Chapter 2
Previous Researches on Lexical Ambiguity and Polysemy

The goal of this study aims to explore all possible senses of the four target words, which are all lexically ambiguous words. In this chapter, I would like to introduce and discuss lexical ambiguity and polysemy. In addition, I also would like to discuss that lexical ambiguity studies are used in the corpus-based and computational and psycholinguistic approaches. Therefore, I will discuss corpus-based and computational models and experimental evaluation of the psycholinguistic perspective for the four target words in this sense prediction study. Of course, I will discuss several hypotheses and outline the related research questions.

2.1 What are Lexical Ambiguity and Polysemy?

In this sense prediction study, the main aim is to explore all possible senses for undefined words; usually, these words have two or more different senses, have more ambiguous interpretations, and have more polysemous explanations. In other words, these words are regarded as lexically ambiguous or polysemous. However, concerning semantic knowledge, there are also some differences between lexical ambiguity and polysemy. Therefore, it is necessary to define both lexical ambiguity and polysemy in order to determine how undefined words are classified. Moreover, in this chapter, I will determine whether the four target words in this study are classified as lexically ambiguous or are classified as polysemous.

2.1.1 Lexical Ambiguity

In lexical semantics, computational linguistics, and psycholinguistics, the issue of lexical ambiguity is often discussed. Many scholars talk about polysemy and lexical ambiguity in their studies because they are related concepts. However, they are different when it comes to vague words versus ambiguous words and polysemous words versus homonymous words.

J.-F. Hong, *Verb Sense Discovery in Mandarin Chinese—A Corpus based Knowledge-Intensive Approach*, DOI 10.1007/978-3-662-44556-3_2

Lexical ambiguity and polysemy both indicate vague, unclear, and indefinite senses; that is to say, lexically ambiguous words and polysemous words can refer to more than two senses at the same time. Because they are so similar, it is necessary to define lexical ambiguity and polysemy as accurately as possible. Therefore, I will discuss lexical ambiguity in this section and will discuss polysemy in the next section.

According to interpretation and comprehension, lexical ambiguity is the property of being ambiguous; that is, a word, term, notation, sign, symbol, phrase, sentence, or any other form used for communication is called ambiguous if it can be interpreted in more than one way. Lexical ambiguity (*bank*) is different from vagueness (*aunt*), which arises when the boundaries of meaning are indistinct. Lexical ambiguity is context-dependent: the same linguistic item (be it a word, phrase, or sentence) may be ambiguous in one context and unambiguous in another context. For a word, lexical ambiguity typically refers to an unclear choice between different definitions as may be found in a dictionary. A sentence, however, may be ambiguous due to different ways of parsing the same sequence of words.

Lexical ambiguity is a linguistic term for a word's capacity to carry two or more obviously different meanings, for example, *bank*. The word "*bank*" has several distinct lexical definitions, including "financial institution" and "edge of a river." The context in which a lexically ambiguous word is used often makes evident which of the meanings is intended. Therefore, if someone uses a multi-defined word, it is sometimes necessary to clarify the context by elaborating on the specific intended meaning (in which case, a less ambiguous term should have been used). Lexical ambiguity arises when a word or concept has an inherently diffuse meaning based on widespread or informal usage. This is often the case, for example, with idiomatic expressions, whose definitions are rarely if ever well defined and are presented in the context of a larger argument that invites a conclusion.

Lexical ambiguity is one of the most difficult problems in language processing studies and thus, not surprisingly, it is at the core of lexical semantics research. Concerning the distinction of lexical ambiguity, Weinreich's (1964) distinction between contrastive lexical ambiguity and complementary ambiguity was illustrative to this point. Contrastive lexical ambiguity is the situation where a lexical item is associated with at least two distinct and unrelated meanings while complementary ambiguity must be distinguished in a full semantic description: a purely formal analysis, without reference to the substance.

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. This definition is also used for polysemy. In other words, WordNet research team members regard polysemy and lexical ambiguity as equivalent.

When determining the sense of a word, it is useful to distinguish the three stages of processing lexical ambiguity: (1) decoding the input and matching it with a lexically ambiguous word; (2) accessing the information about the ambiguous word; and (3) integrating the information with the preceding context (Cottrell 1984).

Therefore, when defining lexically ambiguous senses, it is important to keep in mind 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.

