Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 417-430

The Sinica Sense Management System: Design and Implementation

Chu-Ren Huang*, Chun-Ling Chen*, Cui-Xia Weng*, Hsiang-Ping Lee*,

Yong-Xiang Chen* and Keh-Jiann Chen*

Abstract

A sense-based lexical knowledgebase is a core foundation for language engineering. Two important criteria must be satisfied when constructing a knowledgebase: linguistic felicity and data cohesion. In this paper, we discuss how data cohesion of the sense information collected using the Sinica Sense Management System (SSMS) can be achieved. SSMS manages both lexical entries and word senses, and has been designed and implemented by the Chinese Wordnet Team at Academia Sinica. SSMS contains all the basic information that can be merged with the future Chinese Wordnet. In addition to senses and meaning facets, SSMS also includes the following information: POS, example sentences, corresponding English synset(s) from Princeton WordNet, and lexical semantic relations, such as synonym/antonym and hypernym/hyponym. Moreover, the overarching structure of the system is managed by using a sense serial number, and an inter-entry structure is established by means of cross-references among synsets and homographs. SSMS is not only a versatile development tool and management system for a sense-based lexical knowledgebase. It can also serve as the database backend for both Chinese Wordnet and any sense-based applications for Chinese language processing.

Keywords: Lexical Knowledgebase, Word Senses, Chinese Wordnet, Lexical Semantic Relation

* Institute of Linguistics, Academia Sinica, 128, Section 2, Academia Road, Nankang, Taipei, 115, Taiwan Tel: 886-2-26523108 Fax: 886-2-27856622
E-mail: churen@gate.sinica.edu.tw

[Received April 10, 2005; Revised June 16, 2005; Accepted August 20, 2005]

1. Background and Motivation

A sense-based lexical knowledgebase is a core foundation for language engineering. WordNet and Euro WordNet are two well-known examples. Two important criteria must be satisfied when constructing a knowledgebase: linguistic felicity and data cohesion. Huang *et al*. discussed how linguistic felicity can be achieved when building a comprehensive inventory of Chinese senses from corpus data [Huang *et al*. 2003]. They introduced five criteria as well as operational guidelines for sense distinction. In this paper, we will discuss how data cohesion of the sense information collected using the Sinica Sense Management System (SSMS) can be achieved.

2. Introduction to the Content of the SSMS

The structure of a lexical semantic web can provide not only materials for linguistic research but also an infrastructure for NLP and other applications. Two main tasks are involved in constructing a Wordnet. One is distinguishing synsets which are clustered according to word senses, and the other one is collecting the semantic relations connecting the synsets. A wordnet is formed by taking these synsets as the nodes and then connecting each node by using the semantic relations. Of the two tasks, constructing the synsets is the most fundamental work. Constructing a synset involves classifying a series of synonyms that carry one specified lexical concept. A polysemous word is assigned more than one synset in order to show the variety of word senses. Analyzing and distinguishing word senses are two crucial steps when constructing a Wordnet. In order to lay a solid theoretical foundation of the current work, we developed a series of principles for analyzing Chinese word senses [Huang *et al*. 2003]. These principles not only satisfy the conditions of felicity and cohesion but also serve as guidelines for distinguishing and analyzing large numbers of Chinese word senses. Based on the word-sense identification criteria, the linguistic knowledge can be consistently described, and we can easily map the linguistic knowledge to ontologies or transform it into formal representations.

The SSMS system is designed to store and manage word sense data generated in the analysis stage. SSMS manages both lexical entries and word senses. This system has been designed and implemented by the Chinese Wordnet Team at Academia Sinica. It contains all the basic information that can be merged with the future Chinese Wordnet. SSMS is meaning-driven. Each sense of a lemma is identified specifically and given a separate entry. When further differentiation at the meaning facet level is called for, each facet of a sense is also described in a full entry [Ahrens *et al*. 1998]. In addition to senses and meaning facets, this system also includes the following information: POS, example sentences, corresponding English synset(s) from Princeton WordNet, and lexical semantic relations, such as synonym/antonym and hypernym/hyponym. Moreover, the overarching structure of the

system is managed using a sense serial number, and the inter-entry structure is established by means of cross-references among synsets and homographs.

Currently, the Chinese Wordnet Team is focusing on analyzing middle-frequency words in the Sinica Corpus. Our reason for choosing middle-frequency words as our target ones is that with only three to five senses of a word, we can investigate the senses and meaning facets of each word deeply and accurately, while avoiding the simple situation of one sense for low-frequency words and the complicated situation of numerous senses for high-frequent words. So far, nearly 3,000 more lemmas have been analyzed, and close to 4,000 senses have been identified. These data are presented in a technical report that is updated yearly [Huang *et al*. 2005]. In the near future, these results will be used as a basis for Natural Language Processing or E-learning applications.

3. Design Principle of SSMS

A sense-based lexical knowledgebase with data cohesion must meet three requirements: unique identification of senses, trackability of senses, and consistent sense definitions. SSMS uses four tools to satisfy these requirements.

3.1 Unique Serial Number

Each sense or meaning facet is identified by a unique serial number in SSMS. In Princeton WordNet [Fellbaum 1998], each synset is given a unique offset number. However, the offset number does not have a logical structure. Hence, although it guarantees unique identification, it is not easily trackable. An alternative approach is to set up a base ontology and assign senses to an ontological node with a unique ID. However, this is not feasible since we cannot pre-designate all the possible conceptual and semantic relations. In addition, if a decision is made to encode only certain higher level nodes, the random assignment problem will occur because of the coarse granularity inherent to an upper ontology. In our system, the unique serial number of each sense is composed of three segments: sequential information indicating when the lemma was processed, the lemma form, and the sense classification code for each lemma (including the meaning facet level). Take “bao4 zhi3 (newspaper)” as an example: “bao4 zhi3” has two senses as well as two meaning facets for the first sense. The lexical entry of “bao4 zhi3” is as follows:

Example 3-1: The result of sense distinction for “bao4 zhi3 (newspaper)”

報紙 bao4 zhi3 ㄅㄠˋ ㄓˇ

詞義 1：【名詞，Na】指定期出版，報導新聞、提供各式訊息的出版品。

義面 1：指刊物，尤其指內容部份。{ newspaper, 03039218N }

例句：儘管他出現在**報紙**頭條的頻率極高，被刊登的卻幾乎都是片段性的談話。

義面 2：指定期出版，報導新聞、提供各式訊息的紙張本身。{ newspaper, 04738466N }

例句：他找了一張**報紙**，平鋪在面前，取下身邊掛著的匣子之後就開始自言自語。

詞義 2：【名詞，Na】指定期出版，報導新聞、提供各式訊息出版品的組織。{ newspaper, 06009637N }

例句：**報紙**對他進行專訪的內容將刊登於隔天的頭條新聞上。

The four-level unique serial number shown below contains four segments of the unique serial number for the first meaning facet of the first sense of “bao4 zhi3”. Note that the lemma form ID “0018” can be replaced by the actual lemma form “報紙” or “bao4 zhi3” for processing purposes:

