# From Synergy to Knowledge:
## Corpus as a natural format for integrating multiple educational resources

*Chu-Ren Huang*
*Institute of Linguistics, Academia Sinica*
http://www.ling.stinica.edu.tw/cwn/huang.html

**Abstract**

Integrating information from multiple domains and cross-lingual sources is probably the most important skill to be learned by student these days. In turn, design of a high-performance learning environment needs to incorporate multilingualism and multi-domain information. We demonstrate in this talk how various corpora and language resources can be effectively integrated to provide an infrastructure of synergy for new knowledge. First, by integrating a billion word corpus with in-depth grammatical knowledge, *Chinese WordSketch* is a system which generates linguistic descriptions that can be easily applied in language pedagogy. The most salient usage of each word, as well as how the uses of two near synonyms contrast with each other, can be automatically summarized, and with hundred of supporting example sentences. Second, *Hantology* integrates the conventionalized structure of Chinese characters with the cutting edge knowledge engineering theory of ontology. In terms of language teaching, it provides a meaningful way to breakdown components of Chinese characters to learn to write them, as well as a more explanatory framework to derive their meanings. Lastly, corpora can be integrated for structured learning, as in *Adventures in Wen-Land*. Three different curricula of elementary school Mandarin in Taiwan are converted to corpora and integrated as the main framework of this digital language learning site. Resources dedicated to linguistic skills (*classifier*, *chengyu*), and classical literature (*Tang poem, The Dream of the Red Chamber*, etc.) are linked through a tracked lexical list. This allows cross-curricula and cross-domain learning for both teachers and students.

## I. Introduction:

### 1.1 Motivations

In this talk, I will underscore the importance of integrating multiple resources in corpus-based approaches to education. There are three strong motivations for

taking the integrating approach: that learning is a multi-modal activity; that Asia is a multilingual society, and that new knowledge is created by synergizing old knowledge.

First, it is obvious that we acquire knowledge by seeing, reading, listening, and even touching. Is it also not difficult to see that learning materials may exist in different forms at different locations. Integrated educational resources allow learners to be exposed to the full extent or relevant information and to choose the most accessible approach for each individual learner. Moreover, a simple resource can serve many different learning functions. For instance, the quotation 'The rain in Spain never falls on the plains' can be used to teach phonics, rhyming, meteorology, Spain tourism, and of course, literature. Integrating educational materials from multiple sources make the learning process realistically situated and more productive.

Second, the fact that Asia is multilingual society is often overlooked since we often speak no more than one of the more dominant western languages. However, it is more likely than not that each Asian society support two or more local languages in addition to the gaining popularity of English. Cantonese, Hokkien, Malay, Mandarin, and Tamil are all local languages in Singapore. And Taiwan has Hakka, Mandarin, Taiwanese and more than 10 Formosan languages. It is not practical to create exactly parallel resource for all languages. Hence the need to link and integrate resources from different languages is great.

Third, the English expression 'putting two and two together' illustrates vividly the process of learning and acquiring new knowledge. Learning does not create new entity. It simply gives us new information about existing entities. We put two and two together to get 4, but 4 must exist a priori for the learning process to be valid. Hence the more knowledge source we have and the better we put them together, the more knowledge we can acquire. I believe that fact that corpora and other electronic resources from widely different origins are becoming more readily available makes is possible to integrate them and to create an environment for discovery of new knowledge. Hence this talk is entitled 'From Synergy to Knowledge.'

### I.2. Three Modes of Integration

In the remaining part of this paper, I will further elaborate our approaches to integrating multiple resources by discussing three integrated resources that our Academia Sinica team created. The first is **Chinese WordSketch**, where lexical knowledge of verb subcategorization is added to the powerful

corpus-access platform of Sketch Engine. This powerful tool is then applied to a 1,400 million character corpus to discover rich information such as collocation and synonymy. The second is **Hantology**, where we map the radical based Chinese character database to the conceptual of a upper onotology. This mapping allows us to make the meaning composition of Chinese characters more explicit. We also devise the infrastructure for Chinese textual databases to be mapped to Hantology to show variations in character forms and meaning. The third one is **Adventures in Wen-Land**, a language learning website integrating multiple resources. The Adventures in Wen-Land is anchored on the corpus of elementary school textbook. A wordlist organized by the years of teaching is then linked to various reference resources to enrich teaching. The resources linked include general and special dictionaries, literature texts, as well as learning games. I will end with a short conclusion to summarize.

## II. Chinese WordSketch: Discovering grammar in a gigantic corpus

http://www.sketchengine.co.uk
http://www.ling.sinica.edu.tw/wordsketch (for Taiwan only)

The dilemma of applying corpus to language learning is that many interesting linguistic facts cannot be found when corpus is not big enough, but when corpus is over certain size, it is often very difficult for a human to browse and use. For instance, there are 24,858 instances of the verb ‘to speak’ in the 100 million words British National Corpus. The number is certainly big enough to contain almost all interesting linguistic behaviors of the verb. Yet, it will be a daunting task for a person to go through all these instances and not make any mistakes, not to mention wasting time on so many similar examples. Since it is not practical to expect a researcher or a teacher to sort out so many examples effectively, it would be even more unreasonable to expect a regular learner to do so. The above dilemma can be solved if there is a reliable computational tool to reliably extract relevant grammatical patterns, and hence allowing the rearcher/teaher/learner to go through the patterns without having the tedious job of going through thousands of example sentences. It will be even better if the relevance among the patterns can be automatically predicted and ranked such that human only need to check out the more relevant patterns. The Sketch Engine, designed by Adam Kilgarriff and his colleagues (Kilgarriff et al. 2002) is created to solve this problem.