2.1.2 Polysemy

Polysemy is also a linguistic term for words with two or more meanings, usually multiple and related meanings for a word or words. The words polysemy and polysemous are defined as having or characterized by many meanings; the existence of several meanings for a single word or phrase. When polysemous words are discussed, homonymous words are likely to be discussed at the same time. However, polysemous words present different related meanings while homonymous words present unrelated meanings.

Since the vague concept of relatedness is one test for polysemy, judgments of polysemy can be very difficult to make. Because applying pre-existing words to new situations is a natural process of language change, looking at a word's etymology is helpful in determining polysemy but this is not the only solution; as meanings become lost in etymology, what once was a useful distinction of meaning may no longer be so. Some apparently unrelated words share a common historical origin; however, etymology is not an infallible test for polysemy, and dictionary writers often defer to speakers' intuitions to judge polysemy in cases where it contradicts etymology. Many words in Chinese are polysemous. For example the verb 打 (*da3* "hit") can mean 打手臂 (*da3 shou3bi4* "hit the back of a hand"), 輪胎打氣 (*lun2tai1 da3qi4* "pump gas into tire"), 打針 (*da3 zhen1* "inject"), 把碗打破 (*ba3 wan3 da3 po4* "break a bowl"), etc. (Hong et al. 2007, 2008).

There are several tests for polysemy, but one in particular is zeugma: if one word seems to exhibit define when applied in different contexts, it is likely that the contexts bring out different polysemy of the same word. If two senses of the same word do not seem to match [e.g., 打 (*da3* "hit")], yet they seem related, then it is likely that they are polysemous. The fact that this test again depends on speakers' judgments about relatedness, however, means that this test for polysemy is not infallible but rather is merely a helpful conceptual aid.

The study of polysemy, the multiplicity of meanings of words, has a long history in the philosophy of language, linguistics, psychology, and literature (Ravin and Leacock 2000). Ravin and Leacock (2000) pointed out three major approaches to semantics represented in polysemy study: (1) the Classical Approach (e.g., Goddard 2000); (2) the Prototypical Approach (e.g., Fillmore and Atkins 2000); and (3) the Relational Approach (e.g., Fellbaum 2000).

While the classical theories emphasizes definitions and related meaning to truth conditions, possible words, and states of affairs, prototypical approaches emphasizes meaning as part of a larger cognitive system and relates it to mental representations,

cognitive models, and bodily experiences. It is problematic to represent polysemy within a relational framework, as polysemous word senses can be very distant from each other in the semantic network's conceptual space.

In addition, Geeraerts (1993) emphasized the importance of context when determining the predictions of each of his tests, as he demonstrated that context alters the senses of the words found in it. This emphasis on context is common to all lexical ambiguity studies. In the above section on lexical ambiguity, it was mentioned that senses could be represented independently of the context in which they occur. However, it is very important to focus on context for both lexical ambiguity studies and polysemy studies.

In general, when talking about lexically ambiguous words or polysemous words, word sense disambiguation (WSD) also should be taken into consideration. Karov and Edelman (1998) pointed out the typical construct of WSD as follows:

> Word sense disambiguation (WSD) is the problem of assigning a sense to an ambiguous word, using its context. We assume that different senses of a word correspond to different entries in its dictionary definition. For example, *suit* has two senses listed in a dictionary: 'an action in court,' and 'suit of clothes.' Given the sentence *The union's lawyers are reviewing the suit,* we would like the system to decide automatically that *suit* is used there in its court-related sense (we assume that the part of speech of the polysemous word is known).

In other words, if researchers would like to decide the correct sense for polysemous words automatically based on context, they will refer to computer applications. According to Yael and Leacock (2000), computer applications that handled the content of natural language texts need to come to terms with polysemy. They consider that the study of polysemy in computational linguistics addresses the problem of how to map expressions to their intended meanings automatically. As a matter of course, it is very important to employ computer applications in polysemy studies. More extendedly, corpus-based approaches have been used and machine-readable corpora have also existed.

Finally, using computer applications to deal with word sense disambiguation in polysemy studies is reasonable because corpus-based approaches can provide statistical corpus analyses and the machine-readable corpora can provide large-scale data. In addition, the three approaches (the Classical Approach, the Prototypical Approach, and the Relational Approach) that Ravin and Leacock (2000) pointed out are appropriate when dealing with different problems in polysemy studies.

