Journal of Chinese Language and Computing, 15 (2): (53-71)

XHK: The Grammar-based Lexical Semantic Knowledge base

Wang Hui[1] Su Xinchun [2]

1.National University of Singapore，chswh@nus.edu.sg

2.Xiamen University, China，suxch@jingxian.xmu.edu.cn

Abstract

Although the semantic analysis and the grammatical distribution are treated as separate issues in linguistic theories, there is a close interconnection between the two in that the differences in the lexical meanings are often realized at the levels both of the grammatical function and of the lexical collocation. This is the basic assumption utilized when we design and develop Knowledge Base of Modern Chinese Lexical Semantics (XHK). By incorporating the 61，000 entries and 82,000 semantic items in *Modern Chinese Dictionary*, it presents a detailed package of the information about the lexical items in Chinese, including the basic meaning, the grammatical meaning, the associative meaning and the detailed descriptions of the behaviors of the lexicon based on the 20-million-word Central Corpus of Modern Chinese of Ministry of Education, for example, word frequency, lexical meaning frequency, and numerous cases where the word is used. The major purposes of this research are three-fold: based on the theoretical findings of the modern grammatical research, to conduct a comprehensive lexical-item-based description of the characteristics of the Chinese lexicon and their collocation patterns; to improve the functions of developing the dynamic semantic rules and retrieving the lexical components; and finally, to establish a fresh and feasible theoretical framework which describes the Chinese lexical semantic knowledge.

Keyword:

Lexical Semantics, components, grammatical distribution, semantic category

1. Introduction

For a very long time, the dominant paradigm in the linguistic enterprise is to consider the lexical semantics and syntax as two separate issues: while many scholars in the syntactical camp discard lexical meaning as diffuse, loose phenomena which vary a great deal from language to language, most scholars in the circles of lexical semantics focus on the

description of the lexical and semantic systems, without considering the syntactical factors involved.

This situation continued until the 1970s, when some advocates of the Generative Semantics, Kazt and Fodor incorporated lexical semantic analysis into the syntactical frameworks. After 1980s, the interface between lexical semantics and syntax became the focus of many American and European scholars, who began to employ the lexical semantic approach to analyze the syntactical structures and semantic relations between words and sentences. The issue that attracted the most scholarly attention was how to identify the nuance in the lexical meaning by observing the behaviors displayed by the lexical items in the sentence, and to investigate semantic changes, distribution patterns and collocation restrictions and so on. Lyons (1995) observes that any lexical meaning, whether restricted strictly or not, includes both syntagmatic and paradigmatic relations. Some theoretical endeavors, which emerged in the 1980s and 1990s, for example, situation semantics, cognitive semantics, conceptual semantics and frame semantics, also take the same perspective, that is, to study the lexical meanings by referring to the grammatical distribution.

The recent language engineering also demands the combination of the two strands, both semantical and syntactical. On the one hand, with the lexicographical methods modernized, the analysis of the lexical meanings will become more and more rigorous. In this way, the accuracy and the appropriateness required in the process of lexical segmentation and interpretation should not rely on the intuition of the compliers, instead, but on the collocational characteristics as displayed by the specific behavior of the lexical items.

On the other hand, with the development of the natural language processing (NLP), the analysis of the lexical meaning is becoming more and more important and urgent. The dilemma involved in this process is also a serious question posed to many NLP systems. In order to retrieve enough lexical information, many countries in the world have invested enormously to develop machine semantic dictionaries and many scholars have also tried to retrieve lexical knowledge from machine dictionaries and large corpora（Chodorow, 1985; Ide, 1993; Huang Chu-Ren, 1998）. So far, however, the most serious problem with these lexical semantic systems is the categorization simply based on the static analysis in the absence of the specific theoretical framework, so the utility of these systems are weakened a great deal.

To sum up, as influenced by the theoretical progress and the practical considerations, the analysis of the lexical meanings is required to take a more rigorous and comprehensive approach by combining the lexical components and the syntactical distributions. XHK is just a product of this theoretical assumption. Designed by National University of Singapore and Xiamen University, this knowledge base is aimed to conduct a comprehensive and accurate description of the lexical components of modern Chinese, to improve the mechanisms involved in the dynamic discovery of the semantic rules and the retrieval of the lexical components, and to establish a new framework, both advanced and feasible, within which to describe the lexical knowledge.