	報紙 “bao4 zhi3 (newspaper)”
Lemma processing year	03-
Lemma form ID	-0018-
The first sense	-01-
The first meaning facet	-01

There are four advantages to managing the sense database with unique serial numbers. First, the sequential number not only gives a unique code for each lemma; it also enables a project manager to track work progress more easily. Second, including the lemma in the serial number helps human users to quickly identify the relevant senses. It also facilitates man-machine interaction, such as keyword search for senses. Third, it also provides a logical structure for the sense serial number, since each lemma represents a small number of possible senses. Lastly, four digits are reserved to identify senses and meaning facets belonging to each lemma. The first two digits are reserved for senses and the last two for meaning facets. These four digits also allow the minimal amount of space needed to identify exact sense in the database. For instance, when stipulating a synonym, we can identify it as word0200, which refers to the second sense of a certain lemma. There is no need to repeat the complete sense serial number. The sense serial number enables unique identification and also assists trackability.

3.2 Cross-Reference Device

SSMS automatically prompts all possible cross-references. When a lemma is called up for analysis, all existing records that contain this lemma are prompted. These include either lexical semantic relations, such as synonyms and hyponyms, sense definitions that contains this lemma, or explanatory notes that contain this lemma. In addition to offering rich semantic association information, this feature allows sense relations to be clearly defined, and inconsistencies to be detected. In addition, any anomaly in a definition or expression format can also be discovered. This process also helps us to narrow our focus to a set of control vocabulary for sense definitions. This feature also improves both the trackability of senses and the consistency of sense definitions. For example, if we enter the term "you2 mu4 (nomadic)" in a query, SSMS will automatically prompt two possible relevant references, "man3 (Manchu)" and "meng2 gu3 (Mongol)," both of which refer to areas where nomads live.

3.3 Concurrent Access to the Lexical Knowledgebase and Corpus

In addition, SSMS enables parallel and concurrent access to the lexical knowledgebase and corpus. When a lemma is chosen in the system, all tagged examples of that lemma from the Sinica Corpus are retrieved. This allows closer examination of how the senses are used and distributed. It also enables automatic selection of corpus example sentences. In turn, when sense classification is completed, SSMS allows all the corpus sentences to be sense-tagged and returned to be merged with the original corpus. In other words, a sense-tagged corpus is being processed in parallel. This feature allows each lexical sense to be traced to its actual uses in the corpus. It also allows linguists to examine the data supporting each sense classification. For example, if we enter the term "you2 mu4 (nomadic)" in a query, 9 tagged examples of the lemma from the Sinica Corpus will be retrieved, as shown in *Figure 1*.

Figure 1. Tagged examples of the lemma "you2 mu4 (nomadic)"from Sinica corpus

3.4 Linking to the Sinica BOW

Lastly, SSMS is also linked to the bilingual wordnet information at Sinica BOW. Candidate English synset correspondences, including offset numbers, are shown after a Chinese lemma is chosen. This aids cross-lingual trackability and consistency.

4. The Implementation of SSMS

The implementation of SSMS can be divided into three stages as shown in *Figure 2*. In the lemma analysis stage, based on the criteria and operational guidelines proposed by [Huang 2003], we distinguished senses and meaning facets for each word. At the same time, the Sinica Corpus and Wordnet were referred to for POS, examples, and English translations. Then with the help of dictionary resources or word mapping done by the system, we determined the word relations. The second stage involved two steps. First, we designed the schema of the sense management system database for storing analysis results obtained in the first stage. Then, for the purpose of database management, we developed an interface to help the Chinese Wordnet Team accessing and revising the database. We employed the DELPHI tool to design our system interface. Through the interface, the data in the database can also be exported in lexicon documents. The last stage of this system implementation is the application stage. The aim of our project is to build Chinese Wordnet web sites for user querying. The development languages for these web pages are HTML and ASP. The web pages in these web sites will be viewed through web servers. Through the Internet, people will be able to retrieve data from our sense management database system everywhere, at anytime. The flow of the Sinica Sense Management System is displayed in the following chart. There note that when the initial goals of all three stages are met, work on the first stage will continue to expend the database.

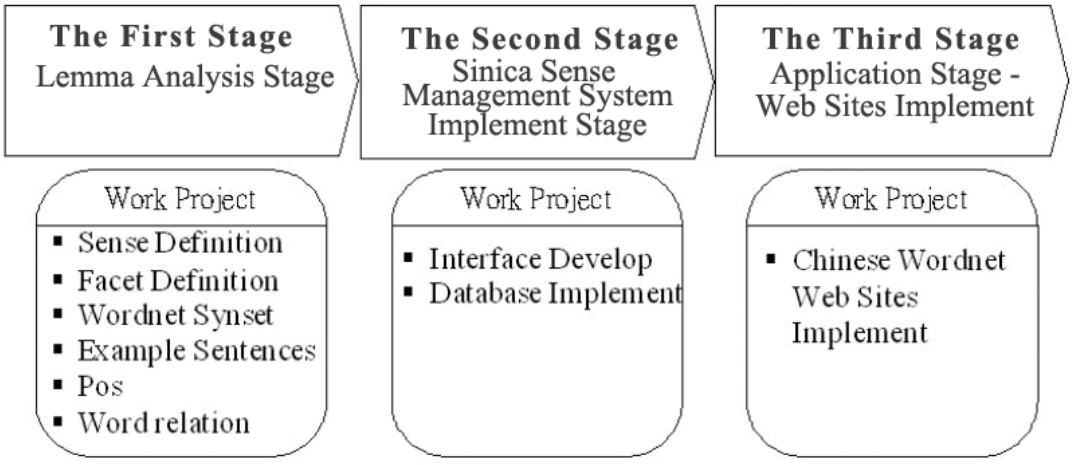


Figure 2. The flow chart of the Sinica Sense Management System

We represent the overall framework of SSMS diagrammatically in *Figure 3*. As the diagram indicates, the Chinese Wordnet Team uses SSMS to access the database and electronic documents as Lexicon reports. Moreover, users on the Internet can browse HTML/ASP pages to query the database through web servers.