2.1.3 The Relationship Between Lexical Ambiguity and Polysemy

In this study, I choose four undefined target words with the intent of finding their correct senses and assigning their appropriate senses based on different contexts. If a word has more than two senses at the same time, then it is usually called a lexically ambiguous word or a polysemous word. However, "lexical ambiguity" or

"polysemy" is presented in several different related researches as referring to the same target. Even though WordNet research team members regard lexical ambiguity and polysemy as synonymous, lexical ambiguity and polysemy also can be used in different contexts to represent two or more different meanings. It is very difficult to differentiate lexically ambiguous words and polysemous words because they have common points: more than two senses, vague senses, related senses, and extended senses at the same time.

In fact, lexical ambiguity and polysemy are concepts used in several perspectives, such as Information Retrieval, computational approaches, natural language processing (NLP), artificial intelligence, semantics, pragmatics, discourse, psycholinguistics, and neuropsychology. That is to say, lexical ambiguity and polysemy are very similar among different fields.

The main aim of this sense prediction study is to predict all possible senses for the four target words—*chi1* "eat", *wan2* "play", *huan4* "change", and *shao1* "burn" rather than to disambiguate word senses. Because I will be using a large-scale corpus as my empirical data in this study, firstly, I will extract the collocation words of the four target words and cluster related collocation words in order to obtain words that have the same senses and then cluster them in the same cluster. I will then attempt to use the collocation words as intermediaries in order to predict all possible senses and to examine all sentences of the four target words. In doing so, I can obtain and predict all possible senses for the four target words from these sentences. In the previous sections, I defined lexical ambiguity and polysemy and introduced the following differences: (1) it is necessary to disambiguate word senses in polysemy studies; and (2) it is necessary to divide word senses in lexical ambiguity studies. Since the main goal of this study is to predict the correct senses for the four target words chosen by dividing word senses rather than word sense disambiguation in this study.

In sum, the main work of this study is to explore, predict, and obtain all possible senses from all sentences for the four target words—*chi1* "eat", *wan2* "play", *huan4* "change", and *shao1* "burn"—rather than disambiguate all possible word senses based on the context of the target words. Finally, I will regard the four target words as lexically ambiguous words and therefore will use the linguistic terms "lexical ambiguity" and "lexically ambiguous words" throughout the remainder of this sense prediction study.

2.2 Corpus-Based and Computational Model

2.2.1 Review of Previous Studies

Computational programming systems are designed to determine the appropriate senses of words as they appear in linguistic contexts. Therefore, it is necessary to review previous corpus-based and computational studies and discuss their models and approaches.

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 2002). Several previous studies concerning lexical ambiguity are well-cited in the literatures. However, the focus of these studies is nearly all on verbs rather than nouns. These studies include mental processing comprehension (Ekaterini 2002), lexicon and WordNet interpretations (McRoy 1992; Heiko 2002; Wu 2003; Buscaldi et al. 2007), context-based analysis (Jos'e et al. 2005; Cyma 2006; Wong et al. 2006), information retrieval and machine translation (Li et al. 2000; Jos'e et al. 2005; Wong et al. 2006; Zhou et al. 2006; Buscaldi et al. 2007), and lexical semantic knowledge representation and frame-based approach (Bolette 1997; Lien 2000; Hsu and Liu 2004; Liu et al. 2005).

In the case of the previously mentioned related lexical ambiguity studies in this chapter (i.e., studies based on the corpus-based and computational perspective, psycholinguistics perspective, and neurolinguistics perspective), I will divide these lexical ambiguity studies into three different categories, list previous studies in these three categories, and point out their significance, as shown in Table 2.1. I will focus on discussing previous corpus-based and computational studies and point out the gaps of these previous studies.

Veronis and Ide (1990) described a means for automatically building very large neural networks (VLNNs) from definition texts in machine-readable dictionaries and demonstrated the use of these networks for WSD. In their model, words were complex units. Each word in the input was represented by a *word node* connected by excitatory links to *sense nodes,* representing the different possible senses for that word in the *Collins English Dictionary*. However, as they noted several improvements can further be made: (1) the parts of speech (POS) for input words and words in definitions can be used to extract only the correct lemmas from the dictionary; (2) the frequency of use for particular senses of each word can be used to help choose among competing senses; and (3) additional knowledge can be extracted from other dictionaries and thesauruses.