2. How to describe the lexical information

Since the lexical meaning is composed mainly of three parts, lexical meaning, associative meaning and syntactical meaning, the information categories in XHK should be designed so as to contain the above three types of meanings so that, on one hand, the grammatical distribution and the nuance in the associative meaning can be displayed, and on the other hand, the computer can dissolve the lexical and structural ambiguities by identifying the lexical meanings in the texts.

Currently, there are eight categories in XHK:

- Lexical orthography: word, homonymy, heterography;
- lexical pronunciation: *pinyin*, homograph, phonological variants;
- basic meaning: entry, interpretation, semantic category;
- grammar: part of speech (POS), overlapping POSs, syntactical function, lexical collocation;
- associative meaning: style, register, emotion;
- origin: new word, foreign word, dialectal word, archaic word;
- instances: specialist gloss, example;
- statistics: character frequency, word frequency and meaning frequency.

In the above categories, (1) and (2) are about the basic lexical meaning. (3) is about the conceptual meaning of the word and (4) about the grammatical meaning. (5) and (6) describe the associated meaning of the word, and (7) and (8) provide the actual examples and the

statistical information about the word.

2.1 Lexical entries in XHK

In XHK which contains 81,899 words, each lexical meaning is listed as a separate entry labeled with a number, with all the related meanings of particular word clustered together.

The following is an example of coffee (咖啡) to illustrate how the items are arranged in XHK.

【咖啡】①常绿小乔木或灌木，叶子长卵形，先端尖，花白色，有香味，结浆果，深红色，内有两颗种子。种子炒熟制成粉，可以作饮料。

A kind of plant called café by westerners

②咖啡种子制成的粉末。

The powder made from the seeds of the cafe

③用咖啡种子的粉末制成的饮料。

The drink made from cafe

Table 1 The lexical entry in XHK

ID	词语	义项	释义
33032	咖啡	1	常绿小乔木或灌木，叶子长卵形，先端尖，花白色，有香味，结浆果，深红色，内两颗种子。种子炒熟制成粉，可以做饮料。产在热带和亚热带地区。
33033	咖啡	2	咖啡种子制成的粉末。
33034	咖啡	3	用咖啡种子的粉末制成的饮料。

This example can illustrate the different way the entries are arranged in *Modern Chinese Dictionary* and XHK. The former is based on the linear sequence of the words, with all its lexical meanings listed under it, so ca1fe1i (咖啡) is considered as a lexical entry which has three lexical interpretations in *Modern Chinese Dictionary*. In XHK, each lexical meaning of a particular lexical item is considered as an independent one, with all the interpretations of the same lexical item clustered together but labeled separately. The compilation of XHK is based on the 61,261 words in *Modern Chinese Dictionary* (1996), together with the 1,205 new words updated in the 2002 edition of *Modern Chinese Dictionary*. Among them, 75.58%

are the words which can be used independently and 15.51% is the linguistic units smaller than word, for example, prefixes, suffixes, non-lexical morphemes, non-morphological characters and so on. The following table indicates the percentages of each kind of linguistic unit in XHK.

Table 2 The Linguistic Units incorporated in XHK

Linguistic Unit	**Number**	**Percentage**	**Examples**
Words	63216	77.19%	来 去 人 白 整洁 老师 金灿灿
Non-lexical morphemes	11678	14.25%	佳 宏 旅 畔 夏 遐 骤 机
Prefixes	18	0.02%	阿 超 非 过 老 小 有 准
Suffixes	48	0.05%	子 儿 头 度 界 论 然 型
Non-morphological characters	981	1.19%	鹌 鹑 龌 龊 垃 圾 蜈 蚣
Abbreviated form	267	0.32%	安检 打假 奥运会 东三省
Idioms	3637	4.44%	安居乐业 百发百中 牵一发而动全身
Frequent collocations	3404	4.15%	爆冷门 车轱辘话 八九不离十 不…不… 一…就… 前…后…

There is another serious problem to be solved. The heterogeneity of the lexical system in Chinese, which may be attributed to the historical evolution, the borrowing from various sources and the stylistic difference, will cause much variation among the same morpheme or word, in terms of orthography, or the linear order within the word.

For example,

【定婚】：同“订婚”。(Betrothal)

【人材】： 同“人才”。 (Talent)

【源源本本】同“原原本本”。(To speak out all the truth)

【成竹在胸】：见 1414 页〖胸有成竹〗 (To be very confident)

To solve this problem, the designers set up a separate section to store the orthographical variations. In this way, the original format of *Modern Chinese Dictionary* is reserved since the variations in the lexical items incorporated in it are the characteristic phenomena in modern Chinese, and the occurrences of these variations are characteristic of the usage of modern Chinese (Su, 2001).

