

# Corpus-driven Linguistic Approaches to Sense Prediction

**Jia-Fei Hong**

Graduate Institute of  
Linguistics, National  
Taiwan University, Taiwan  
jiafei@gate.sinica.edu.tw

**Sue-Jin Ker**

Department of  
Computer Science  
and Information  
Management,  
Soochow University,  
Taiwan  
ksj@cis.scu.edu.tw

**Kathleen Ahrens**

Graduate Institute of  
Linguistics, National  
Taiwan University, Taiwan  
Language Centre, Hong  
Kong Baptist University,  
Hong Kong  
kathleenahrens@yahoo.com

**Chu-Ren Huang**

Institute of Linguistics,  
Academia Sinica, Taiwan  
Faculty of Humanities,  
The Hong Kong  
Polytechnic University,  
Hong Kong  
churenhuang@gmail.com

## Abstract

In this study, we propose to use corpus-driven linguistic approaches for a sense prediction study. We will concentrate on individual semantic features to predict the senses of non-assigned words by using corpora and tools, such as Chinese Gigaword Corpus. In this study, we would like to explore some related issues of the sense prediction study by using the corpus-based analysis and the experimental evaluation in order to achieve automatic prediction in machine programming. Using these corpora, we will determine the collocation clusters of our target word *chi1* “eat” through semantic features and concepts. This requirement will demonstrate the visibility of the corpus-based approaches.

**Keywords:** Lexical ambiguity, sense prediction, corpus-based approach, experimental evaluation

## 1 Introduction

In this study we run an undefined sense prediction study to generate solutions for lexical ambiguity resolution. In particular, we will be looking at words without pre-assigned lexical meanings and try to predict the range of senses a word form may have. Sense predicting should be even more challenging and shed more light on how human process meaning. In this study, we would like to

explore some related issues of the sense prediction study by using the corpus-based analysis and the experimental evaluation in order to achieve automatic prediction in machine programming.

Although a suite of heuristical methods are presented for word sense disambiguation of Chinese Wordnet glosses, unfortunately we know of several researchers who use only manual analysis to find out the argumentative roles and predict their semantic features to determine their senses. Therefore, they can't deal with more quantities of lexically ambiguous words at the same time. We consulted Fujii and Croft's study (1993) to collect relevant collocations to categorize different clusters for achieving automatic sense prediction.

In this study, we first review some related previous studies of sense prediction and lexical ambiguity resolution, especially in corpus and computational and psycholinguistic perspectives. Second, we point out lexical ambiguity determination, hypotheses, research questions, and the goal of this study. Further, data collection and two main research methods, corpus-based analysis and experimental evaluation, will be shown and explained in the cause of analyzing them. Finally, we will show some results of combining, comparing, and discussing these two methods in this study.

## 2 Previous studies

What is “lexical ambiguity”? Lexical ambiguity indicates vague, unclear, and indefinite senses; that is to say, lexically ambiguous words can refer to

more than two senses at the same time. Lexical ambiguity is a linguistic term for a word's capacity to carry two or more distinct meanings, for example, *bank*. In some modern linguistic and literary theories, it is argued that all signs are polysemous, and the term has been extended to larger units, including entire literary works. In WordNet, the definition of a lexically ambiguous word is the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings. WordNet researchers also regard polysemy and lexical ambiguity as synonym.

In the case of the previously mentioned related lexical ambiguity or polysemy studies, I categorize them into three different sections. Several studies may concentrate on corpus and computational perspective (Ker and Chen, 2004; Moldovan and Novischi, 2004; Xue et al., 2006, Peng et al., 2007), some studies focus on psycholinguistic studies (Ahrens, 1998, 2001, 2006; Lin and Ahrens, 2000), while others emphasize neurolinguistic studies (Mason and Just, 2007; Zemleni et al. 2007).

Overall, regarding these previous corpus and computational studies, these scholars proposed corpus-based, algorithm, automatic programming system, and collocation approaches to analyze sense prediction studies or WSD studies. Unfortunately, they only employed one corpus in their study and got less information of lexical ambiguity for their sense prediction study; they also did not combine these approaches. With regard to previous psycholinguistic and neurolinguistic studies, there is no direct influence or hint for our study; however, the inspiration is that we can do experimental evaluation to verify some results from corpus-based analyses and demonstrate the visibility of the corpus-based approaches.