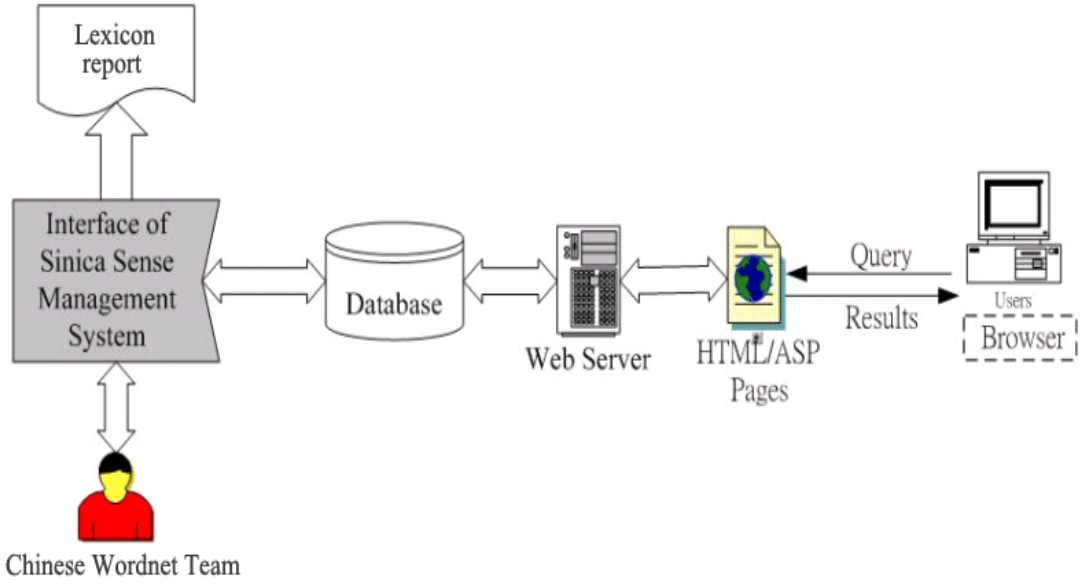


Figure 3. The overall structure of SSMS

4.1 The Schema of the SSMS Database in Class Diagrams

In the section, we will discuss the schema of the SSMS Database. The Unified Modeling Language (UML) [Booch *et al.*1999; Oestereich 1999] is a graphical notation that provides a conceptual foundation for assembling a system out of components from 4+1 views and nine diagrams. Each view is a projection into the organization and structure of the system, focusing on a particular aspect of that system.

We employ the class diagram notations in UML to provide a static view of application concepts in terms of classes and their relationships including generalization and association. Therefore, we will only provide some details about class diagrams in the following.

Class diagrams [Booch *et al.*1999; Oestereich 1999; Mullar 1999] commonly contain the following features:

1. A class diagram shows a set of classes and their relationships. For example, the class diagram of the Suppliers-and-Parts database is shown in *Figure 4.* The terms in italics in *Figure 4* indicate concepts concerning class diagrams.

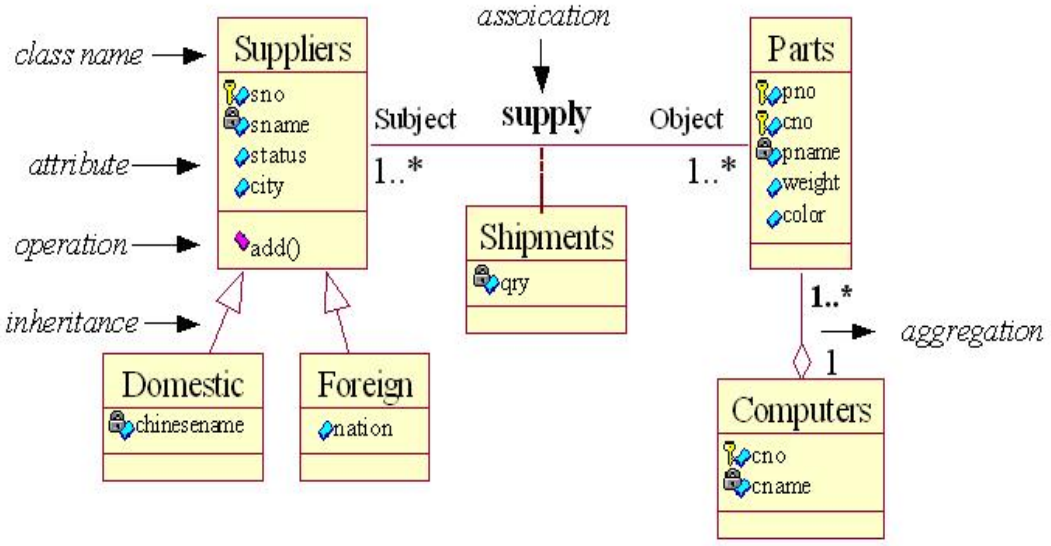


Figure 4. A class diagram for the Suppliers-and- Parts Database

2. A class is a description of a set of objects that share the same attributes, operations, relationships, and semantics. A class mainly contains three important parts: its name, attributes, and operations. We will explain these terms in the following:

(a) Class name: every class must have a name to distinguish it from other classes. For example, **Suppliers** and **Parts** are class names.

(b) Attribute: an attribute represents some property that is shared by all objects of that class. A class may have any number of attributes or no attributes at all. For example, in *Figure 4,* the class **Suppliers** has some attributes, such as *sno, sname, city.*

(c) Operation: an operation is the implementation of a service that can be requested from any object of the class in order to influence behavior. A class may have any number of operations or no operations at all. For example, in *Figure 4,* the class **Suppliers** has one operation: *add().*

3. There are three kinds of relationships between classes:

(a) Association: an association is a structural relationship that specifies objects of one thing to be connected to objects of another. For example, in *Figure 4,* a line drawn between the involved classes (**Suppliers** and **Parts**) represents an association named *supply.*

(b) Aggregation: an aggregation is a "whole/part" relationship, in which one class represents a larger thing (the "whole" class), which consists of smaller things (the "parts" class). Moreover, an aggregation represents a "has-a" relationship, which means that an object of the "whole" class has objects of the "part" class. To represent an aggregation, an empty diamond will be drawn at the "whole" class end of the line linking two classes.

(c) Inheritance: An inheritance relationship can be regarded as a generalization (or specialization), which is a taxonomic relationship between a general (super class) and a special (subclass) element, where the special element adds properties to the general one and behaves in a way that is compatible with it. Therefore, it is sometimes called an "is-a-kind-of" relationship. An inheritance relation is represented by means of a large empty arrow pointing from the subclass to the super class. For example, in *Figure 4*, **Domestic** and **Foreign** suppliers (two subclasses) together are a kind of supplier (the super class).

Based on the need for SSMS content and design principle, *Figure 5* shows the schema of SSMS database using the concepts of class diagrams.