The compiling methodology underlying XHK, i.e. by incorporating both lexical and grammatical analysis, has several advantages. While the lexical system of modern Chinese can be analyzed at a general level, a specific item in the sub-system can also be analyzed in

great detail. And on the other hand, the efficiency can also be improved a lot, because the linguistic units incorporated in XHK have very frequent occurrences. By sorting all the orthographical variations, for example, 人材 and 人才, all the relevant information concerning the target lexical item can be exhausted.

2.2 Pronunciation

The number of the syllables in the words listed in XHK varies a great deal, as is indicated in the following table. The majority of the words have two syllables 61.38%, 23.57% of them have only one syllable.

Table 3 The number of the syllables of the words in XHK

Number of the syllables	Occurrences	Percentage	Examples
One	19310	23.57%	机 去 球 人 啊 子 阿 蜈
Two	50277	61.38%	阿姨 安稳 新春 吧台 把持 盖子 电大 录像 翘望 誓约
Three	6181	7.54%	安全门 奥运会 绿油油 鼠标器 出风头 打先锋 发烧友 犯不着
Four	5357	6.54%	阿猫阿狗 基础教育 新闻公报 礼尚往来 新陈代谢 爱不释手
Five	398	0.48%	国际儿童节 电子计算机 三下五除二 坐山观虎斗
Six	218	0.26%	汉语拼音方案 个人数字助理 有一搭没一搭 五十步笑百步
Seven or more	158	0.19%	计算机断层扫描 一个萝卜一个坑儿 只要功夫深，铁杵磨成针 只许州官放火，不许百姓点灯

2.3 Lexical meanings

When analyzing the conceptual meaning, XHK adopts the same analytical system as that used in *Modern Chinese Dictionary*, but at the same time, the analytic format in XHK is more formal and rigorous in order to facilitate computer-aided processing.

Among the 62,466 lexical items in *Modern Chinese Dictionary*, there are 46,797 homonymical words, amounting to 74.92% and there are only 15,669 polysemous

words(including homographical words), amounting to 25.08%.

This group of figures seems to indicate that the cases of the polysemous words in Chinese are only sporadic, but the reality is just the opposite. A random investigation of a 46,760-word corpus in *People's Daily* by Li (1999) indicates that words with a specific clear-cut meaning occupy only 58%, while words with ambiguous meanings occupy as much as 42%. This finding proves that polysemy is a frequent phenomenon in modern Chinese. Our investigation of the XHK data base also come to the similar result: 15,669 words in XHK have as many as 35,100 lexical meanings, amounting to 42.86% of all the lexical meanings. And on the other hand, this fact poses another tricky issue, that is, with every word carrying 2.24 lexical meanings on average, how can we distinguish these lexical meanings?

2.4 Part of speech (POS)

Although theoretically, the lexical semantic analysis and grammatical analysis are separate issues in linguistics, there exists a close interconnection between the lexical meaning and the part of speech. The change in either one will lead to the change in the other one. There are many cases where the shift in the part of speech leads to the shift in the lexical meaning, for example, in nouns, verbs, adjectives and functional words.

Based on this fact, XHK gets every one of the 82 thousand words tagged with the appropriate part of speech by employing the encoding systems provided by Beijing University (Yu, 1998; Guo, 2003). And there are 18 basic parts of speech and 7 attached ones (Appendix 1).

In all the parts of speech in XHK, there are 24.42%（§2.1）attached parts of speech and 75.58% basic parts of speech. Table 4 shows that nouns, verbs and adjectives occupy 68.38% of the total words in Chinese. And Table 5 indicates furthermore that these three categories occupy 82.79% among all the 35,100 lexical meanings.