### 3 Research Question

The focus of this study aims to look for a unified analysis of lexical ambiguity, as the problem of lexical ambiguity often poses theoretical and computational problems in lexical semantic studies (cf. Ravin and Leacock, 2000). When we seek to define lexically ambiguous senses, we need to notice that (1) senses are represented as sets of necessary and sufficient conditions that fully capture the conceptual content conveyed by words; (2) there are as many particular senses for a word as there are differences in these conditions; and (3)

senses can be represented independently of the context in which they occur.

Regarding lexical ambiguity, there are three hypotheses in this study. First, lexical ambiguity is the property of some words to have multiple meanings or senses (Moldovan and Novischi, 2004). We can use the senses of the words to determine their semantic relations in Chinese Wordnet (Huang et al., 2003). Second, collocation is a combination of words that has a certain tendency to be used together, and it is used widely to deal with the WSD task. We take a collocation-based approach (Peng et al., 2007) Therefore, using the argument role, labeling information can help us to extract some types of semantic features. Third, if we regard different word senses as occurring in different domains, in an experimental task, we hypothesize that these different word senses are presented with different expressions or effects. We may also obtain some useful evidence in psycholinguistic experimental studies. For this reason, there are three research questions in this study: (1) How do we predict the word senses of a lexically ambiguous word to present different interpretations in different contexts or domains? (2) How do we use a corpus as the database to support a word sense prediction study? And (3) Can we use other approaches to certify the analysis of the written corpus approach for the word sense prediction study?

### 4 Methodology

In order to collect large data to explore our sense prediction study, we focus on Taiwan's Central News Agency Gigaword Corpus (Traditional Gigaword Corpus). The Chinese Gigaword Corpus contains about 1.4 billion Chinese characters, including about 800 million characters from Taiwan's Central News Agency (from 1991 to 2004), nearly 500 million characters from China's Xinhua News Agency, and approximately 30 million characters from Singapore Zaobao. In this study, we use the corpus-based analysis and the experimental evaluation to deal with the sense prediction and determine which target word (*chil* "eat") in the different sentences belong to the same sense. We use the corpus-based analysis and the experimental evaluation to deal with the sense prediction and determine which target word (*chil*

“eat”) in the different sentences belong to the same sense.

In order to gain these relative collocations, from Taiwan’s Central News Agency Gigaword Corpus, we use three different ways to collect them: (1) the noun after the target verb; (2) the head noun of the first noun phrase after the target verb; and (3) the head noun after the first punctuation mark of the target verb. They are shown following, in Table 1 through Table 3.

Table 1: The noun after the target verb

Connected Sentence	Relative Collocation
民众除了多食用蔬菜, 多吃{VC}鱼{Na}也有益健康。	鱼{Na}

Table 2: The head noun of the first noun phrase after the target verb

Connected Sentence	Relative Collocation
每天特别搜购温体猪肉, 也就是吃{VC}残羹{Na}剩菜{Na}长大的黑毛猪。	剩菜{Na}

Table 3: The head noun after the first punctuation mark of the target verb

Connected Sentence	Relative Collocation
希望民众能多吃{VC}对健康有益的芒果{Na}。	芒果{Na}

Among all large databases of Taiwan’s Central News Agency, we collected 22,906 sentences for our target verb *chil* “eat” and examined these collocations. It’s very important that we obtain 4032 types from these 22,906 sentences.

For experimental evaluation, different collocations will affect the interpretations of the target verb, such as *chil* “eat”. If we can demonstrate that there are several clusters of the related collocations for *chil* “eat” by the experimental evaluation, we can certainly predict several different senses for the target verb.

In order to test the related collocations for the target word of lexical ambiguity or polysemy, we will run a multiple-choice task experiment to demonstrate which target word (*chil* “eat”) in the different sentences belongs to the same sense cluster in this study. For that reason, we will utilize

the multiple-choice task and obtain some experimental data from the related collocations for *chil* “eat” of the corpus-based analysis.

In this multiple-choice task, we will let my participants choose one word/one item that is different from three other words/items. This multiple-choice task is a test without context. In other words, in this multiple-choice task, whether some conceptual words belong to the same sense cluster helps me prepare some materials in my sense prediction study.

## 5 Analysis

### 5.1 Corpus-based analysis

There are two important steps for the sense prediction experiment in this study: (1) collect appropriate collocations and rank them by their frequencies; and (2) group some similar collections into the same cluster. Among these 22,906 sentences, we collected their appropriate collocations and ranked them by their frequencies. It’s very important that we obtain 4032 types from these 22,906 sentences. These partial collocation words of higher frequency of these sentences are shown following, in Table 4.