The original goal of corpus-based studies was to provide ‘a body of

evidence' for more theoretical linguistic studies (Francis and Kucera 1965). However, corpus-based studies evolved with the improvements made in electronic data manipulation as well as the popularity of new applications in corpus-based learning, highlights the emergent need of tools to automatically acquire grammatical information. Previous works that made significant contribution to the study of automatic extraction of grammatical relation includes Sinclair's (1987) work on KWIC, Church and Hanks' (1989) introduction of Mutual Information, and Lin's (1998) introduction of relevance measurement. Kilgarriff and colleagues' work on Word Sketch Engine (WSE) makes a bold step forwards in automatic linguistic knowledge acquisition (Kilgarriff and Tudgell 2002, Kilgarriff et al. 2004). The main claim is that a 'gargantuan' corpus[1] contains enough distributional information about most grammatical dependencies in a language such that the set of simple collocational patterns will allow automatic extraction of grammatical relations and other grammatical information. Crucially, the validity of the extracted information does not rely on the preciseness of the rules or the perfect grammaticality of the data. Instead, WSE allows the presence of ungrammatical examples in the corpus and the possibility for collocational patterns to occasionally identify the wrong lexical pairs. WSE assumes that these anomalies will be statistically insignificant, especially when there are enough examples instantiating the intended grammatical information. In addition, WSE relies on a salience measurement to rank the significance of all attested relations

### II.1. The Sketch Engine

The Sketch Engine is a corpus processing system developed in 2002 (Kilgarriff et al. 2004). The main components of the Sketch Engine are KWIC concordances, word sketches, grammatical relations, and a distributional thesaurus. In its first implementation, it takes as input basic BNC (British National Corpus) data: the annotated corpus, as well as list of lemmas with frequencies. The Sketch Engine has a versatile query system. Users can restrict their query in any sub-corpus of BNC. A query string may be a word (with or without POS specification), or a phrasal segment. A query can also be performed using Corpus Query Language (CQL). The output display format can be adjusted, and the displayed window of a specific item can be freely expanded left and right. The most relevant feature is that the Sketch Engine

[1] The required corpus size was not specified in WSE literature. However, we estimate from existing work that for WSE to be efficient, corpus scale must be 100 millions words or above.

produces a word sketch that is an automatically generated grammatical description of a lemma in terms of corpus collocations (Kilgarriff et al. 2002). All items in each collocation are linked back to the original corpus data.

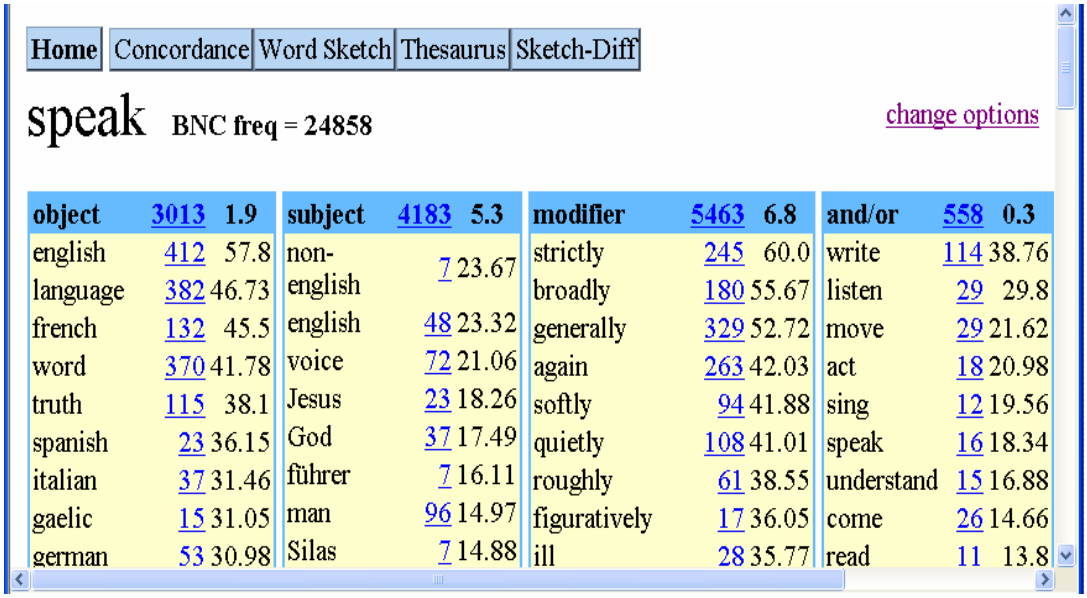


Figure 1, WordSketch of ‘To Speak’ based on BNC

A Word Sketch is a one-page list of a keyword’s functional distribution and collocation in the corpus. The functional distribution includes: subject, object, prepositional object, and modifier. Its collocations are described by a list of linguistically significant patterns in the language. Word Sketch uses regular expressions over POS-tags to formalize rules of collocation patterns. For instance, (1) is used to retrieve the verb-object relation in English:

(1) . 1:”V” “(DET|NUM|ADJ|ADV|N)”* 2:”N”

The expression in (1) states that: extract the data containing a verb followed by a noun regardless of how many determiners, numerals, adjectives, adverbs and nouns preceding the noun. It can extract data containing *cook meals* and *cooking a five-course gala dinner*, and *cooked the/his/two surprisingly good meals* etc.

The Sketch Engine also produces thesaurus lists, for an adjective, a noun or a verb, the other words most similar to it in their use in the language. For instance, the top five synonym candidates for the verb *kill* are *shoot* (0.249), *murder* (0.23), *injure* (0.229), *attack* (0.223), and *die* (0.212). It also provides direct links to the Sketch Differences which lists the similar and different

patterns between a keyword and its similar word. For example, both *kill* and *murder* can occur with objects such as *people* and *wife*, but *murder* usually occurs with personal proper names and seldom selects animal nouns as complement whereas *kill* can take *fox, whale, dolphin*, and *guerrilla,* etc. as its object.

**II.2. Implementing Chinese WordSketch**

In order to show the cross-lingual robustness of the Sketch Engine as well as to construct a powerful tool for grammar discovery in Chinese; we created Chinese WordSketch (CWS) by loading the Chinese Gigaword to the Sketch Engine. The Chinese Gigaword Corpus (CGW) 2.0. contains about 1.29 billion Chinese characters, including 792 million characters from Taiwan's Central News Agency (CNA), 471 million characters from China's Xinhua News Agency (Xinhua), and 28 million characters from Singapore's Lianhe Zaobao (Zaobao). In the original data collected in version 1.0. Taiwan's CNA is from 1991 to 2002, and Mainland China's XIN is from 1990 to 2002. Each file contains all documents for the given month from the given news source. Version 2.0. added data from Singapore and extended the coverage to 2003.