An up-to-date sampling of a wide range of methods can be found in a special issue of Computational Linguistics on WSD (Philip and Yarowsky 2000). Annotated data has since facilitated recent advances in POS-tagging, parsing, and other language processing sub-problems. They also presented a substantial exploration of the relationship between monolingual sense inventories and translation distinctions across languages.

Regarding Canas et al.'s (2003) study, they proposed using an algorithm to (a) enhance the "understanding of the concept map by modules in the CmapTools software that aide the user during map construction"; and (b) sort the meanings of a word selected from a concept map according to their relevance within the map when the user navigates through WordNet's hierarchies, searching for more appropriate terms. They presented the possibility of using an algorithm that exploits WordNet to disambiguate the sense of a word that is part of a concept or linking phrase in a concept map. The results shown were encouraging and suggest more research should be done to improve the algorithm.

Table 2.1 Three categories of lexical ambiguity studies

Category	Previous related studies	Significant points
Corpus-based and computational	Veronis and Ide (1990)	1. Uses the corpus-based approach 2. An adaptive system 3. Based on context 4. Divides the sense of lexically ambiguous words 5. Finds the possible senses of a word
	Philip and Yarowsky (2000)	
	Canas et al. (2003)	
	Ganesh and Prithviraj (2004)	
	Ker and Chen (2004)	
	Moldovan and Novischi (2004)	
	Chen et al. (2005)	
	Zhang et al. (2005)	
	Martinez et al. (2006)	
	Xue et al. (2006)	
	Peng et al. (2007)	
	Kipper et al. (2008)	
	Chen and Palmer (2009)	
	Pitler et al. (2009)	
Psycholinguistic	Tabossi and Zardon (1993)	1. Experimentally-based 2. Determines literal bias meanings or metaphorical bias meanings 3. Context influences lexical access 4. Conceptual domains and the linguistic context
	Li and Yip (1996, 1998)	
	Li (1998)	
	Ahrens (1998, 2001, 2006)	
	Lin and Ahrens (2000)	
Neurolinguistic	Gunter et al. (2003)	1. Takes lexically ambiguous words to examine comprehensions of different senses 2. Processes ambiguous words that can occur both as nouns and as verbs 3. Examines lexical ambiguity comprehension in order to determine the meanings of literal bias or metaphoric bias
	Li et al. (2004)	
	Elston-Guttler et al. (2006)	
	Mason and Just (2007)	
	Zempleni et al. (2007)	

Ganesh and Prithviraj (2004) introduced the notion of soft WSD, which states that given a word, the sense disambiguation system should not commit to a particular sense but, rather, should commit to a set of senses that are not necessarily orthogonal or mutually exclusive. In their work, WordNet gave multiple senses for a word, which were related and which helped connect other words in the text. They defined soft WSD as the process of enumerating the senses of a word in a ranked order. This could be an end in itself or an interim process in an IR task, such as question answering. They also found a Bayesian belief network (BBN), a natural structure to encode such combined knowledge from WordNet corpus for training.

Ker and Chen (2004) described a general framework for adaptive WSD. Three issues must be addressed in a lexicalized statistical WSD model: (1) data sparseness; (2) lack of abstraction; and (3) static learning. They also mentioned that an adaptive system is superior in two ways to static word-based models trained on a

corpus. Through this learning strategy, an initial knowledge set for WSD was first built based on the sense definition in training data.

Moldovan and Novischi (2004) showed how lexical chains and other applications could be built on this richly connected WordNet. They used the senses of the words and defined them in WordNet. In order to overcome the data sparsity problem, they relied on a set of methods that showed that disambiguation classes of words share a common property. A suite of heuristical methods was presented for the disambiguation of WordNet glosses. Moldovan and Novischi have used lexical chains successfully to link question keywords with answer texts, providing axioms to a Question-Answering logic prover.

In a different approach, the contexts that include ambiguous words are converted into vectors by means of a second-order context method, and these context vectors are then clustered by the k-means clustering algorithm (Chen et al. 2005). The k-means clustering approach is an important method for data mining and knowledge discovery, as it has the characteristics of simplicity and fast convergence.