Table 4 Part of speech and lexical meaning in XHK

	Pinyin	Sense No.	Lexical Meaning	Example
新	xin1	1	以刚出现的或刚经验到的（跟‘旧’或‘老’相对）。	～风气丨～品种丨～的工作岗位。
新	xin1	2	性质上改变得更好的；使变成新的（跟‘旧’相对）。	～社会丨～文艺丨改过自～丨一～耳目丨粉刷一～。
新	xin1	3	没有用过的（跟‘旧’相对）	～笔丨～锄头丨这衣服是全～的
新	xin1	4	指新的人或事物。	尝～丨花样翻～丨推陈出～
新	xin1	5	结婚的或结婚不久的。	～女婿丨～媳妇。
新	xin1	6	新近；刚。	我是～来的丨这本书是～买的
新	xin1	7	(Xin)姓。	
新春	xin1chun1		指春节以后的一二十天。	欢度～|～佳节|辞旧岁，迎～

Table 5 The raw number and frequency of the basic parts of speech

Part of speech	Number	Percentage
Noun	30112	36.76%
Verb	21021	25.66%
Adjective	4886	5.96%
Adverb	1352	1.65%
Word of state	976	1.19%
Word of distinction	700	0.85%
Time noun	652	0.79%
Classifier	449	0.54%
onomatopoeia	289	0.35%
Pronoun	282	0.34%
Space noun	243	0.29%
conj	214	0.26%
Numerical	193	0.23%
Location noun	172	0.21%
Preposition	120	0.14%
Particle	62	0.07%
Interjection	57	0.06%

Table 6 The percentages of nouns, verbs and adjectives among the polysemous words

Part of speech	**Number**	**centage**
Noun	15413	91%
Verb	11314	23%
Adjective	3912	15%
Total	30639	29%

Since nouns, verbs and adjectives are the most frequently used words in Chinese, we will focus on these three kinds of words, particularly how to describe and retrieve their lexical components an polysemous words.

2.5 Words with overlapping POSs

Closely associated with the assignment of POS is the phenomenon that some words can function as different POSs. More specifically, this refers to those polysemous words the lexical meanings of which belong to different parts of speech.

The following is a case in point, where 包装 (ba1ozhua1ng) has three lexical meanings, among which, the first and third belong to the verb while the second belongs to the noun. So in this sense, 包装 (ba1ozhua1ng) can function both as the verb and as the noun.

【包装】(1) 在商品外面用纸包裹或把商品装进纸盒、瓶子等：定量～｜～商品要注意质量。

To get things packed

(2) 指包装商品的东西，如纸、盒子、瓶子等：～美观｜运输不慎，～ 破损严重。

Things that are used for packing

(3) 比喻对人或事物从形象上装扮、美化，使更具吸引力或商业价值：～歌星｜～体育比赛。

Metaphorically, to make people or things more attractive

XHK will treat all those words which bears the same orthography but whose lexical meanings fall into different parts of speech as words with overlapping parts of speech since both homographs and polysemous words are considered as the same when processed in the computer and have no clear boundary (Su, 2001). Table 7 below shows that bēi (a classifier), bèi (a verb) and bèi (an adjective) under the same orthographical item 背 (bei) have different lexical meanings and parts of speech.

Table 7 How words with overlapping parts of speech are displayed in XHK

Word	Pinyin	Orthography	Sense No.	POS	Other POS	Lexical meaning	Examples
背	bei1	A	3	q	vna	The amount that one can carry once	一～麦子丨一～柴火
背	bei4	B1	1	n	qva	One's back	后～丨～影。
背	bei4	B2	4	v	qna	Rehearse	～台词丨书～熟了。
背	bei4	B2	6	v	qna	Facing the opposite direction	他把脸～过去，装着没看见。
背	bei4	B2	7	a	qnv	Very remote	～静|～街小巷|深山小路很～。

A preliminary statistical result shows that there are 14,246 words with overlapping parts of speech, amounting to 39.5% among all the polysemous words in XHK. And at the same time, this statistics also proves that to assign part of speech is a very important and effective means to identify the lexical meanings and to distinguish the polysemous words.

2.6 Syntactical Functions

In many cases, the lexical meanings of a particular word belong to the same part of speech (60.5%), and so the lexical meanings of this type must be distinguished with reference to the syntactical collocations in which the word occurs. The following is a case in point.

【请】(1) 邀请：～客｜～老李做报告。

To invite

(2) 敬辞，用于希望对方做某事：您～坐丨～准时出席。

An honorific term

Although both of the above meanings belong to the verb, there is still some minor difference between these two meanings. While the first one always takes the noun or the noun phrase as its object, the second takes only the verb as its object. According to this criterion, the meanings of the following tokens of "qing" can be identified very easily.