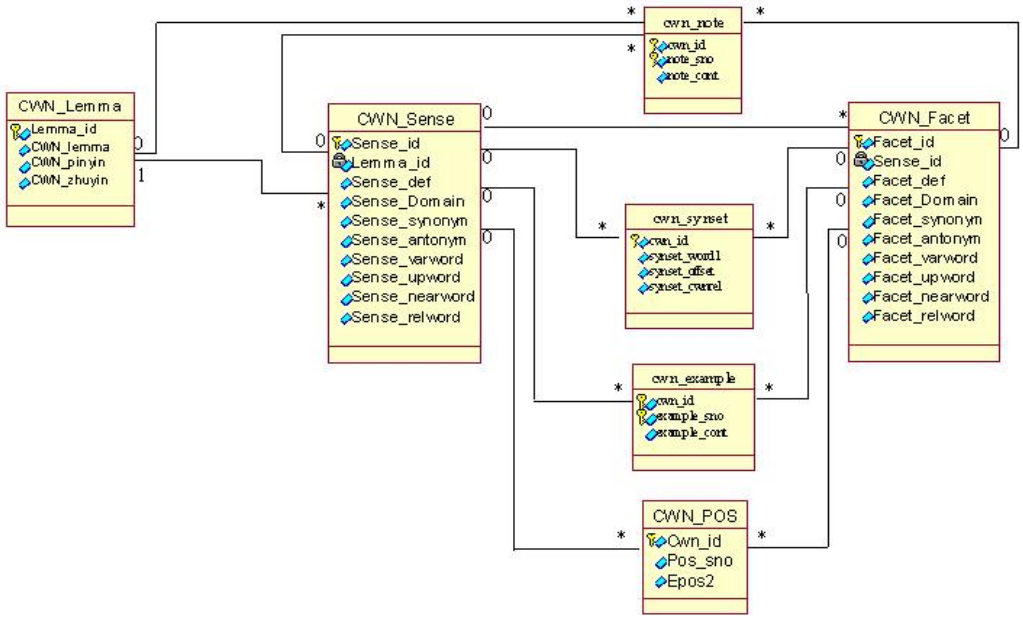


Figure 5. The schema of the Sinica Sense database

4.2 The Function of SSMS

In this section, we will discuss the interface marking for SSMS. Our task orientation is to design a clear and easy interface because we are trying to provide a helpful tool that the members of the Chinese Wordnet team will be able to easily use to access their analyzed records and query words, and compare senses with senses.

The development language for the SSMS interface is DELPHI 7.0. Based on the need for program execution, the architecture of SSMS is shown in *Figure 6.* The SSMS has different functions, and these functions have been represented in different sub-windows of main-windows interface and with inter-linked ASP pages. Sense management and sense visualization are two major functions in SSMS. With the Sense management function, the Chinese Wordnet term can insert, update, and delete datas including lexical entries, word senses, meaning facets, POS, example sentences, English synset(s), and lexical semantic relations. The sense visualization function is the SSMS interface and can be divided into two parts: sense query and lexicon report. The main interface of SSMS is shown in *Figure 7.* The SSMS interface provides a user-friendly interface for operation and maintenance. With the sense query function, users can enter a serial number or a lexical entry for sense querying using the SSMS interface. *Figure 8* shows the system view of the query sense function. Various information about individual words can be shown in the working window, include synonyms, antonyms, hyponyms, and variants. Through the clear presentation, lexical semantic relations can be understood and compared. Another function, the lexicon report, uses the development software Crystal Report9 to produce electronic documents as shown as *Figure 9.*