Zhang et al. (2005) proposed a corpus-based Chinese WSD approach using HowNet. The approach is based on the following observation: The different senses of a polysemous word tend to appear in cognizably different contexts. They described a method that performs Chinese WSD by combining lexical co-occurrence knowledge, semantic knowledge, and domain knowledge. By this approach, the experimental results showed that the method is very promising for Chinese WSD in that study.

Martinez et al. (2006) observed that each algorithm, based on Leacock et al. (1998), performed better for different types of words and each of them failed for particular words. They observed a similar performance in preliminary experiments when using an ML method or applying a heuristic on the different factors. They also built a disambiguation algorithm that can be explained in four steps. The results showed that the new method clearly outperforms the monosemous relatives in that dataset. However, they also noticed that this improvement does not happen for all the words in the set.

Concerning computational systems, Xue et al. (2006) devised a WSD system to analyze ten highly polysemous verbs in Chinese. They compared the features they used for Chinese with those used in a similar English WSD system. In that study, they demonstrated that rich linguistic features, specifically features based on syntactic and semantic role information, are useful for the WSD of Chinese verbs.

Peng et al. (2007) mentioned that collocation was a combination of words that has a certain tendency to be used together—and this was used widely to attack the WSD task—and word classes were often used to alleviate the data sparseness in NLP. They claimed that the algorithm of extending the collocation list that was constructed from the sense-tagged corpus was quite straightforward. In their experiments, the precision was proportional to the number of word classes. The results of these experiments have shown that the average F-measure improved to 70.81 % compared to 54.02 % of the baseline system where the word classes were not considered, although the precision decreased slightly.

Several scholars are still devoted to related works of the sense prediction study or sense distinction performance study, such as Kipper et al. (2008), Pitler et al. (2009), and Chen and Palmer (2009). Kipper et al. (2008) mentioned that lexical classifications have proved useful in supporting various NLP tasks and some information of VerbNet (VN). VerbNet is an extensive on-line lexicon for English verbs, providing detailed syntactic-semantic descriptions and a hierarchical domain-independent, broad-coverage verb lexicon with mappings to several widely used verb resources. They integrated two extensions into VN and incorporated the new classes into VN. Therefore, these steps were syntactic descriptions, thematic roles, and semantic descriptions of classes, such as entirely novel classes, novel sub-classes, and classes where restructuring was necessary. Many uses of verb classes in VN were being attested in a variety of applications, such as automatic verb acquisition, semantic role labeling, robust semantic parsing, word sense disambiguation, building conceptual graphs, and creating a unified lexical resource for knowledge extraction.

In another recent automatic sense prediction study, Pitler et al. (2009) worked with a corpus of implicit relations present in newspaper text and reported results on a test set. They used several linguistically informed features: polarity tags, Levin verb classes, length of verb phrases, modality, context and lexical features, and used the Penn Discourse Treebank (PDTB). They examined the most informative word pair features and found that they were not the semantically related pairs that researchers had hoped. In order to identify features useful for classifying comparison and other relations, they chose a random sample of 5,000 examples for contrast and 5,000 other relations. They used experiments to demonstrate that features developed to capture word polarity, verb classes, and orientation and found that several lexical features were strong indicators of this type of discourse relation.

In the case of Chen and Palmer (2009), they discussed a high-performance, broad-coverage supervised WSD system for English verbs that used linguistically motivated features and a smoothed maximum entropy machine-learning model. In their work, there were three major aspects: (1) developing a high-performance WSD system for English verbs by using linguistically motivated features; (2) applying this system to the first large-scale annotation effort aimed specifically at providing suitable training data for high-performance WSD, followed by discussion and analysis of these results; and (3) discussing potential future research areas for large-scale, high-performance supervised WSD. In fact, their analysis showed that using linguistically motivated features, such as semantic features, helped to relieve the data sparseness problem. In addition, their experimental results on the larger set suggested several areas they can explore in the future for improving high-performance WSD.

Some related previous studies were involved WSD in the sense prediction, I consider that word sense induction (WSI) maybe more related for this sense prediction study. Navigli (2009) mentioned that a different approach to the induction of word senses consisted of *word clustering* techniques, that was, methods which aimed at clustering words which were semantically similar and could thus convey a

specific meaning. Navigli (2009) also mentioned that word sense induction was performed with high precision (recall varies depending on part of speech and frequency).

