Language, Information and Computation(PACLIC 11),1996, 167-176

SINICA CORPUS: Design Methodology for Balanced Corpora

*Keh-Jiann Chen, †Chu-Ren Huang, *Li-Ping Chang, *Hui-Li Hsu

*Institute of Information Science, Academia Sinica
†Institute of History & Philology, Academia Sinica
*kchen@iis.sinica.edu.tw , †hschuren@ccvax.sinica.edu.tw

Abstract

The Academia Sinica Balanced Corpus (Sinica Corpus) is the first balanced Chinese corpus with part-of-speech tagging. The corpus (Sinica 2.0) is open to the research community through the WWW (http://www.sinica.edu.tw/ftms-bin/kiwi.sh). Current size of the corpus is 3.5 million words, and the immediate expansion target is five million words. Each text in the corpus is classified and marked according to five criteria: genre, style, mode, topic, and source. The feature values of these classifications are assigned in a hierarchy. Subcorpora can be defined with a specific set of attributes to serve different research purposes. Texts in the corpus are segmented according to the word segmentation standard proposed by the ROC Computational Linguistic Society. Each segmented word is tagged with its part-of-speech. Linguistic patterns and language structures can be extracted from the tagged corpus via a corpus inspection program which has the functions of KWIC searching, filtering, statistics, printing, and collocation.

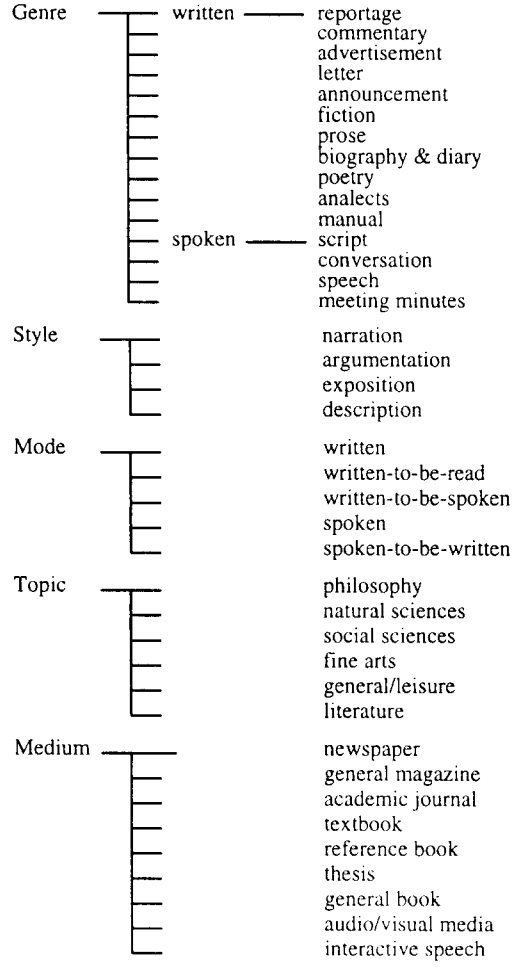
1. Introduction

Corpus-based approaches are fast becoming the most essential and productive technique for theoretical and computational linguistics research (Svartvik 92, Church & Mercer 93). Their impact reaches almost all areas of natural language studies, such as speech processing, information retrieval, lexicography, character recognition etc. Version 2.0 of the Academia Sinica Balanced Corpus (Sinica Corpus) contains 5,345,871 characters, equivalent to 3.5 million words. The Sinica Corpus is the first balanced Chinese corpus with part-of-speech tagging. The following issues have been the major concerns in designing the Sinica Corpus: 1. organization of the corpus, 2. preparation of the corpus, and 3. the use of the corpus. Since a corpus is a sampling of a particular language or sublanguage, which contains an infinite amount of data, it must be representative and balanced if it claims to faithfully represent the facts in that language or sublanguage (Sinclair 87). However there is no reliable criteria for measuring the balance and representation of a corpus. The Brown corpus is the first corpus that claimed to be balanced. It takes the topic domain distribution as the only balancing criterion. In the Sinica Corpus, we explore the possibilities of multi-dimensional attributes and try to balance the corpus in each dimension. The detailed organization of the Sinica Corpus is discussed in section 2. The Sinica Corpus is a word-based Chinese corpus with part-of-speech tagging. Word segmentation, automatic part-of-speech tagging and quality assurance are major concerns after text selection. They are discussed in section 3. The tools for using a tagged corpus are illustrated in section 4.