CGW Corpus, as prepared and released by LDC, however, is pre-processed for data uniformity, but is neither word-segmented nor tagged with POS information. Segmentation and tagging was performed adopting the Academia Sinica segmentation and tagging system (Ma and Huang 2006). The actual running time of segmentation and tagging took over 3 days to perform. The corpus after processing contains more than 825 million words, each tagged and linked to the 2,803 documents originally marked-up with sources and topics. The distributional information is given in Table 1.

| | Source | Character | Word | Doc. |
|---|---|---|---|---|
| **First Edition** | **CNA** | 735 | 462 | 1,649 |
| | **Xinhua** | 382 | 252 | 817 |
| | **TOTAL** | 1,118 | 714 | 2,466 |
| **Second Edition** | **CNA** | 792 | 497 | 1,769 |
| | **Xinhua** | 471 | 310 | 992 |
| | **Zaobao** | 28 | 18 | 41 |
| | **TOTAL** | 1,291 | 825 | 2,803 |

Table 1: Tagged CGW Corpus, Edition 1.0 and 2.0

All components of the Sketch Engine were implemented and applied to the tagged CGW Corpus,[2] including Concordance, Word Sketch, Thesaurus and Sketch Differences. Screens captured directly form the CWS system are given below to illustrate the kind of information available in WordSketch (Fig. 1), linking from WordSketch to concordance (Fig. 2), SketchDifference (Fig. 3), and Thesaurus (Fig. 4).

[2] The segmented and tagged version of CGW Corpus will be available from LDC later in 2007.

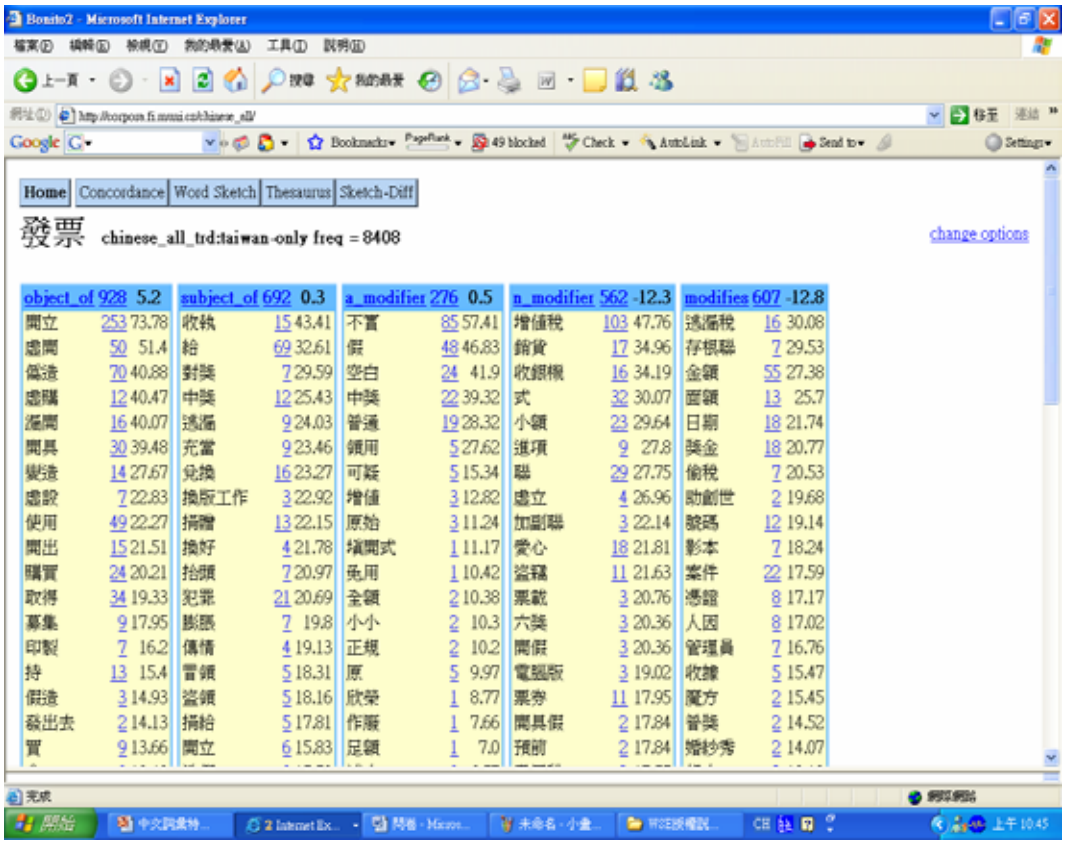


Figure 1. Chinese WordSketch of 發票 'Invoice'