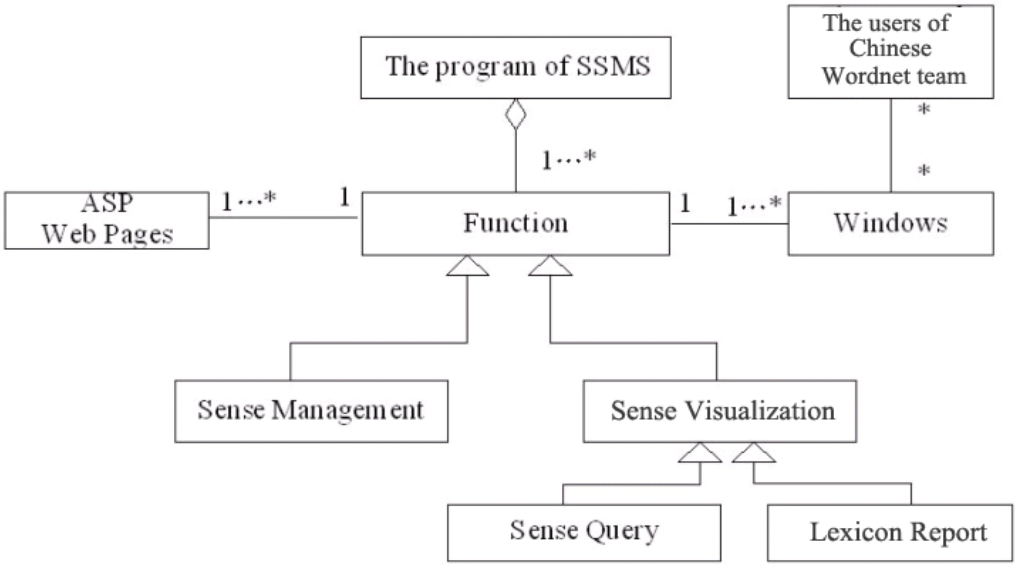


Figure 6. The class diagram of an SSMS function description

Figure 7. The main interface of SSMS

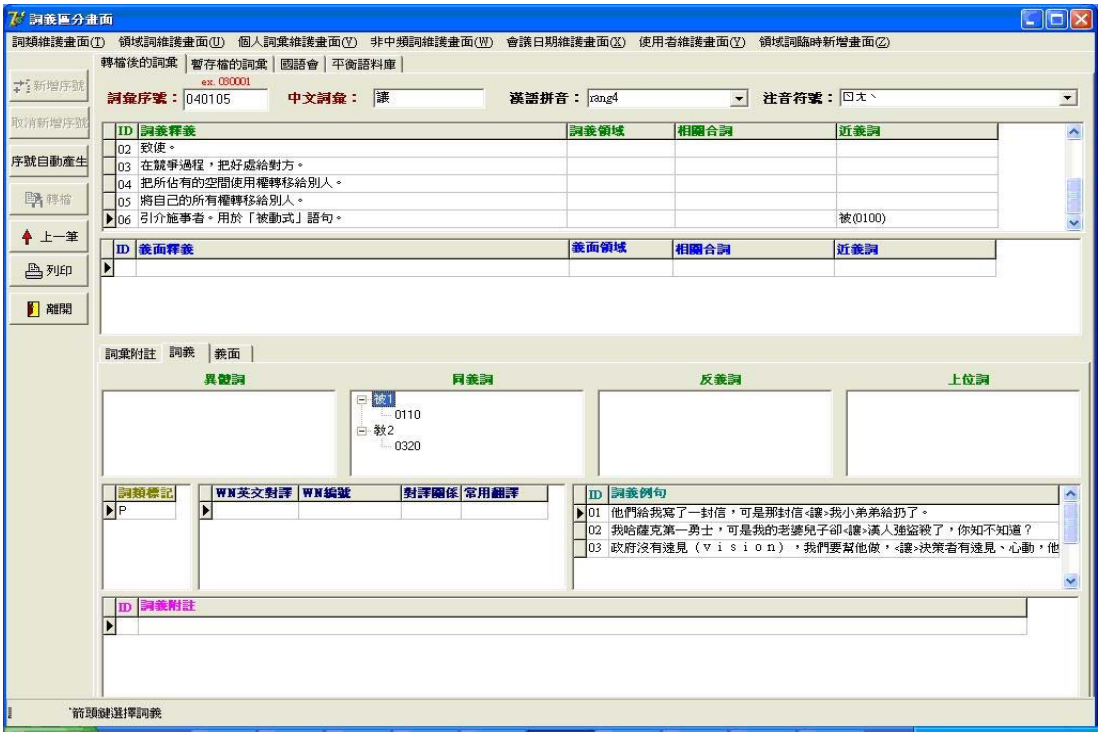


Figure 8. The system view of the query sense function

拌 ban4 ㄅㄢˋ
詞義 1：【及物動詞，VC】將任何材料混和、攪在一起。異體詞「伴」。{ blend, 00274169V }
義面 1：將食材和食材，或食材和醬汁混和在一起。
例句：小女孩剛開始是給小黃吃鮮魚拌飯，小黃吃得津津有味。
例句：植物油及動物油的含量如何能攝取到最低的程度，通常拌沙拉以植物油處理。
例句：他把牛肉切薄片，並舀一點湯燙一下就上桌，而乾麵則只加一些醬油拌一拌，吃不到師傅的手藝。

義面 2：將一般液態和固態的材料混和在一起，使成爲漿狀。
例句：他交給我一份以石灰水拌製海砂混凝土的研究計劃報告文件。
例句：牛屎伯公口裡從來沒停止過嚼檳榔，雙手則忙著拌紅灰、包荖葉。

附註：
1.分義面的主要原因在於「拌」這個詞用在食物上的用法頻率上相當高，已經形成特殊用法，所以以義面方式處理，以標誌出這種情形。

Figure 9. The format of a lexicon report

5. Results and Discussion

There is no question that semantic knowledge is needed to solve many problems in natural language processing. Building a sense-based lexical knowledgebase will be a subjective process, and the quality of this lexical knowledgebase will be highly dependent on the criteria which it must satisfy. In this paper, we introduced five criteria as well as operational guidelines for sense distinction and described how the Sinica Sense Management System can be used to manage both lexical entries and word senses.

More than 10 members of Chinese Wordnet Team were involved in the project. When we began this work, middle-frequency words were selected as the main terms to be analyzed. There are 2,018 middle-frequency words in Sinica Corpus. To date, 3,344 lemmas have been analyzed, and 5,914 senses have been identified. Among these records, the word “xia4” includes 42 senses and is the most complex term analyzed so far.

While more work needs to be done to improve the quality of this sense-based lexical knowledgebase, the goal of refining the procedure for word sense distinction needs to be accomplished at the same time. We believe that is a well-designed sense-based knowledgebase that it will serve as an important tool in Chinese knowledge processing applications.

References

Ahrens, K., L. Chang, K. Chen, and C. Huang, "Meaning Representation and Meaning Instantiation for Chinese Nominals," *Computational Linguistics and Chinese Lnaguage Processing,* 3, 1998, pp.45-60.

Booch, G., J. Rumbaugh, and I. Jacobson, *The Unified Modeling Language User Guide,* Addison-Wesley, 1999.

Fellbaum, C., et al., *WordNet: An Electronic Lexical Database,* MA: MIT Press, Cambridge, 1998.

Huang, C., et al., *Meaning and Sense in Chinese, 2nd Edition Technical Report 05-01.* CKIP, Academia Sinica, Taipei, 2005.

Huang, C., "Sense and Meaning Facet: Criteria and Operational Guidelines for Chinese Sense Distinction," Presented at *The Fourth Chinese Lexical Semantics Workshops,* 2003, Hong Kong City University, Hong Kong.

Muller, R.J., *Database Design for Smarties: Using UML for Data Modeling,* Morgan Kaufmann, 1999.

Oestereich, B., *Developing Software with UML Object-Oriented Analysis and Design in Practice,* Addison-Wesley, 1999.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/

Sinica Corpus: http://www.sinica.edu.tw/SinicaCorpus/

WordNet: http://www.cogsci.princeton.edu/~wn/

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 431-444

From Frame to Subframe: Collocational Asymmetry in Mandarin Verbs of Conversation

Mei-Chun Liu* **and Chun Edison Chang+**

Abstract

This paper examines the collocational patterns of Mandarin verbs of conversation and proposes that a finer classification scheme than the flat structure of 'frames' [cf. Fillmore and Atkins 1992; Baker *et al.* 2003] is needed to capture the semantic granularity of verb types. The notion of a 'subframe' is introduced and utilized to explain the syntactic-semantic interdependencies among different groups of verbs in the Conversation Frame. The paper aims to provide detailed linguistic motivations for distinguishing subframes within a frame as a *semantic anchor* for further defining near-synonym sets.

Keywords: Mandarin Verbs of Conversation, Semantic Frame, Subframe, Collocational Association

1. Introduction[1]

As the importance of lexical semantic research has grown with the need to represent human knowledge, various lexically-based information networks have been proposed. This includes the comprehensive work on differentiating word senses and sense relations in WordNet [Miller *et al.* 1990], the ontological hierarchy in SUMO [Niles and Pease 2003], and the more linguistically-motivated model of FrameNet [Baker *et al*. 2003]. While all these databases provide valuable information regarding word senses, the first two are constructed in a more

* Graduate Institute of Foreign Literatures and linguistics, National Chiao Tung University, Taiwan
E-mail: mliu@mail.nctu.edu.tw
+ Graduate Institute of Linguistics, National Tsing Hua University, Taiwan
E-mail: d948704@oz.nthu.edu.tw

[1] This paper is a part of a research project funded by the National Science Council (NSC 92-2411-H-009-020-ME). An earlier version was presented at the 5th Chinese Lexical Semantics Workshop at National Singapore University. We thank Chu-Ren Huang, Kathleen Ahrens, and two anonymous reviewers for their valuable comments. Of course, the authors are responsible for any possible errors.

[Received April 13, 2005; Revised June 14, 2005; Accepted August 14, 2005]

intuitive and pre-theoretical manner without detailing the linguistic evidence for sense distinctions. FrameNet, on the other hand, is based on the theory of Frame Semantics [Fillmore and Atkins 1992] and attempts to define meaning within a set of shared knowledge or background information, that is, a *frame*. However, as pointed out by Liu and Wu [2003], if meaning is anchored in the notion of a 'frame', then we need independent motivations for postulating different frames. What seems to be lacking in the current framework is a cognitive linguistic explanation as to how individual 'frames' are distinguished and interrelated. To answer this question, we will show that within a frame, a more elaborate classification system is needed to account for the variety of verb behaviors. The notion of a 'subframe' is introduced and utilized to capture the syntactic-semantic interdependencies observed in corpus data[2].