2. Organization of the Sinica Corpus — Texts Selection and Classification

We set up a systematic design in order to use and maintain a large amount of texts. The texts are classified according to five attributes: source, mode, style, topic, and genre. Every text is marked with five attribute values. The five attributes are from five independent, though possibly interactive, hierarchies, as shown in Figure 1 (Hsu & Huang 95).

Figure 1.



The attribute values for classifying texts are established by consulting the Lancaster-Oslo/Bergen (LOB) corpus(Atwell 84), the Brown Corpus (Ellegard 78), the Cobuild Project (Sinclair 87), and the Chinese library topic classification system (Lai 89). The topic attribute is self-explanatory and indicates what the text is about. The attributes of genre, style, and mode deal with how the text is presented. Lastly the source attributes explicate the medium, information about the author, and publication type. Figure 2 is an instance of textual mark-up.

Figure 2.

```
%% genre = prose
%% style = description
%% mode = written
%% topic = literature—children
%% medium = textbook
%% author =
%% sex =
%% nationality = Taiwan, ROC
%% native-language = Mandarin
%% publisher = National Compilation Bureau
%% location = Taiwan
%% date =
%% edition =
%% title = Starlight
I will never forget when I was little, The moments when I lean
close to my mother, Ah! Recollections of my childhood
        ⋮
```

While balancing the corpus, we take the attribute of topic as the primary consideration. Each topic area is assigned a certain target proportion. For the Sinica Corpus, the following balance is targeted and achieved.

philosophy	10%
natural sciences	10%
social sciences	35%
arts	5%
general/leisure	20%
literature	20%

In addition, distribution according to the four other classificatory attributes (i.e. genre, medium, style, and mode), are monitored and checked to meet respective requirements. The feature values of each attribute are represented hierarchically.

The length of texts in the Sinica Corpus Version 2.0 varies individually to keep the structural completeness of each text. The length ranges from 327 characters (an elementary school textbook article) to 1,941 characters (a magazine article). The average length of a news article is 415 characters.

What is a balanced corpus? With the help of five major attributes, the Sinica Corpus is quite different from a corpus which controls only one parameter. With variant parameters, we may adjust our proportions in different attributes to achieve an ideally balanced corpus. Another benefit of the hierarchical attribute assignment is that we can control our proportions of value according to different usages of the corpus. This design allows flexibility for on-line composition of subcorpora as well as for quick comparison of different subcorpora.

3. Preparation of Sinica Corpus—Word Segmentation and Part-of-speech Tagging

To prepare a tagged Chinese corpus, word segmentation and automatic tagging are two major processes after text selection. Word segmentation for Chinese is a

difficult task due to the lack of delimiters to mark word boundaries. Simply looking up a word dictionary to identify words is not sufficient to solve the problem, because of the existence of unknown words, such as proper names, compounds, and new words. Basically an automatic word segmentation system for Chinese works as follows: an electronic dictionary provides a list of common words, and a set of morphological rules to generate/identify a variety of derived words and compounds, such as determiner-measure compounds, reduplication etc., as supplement. An algorithm will resolve ambiguous segmentation by either heuristic or statistic methods (Chen & Liu 92). The remaining segmentation errors (as well as tagging errors) are fixed by human post editing.