(1) 于是 c ，知心 a 姐姐 n **请 v** 她 r 来 vd 参加 v 今天 nt 这个 r 活动 n 。

(2) 财务科 n 同志 n **请 v** 厂 n 领导 n 出面 v 做工作 v ，副 h 厂长 n 、财务 n 科长 n 亲自 d 把 p 钱 n 送到 v 老 h 刘 nhf 的 u 家里 nl ，说服 v 他 r 收下 v

(3) 宜兴 ns 市委 j 、 市政府 ni **请 v** 市 n 保险 n 部门 n 抓紧 v 调查 v 核实 v ，迅速 a 理赔 v 兑现 v 补偿 v 一些 m。

(4) **请 v** 收回 v 你 r 的 u 一半 m 吧 u 、 我 r 亦 d 收回 v 我 r 的 u 一半 m .

(5) **请 v** 记住 v ， 你 r 本人 n 的 u 到场 v 本身 n 就是 d 一 m 件 q 十分 d 宝贵 a 的 u 礼物 n ， 这 r 是 vl 其它 r 任何 r 礼物 n 所 u 不能 vu 比拟 v 的 u

(6) 我方 n 坚守 v 信用 n ， 绝对 a 保密 v ， **请 v** 放心 v

(7) “咳 e ， **请 v** 喝 v 大碗茶 n ， 二 m 分 q 钱 n 一 m 碗 n … ”一阵 mq 清亮 a 的 u 吆喝声 n ， 从 p 前门 ns 附近 n 的 u 打磨 v 厂 n 胡同口 n 传来 v 。

All the above examples are taken from the Central Corpus of Modern Chinese of Ministry of Education. The tokens of “qing” in (1), (2) and (3) are followed by the noun or the noun phrase, so these tokens of “qing” mean “to invite”. The tokens of “qing” in (4),(5),(6) and (7), which are followed by the verb or the verb phrase, are used as the honorific term.

XHK will provide a detailed description of the characteristics of all the lexical items by setting up separate data bases for the lexicon.

2.7 Collocational Pattern

There are some cases where some lexical meanings of a polysemous word bear much similarity in terms both of the part of speech and of the syntactical function. In this case, the difference in the collocational patterns displayed by the word will be considered.

The following is an example, in which 架子 (jia4zi) has two lexical meanings, which can be decided only with reference to the collocational environment in which it appears.

【架子】 (1) 由若干材料纵横交叉地构成的东西，用来放置器物、支撑物体或安装工具等：花瓶～｜骨头～｜保险刀的～

A frame that supports something

(2)自高自大、装腔作势的作风：官～｜拿～｜那位局长一点～都没有。

To be arrogant

At first sight, both the above meanings can be used as nouns, which can function as a subject, an object or a sentence component to be modified, except as a modifier. The syntactical environment in which they appear, however, are different.

Table 8 The comparison between the collocational patterns of the two meanings of 架子

Grammatical function	A frame that supports something	To be arrogant
As the subject	～搭好了/～倒了/～摇晃了 ～不结实/～松了/～很精致	～放下来了 / ～摆得很足～太大 / ～不小
As the object	做好～/搭好～ / 爬上～	摆～/放下～/拿～/端～
As a modifier	/	/
To be modified by the noun	材料＋～（表示材料） 木头～/紫檀～/铁～/塑料～ 物品＋～（表示用途） 货物～/行李～/葡萄～/烟～ 物品＋～（整体～部分关系） 伞～/床～/售货亭～/鸡～	身份＋～： 官僚～/官～/明星～
To be modified by the classifier	一个～ / 一种～ / 一排～	一点～

The above table shows that it is the different lexical meaning that determines the different collocational behaviors displayed by the word 架子.

For example, the first meaning of 架子 indicates that this word can be modified by the nouns which indicate materials, for example, wood, steel, iron and so on in order to indicate what the frame is made of. This can be proved by the facts that this word can be modified by such words as goods, luggage book and so on and that this word can modify such words as car, bed and so on as indicating its supporting function.

In contrast, the second meaning (the arrogant attitude) can be modified only by the nouns indicating the people of rather high social position.

Based on the semantic classification, XHK describes the collocational possibilities that the lexical item may have, in order to combine the lexical collocation and the systematical grammatical description.

2.8 Semantic Category

It is assumed that the semantic categorization should be combined with the syntactical knowledge so that some tricky problems which can not be solved within the syntactical framework can be dealt with. So far, only nouns (words indicating time, location and space), verbs and adjectives are assigned the appropriate semantic category, while those function

words have not been tackled in this way.