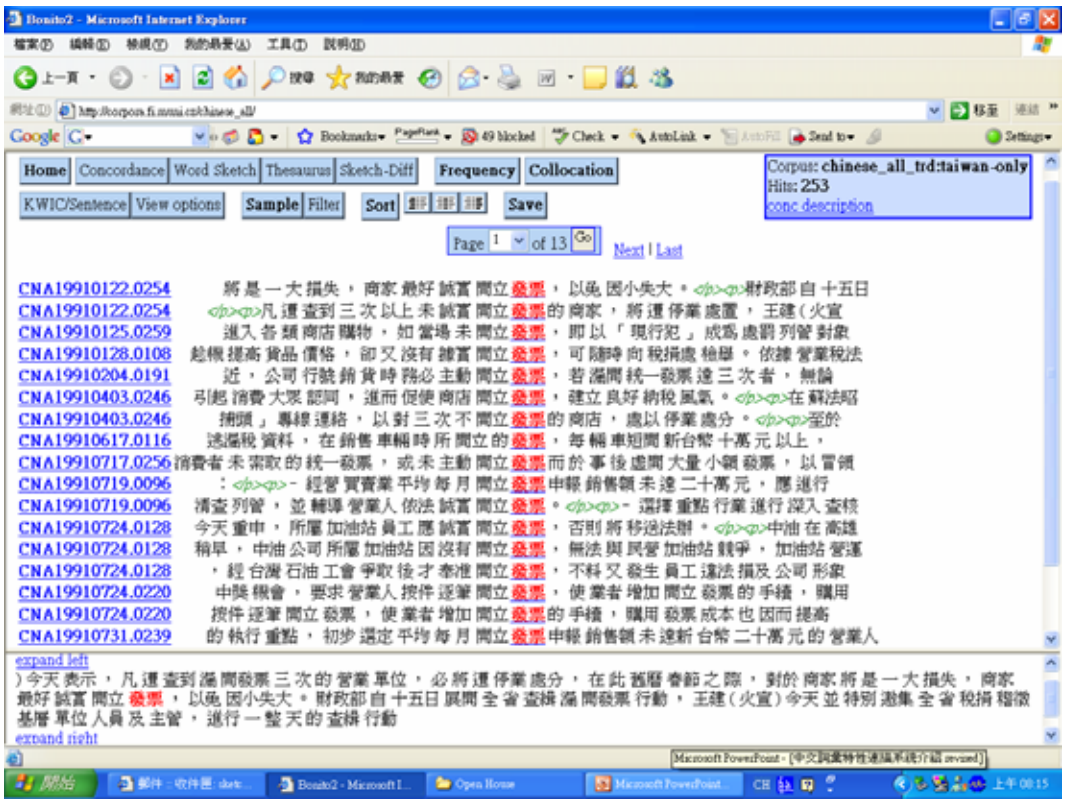


Figure 2. Concordance Paged Linked from CWS

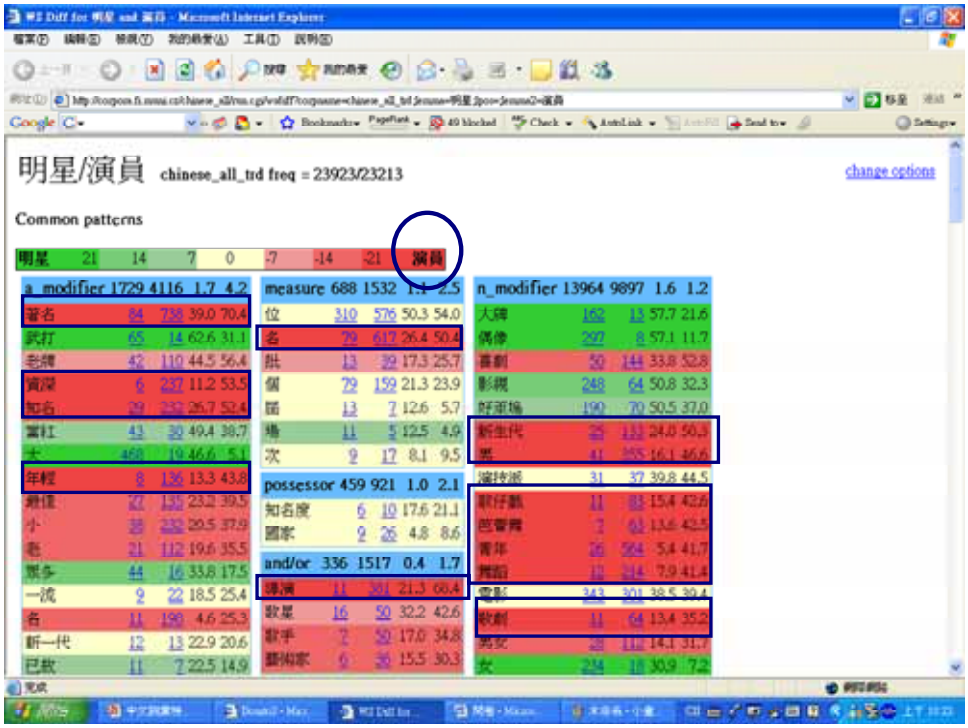


Figure 3. Sketch Difference between 明星'star' and 演員 'actor'

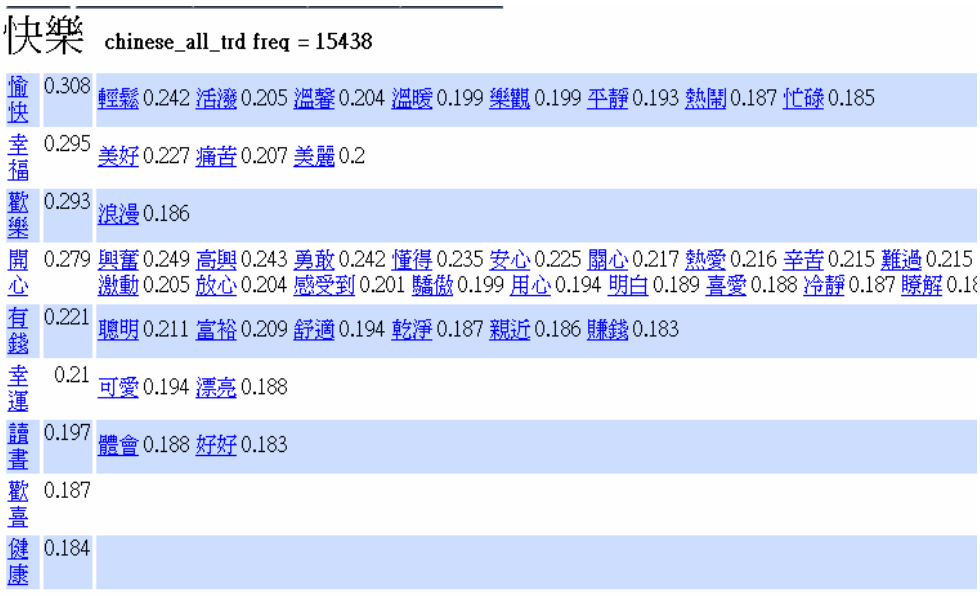


Figure 4. Thesaurus of 快樂 'be happy'

### II.3. How Grammatical Knowledge was Integrated to Generate More and Better Knowledge

The most crucial issue involved in the implementation of CWS was what kind of information is needed to ensure effective and reliable extraction of

grammatical information of Chinese. A first and naïve approach towards implementation using English like grammar templates were promising yet unsatisfactory (Huang et al. 2005). After error analysis, our conclusion was that the original templates inadequately represented Chinese grammar. Hence we adopted the rich lexico-grammar of ICG (Chen and Huang 1992), and implemented CWS Version 2 (Huang and Hong 2007). We show that when better grammatical resources are used, the recall, precision, and the coverage of different verb subcategories have improved drastically. Hence we showed that it is worthwhile to pay attention to linguistic issues in NLP.