3.1 Word Segmentation Standard

The form and content of a correct word segmentation criteria has been discussed and disputed in the field. Different word segmentation systems have been designed and they all follow their own idiosyncratic guidelines (Chen & Liu 92, Sproat et al. 94). While the Sinica Corpus was being developed, a standard for Chinese word segmentation was also drafted and proposed by the ROC Computational Linguistic Society. The word segmentation standard project fully utilized the variety of actual examples encountered in corpus tagging in order to ensure better coverage on the definition of words. The word segmentation of Sinica Corpus follows this standard. Therefore it also became the best testing data for the proposed segmentation standard. The segmentation standard is composed of two parts: a set of segmentation criteria and a standard lexicon. The segmentation criteria can be further divided into the lexicon-independent and the lexicon-dependent parts. The lexicon-independent parts include the definition of a segmentation unit and two segmentation principles.

(1) Segmentation Unit_{def} is the smallest string of character(s) that has both an independent meaning and a fixed grammatical category.

(2) Segmentation Principles
   (a) A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.
   (b) A string whose structural composition is not determined by the grammatical requirements of its components, or a string which has a grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

The definition of a segmentation unit is an instantiation of the ideal definition of a word. The two segmentation principles are a functional definition of the segmentation unit as well as a procedural algorithm of how to identify segmentation units. In addition, the segmentation guidelines are lexicon-dependent and give instructions on how each lexical class should be treated.

(3) Segmentation Guidelines
   (a) Bound morphemes should be attached to neighboring words to form a segmentation unit when possible.
   (b) A string of characters that has a high frequency in the language or high co-occurrence frequency among the components should be treated as a segmentation unit when possible.
   (c) A string separated by overt segmentation markers should be segmented.
   (d) A string with complex internal structures should be segmented when possible.

Lastly, the lexicon contains a standard list of words as well as productive morphological suffixes, and obligatory segment markers. For more detail on the word segmentation standard, please refer to Huang et al. (1996).

3.2 Part-of-speech Tagging

The possible part-of-speech (abbr:pos) of each word was given after segmentation process. To resolve the ambiguities, a two-stage automatic tagging process was designed to disambiguate multi-category words. In the first stage, a small portion of the corpus was resolved by a hybrid method which combines rule-based and relaxation methods to select the most plausible pos tag for each multi-category word (Liu et al. 95). This initial corpus was post-edited manually and then became training data for the second stage statistical tagging model adapted from (Church 88). This statistical tagging model selects the pos sequence with the highest probability among all possible pos sequences P for each input sentence W, i.e. $\arg\max_P \Pr(P|W)=\arg\max_P \Pr(P)\Pr(W|P)$ and the probability of a pos sequence was approximated by pos-bigram statistics, i.e. $\Pr(P) = \Pr(P1,P2,...,Pn) \approx \prod_i \Pr(Pi|Pi-1)$, and $\Pr(W|P) \approx \prod_i \Pr(Wi|Pi)$

After the automatic assignment of pos tags, the remained segmentation errors and tag errors need manual post-editing. An on-line editing tool--TAGTOOL was designed to provide functions of (1) on-line dictionary look-up, (2) short term memory and recall ability, and (3) new word collection (Chang & Chen 95). The function of an on-line dictionary to provide the possible pos and their respective examples for each word. Human taggers may examine examples of different word uses to help them determine a correct tag. The most recent correction of segmentation as well as tag assignments are recorded by TAGTOOL such that the same type of errors will be automatically corrected. The function of new word collection will report any new word, which is not listed in the lexicon, such that the lexicon can be augmented for future tagging. TAGTOOL not only speeds up the post-editing process, but also make the results more consistent when providing on-line consulting functions. However, human-induced inconsistencies are inevitably exist after post-editing. The last process to improve the tagging quality is by checking the KWIC file for each multi-tagged word in the corpus. If we sort each KWIC file according to the context around the key word, it is easy to find the inconsistent results. After sequentially proof-reading the corpus via TAGTOOL and selectively examining KWIC files for multi-category words, the SINICA corpus is able to maintain a high quality of standard both on word segmentation and on the pos tagging.