2. Defining the Conversation Frame

Compared with the other communication frames[3], the conversation frame is unique in that it profiles the property of *reciprocality* or two-way communication. Mandarin verbs in the conversation frame encode reciprocal communicative events, where participants are involved as Interlocutors, such as *tan* 談 'talk', *taolun* 討論 'discuss', *xietiao* 協調 'negotiate', *chaojia* 吵架 'quarrel', *xianliao* 閒聊 'chat', etc[4]. These conversation events highlight a particular subpart of the communication schema with its core being frame elements, as proposed in Liu and Wu's conceptual schema for the conversation frame [2003]. We can further characterize the conversation frame as follows:

(1) The conversation frame:

a) Cognitive Schema (grayed and bolded area):

b) Domain: communication;

c) Definition: a two-way communicative process between Interlocutors (Intl1, Intl2, or Intls) via a certain Medium, on a given Topic;

d) Semantic profile: reciprocality between Interlocuters;

e) Core frame elements (FEs): Intl1/Intl2 or Intls, Medium, Topic;

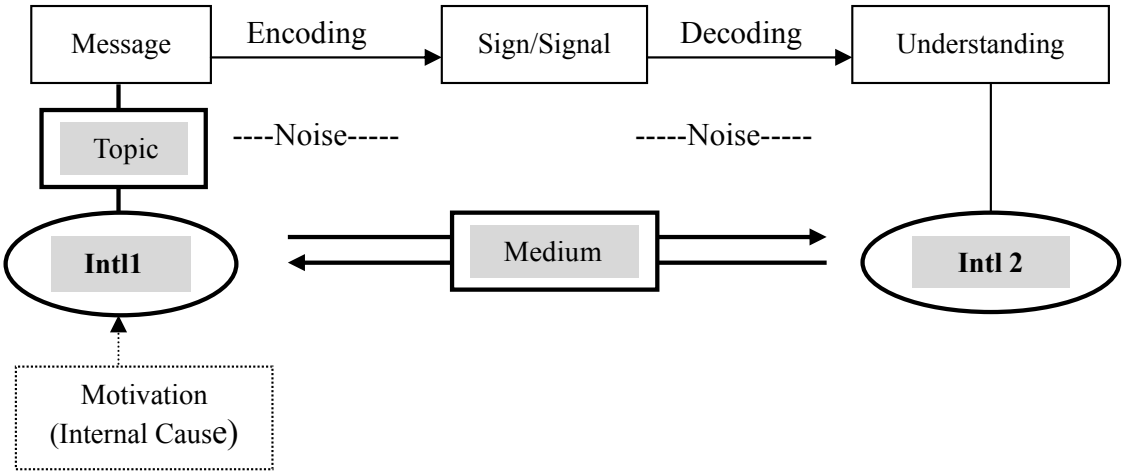
---

[2] A preliminary model of the Mandarin VerbNet (http://140.114.75.18/verbnet/webform1.aspx/) has been constructed by researchers from National Chiao Tung University and National Tsing Hua University.

[3] For details, please see [Liu *et al*. 2004], where the communication domain is divided into 14 frames and provides conceptual motivations for the cognitive bases of individual frames.

[4] The lemmas discussed in this paper are high-frequency words of the conversation frame used in Taiwan.

f) Basic syntactic patterns: 他們[Intls]面對面[Medium]談政治[Topic].
我[Intl1]和/與/跟他[Intl2]談政治[Topic].



By specifying its cognitive schema, definitional description, semantic profile, a distinct set of frame elements, and the basic syntactic patterns, the conversation frame can be uniquely defined and proved to be well-motivated in relation to other communication frames. However, given the diverse range of lemmas included in the frame, there is a fundamental question to be answered: within the conversation frame, are there semantic subtypes that need to be captured? This paper aims to show that corpus-based analyses of verb behavior render clear evidence for further distinguishing 'subframes'[5].

3. Motivations for Distinguishing Subframes

As mentioned above, verbs of conversation share a set of core frame elements: Interlocutor 1, Interlocutor 2 (or combined as Interlocutors), Topic, and Medium. However, a wide range of verbs are found in the conversation frame, such as *tan/tanlun* 談/談論 'converse, talk', 討論/商量 'discuss', *xietiao/goutong* 協調/溝通 'negotiate', *chao/chaojia* 吵/吵架 'quarrel', and *lia/liaotian* 聊/聊天 'chat'. One inevitably wonders how these verbs differ from each other. Although sharing the same basic pattern, these lemmas differ obviously in terms of manner, register, and purpose. Consider (2):

[5] Although Fillmore *et al.* [2005] put the label *'subframe'* under the description of frames in the new version of FrameNet (http://www.icsi.berkeley.edu/~framenet), they did not explicitly discuss how to distinguish *subframes*, that is, what the linguistic motivations for distinguishing and defining *subframes* are.

(2) 他們在 談/討論/溝通/吵/聊 人生的意義。

The five verbs in (2) may share the same topic, but encode distinct types of conversational events. We would like to know: what exactly are the distinctions in their lexical meanings and what grammatical correlations can be found to such lexical semantic distinctions? In the following, a detailed discussion regarding the coding of their core frame elements (FEs) and the various collocational associations will be given to capture the grammatical motivations for distinguishing subframes.

3.1 Realization and Profiling of Core Frame Elements

As mentioned before, a frame is distinguished according to syntactically expressed core frame elements (FEs). Subgroups of a frame may be further distinguished based on frequent foregrounding or backgrounding of certain frame elements observed in the corpus data. Among the core frame elements of the conversation frame, Intl1/Intl2, Intls, Medium, and Topic, the first two are commonly shared by all conversation verbs, but Medium is most frequently found with the 'converse' verbs, *tan/tanlun* 談/談論 (4.3%) but not common with other verbs (less than 1.2%):

(3) 大家[Interlocutors]面對面[Medium_Means]談[Conversation-Converse]，可直接爭取訂單。

Although Topic is a default FE for conversation events, only the 'discuss' verbs occur most frequently with a topic (73%) and allow a Topic to be preposed, as in (4):

(4) 宣傳的事[Topic]你們[Interlocutors]再討論[Conversation-Discuss]。

In addition, Topic tends to be absent in a chatting event. In the Sinica Corpus, Topic is simply not found with the verbs *liaotian* 聊天 and *xianliao* 閒聊 'chat' (0%). The suppression of Topic indicates that Topic may not be important in 'chat' events and, thus, tends to be backgrounded:

(5) 旅行回來…毛空去看望他，兩人[Interlocutors]就閒聊[Conversation-Chat]起來。

Another interesting observation regarding the coding of Topic is that it may overlap with the marking of Cause; the 'quarrel' verbs tend to take a Topic-Cause as one important role:

(6) 他們 爲/因爲 錢/感情[Topic_Cause]吵架[Conversation-Quarrel]。

The overt marking of Topic-Cause with the marker *wei/yinwei* 'because of' (6% with *chaojia/zhenglun* 吵架/爭論) seems to indicate that the 'quarrel' events, coding a highly *marked* manner of communication, require an explanation for their occurrence. The overt Cause in the above examples also fulfills the role of Topic, since what is being argued about has to do with the cause of the argument.