## III. Hantology: What can character components tell us?

http://www.ling.sinica.edu.tw/hantology

Hantology as a character-based Chinese language resource focuses on how to represent the implicit conceptual structure conventionalize by the Chinese writing system. Unlike alphabetic or syllabic writing systems, the ideographic writing system of Chinese poses both a challenge and an opportunity. The challenge is that a totally different resources structure must be created to represent and process speaker's conventionalization of the language. The rare opportunity is that the structure itself is enriched with conceptual classification and can be utilized for ontology building. There are two important areas of significance with regard to this work. The first, in light of the emergence of the Semantic Web and the requirement of ontology as most important media to represent knowledge on the web, is that Hantology may be the key to allow Chinese based resources to thrive in the Semantic Web. The second, it terms of integrating language resources, is that this work shows how implicit human knowledge structure can be made explicit by mapping to a well-formed common knowledge representation.

### III.1 Character and Chinese Processing

One of the recalcitrant problems in Chinese text processing is the computers can not process variants correctly. Chinese characters have lots of variants which are the different glyphs with the same word or morpheme. For instance, both characters '体' and '體' are the same word and morpheme with different glyphs. Actually, they are variants and can be replaced each other. These problems also cause information retrieval and interchange problems. For computer systems, it is very important to know the meaning and concept carried by this different form. For English, each character is just a writing unit

without carrying any concept. Therefore, it does not have the requirement to build resources for alphabetic characters. However, each Chinese character is not only a writing unit but also a concept unit. Because there are lots of relations among concepts, the characters are not independent of each other.

There are several studies on the creation of Chinese characters database. One important study is Chinese glyph expression database which consists of 59000 glyph structures (Juang & Hsieh, 2005). The glyphs of Chinese characters are decomposed into 4766 basic components. Each Chinese character can be expressed by the basic components. Chinese glyphs database also contains oracle bone, bronze, greater seal and lesser seal scripts. The largest Chinese characters database is Mojikyo font database which contains more than 110000 characters (Ishikawa, 1999). Both Chinese glyph expression database and Mojikyo font database contain only glyph knowledge. Yung created an ancient pronunciations database for Chinese characters (Yung, 2003). Hsieh proposed a HanziNet which represent Characters by 16 bits binary code (Hsieh, 2005). Chinese characters are classified into hierarchy categories. HanziNet can describe the upper layer concept of a character. These previous studies only conceded on one dimension of Chinese characters. However, each Chinese character consists of glyphs, scripts, pronunciations, senses, and variants dimensions. The previous studies can not provide enough knowledge for computer applications and researchers. Chou and Huang propose an ontology named Hantology to provide glyph, script, pronunciation, sense, and variants of Chinese characters (Chou, 2005; Chou & Hung, 2005)

**III. 2 The Contents of Hantology**

Hantology describes orthographic forms, phonological forms, senses, variants, variation and lexicalization of Chinese writing system. The orthographic composition of a Chinese word is either ideographic or radical-phonetic pairing. In general, each Chinese character is not only a writing unit but also itself a word or morpheme. The most important feature of Chinese writing system is that orthographic forms and senses are extensions of semantic symbols, so the concepts indicated by semantic symbols become the core of Chinese writing system. In this study, we use 540 radicals of ShuoWen as basic semantic symbols. To enable the conceptualization and relation of semantic symbols to be processed by computer systems, the concepts indicated by each radical are analyzed and mapped into IEEE Suggested Upper Merged Ontology (SUMO). In addition, adopting SUMO allows Hantology to integrate

and share with other ontologies like WordNet or the Academia Sinica Bilingual Ontological WordNet (Sinica BOW, Huang, et al., 2004).

The senses of each Chinese character also adopt SUMO to represent the conception and relation among various senses. The lexicons generated by different senses are constructed to express the morphological context. Since the senses depend on pronunciations, the relation between pronunciations and senses are described by Hantology. In Chinese writing, there are lots of variants which are different orthographic forms of the same word or morpheme. A linguistic context is proposed to describe the relations of variants.

To illustrate the major contents of Hantology, we use the character '臭' as an example. Figure 5 shows the glyphs, pronunciations and variants for '臭'. The principle of formation is '會意'(ideographic compound). Glyph evolution shows the ancient glyph of '臭' is '臭'. '臭' is a verb when it used as to smell by nose. There are four variants for the sense of smell. All senses are mapped into SUMO. '後作' in the Figure 5 means '臭' is replaced by '嗅' to express the sense of smell. The first citation appears in period of '唐' (Ton dynasty). For instance, the sense of '臭名' (a bad reputation) is mapped into the concept '主觀評價屬性'(Subjective Assessment Attribute) in SUMO. The generated word from sense of '臭名' is '遺臭萬年' which appeared in period of '元' (Yuan dynasty).