Table 4: Partial collocation words by ranking their frequencies

Feature	Frequency
食物	720
药	683
东西	525
饭	411
人	360
案	243
早餐	238
便当	228
水果	227
槟榔	226

Similar features are often synonymous compounds that share a common morpheme. For instance, [饭 (*fan4* “rice”), 米饭 (*mi3 fan4* “rice”)] and [案 (*an4* “case”), 案件 (*an4 jian4* “case”)],

respectively, share a common morpheme [饭 (*fan4* “rice”)] and [案 (*an4* “case”)]. Fujii and Croft (1993) also pointed out a similar thesaurus effect of Chinese characters in Japanese Information Retrieval. In the cluster step, there are two sub-steps here: (1) character similarity comparison between words; and (2) group similarity comparison between words. Two formulas for these sub-steps are presented as the following:

Formula 1: Character similarity comparison between words

$$dice(x, y) = \frac{2|x \cap y|}{|x| + |y|}$$

By using this formula, we will obtain some collocations and regard 药 (*yao4* “medicine”), 减肥药 (*jian3 fei2 yao4* “reducing weight medicine”), and 中药 (*zhong1 yao4* “traditional Chinese medicine”) as the same cluster.

Formula 2: Group similarity comparison between words

$$sim(x, Y) = \frac{\sum_{y \in Y} dice(x, y)}{|Y|}$$

In Formula 2, we take one undefined word (*x*) to compare with each word (*y*) of one certain cluster (*Y*), calculate their average similarities, gain the maximum, and then this undefined word (*x*) will belong to this certain cluster (*Y*). After comparing some cluster similarities’ comparisons between them, we can put 败绩 (*bai4 ji1* “defeat”) and 败仗 (*bai4 zhang4* “defeat”) in the same cluster.

From these two different experimental steps, we reduce 4032 types from these 22,906 sentences to 348 clusters. Among 348 clusters, we further categorize the results into two sub-groups, the physical sense group and the metaphorical sense group, such as shown in both Table 5 and Table 6.

Table 5: Physical senses clusters

[ <i>Yao4 Wu4</i> “medicine”]	E.g.1: 一名女子吃 <b>减肥药</b> 吃到被判刑。
[ <i>Fan4</i> “boiled rice”]	E.g.1: 捐十万元者, 可以跟柯林顿吃 <b>两顿饭</b> 。
[ <i>Can1</i> “meal”]	E.g.1: 到了晚间, 总统则与参访团员同样吃 <b>飞机餐</b> 。
[ <i>Rou4</i> “meat”]	E.g.1: 北韩难民为了活命, 吃 <b>鼠肉</b> 、假结婚, 什么手段都使用了。

Table 6: Metaphorical sense clusters

[ <i>An4 jian4</i> “case”]	E.g.1: 警察吃 <b>案</b> 一向最为民众诟病。
[ <i>Bai4 zhang4</i> “defeat”]	E.g.1: 柯林顿说, 巴格达自从一九九一年在波斯湾战争吃 <b>败仗</b> 以来, 这是首次同意开放所有地点。
[ <i>Bi4 men2 geng1</i> “to slam the door in one’s face”]	E.g.1: 乡镇公所承办人或村里干事指称前往办理的残障同胞吃 <b>闭门羹</b> , 他将诉诸法律追究相关人员的责任。

From Table 5 and Table 6, I confidently assert that we can predict some sense clusters for physical senses and metaphorical senses before doing sense division work.

Regarding the accuracy of these clusters by automatic programming selection, we randomly selected four clusters—the *yao4* “medicine” cluster, the *fan4* “rice” cluster, the *can1* “meal” cluster, and the *rou4* “meat” cluster and the accuracy is shown as:

Table 7: The accuracy for four clusters

Cluster	Right	Wrong	Total	Remark
<i>yao4</i> “medicine”	66 (75%)	22 (25%)	88	The accuracy: All are over 75%.
<i>fan4</i> “rice”	59 (83.1%)	12 (16.9%)	71	
<i>can1</i> “meal”	50 (86.2%)	8 (13.8%)	58	
<i>rou4</i> “meat”	93 (90.3%)	10 (9.7%)	103	

Therefore, it is assured that we can obtain high accuracy by automatic programming selection for this sense prediction study.

## 5.2 Experimental evaluation

If we focus only on the morpheme, perhaps many non-related collocations will be assigned to the same cluster, or perhaps many related collocations will be assigned to different clusters. For example, *shan1 yao4* “Chinese yam” and *yao4* “medicine” are in the same cluster. *Han4 bao3 rou4* “hamburger meat” is categorized into *han4 bao3* “hamburger” rather than *rou4* “meat”.