In addition to the above previous studies, I have also investigated some representative studies concerning lexical ambiguity in lexical semantics. These include Lexical Semantics and Meaning in Language (Cruse 1986, 2004), WordNet (Fellbaum 1998), and Generative Lexicon (Pustejovsky 1991, 1995). From these previous studies, I observed that lexically ambiguous word senses might include several cases illustrating the relation of indefiniteness, in which the significant part is more predominant than the overlapping semantic element.

In Pustejovsky's (1995) generative lexicon study, especially, he discussed the logical problem of polysemy and pointed out two types of ambiguity—contrastive ambiguity and complementary polysemy—by following Weinreich (1964). Concerning contrastive ambiguity, Pustejovsky mentioned that given the current representational techniques and strategies for differentiating word senses, there would appear to be no reason to make a logical distinction between these two types of ambiguity. A dictionary called a *Sense Enumeration Lexicon* (*SEL*) was introduced, and it appeared at first to handle adequately the sense differentiation for both ambiguity types. From the theoretical perspective, the major problems posed by contrastive ambiguity involved issues of discourse inference and the correct integration of contextual information into processing. Therefore, Pustejovsky brought up the elementary lexical semantic theory and considered that the major part of semantic research had been on logical form and the mapping from a sentence-level syntactic representation to a logical representation language. In addition, regarding the *Sense Enumeration Lexicon*, he characterized it directly as follows:

> A lexicon L is a *Sense Enumeration Lexicon* if and only if for every word w in L, having multiple senses s_1, …, s_n, associated with that word, then the lexical entries expressing these senses are stored as $\{w_{s1}, \ldots, w_{sn}\}$.

As in the example *bank* in above I mentioned, two contrastive senses could be listed in a straightforward fashion as shown in (2.1) and (2.2), using a fairly standard lexical data structure of category type (CAT) and a basic specification of the genus term (GENUS), which locates the concept within the taxonomic structure of the dictionary

$$\begin{pmatrix} \text{bank}_1 \\ \text{CAT} = \text{count\_noun} \\ \text{GENIUS} = \text{financial\_institution} \end{pmatrix} \tag{2.1}$$

$$\begin{pmatrix} \text{bank}_2 \\ \text{CAT} = \text{count\_noun} \\ \text{GENIUS} = \text{store} \end{pmatrix} \tag{2.2}$$

All possible selectional requirements of verbs are defined from the features or types as the genus terms, and disambiguation would appear to be merely the process

of correctly matching the features of functor and arguments from the available set of lexical entries.

Although this approach was taken by many researchers within both theoretical and computational traditions, Pustejovsky presented three arguments against using the *SEL* as a model of lexical semantics: (1) "The Creative Use of Words"—the *SEL* cannot capture the full range of word usages; (2) "Permeability of Word Senses"—the *SEL* cannot capture the relationship between senses; and (3) "Difference in Syntactic Forms"—the *SEL* cannot allow senses to have an adequate range of syntactic realizations.

These arguments present problems in defining a set of features or types for contrastive senses of the verbs in the *Sense Enumeration Lexicon*. It is necessary to improve this approach and then make a useful model to deal with lexically ambiguous words, which is the aim of this sense prediction study. I will also use the corpus-based and computational approach but with two different strategies—character similarity clustering analysis and concept similarity clustering analysis. I expect the results to be better than the results using the *Sense Enumeration Lexicon*.

2.2.2 Gap of Previous Studies

Overall, regarding these previous corpus-based and computational studies, these scholars proposed corpus-based, algorithm, automatically computational programming system, and collocation approaches to analyze sense prediction studies or WSD studies. Moreover, they also recommended that using large-scale corpus, context, semantic features, and concepts could achieve high performance for sense prediction studies. In the above studies, I found that they generally employed only one corpus in their studies, which resulted in less information of lexical ambiguity for their sense prediction studies; they also did not combine the various approaches available.

Focusing on some previous studies of them, I consider that specific research gaps existed and these research gaps were easily observed. For example, they were observed in Ker and Chen (2004), Chen et al. (2005) and Peng et al. (2007).

Ker and Chen (2004) mentioned that the first step of their study was to construct an initial knowledge from training corpus. However, they did not point out how many training data which they selected can let them obtain better performance under the adaptive sense disambiguation approach. In this study, I will point out the number of predicting clusters as my default target for the four target words to present the best results.