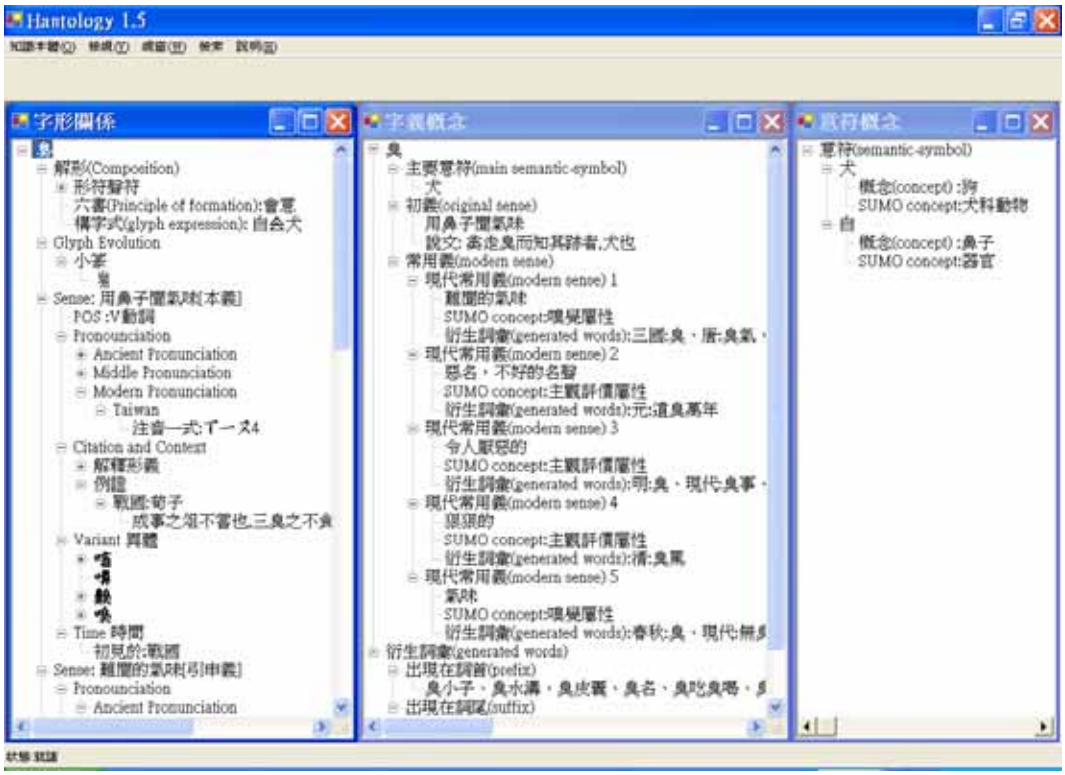


Figure 5. Browsing Interface of Hantology

### III.3The Ontology of a Semantic Radical

An important discovery by in studies based on Hantology is that each Chinese radical represents a small conceptual system. Contrary to previous observations, the cluster of concepts realized by a radical can be associated to the core meaning by a set of rules. We illustrate with the radical 艸 cao3 ‘grass’, which is generally considered to represent the concept ‘plants’.

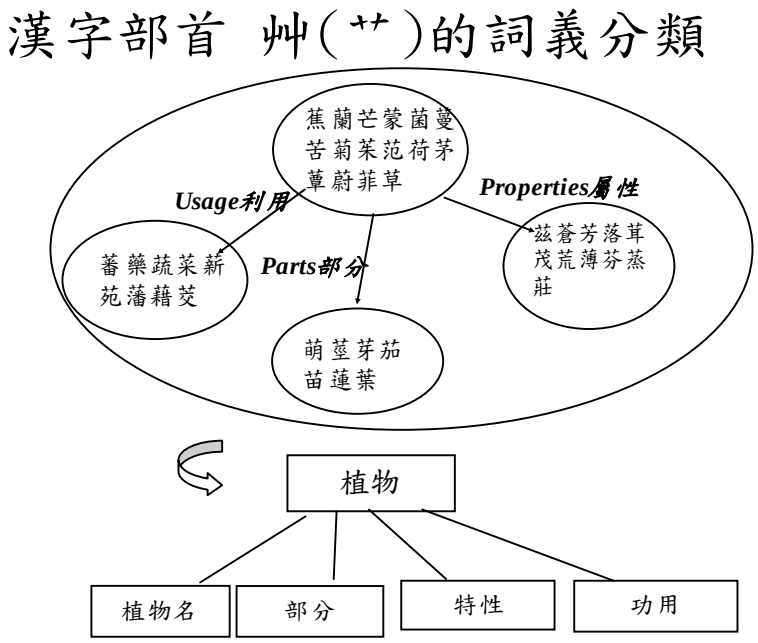


Diagram 1. Conceptual Classes Represented by the Plant Radical *CAO3*

Of the 444 characters sharing the radical 艸, what is surprising is that the conceptual clustering is not simply of taxonomy classification. As seen in Diagram 1, there are four productive relations described by the radical: being a kind of plant (e.g. orchid), being a part of a plant (e.g. leaves), being a description of a plant (e.g. fallen (leaves)), and being the usage of a plant (e.g. medicine). We observe that this actually attests to Pustejovsky’s (1995) recent theory of generative lexicon, where formal, constitutive, descriptive, and telic are the main motivations for semantic changes and coercions. The fact that these are the same principles used for deriving Chinese characters 3000 years ago suggests that there is cognitive validity.

The most important implications of this discovery, is that there is an underlying conceptual structure of the Chinese character writing system which can be utilized in reading and writing. This is not only promising for Chinese in future semantic web applications, but also offers a potentially useful

conceptual representation shared by several language in the Sino-Sphere. It is also likely that the explicit concept-based structure will help students learning writing Chinese characters in the future.

## IV. Adventures in Wen-Land: Synergizing language resources for language learning

http://www.sinica.edu.tw/Wen/

The two different ways to integrate corpus and language resources introduced in the last two sections led to the natural question of whether such synergy is directly possible in education and how can it be done. I answer this question by introducing the integrated language learning resources: Adventures in Wen-Land (*Wenguo* here after), created in 2001.

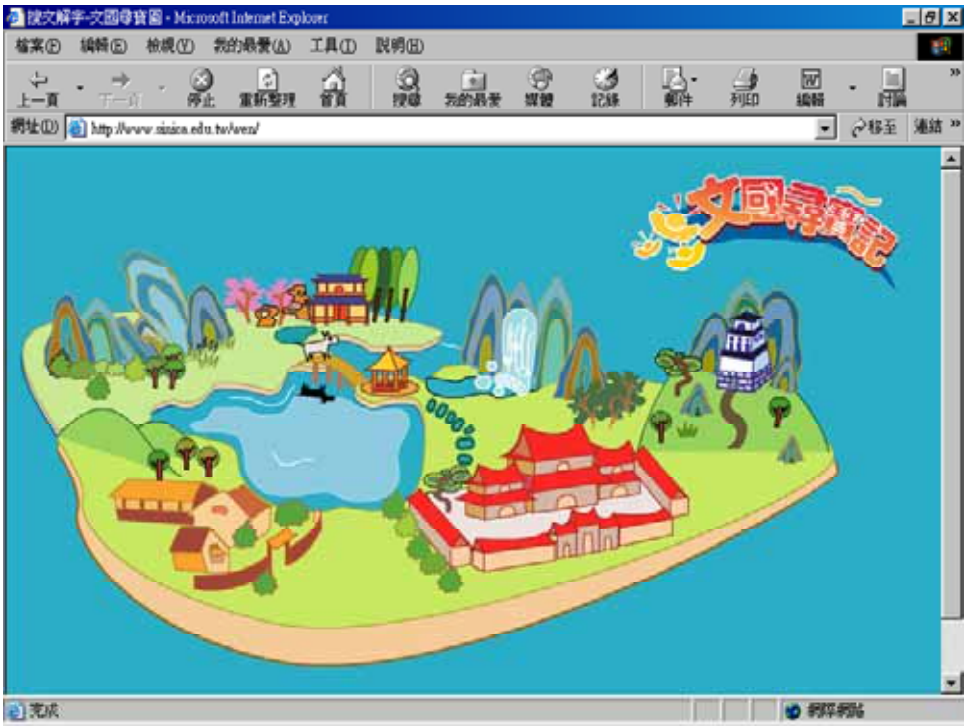

Figure 6. Homepage of Adventures in Wen-land

### IV.1 Overview

*Wenguo* Is a virtual theme park for on-line Chinese language learning and teaching. It s the end product of a National Digital Museum Project sponsored by the National Science Council, ROC (A Linguistic and Literary KnowledgetNet for Elementary School Children). It was completed in spring, 2001 by a team of researchers include Chu-Ren Huang (a linguist), Feng-ju Lo (a literary scholar), Hui-chun Hsiao (an art-designer specialized on