Compared with several current semantic categorization systems, for example, Mei (1983), Lin (1987, 1998), Chen (1996), Chen (1998), Dong (1998) and Dong (1998, 1999), the most obvious advantage of XHK is that closely correlated to the syntactical analysis, the depth and width of the semantic categorization is subject to the lexical semantic analysis. The nouns are given the most detailed semantic categorization with an ordered hierarchy. The top five broad categories are concrete things, abstract things, process, time and space, and furthermore, concrete things are divided into living things and non-living things, with the former divided into human beings, animals, plants and bacteria and so on. And finally, the non-living things are divided into man-made things and natural things and so on. Verbs and adjectives, however, are given a more general categorization, but their collocational relationships with nouns are the major analytical focus.

Table 9 Semantic categories in XK

Word	Part of speech	Semantic category
Plain poem	n	Created work
Green vegetable	n	Plant/vegetable
Like	v	Psychological movements
Thunder	v	Cosmological phenomena
Red	a	Color
Honesty	a	Chracter

2.9 Associative meaning

Although XHK focuses on the conceptual meaning and the grammatical functions of the lexical items, it does not ignore the associative meanings of the lexical items, which are displayed in such aspects as register, emotion, context, origin and so on. For example, the words “material” and “software” have different associative meanings when used in different registers. There will be another paper which introduces how XHK describes the associative meaning.

Word	Register	Sense No.	Lexical Interpretation
质量	物理	1	量度物体惯性大小的物理量。数值上等于物体所受外力和它获得的加速度的比值。有时也指物体中所含物质的量。质量是常量，不因高度或纬度变化而改变。
质量		2	产品或工作的优劣程度。
软件	计算机	1	计算机系统的组成部分，是指计算机进行计算、判断、处理信息的程序系统或设备。包括汇编程序、操作系统、编译程序、诊断程序、控制程序、数据管理系统等。
软件		2	借指生产、科研、经营等过程中的人员素质、管理水平、服务质量等。

3.Automatic retrieval and frequency calculation

3.1 Corpus reviewing

In order to display the concrete usage of the every word, XHK incorporates the illustrating examples in *Modern Chinese Dictionary* and a large number of natural examples in which the target word occurs. And the compilers also design a corpus retrieving system so that the examples and their frequencies can be obtained automatically given any command at any domain. The following is an example with the word 科学（ke1xu2e）.

【科学】(1) 反映自然、社会、思维等的客观规律的分科的知识体系。

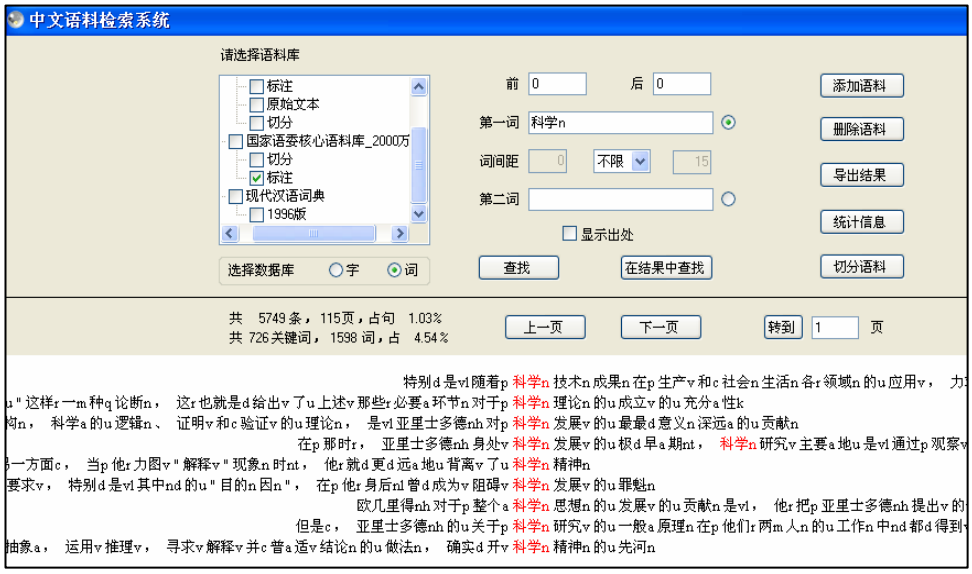
The knowledge system which reflects the truth

