ge4di4 lu3xing2* (s/he go world everywhere travel)'. Second, N-gram in this module now can include both sequences of characters and sequences of words.

Two additional tools directly utilize grammatical tags. The first one is the computing of the distribution of grammatical categories in the context. The second is contextual filtering in terms of grammatical categories. One caution needs to be mentioned here. As mentioned

earlier, we do not have a highly reliable automatic tagger yet because there is no dependably tagged large Chinese corpus. Hence our automatic segmentation program looks up the categories of the words but do not attempt to resolve ambiguity. Since categorically ambiguous words make up only around 20% of the texts (Chen and Liu 1992, Chen et al In Preparation), keeping all possible tags seem to be an acceptable compromise for the moment. But this also means that a user must be on the lookout for possible errors caused by multiple tags. Our system allows the user to view the categorical distribution of the whole context, as well as to focus on a smaller context and specific categories. Diagram 3 shows the categorical collocation of the head of the post-verbal argument of huo4de2 'to get/ receive.' We obtained this information by first perform the discontinuous KWIC on huo4de2 and the relative clause head marker de. After segmentation and collocation, we restrict the display to the first position to the right of de, and to the two major categories of N and V. The result shows that this verb typically take subclasses of common noun and (nominalized) transitive verbs as arguments.

## Diagram 3. Categorical Collocation Heads of Relative Clause Arguments of *huo4de2*

r1	frequency	rl	frequency
Nab	63	Vc2	62
Nac	40	Vh11	28
Nad	27	Vkl	20
Nca	26	Vcl	19
Ncb	13	Vc2	16

Last, the word-based collocation system is the part of our system that will take the most processing capacity. This is also the only part of our system that is still being tested at this moment. Word frequencies of our corpus have already been calculated and stored. The automatically segmented word-based collocation module should be available for linguistic research within weeks.

## VI. Conclusion

In this paper, we described a collocation system that works on text corpora without word marks. This system has the advantage of extracting sub-lexical information. This is also particularly useful in studying Chinese language corpora since sociological words are distinct from syntactic words in Chinese (Chao 1968). Thus in linguistic and literary computing, it is often necessary to formulate generalizations based on zi4, the sociological word. The techniques reported in this paper should also find applications in two aspects of future computational linguistic research. First, it can be applied to other language text corpora for extraction of sub-lexical collocation. Second, it can be applied to text corpora which do not come with clear word demarcation: including corpora in languages in which sociological words and syntactic words do not coincide and spoken corpora.

#### Acknowledgements

Research of this project was partially funded by the Chiang Ching-kuo Foundation for International Scholarly Exchanges, the National Science Council of R.O.C., and Academia Sinica. We would like to thank Li-ping Chang and other colleagues at CKIP for their help and comments. Responsibility of any remaining errors is ours alone.

## Bibliography

Biber, D. 1993. Co-occurrence Patterns among

Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition. Computational Linguistics. 19.3:531-538.

- Chao, Y. R. 1968. A Grammar of Spoken Chinese. Berkeley: University of California Press.
- Chen, K.-j. 1994. Linguistic Information and Lexi- cal Data Management in Dictionary Research. Invited Paper presented are the Internal Conference on Computer Processing of Oriental Languages. Daejon, Korea. May 10-14.

\_\_\_\_, and S.-II. Liu. 1992. Word Identification for Mandarin Chinese Sentences. COLING-92. 101-105. Nantes, France.

\_\_\_\_, S.-h. Liu, L.-p. Chang, and Y.-h. Chin. In Preparation. A Practical Tagger for Chinese Corpora. Nankang: Academia Sinica.

Chinese Knowledge Information Processing Group. 1993. The CKIP Categorical Classification of Mandarin Chinese (In Chinese). CKIP Technical Report no. 93-05. Taipei: Academia Sinica.

- \_\_\_ 1994. A Frequency Dictionary of Written Chinese. CKIP Technical Report no. 94-01. Taipei: Academia Sinica.
- Church, K., and P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics. 16.1:22-29.

\_\_\_\_, and R. L. Mercer. 1993. Introduction to

the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics. 19.1:1-24.

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

Huang, C.-R. In Press. Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In M. Y. Chen and O. J-L. Tzeng Eds. 1993. Linguistic Essays in Honor of William S.-Y. Wang, Taipei: Pyramid. \_\_\_\_, Kathleen Ahrens, and Keh-jiann Chen. 1993. A Data-driven Approach to Psychological Reality of the Mental Lexicon: Two Studies in Chinese Corpus Linguistics. Proceedings of the International Conference on the Biological Basis of Language. 53-68. Chiayi: Center of Cognitive Science, National Chung Cheng University.

\_\_, L.-p. Chang., and M.-L. Yeh. A Corpus-

- based Study of Nominalization and Verbal Semantics: Two Light Verbs in Mandarin Chinese. Paper presented at the Sixth North American Conference on Chinese Linguistics. May 13-15, USC.
- \_\_\_\_, and K.-j. Chen. 1992. A Chinese Corpus for Linguistic Research. COLING-92. 1214-1217. Nantes, France.
- Sinclair, J. M. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sproat, R., and C. Shih. 1990. A Statistical Method for Finding Word Boundaries in Chinese Text.

Computer Processing of Chinese and Oriental Languages. 4.4:336-351.

Svartvik, J. 1992. Ed. Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, 4-8 August 1991. Trends in Linguistics Studies and Monographs 65. Berlin: Mouton.

Wang, M.-C., C.-R. Huang, and K.-j. Chen. 1994.

The Identification and Classification of Unknown Words in Chinese: A N-gram-Based Approach. Manuscript. Academia Sinica.