Discussing all materials from the related collocations for *chi1* “eat” of the corpus-based analysis, we will attempt to focus on four different sense clusters: 药 (*yao4* “medicine”), 饭 (*fan4*

“rice”), 餐 (can1 “meal”), and 肉 (rou4 “meat”). There are sixty totally different collocation items for *chi1* “eat” of the corpus-based analysis in Mandarin Chinese, such as 中药 (zhong1 yao4 “traditional Chinese medicine”), 米饭 (mi3 fan4 “rice”), 早餐 (zao3 can1 “breakfast”), 鱼肉 (yu2 rou4 “fish”)... and so on.

In order to test the related collocations for the target word of lexical ambiguity, we will run a multiple-choice task experiment to demonstrate which target word (*chi1* “eat”) in the different sentences belongs to the same sense cluster in this study. We use this task and obtain some experimental data from the related collocations for *chi1* “eat” of the corpus-based analysis. In our question, for example, the candidates are 餐刀 (*can1 dao1* “knife”), 减肥餐 (*jian3 fei2 can1* “reducing weight meal”), 早餐 (*zao3 can1* “breakfast”), and 午餐 (*wu3 can1* “lunch”). Obviously, 餐刀 (*can1 dao1* “knife”) is a different conceptual word. We will let our participants choose one word/one item that is different from three other words/items. In other words, the concept of this selected word/item is obviously different from the concept of the other three words/items. This multiple-choice task is a test without context.

Twenty undergraduate students from National Taiwan University participated in this production test (mean age = 20.9 years; SD = 1.2 years; range = 19 to 23 years). There were ten males and ten females, all native Mandarin Chinese speakers and all right-handed, with no linguistic background knowledge.

All of our participants are asked to choose the most appropriate answer for each question, such as the following instruction 1 and Table 8:

Instruction 1: Instruction in the multiple-choice task

问卷一共有 60 题, 里面都是中文的句子。填写问卷时, 你要做的就是, 先读完每个句子中的四个词组的概念。请你在看过词组之后, 根据你的语感, 决定每题里的概念与概念之间, 哪一个概念与其它三个概念不同, 然后圈选作答。请你一定要在这四个词组之间圈选一个你觉得最适当的答案。举例如下:

1 请问下列哪一个概念与其它三个概念不同?

a) 炒菜锅;	b) 电饭锅;
c) 罗锅;	d) 闷烧锅
2 请问下列哪一个概念与其它三个概念不同?	
a) 洗碗精;	b) 糖精
c) 洗衣精;	d) 洗发精

Table 8: Multiple-choice task for sense prediction study

- |  |
|--|
| (1) 请问下列哪一个概念与其它三个概念不同?<br>a) 喷饭; b) 午饭; c) 团圆饭; d) 中饭 |
| (2) 请问下列哪一个概念与其它三个概念不同?<br>a) 晚餐; b) 佐餐; c) 餐点; d) 自助餐 |
| (3) 请问下列哪一个概念与其它三个概念不同?<br>a) 皮肉; b) 鸡肉; c) 甲鱼肉; d) 鱼肉 |
| (4) 请问下列哪一个概念与其它三个概念不同?<br>a) 药品; b) 农药; c) 乌药; d) 止痛药 |

Following this multiple-choice task, which includes sixty questions for our sense prediction study, we distinguished these sixty questions of the questionnaire into a YES group and a NO group for each participant. Regarding all candidates in the YES group, we selected them from the *yao4* “medicine” cluster, the *fan4* “boiled rice” cluster, the *can1* “meal” cluster, and the *rou4* “meat” cluster; all candidates had the same morpheme in each cluster, for example, *an1 mi2 yao4* “sleeping pill”, *xie4 yao4* “laxative”, and *cheng2 yao4* “patent medicine”. Regarding all candidates in the NO group, we chose for them to have the same morpheme in each cluster based on the new dictionary of Ministry of Education, R.O.C; for example, *ye2 rou4* “coconut (meat)” in the *rou4* “meat” cluster.

While we distinguished and analyzed the YES group and the NO group by each participant, this multiple-choice task also demonstrated other related analysis by each item. They are shown in Table 9 and Table 10:

Table 9: Multiple-choice task by subject for sense prediction study

	YES	NO
Average	15.03 (75.17%)	4.97 (24.83%)
T-test	P = 4.52119E-20 (P < 0.05), Significant	

We used a T-test to compare the YES group with the NO group by subject in this multiple-choice task. We found that the p value is 4.52119E-20 (p < 0.05), and it is significant, which

means we controlled all situations for the YES group and the No group.