web-design), as well as a team of consulting elementary school teachers. The implementation work is done by a team of research associates with either computer science of linguistics background.

### IV.2. Design Criteria and Implementation

The design criterion of *Wenguo* is to use a lexicon as the threads that hold and pull together the different information that exist among different resources. A lexical needle picks up and connects only the textual materials that matches its description, and creates hyperlinks among the relevant data. In particular, we assume users will be using textbook vocabulary to guide them in *WenGuo.* Hence a reference lexicon based on the textbooks of elementary school on Chinese is build to serve as both a lexical thread and textual filter.


Figure 7. Integrated Resources as Learning Background in *Wenguo*

The Lexcial knowledgebase we build using the integrated corpora of three different editions of text books provides both chronological (*e.g. when a word is first taught/learned)* and distributional (e.g. frequency) feature of each word. Given this lexical chronology, a learner/teacher can easily gauge the progress of learning, as well as acquire related learning materials from other sources. An example of this integrated lexical entry is given in Figure 8.

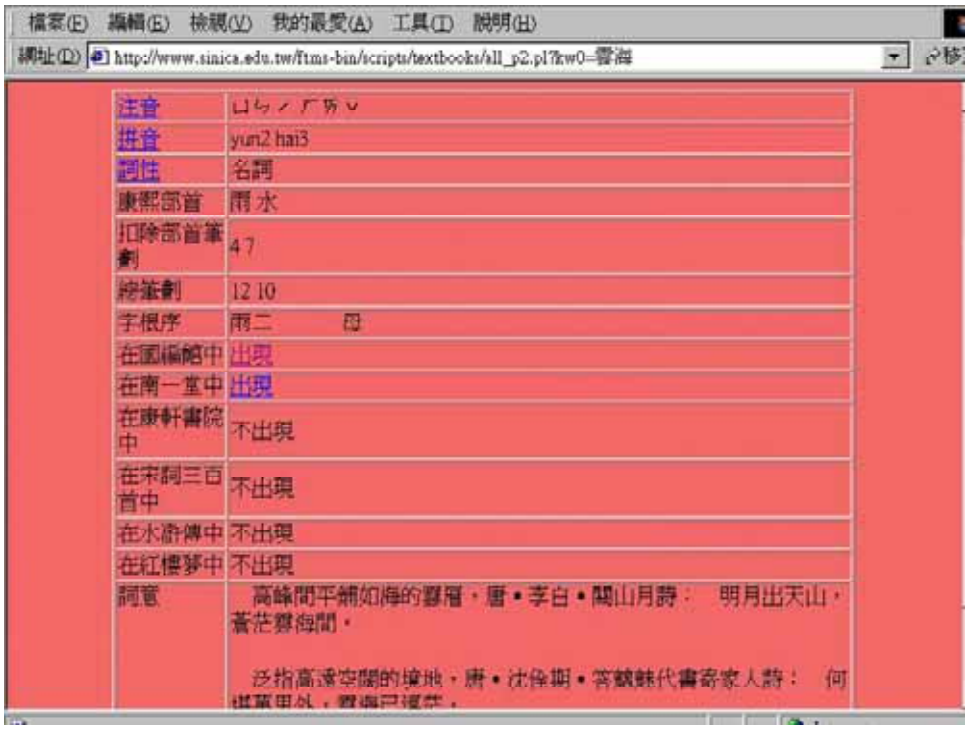


Figure 8. Integrated Lexical KnowledgeBase entry of 雲海 yun2hai3

The 雲海 yun2hai3 ‘Sea of Clouds’ entry that was given is an entry based on the three textbook corpora combined. We can see that the information given include two types of spelling (national phonetic alphabets (bpmf) and Pinyin romanization), part of speech, semantic radical, stroke numbers excluding radical (for each component character), total stroke number (for each character), *BuJian* component order for each character, cross-reference in three textbooks, cross-reference in three Chinese classics (Song Poetry, On Water Margin, and Dream of the Red Chamber), and lastly, a citation of the entry from Ministry of Education dictionary.

*Wenguo* takes full advantage of the rich lexical knowledge as well as the power of electronic media. Hence it is possible to search for a word by Chinese character,bpmt, Pinyin, or *Bujian* character components. It also allows search to specify initial consonant, syllable length of word, reduplication patterns. Once a search result is returned in the form of an entry, hyperlinks are already established. For instance, the screen shown in Figure 8 can be linked directly to the textbook sentences containing 雲海 yun2hai3, or its use in one of the literary classics (none in this case).

The citation of Bu4jian4 perhaps needs a little elaboration. The part represents shared knowledge of the glyph composition of Chinese characters, which is now represented by a limited number of bujian inventories. Note also

that since bujian are not necessarily words, it is not usual for a normal computer to not to be able to display them all, as given in Figure 8.