Finally, a unique pattern is observed regarding the coding of interlocutors: with the 'negotiate' verbs, *xietiao/goutong* 協調/溝通, there may be a third participant, Interlocutor 3, since a negotiating event is often conducted between Interlocutor 2 and Interlocutor 3 with a Mediator, Interlocutor 1. Syntactically, Intl2 and Intl3 may be foregrounded as the Direct Object, as exemplified below:

(7) 中東戰爭爆發，省府[Interlocutor_1]已協調[Conversation-Negotiate]台汽[Interlocutor_2]和台鐵[Interlocutor_3]，完成預期中東供油減少時的各項因應措施。

The additional participant, Interlocutor 3, is only found in the 'negotiate' events (3.1%) but completely absent with other verbs. Although most verbs only take the Topic as the direct object, the 'negotiate' verbs may encode Interlocutor 2 (or with Interlocutor 3) as the direct object without adding the associative marker *han/yu/gen* 和/與/跟 'and', as further exemplified below:

(8) a. [執行秘書]$_{Intl1}$ 已協調 [相關單位]$_{Intl2}$，　　b. 由[他]$_{Intl1}$ 負責溝通 [校方]$_{Intl2}$。

This observation suggests that with the 'negotiate' verbs, the co-participant, i.e., Interlocutor2 (or/and Interlocutor 3) may be viewed as the undergoer or the affected target of the event, which is a unique pattern that sets the 'negotiate' verbs apart from other conversation verbs.

The distribution of core frame elements across different conversation verb types is summarized in the following table:

(9) Distribution of Core Frame Elements among Conversation Verbs:

FEs / Verbs	Intls	Intl1/Intl2	Intl3	Intl2/Intl3 as DO	Topic	Topic as Cause	Medium
Converse 談/談論	40% (441/1103)	14 % (156/1103)	0% (0/1103)	0% (0/1103)	17.3% (191/1103)	0% (0/1103)	4.3% (47/1103)
Discuss 討論/商量	62% (129/208)	29% (60/208)	0% (0/208)	0% (0/208)	73% (152/208)	0% (0/208)	0% (0/208)
Negotiate 協調/溝通	15 % (74/485)	29.5% (143/485)	3.1% (15/485)	7% (33/485)	16.7% (81/485)	0% (0/485)	0.9% (4/485)
Quarrel 吵架/爭論	48% (101/211)	11.4% (24/211)	0% (0/211)	0% (0/211)	10.4% (22/211)	6% (13/211)	0% (0/211)
Chat 聊天/閒聊	38% (66/174)	46 % (81/174)	0% (0/174)	0% (0/174)	0% (0/174)	0% (0/174)	1.1% (2/174)

It is clear from the table that each verb type displays a distinct pattern in coding the core frame elements through either the foregrounding or backgrounding of certain participant roles.

3.2 Lexical Collocations and Grammatical Functions

Besides the realization of core frame elements, conversation verbs also differ in terms of lexical collocation and grammatical function. Looking closely at their collocational associations in the Sinica Corpus, we found that there are asymmetrical distributions in four respects: 1) V+V pattern: some verbs may occur with a preceding light verb, such as *jinxing* 進行 'proceed' or *dacheng* 達成 'achieve'; 2) metonymic subject: the subject of some, but not all, verbs of conversation may be inanimate entities taking the role of Interlocutor based on the principle of metonymy; 3) V+ Complement pattern: some verbs take a postverbal complement or adverbial adjunct denoting 'result evaluation', such as *chenggong* 成功 'successful' or *shibai* 失敗 'failing'; 4) in terms of the distribution of grammatical functions, conversation verbs exhibit different frequencies of nominalization. Based on the above four criteria, verbs of conversation can be further divided into 5 subgroups with a corresponding set of unique behaviors. We will address the syntactic-semantic interdependencies manifested by each association pattern in the following sections.

3.2.1 V+V Pattern: with Light Verbs *Jinxing* 進行 or *Dacheng* 達成

The use of the light verb *jinxing* 進行 'proceed' entails a formal register and encodes a procedural process or *atelic* event, according to Huang *et al*. [1995][6]. It tends to occur with an activity verb that is compatible with the formal register and involves a durative process, as shown in (10):

[6] Huang *et al.* [1995] claimed that *jinxing* 'to proceed' takes a processed type argument, which denotes the process of an event. In other words, the verb *jinxing* implies a non-punctual and non-telic event.

(10) a. 研究小組[Intls]針對口訪前置作業階段之相關事宜進行討論[Conversation-Discuss]。
b. *進行 談論/吵架/聊天。

Below is the distributional tendency of *jingxing* in the Sinica Corpus:

(11) Distribution with *jingxing* 進行 'proceed'

V1 \ V2	*taolun* 討論	*goutong* 溝通	Other Verb Types *tanlun/chaojia/liaotian* 談論/吵架/聊天
Jingxing 進行	4% (3/83)	11% (25/231)	0%

Another verb, *dacheng* 達成 'achieve', is also found with some conversation verbs; it requires a formal register but encodes a *telic* event. Denoting goal-orientation, *dacheng* 達成 is compatible with activity verbs entailing a semantic endpoint with an incremental theme, and it occurs mostly with the nominalized forms of 'negotiate' verbs, such as *goutong/xietiao/xieyi* 溝通/協調/協議, as in (12-13):

(12) a. 雙方最後達成協議[conversation-negotiate]。
b. *達成 談/談論/討論/吵架/爭論/聊天/閒聊。

Additional examples are found with *goutong* 溝通 and *xietiao* 協調 at the Kimo website:

(13) a. 要耗費相當長的時間來達成溝通[conversation-negotiate] (yahoo 2005/07/01)。
b. 各方股東達成協調[conversation-negotiate] (yahoo 2005/06/13)。

(14) Distribution with *dacheng* 達成

V1 \ V2	*xieyi* 協議	Other Verb Types *tanlun/taolun/chaojia/liaotian* 談論/討論/吵架/聊天
Dacheng 達成	23% (36/154)	0%

The co-occurrence with the preceding verb *jinxing* 進行 'proceed' or *dacheng* 達成 'achieve' serves to distinguish a conversation event in terms of its pragmatic mode (formal vs. informal) and event type (telic vs. atelic).