The tagset of the Sinica Corpus is reduced from the syntactic categories of the CKIP lexicon (CKIP 93). Appendix 1 lists the Sinica Corpus tagset and its interpretation. However, the reduced tags can still predict the finer-grained grammatical categories unambiguously, by fixing the domain of the mapping to the set of categories from each individual word.

In addition to the pos tags, we adopts 8 feature tags, in order to preserve important morpho-syntactic information, such as separable VO compounds, VR compounds, and abbreviated conjunct words, as shown in Fig. 3. The separable VO and VR features are essential since they identify discontinuous parts of a word.

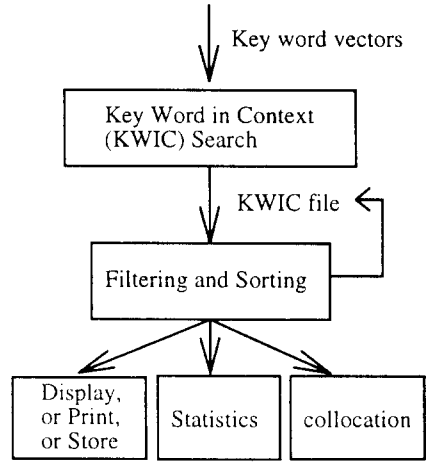
Figure 3. Table of Attributes

Attr.	Explanation	Example
+vrv	V of a separable VR compound	jiao(Vc)[+vrv] bu xin call-NEG-wake 'cannot awaken'
+vrr	R of a separable VR compound	jiao bu xin(Vc)[+vrr] call-NEG-wake 'cannot awaken'
+vov	V of a separable VO compound	chi(Vc)[+vov] le ta de kui eat-PERF-his/her-DE-vacancy 'be taken advantage of by him/her'
+voo	O of a separable VO compound	chi le ta de kui(Vc)[+voo] eat-PERF-his/her-DE-vacancy 'be taken advantage of by him/her'
+p1	the first part of a separated compound	chu(Nc)[+p1] , gaozhong junior+senior-high-school 'junior and senior high schools'
+p2	the second part of a separated compound	xinqiliu , ri(Nd)[+p2] Saturday-Sun 'Saturday and Sunday'
+fw	foreign word	kala-OK(Na)[+fw] 'kareoke'
+nom	nominalized verbs	ta de bu jiangli(Va)[+nom] s/he-DE-NEG-rationalize 'his/her being irrational'

4. Using Sinica Corpus-the Inspection System

A corpus inspection tool was designed for the purposes of observing and statistically analyzing texts with key-word-in-context (KWIC). The inspection system has the functions of (a) KWIC searching, (b) filtering, (c) statistics, (d) displaying, printing, and storing, (e) collocation finding by mutual information. Figure 4 shows the system flow diagram.

Figure 4.



(a) KWIC search

The function of the KWIC search provides users a way to search key words to create a key-word-in-context file for further manipulation or inspection. A key word is

defined as a vector of four components: 1) word, prefix, suffix, or stem, 2) part-of-speech, 3) features, 4) number of syllables. Each component may be under-specified or empty. The process of the KWIC search will match words in the corpus with the specified key word vector (or vectors) and produce key-word-in-context files. For example,

	Key word vector	what is matched
(1)	[代表, N, ϕ, ϕ]	every word 代表 *daibiao* tagged with the pos noun (i.e. 'a representative but not ' to represent'
(2)	[ϕ,VA, ϕ, 1]	all monosyllabic intransitive verb(VA)
(3)	[ϕ, ϕ,+fw,ϕ]	all foreign words
(4)	[..化,V, ϕ, 3]	all tri-syllabic verbs with the suffix 化 *hua* '-ize'

(b) Filtering

The result of the KWIC search may produce a large amount of text containing key words. Users can filter out redundant or irrelevant data through successive applications of the filtering functions. The filtering methods include 1) random sampling, 2) removing redundant samples, 3) removing irrelevant samples by restricting the content in the window of key words. For instance, if we are interested in the cases of the verb 乾淨 *ganjing* 'to cleanse' which functions as the result complement of another verb, both KWIC and filtering are necessary. First, we do the KWIC search by setting key word vector [乾淨, ϕ, ϕ, ϕ]. The result is a KWIC file that contains all of the samples with the key word 乾淨. Second, we apply the filtering step by restricting the first word to the left of the key word to be a verb, i.e. to set the restriction vector on left position to be [ϕ, V, ϕ, ϕ].