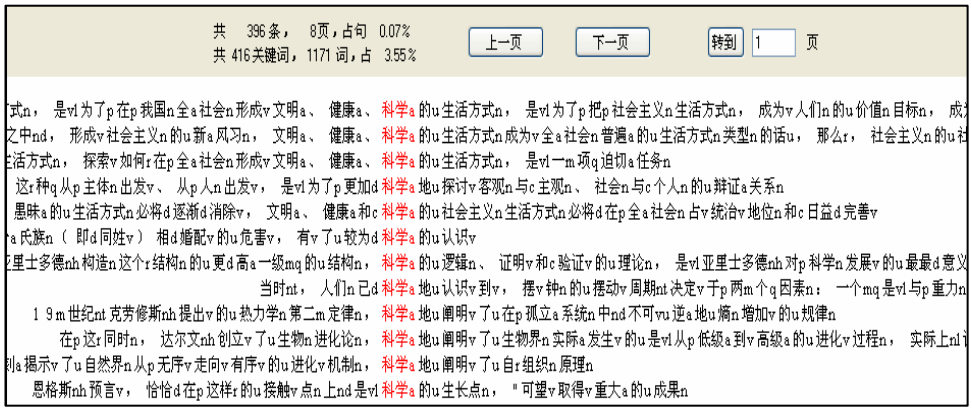
(2)合乎科学的：～种田|这种说法不～丨革命精神和～态度相结合

To be scientific

While (1) is used as a noun，(2) is used as an adjective. By searching all the sections related to the meanings of 科学, the corpus reviewing system in XHK will produce all the examples in which the two meanings of 科学 are used.

Diagram 1 The corpus reviewing system in XHK





By pressing the Output key in the first diagram, you can input the relevant examples into the section of corpus examples in XHK.

3.2 Frequency calculation

In order to provide more detailed information about the lexical meanings and to increase the feasibility of the knowledge base, the compilers have also designed XHK to make it calculate the character frequency and word frequency in the Central Corpus of the National Chinese Committee.

Table 11 The 20 characters with the highest frequency

No.	Character	Frequency
1	的	805583
2	一	257929
3	是	244199
4	了	191005
5	不	188311
6	在	176159
7	有	165198
8	人	143537
9	这	130287
10	和	117509
11	我	115433
12	个	104942
13	上	103917
14	大	100832
15	中	98966
16	为	97309
17	他	96943
18	来	96613
19	地	96128
20	生	94517

Table 12 The top 20 words with the highest frequency

No.	Word	Frequency
1	的 u	694848
2	了 u	143331
3	是 v	129042
4	在 p	122993
5	和 c	95259
6	一 m	83762
7	这 r	59957
8	他 r	59058
9	有 v	56220
10	我 r	51671
11	不 d	47257
12	也 d	44665
13	中 f	43232
14	着 u	42156
15	就 d	41315
16	地 u	37787
17	人 n	36444
18	上 f	34691
19	个 q	33367
20	都 d	32621

One important XHK research project focuses on the frequency with which the lexical meanings of a particular polysemous word is used. Table 1 shows how to use the POS assignment method to calculate the frequency with which the meanings of the polysemous word function as different POSs . Take the case of 科学 as an example. In the entire data base there are altogether 5749 sentences where its first meaning appears, while there are only 396 ones in which its second meaning appears. The conclusion is that the frequency of its first meaning is much higher than that of its second meaning.

Table 13 shows more examples of this kind.

Table 13 The frequency of those words with POS labels

No.	Word	Frequency
21500	补贴 v	31
8801	补贴 n	107
2446	把握 v	498
21757	把握 n	31
8208	开阔 a	117
34710	开阔 v	15
34095	内行 n	16
20131	内行 a	34
1367	把 q	920
27	把 p	27075
1049	白 a	1217
9899	白 d	91
22776	白 v	29
16946	左右 v	44
98389	左右 m	2
912	左右 f	1387

From the above table we can come to the conclusion that the frequencies of the meanings of the polysemous words vary a lot according to the lexical meanings which are used in the specific context.

There are two steps involved in order to achieve an exact statistical result of the frequencies of the cases where the lexical meanings of the polysemous word belong to the same POS. The first is to identify the specific lexical meaning displayed in each specific context by the word in the 20-million-word corpus by using the semantic and grammatical knowledge, which involves Herculean labor. The second step is to obtain the frequencies of the lexical meanings. This project will pave the way for the Chinese language processing in many aspects. Both the identification of the lexical meaning and the calculation of the meaning frequency are conducive to the quantitative process of the Chinese lexical system, and furthermore, to the researches in such areas as the core lexicon, frequently used words, derived meanings and metaphorical meanings. And the product of this project will shed new light on the Chinese lexicography and Chinese information processing.