The cross-reference function can be exemplified by the ge5, the default classifier of Mandarin Chinese. The following search result can be found either at xue2tang2 (classroom, 學堂), the central lexical link to all learning module, or at hei1bai2gong1 (Black and White Castle, 黑白宮), the module devoted to learning classifiers.[3]


Figure 9. Citation of ge5 in Three Textbooks

We can see that even though the neutral classifier in *ge5* 個 is introduced very early in all three textbooks, there are still temporal discrepancies. It is taught as early as in the fifth lesson of the first semester or as late as the first lesson of the second semester. We have already seen that a it is even possible for a word to be present in one textbook and missing in another, as shown with yun2hai3 in Figure 8. Our design turns the necessary evil of having severl non-standardized textbooks into an advantage. On one hand, by allowiing for a user's to specify his/her grade level at school, *Wenguo* can deduce immediate whether a word is new vocabulary for the learner or not. One the other hand, the same mechanism can be used by a teacher or a gifted student to find supplementary learning materials without being constrained by the textbook

[3] I apologize for the typo that appears in Figure 8. The tone mark of *ge5* was incorrect in an earlier implementation and remained on this earlier screen shot. I did not find it out in time to change.

their school happens to choose. Please note that all these extra functions are possible because our team tagged the three versions of textbooks as three corpora: including the textual mark-up of year and lesson and the content mark-up of part-of-speech.


Figure 10, Nouns Selected by the Classifier 個 ge5

Lastly, we illustrates how this integrated corpora approach can be applied to learning of a specific grammatical skill in *Wenguo*. Noun-classifier collocation selection is one of the most difficult linguistic properties of Chinese. Huang et al.'s (1997) 'A Noun-Classifier Collocation Dictionary of Chinese' gives all possible nominal patterns that can be selected by a classifier. Their approach is corpus-based since all possible noun-classifier pairs were extracted from Sinica Corpus and analyzed. It is also utilizes knowledge of Chinese linguistics as a noun is selected by a classifier according to its last character (i.e. the head in linguistic terms). For instance, 張 zhang1 selects 床 chuang2, therefore there is no need to duplicate the possibly infinite list of 彈簧床 tan2huang2chuang2， 眠床 mian2chuang2，水床 etc. Note that this approach also takes care of neologism, which was not possible without a new edition of dictionary in the traditional method. *WenGuo* adopts this dictionary and adds versatility to it by linking to all Sinica Corpus examples, as well as attested instances from the textbooks. Hence a student can take both a rule-based

(dictionary) or an example-based (corpus) approach to learn classifiers.

### IV.3 Summary

In sum, *Weguo* integrated many language resources to created synergy for learners. It harmonized three different versions of textbooks with three Chinese classics. In addition, special modules are created to deal with specific linguistic skills. The design depend crucially on a synchronized lexical knowlegebase constructed based on more than 10 tagged corpora.

## V. Conclusion

In this paper, the focus is on how to create synergy and acquire new knowledge from different language resources, especially corpora. We illustrated how grammatical information can be included in a corpus-processing platform to enable automatic discovery of more comprehensive linguistic knowledge. We also demonstrated how a conventionalized and implicit linguistic ontology can be merged with a formal and explicit ontology.

Last, and perhaps most directly relevant to corpus application in education, we showed that several annotated corpora can be threaded together with a structure knowledgebase. This then can be converted to a versatile online language learning site, such as *Adventures in Wen-Land.* The fact that versatility and multi-functionality were attained attested to the corpus-driven approach as well as to the central role corpus annotation plays.

**Acknowledgment**:

I would like to thank all my collaborators in the projects reported, especially Keh-jiann Chen, Feng-ju Lo, Ya-min Chou, as well as all colleagues at CKIP and the Chinese WordNet group at Academia Sinica. Any remaining errors are, of course, mine.

## Bibliography and Websites

Academia Sinica Balanced Corpus for Mandarin Chinese 中央研究院現代漢語平衡語料庫. http://www.sinica.edu.tw/SinicaCorpus

Academia Sinica Bilingual Ontological Wordnet 中央研究院中英雙語知識本體詞網. October 2003. http://BOW.sinica.edu.tw

Chinese Gigaword Corpus.http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09

Chinese Knowledge Information Processing. (1993). The Categorical Classification of Chinese. 3rd Edition. [In Chinese] CKIP Technical Report 93-05. Nankang, Academia Sinica.

Chou, Y. and C. Huang. (2006). Hantology-A Linguistic Resource for Chinese Language Processing and Studying. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Itlay. 24-28 May, 2006.

Fillmore, C. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In Jan Svartvik (ed.) *Directions in Corpus Linguistics*. (Proceedings of Nobel Symposium 82), Berlin: Mouton de Gruyter

Huang, C., Chen, K., Chang, L. and Hsu, H. (1995). An Introduction to Academia Sinica Balanced Corpus. [In Chinese]. *Proceedings of ROCLING VIII.* 81-99.

Huang, C. and Chen, K. (1992). A Chinese Corpus for Linguistic Research. *Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92).* 1214-1217 Nantes, France.

Huang, C., Chen, K., and Chang, L. (1997). Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing.* 2.2.47-62.

Huang, C., Kilgarriff, A., Wu, Y., Chiu, C., Smith, S., Rychly, P., Bai, M., and Chen, K. (2005). Chinese Sketch Engine and the Extraction of Collocations. Proceedings of the Fourth SigHan Workshop on Chinese Language Processing. October 14-15. Jeju, Korea.)

Huang, C. W. Ma, Y. Wu and C. Chiu. (2006). Knowledge-Rich Approach to Automatic Grammatical Information Acquisition: Enriching Chinese Sketch Engine with a Lexical Grammar. Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. November 1-3, Wuhan, China.

黃居仁，陳克健，賴慶雄. （1997）編著.國語日報量辭典. 台北： 國語日報出版社。

**黃居仁**，盧秋蓉，張如瑩. (2004).語言知識網路與未來的數位學習—以「文國尋寶記」為例. 羅鳳珠主編。語言，文學與資訊。頁 487-536。新竹：清華大學出版社。

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. Proceedings of EURALEX, Lorient, France (http://www.sketchengine.co.uk/)

Kilgarriff, Adam and Tugwell, David. Sketching Words. (2002). In Marie-Hélène Corréard (ed.): Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Euralex

Kilgarriff, A., Huang, C., Rychly, P., Smith, S., and Tugwell, D. (2005). Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.

Ma, Wei-Yun and Huang Chu-Ren. (2006). Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Proceedings of the Fifth International Conference on Language Resources and Evaluation. Genoa, Italy.

The Open Language Archives Community, http://www.language-archives.org, or http://linguistlist.org/olac/

Simons, G. and Bird, S. (2003). The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing*, 8(4), 259-65.

*SouWenJieZi* - A Linguistic KnowledgeNet. August 1999. http://words.sinica.edu.tw/