Academia Sinica Balanced Corpus

Ms. Yun-Chin Chou
Institute of **Information Science**

Academia Sinica Balanced Corpus (Sinica Corpus) is the first tagged Chinese corpus in the world. It is designed to provide a large database for linguistic analysis. Texts in Sinica Corpus are segmented and each word is tagged with grammatical categories. The texts are of various genres and topics, and are maintained and balanced by a systematic design in order to provide a valuable tool for a wide range of research purposes.

Sinica Corpus is the newest addition to Corpora Sinica--the Chinese corpora developed and maintained by Chinese Knowledge Information Processing Laboratory in Academia Sinica (Huang&Chen 1992). The CKIP group directed by Keh-Jiann Chen and Chu-Ren Huang began collecting chinese texts since 1990. In the past years, this projects has been funded by the CCK Foundational for International Scholars Exchange, the National Science Council of ROC, and Academia Sinica at various staffs. The Sinica Corpus version 1.0 was completed in July 1995, in which two million words were included. Sinica Corpus is available on Web since November 1996.

In order to maintain and balance a corpus with a large amount of database, the texts in Sinica Corpus are classified according to five attributes: sources, topics, genres, medium, style and mode(Hsu&Huang, 1995). The attribute values for classifying texts are established by consulting the Lancaster-Oslo/Bergen(LOB) Corpus(Atwell 1984), the Brown Corpus(Ellegard 1978), the Cobuild Project (Sinclair 1987), and Chinese library topic classification(Lai 1989). The attribute of topic is the primary consideration to balance the corpus, while distribution of the other four attributes are monitored and checked at least once a month. With five different attributes, Sinica Corpus is versatilely balanced and a subcorpora can be difined with a specific set of attributes based on different research purposes (Hsu&Huang, 1995) .

The corpus browsing system offers an useful tool for linguistic reasearch The corpus retrieval system locates keywords and categories specified by users. Other functions such as filtering, sorting, printing, saving, statistic and collocation are also available. The WWW version of Sinica Corpus was implemented by Lin Shi of Academia Sinica Computing Center in 1996. For references, relevant CKIP technical reports can be purchased from ROCLING:

(E-mail: rocIing @ bdc.com.twFax:886-3-577-0459)