(c) Displaying, printing, and storing

The resulting KWIC files can be displayed on screen, or printed, or stored for future processing.

(d) Statistics

Statistic functions provide statistical distributions of words and categories occurring within the context window of key words. For instance, if we want to know the category distribution of the word 把 *ba*, the statistical function produces the following results.

	Category	Frequency	%
1. preposition	P	2704	92.57
2. measure	Nf	211	7.22
3. transitive verb	Vc	3	0.10
4. determiner	Neqb	2	0.07
5. noun	Na	1	0.03

Surprisingly, other than preposition and measure functions, the word '把' also functions as a transitive verb, determiner, and noun, although these usages are extremely rare.

(e) Collocation finding

The system finds collocations of the key words by computing the mutual

information (Church & Hanks 90) of the key words with the words or parts-of-speech in a user defined window. The resulting word collocations or category collocations will be sorted and displayed according to either their values of mutual information or their frequency.

5. Conclusion

The Sinica corpus is the first balanced Chinese corpus with part-of-speech tagging available to public. The major design features of the Sinica corpus are summarized below. With five variant textual attribute parameters, we may adjust the proportions of text in different attributes to achieve an ideally balanced corpus. Another benefit of the hierarchical attribute assignment is that we can control our proportions of values according to different usages of the corpus. It is easy to establish subcorpora on the basis of our classifications. There are many different dimensions for us to compare all subcorpora from different viewpoints. We hope our work will lead to ideal criteria for a balanced corpus in the future. As for the word segmentation, we followed the draft standard of the ROC Computational Linguistic Society which might be the future national standard. The tag set is reduced from the category set of the CKIP Chinese lexicon under the criterion of keeping an unambiguous mapping between the word tag and syntactic category for each word. The resulting tagged corpus will benefit future tree bank construction, for the unique tag retains the information of the syntactic function and category for each word. The inspection system provides convenient tools for extracting and observing information hidden in the corpus by allowing the user to specify various linguistic and contextual conditions on the key word and the window.

6. References

Atwell, E. S., G. N. Leech, and R. G. Garside. 1984. Analysis of the LOB Corpus: progress and prospects. in Aarts and Meijs (eds.) Corpus Linguistics, 41-52.

Chang, Li-ping and Chen Keh-jiann. 1995. The CKIP Part-of-speech Tagging System for Modern Chinese Texts. Proceedings of ICCPOL '95 Conference, Hawaii.

Chen, Keh-jiann and Shing-huan Liu. 1992. Word Identification for Mandarin Chinese Sentences. Proceedings of COLING'92, pp. 54-59.

Chen, Keh-jiann, Shing-huan Liu, Li-ping Chang and Yeh-Hao Chin. 1994. A Practical Tagger for Chinese Corpora. Proceedings of ROCLING VII, pp. 111-126.

Church, K. W. 1988. A Stochastic Parts Program and Noun Phrase for Unrestricted Text. In Proceedings of 2nd Applied Natural Language Processing, pp. 136-143.

Church, K. & P. Hanks. 1990. Word Association Norms, Mutual Informatoin, and Lexicography. *Computational Linguistics*. 16.1:22-29.

Church, K. W. and R. L. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, Vol.19, No.1, pp.1-24.

CKIP. 1993. Analysis of Syntactic Categories for Chinese. CKIP Tech. Report #93-05, Institute of Information Science, Taipei.