Table 10: Multiple-choice task by item for sense prediction study

	YES	NO
Average	45.10 (75.17%)	14.9 (24.83%)
T-test	P = 2.84874E-16 (P < 0.05), Significant	

With the item test, we also used a T-test to compare the YES group with the NO group. We obtained a p value of 2.84874E-16 ( $p < 0.05$ ), and it is also significant. The result shows that we controlled all situations for the YES group and the No group.

It is interesting to note that in corpus-based analysis, the accuracy rates are over 75%. While in experimental evaluation, the accuracy is 75.17%. The automatic prediction method is very successful in approaching the same level of accuracy as human in this difficult task.

## 6 Conclusion

The aim of this sense prediction study is to explore all possible senses of lexical ambiguity in Mandarin Chinese by automatic prediction in machine programming. We use corpus-based analysis and experimental evaluation to examine and determine the collocation clusters of our target word *chi* "eat". Based on corpus-driven linguistic approaches, we find the precision is over 75% through the corpus-based analysis and the experimental evaluation. These results demonstrate the visibility of the corpus-based approaches.

## Acknowledgements

This work is partly supported by National Science Council, the ROC, under the contract number, NSC 97-2221-E-031 -003 and 2004 Academia Sinica Investigator Award. The authors would like to thank the reviewers for their valuable comments.

## References

- Ahrens, Kathleen. 2006. "The Effect of Visual Target Presentation Times on Lexical Ambiguity Resolution." *Language and Linguistics*, 7(3): 677-696.
- Ahrens, Kathleen. 2001. "On-line Sentence Comprehension of Ambiguous Verbs in Mandarin." *Journal of East Asian Linguistics*, 10/4, pp. 337-358.
- Ahrens, Kathleen. 1998. "Lexical Ambiguity Resolution: Languages, Tasks and Timing." In *Sentence Processing: A Cross-linguistic Perspective*. (Ed.) Dieter Hillert. Academic Press, pp.11-31.
- Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Fujii, Hukari and W. Bruce Croft. 1993. A Comparison of Indexing Techniques for Japanese Text Retrieval, In *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 237-246.
- Goddard, Cliff. 2002. "The search for the shared semantic core of all languages". In C. Goddard and A. Wierzbicka (eds). *Meaning and Universal Grammar -Theory and Empirical Findings*. Volume I. Amsterdam: John Benjamins. pp. 5-40.
- Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy. 2003. "Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations." *Language and Linguistics*. 4.3: 509-532.
- Ker, Sue-Jin and Jen-Nan Chen. 2004. "Adaptive Word Sense Tagging on Chinese Corpus." PACLIC 18, Dec. 8-10, 2004, Waseda University, Tokyo, pp. 267-273.
- Lin, Charles and Kathleen Ahrens. 2000. "Calculating the Number of Senses: Implications for Ambiguity Advantage Effect During Lexical Access." In H. Y. Tai and Chang Y. L. (eds.) *Proceedings of the Seventh International Symposium on Chinese Languages and Linguistics*. Chai-yi: National Chung-Cheng University, pp. 141-155.
- Mason, Robert A. and Marcel Adam Just. 2007. "Lexical ambiguity in sentence comprehension." *Brain Research*, 1146: 115-27.
- Moldovan, Dan, Adrian Novischi. 2004. Word sense disambiguation of WordNet glosses. *Computer Speech and language*, 18: 301-317.
- Peng, Jin, Xu Sun, Yunfang Wu, and Shiwen Yu. 2007. "Word Clustering for Collocation-Based Word Sense Disambiguation." *The Eighth International Conference on Intelligent Text Processing & Computational Linguistics (CICLing 2007)*, LNCS 4394, pp. 267-274.
- Ravin, Yael and Claudia Leacock. 2000. Polysemy: Theoretical and Computational Approaches. *Computational Linguistics*. 28.1: 90-95.

Stevenson, Mark. 2003. "Word sense disambiguation: the case for combinations of knowledge sources." Stanford, California: Center for the Study of Language and Information.

Xue, Nianwen Jinying Chen, and Martha Palmer. 2006. "Aligning Features with Sense Distinction Dimensions." Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 921–928. Sydney, July 2006.

Zempleni, Monika-Zita, Remco Renken, John C.J. Hoeks, Johannes M. Hoogduin, and Laurie A. Stowe. 2007. "Semantic ambiguity processing in sentence context: Evidence from event-related fMRI." *Neuroimage*, 34: 1270–79.