In Chen et al. (2005), they didn't explain what was about second-order context and why this approach could provide more information about word senses in contexts. In addition, they mentioned that the whole process was completed automatically, so a sense-labeled corpus was not need. In my study, I will not only predict all possible senses of the four target words by automatically computational programming, but I will also examine these clusters whether they can be predicted

appropriate senses in manual. In addition, I will use senses divisions in Chinese Wordnet and *Xiandai Hanyu Cidian* (*Xian Han*) to estimate the evaluations of the four target words by my own intuition.

Peng et al. (2007) took the target verb 吃 *chi1* "eat" as the illustration and selected the number of word classes. In their study, they only talked about the concrete objects for 吃 *chi1* "eat" but no abstract objects for 吃 *chi1* "eat". In my corpus-based and computational analysis, I will predict physical senses and metaphorical senses of the four target words.

In consideration of several research gaps presented by these previous studies, this study utilized four corpora in order to obtain richer information, automatically computational programming to gather related collocation words of the four target words—*chi1* "eat", *wan2* "play", *huan4* "change", and *shao1* "burn", and used HowNet in an attempt to identify their semantic features and elements. In addition, this study adopted the same morpheme contrast and concept contrasts by automatically computational programming in the corpus-based and computational approach in order to divide the sense clusters of the four target words.

2.3 Hypotheses and Research Questions

With respect to the sense prediction study of lexical ambiguity, there are three hypotheses in this study. Lexical ambiguity means some words have multiple meanings or senses (Moldovan and Novischi 2004). In the *SEL* model, although lexically ambiguous words list all possible selectional requirements that are associated with those words and then lexically ambiguous words express these senses, the *SEL* model can not capture the full range of word usages. Therefore, the first hypothesis is that words with similar morpheme-character components and concept elements are similar in sense. I will follow Fujii and Croft (1993) to observe character similarity and refer to Li et al. (2003) and Dai et al. (2008) to explore concept similarity via HowNet.

Peng et al. (2007) mentioned that a corpus was divided into five equal parts which one part was used as the test corpus and the collocation list was constructed from the other four parts of corpus. In this study, the second hypothesis is that different corpora with particular functions which provide different lexical knowledge bases. I will use Chinese Gigaword Corpus to select related collocations for the four target words; I will use HowNet to assign all possible concepts to ambiguous senses of the four target words; I will use Chinese Wordnet to estimate the evaluations for the four target words and I will also use *Xian Han* to estimate the evaluations for the four target words.

According to Ahrens' studies (1998, 2001, 2006), she considered that sentential context and meaning frequency could influence the lexical ambiguity resolution and access. Owing to all possible clusters of all collocation words of the four target words being selected from the character similarity clustering analysis by examining their contexts, I consider these collocation words, which are stimuli for the

experimental evaluations, to have been identified by their frequencies and their senses in different contexts. Hence, the third hypothesis is that in the off-line multiple-choice task, subject uses conceptual difference to identify the choice. In this study, I will use an off-line multiple-choice task involving experimental evaluation in order to examine which concept of one selected word/item is obviously different from the concepts of the other three words/items. Stimuli are selected from the character similarity clustering analysis by examining their contexts and they are controlled the frequencies. And then, I can demonstrate other approaches which can verify the analysis of the corpus-based and computational approach.

For this reason, there are three research questions in this study: (1) How do I predict the word senses of a lexically ambiguous word in order to present different interpretations in different contexts or domains? (2) How do I use more than two corpora as the database to support this word sense prediction study? (3) Can I use other approaches to verify the analysis of the corpus-based and computational approach for this word sense prediction study?

I will make use of the lexical semantics, lexical features, concepts, and collocation words to examine these research questions using Chinese Gigaword Corpus, HowNet, Chinese Wordnet, and *Xiandai Hanyu Cidian*. Therefore, I will attempt to utilize the corpus-driven linguistic approach as my main method for this sense prediction study.

http://www.springer.com/978-3-662-44555-6

Verb Sense Discovery in Mandarin Chinese—A Corpus based Knowledge-Intensive Approach

Hong, J.-F.

2015, XIV, 249 p. 7 illus., 3 illus. in color., Hardcover
ISBN: 978-3-662-44555-6