CKIP. 1996. Sow Wen Jie Ji — A Study of Chinese Words and Segmentation Standard. CKIP Tech. Report #96-01, Institute of Information Science, Taipei.

Ellegard, A. 1978. The Syntactic Structure of English Texts: A Computer-based study of four kinds of text in the Brown University Corpus. *Gothenburg Studies in English*. 43.

Huang, Chu-Ren. 1994. Corpus-based Studies of Mandarin Chinese: Foundation Issues and Preliminary Results. In Matthew Chen and Ovid Tzeng (Eds.) In *Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*. pp. 165-186. Taipei: Pyramid.

Huang, Chu-Ren and Keh-jiann Chen. 1992. A Chinese Corpus for Linguistics Research. Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214.--1217. Nantes, France.

Huang, Chu-Ren et al., 1995. The Introduction of Sinica Corpus. Proc. of ROCLING VIII. pp. 81-99.

Huang, Chu-Ren, K. J. Chen and L. L. Chang. 1996. Segmentation Standard for Chinese Natural Language Processing. In Proceedings of COLING-96. pp. 1045-1048. Copenhager, Denmark.

Hsu, Hui-li and Chu-Ren Huang. 1995. Design Criteria for a Balanced Chinese Corpus. Proceedings of ICCPOL'95, Hawaii. pp. 319-322.

Kucera, H. and W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Lai, Yung-Hsiang. 1989. New Classification Scheme for Chinese Libraries. *Modern Library Science Series*. No.1

Liu, Shing-huan, K. J. Chen, L. P. Chang & Y. H. Chin. 1995. Automatic Part-of-speech Tagging for Chinese Corpora, Computer Processing of Chinese and Oreintal Languages, Vol. 9, No.1 pp.31-47.

Sinclair, John, 1987. *Looking Up —An account of the COBUILD Project in Lexical Computing*. London: Collins.

Sproat, R. and C. Shih, W. Gale & N. Chang. 1994. A Stochestic Finite-State Word-Segmentation Algorithm for Chinese. Proceedings of ACL 94, pp.66-73.

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

Appendix 1. Sinica Corpus Tagset

C=conjunction, D=adverbial, N=noun, P=preposition, V=verb, A=adjective, I=Interjection, T=particle, Str=Strings

Caa |/* 和、跟*/

Cab	/*等等*/
Cba	/*的話*/
Cbb	/*following a subject*/
Cbc	/*sentence initial*/
Da	/*possibly preceding a noun*/
Dfa	/*preceding VH through VL*/
Dfb	/*following a V*/
Di	/*post-verbal*/
Dk	/*sentence-initial*/
D	/*adverbial*/
Na	/*common noun*/
Nb	/*proper noun*/
Nc	/*location noun*/
Ncd	/*localizer*/
Nd	/*time noun*/
Neu	/*numeral determiner*/
Nes	/*specific determiner*/
Nep	/*anaphoric determiner*/
Neqa	/*classifier determiner*/
Neqb	/*postposed classifier determiner*/
Nf	/*classifier*/
Ng	/*postposition*/
Nh	/*pronoun*/
P	/*preposition*/
VA	/*active intransitive verb*/
VB	/*active pseudo-transitive verb*/
VC	/*active transitive verb*/
VD	/*ditransitive verb*/
VE	/*active transitive verb with sentential object*/
VF	/*active transitive verb with VP object*/
VG	/*classificatory verb*/
VH	/*stative intransitive verb*/
VHC	/*stative causative verb*/
VI	/*stative pseudo-transitive verb*/
VJ	/*stative transitive verb*/
VK	/*stative transitive verb with sentential object*/
VL	/*stative transitive verb with VP object*/
A	/*non-predicative adjective*/
I	/*interjection*/
T	/*particle*/
Str	/*string*/
DE	/*的, 之, 得, 地*/
SHI	/*是*/
YOU	/*有*/'
FW	/*foreign words*/