4. Conclusion

To sum up, convenient for both human and computer-aided processing, XHK is a knowledge data base, which is aimed at an advanced and feasible analysis, both semantic and syntactical, of the modern Chinese lexical semantics. By incorporating the 61,000 lexical items, with 82,000 lexical meanings from *Modern Chinese Dictionary*, and a large quantity of natural examples from the Central Corpus of the National Chinese Committee which contains 20 million words, XHK provides a comprehensive profile of the information about each lexical meaning of all the words, including *pinyin*, lexical meaning, part of speech, semantic category, syntactical function, collocational pattern, associative meaning, frequency and so on. Besides, XHK contains a feasible search engine which is of great help to the various stages in the Chinese processing, for example, in the analysis of the Chinese lexicology, statistical calculation, Chinese lexicography and so on.

XHK will be useful in many other areas, for example, in machine-translation, document search, information retrieval, corpus processing, the application of semantic information and collocational patterns to Chinese analysis, WSD, and the identification of the syntactical relationships and of the semantic relationships in the phrase, and so on.

Subsequently, some research projects still need to be conducted. The first one is to describe the syntactical and collacational patterns of the polysemous words in order to establish a syntax-based theoretical framework which will facilitate the function of the semantic component retrieval. And at the same time, in order to modify the function of describing the semantic components and to retrieve of large quantity of information from annotated texts, we will try to improve some functions of XHK, for example, the dynamic connection between XHK and the current electronic dictionaries and large language corpora, the automatic extraction of the lexical collocations and natural examples.

References

Bake, C.F, Fillmore, C. J. and Lowe, J. B. 1998. The Berkeley Frame Net Project. In *Proceedings of COLING'98* . 86-90.

Chodorow, M., Byrd, R. and Heidorn, G. 1985.Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the ACL*. 299-304.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Mass: MIT Press.

Huang Chu-Ren, Chen Keh-Jiann, and Gao Zhao-Ming.1998. Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. In Benjamin K. T'sou et al.(eds.). *Quantitative and Computational Studies on the Chinese Language*. 339-352.

Ide, N. and Veronis, J. 1993. Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? In *Proceedings of KB & KS'93* (Tokyo). 257-266.

Kazt, J. 1972. *Semantic Theory*. New York: Harper & Row.

Kazt, J and Fodor, J. A. 1963. Structure of a Semantic Theory. *Language* (39): 170-210.

Lyons, J. 1995. *Linguistic Semantics: An Introduction*. Oxford: Cambridge University Press.

Richardson, S. D. 1998. MindNet: acquiring and structuring semantic information from text. In *Coling'98*. 1098-1102.

陈力为，袁琦主编.1995.《中文信息处理应用平台工程》. 北京:电子工业出版社

陈群秀. 1996. 信息处理用现代汉语语义分类体系的设计思想. 见：罗振声、袁毓林主编《计算机时代的汉语和汉字研究》，北京: 清华大学出版社

陈晓荷，1998，一个面向工程的语义分类体系. *语言文字应用*，第2期

董大年. 1998. 《现代汉语分类词典》. 北京: 汉语大词典出版社

董振东. 1998. 语义关系的表达和知识系统的建造. *语言文字应用*，第3期

董振东. 1999.“知网”（Hownet）介绍. http:// www.keenage.com

郭 锐. 2002. 《现代汉语词类研究》. 北京: 商务印书馆

黄居仁主编.2004.《中文的意义与词义》，CKIP 中文词知识库小组技术报告《意义与词义》系列.台湾中央研究院资讯科学研究所&语言学研究所筹备处.

李涓子. 1999. 汉语词义排歧方法研究. 清华大学计算机科学系[博士学位论文]

林杏光. 1987.《简明汉语义类词典》. 北京: 商务印书馆

林杏光. 1998. 中文信息界的语义研究谭要. *语言文字应用*，第3期

梅家驹. 1983.《同义词词林》. 上海: 上海辞书出版社

苏新春. 2001.《汉语词汇计量研究》. 厦门: 厦门大学出版社

苏新春. 2001. 关于《现代汉语词典》词汇计量研究的思考. 《世界汉语教学》. No.4

王 惠.2004.《现代汉语名词词义组合分析》北京: 北京大学出版社.

于江生，俞士汶. 2002. “CCD的结构与设计思想”. 《中文信息学报》. No.4. pp 12-20

俞士汶，朱学锋，王惠，张芸芸. 1998.《现代汉语语法信息词典详解》. 北京：清华大学出版社. 第 2 版, 2003.