3.2.2 Use of an Inanimate Subject

Interlocutors in conversation events are, by default, human participants. However, unlike the 'chat' verbs, other verbs may all take inanimate subjects (place or institute names) as Interlocutors via metonymic extension from institute/building to human organization:

(15) a. 台北和北京 [Intl1 and Intl2] 談/談論/討論/溝通 了很久。
b. 台北和北京 *聊天/*閒聊 了很久。

Metonymy tends to be associated with verbs that are formal in register and also require an official or non-personal topic (e.g., public affairs). The Sinica Corpus shows that verbs ecoding informal events, such as the 'chat' verbs, seldom occur with metonymic non-human subjects:

(16) Distribution with inanimate subjects

V / Subj. type	*tanlun* 談論	*taolun* 討論	*goutong* 溝通	Other Verbs *Liaotian/xianliao* 聊天/閒聊
Inanimate Subject	3% (8/262)	6% (5/83)	9% (21/231)	0%

3.2.3 Postverbal Complement with Result Evaluation

Among the conversation verbs, only the 'negotiate' verbs (e.g., *xietiao* 協調 and *goutong* 溝通) may collocate with result-evaluating complements, such as *chenggong* 成功 'successfully' and *shibai* 失敗 'failingly', as shown below with examples and percentage rates for the Sinica Corpus:

(17) a. 國防部和兩廳院已初步協調成功。
b. 在協調失敗後，水公司終於昨天宣布放棄。

(18) Distribution of result evaluating complements

V / Result Comp.	*xietiao* 協調	Other Verb Types *tanlun/taolun/chaojia/liaotian* 談論/討論/吵架/聊天
chenggong/shibai 成功/失敗	4 % (3/66)	0%

Additional data obtained from the internet show that the verb *goutong* 溝通 behaves in the same way:

(19) a. 黨團也隨即召開記者會，強調有信心溝通成功 (yahoo 2005/03/20)。
b. 雙方溝通失敗，因此反目成仇 (yahoo 2005/03/17)。

The co-occurrence with effect-evaluating complements indicates that the two-way communicative events with *xietiao* 'negotiate' or *goutong* 'communicate' involve a solution-seeking process, which is semantically bounded and may be *evaluated* as to whether the solution has been achieved.

3.2.4 Frequency of Nominalization

With regard to grammatical functions, some groups of verbs tend to be nominalized more frequently than the others. Comparing high-frequency verbs and their distributions over grammatical functions, we see clear skewing in nominal uses based on the Sinica Corpus:

(20) Distribution of predicate vs. nominal uses

V / Function	*tanlun* 談論	*taolun* 討論	*goutong* 溝通	*chaojia* 吵架	*liaotian* 聊天
Predicate	77% (202/262)	**52%** (83/161)	**55%** (231/419)	76% (123/162)	94% (134/142)
Nominalization	23% (60/262)	**48%** (78/161)	**45%** (188/419)	24% (39/162)	6% (8/142)

Nominalization serves to mark activities as event nominals that may be referred to as quantifiable entities and nominalization is also highly correlated with written texts that are formal in register.

The following table summaries the collocational patterns discussed above:

(21) Distribution of collocational variations

Collocational Variation / Subtype	進行+V	達成+V	V+Result [成功/失敗]	[+Nom]
Converse 談/談論	No	No	No	Low
Discuss 討論/商量	**Yes**	No	No	High (with 討論)
Negotiate 協調/溝通	**Yes**	**Yes**	**Yes**	High
Quarrel 吵架/爭論	No	No	No	Mid
Chat 聊天/閒聊	No	No	No	Low

The syntactic patterns of coding frame elements and collocational variations discussed above provide crucial evidence and support for categorizing conversation verbs into syntactically motivated subgroups, termed 'subframes'. The division of subframes serves to provide the semantic ground for further distinguishing near-synonyms, as one important layer in the hierarchical structure of frame-based semantic classification.

4. Subframes: A Semantic Anchor for Near-synonyms

4.1 Defining Conversation Subframes

As mentioned above, the asymmetrical distributions of conversation verbs over different collocational associations clearly suggest that verbs can be further divided into subtypes. These subtypes may be viewed as different *subframes* as they further characterize the semantic distinctions within a *frame*. Each subframe displays unique patterns of its basic structure (Basic Patterns) and collocational skewings (Collocational Associations) that manifest its unique semantic properties.

The table shown below recaptures all the crucial syntactic distinctions for the 5 subframes:

(22) Collocational patterns associated with conversation subframes

CP / Subframe	Intl2 as DO	Inanimate Subj.	Topic as Cause	Absence of Topic	進行+V	達成+V	V-R 成功/失敗	[+Nom]
1.Converse 談/談論	No	**Yes**	No	No	No	No	No	Low
2.Discuss 討論/商量	No	**Yes**	No	No	**Yes**	No	No	High
3.Negotiate 協調/溝通	**Yes**	**Yes**	No	No	**Yes**	**Yes**	**Yes**	High
4. Quarrel 吵架/爭論	No	**Yes**	**Yes**	No	No	No	No	Mid
5. Chat 聊天/閒聊	No	**No**	No	**Yes**	No	No	No	Low

Based on the above syntactic behaviors, we attempt to differentiate and define five subframes within the conversation frame, each with a clear definition, a set of foregrounded or backgrounded FEs, a unique set of semantic features, and representative lemmas.

(23) Conversation Subframes: definition, FEs, semantic attributes and lemmas

a) Subframe 1_Converse

Definition: Interlocutors exchange ideas on a given Topic, via a Medium (including a means or language) -- the most generic subtype of conversation;

FEs: (the default set) Interlocutors, Topic, Medium;

Semantic attributes: [event type: process], [register: unmarked], [manner: unmarked], [purpose: underspecified];

Lemmas: *tan/tanlun/jiaotan/jiaoliu* 談/談論/交談/交流, etc.

b) Subframe 2_Discuss

Definition: Interlocutors exchange opinions on a given Topic to establish pros and cons in a serious manner;

Foregrounded FE: Topic;

Semantic attributes: [event type: procedural process], [register: formal], [manner: serious], [purpose: establish pros and cons];

Lemmas: *taolun/shangliang/*

shangtao/shangtan 討論/商量/商討/商談, etc.

c) Subframe 3_Negotiate

Definition: Interlocutors confer and negotiate on a certain Topic-Purpose (Topic may overlap with Purpose) in order to reach a consensus or settlement;

Foregrounded FE: Topic-Purpose;

Semantic attributes: [event type: procedural process], [register: formal], [manner: serious], [purpose: achieve consensus];

Lemmas: *xietiao/xieyi/goutong/xieshang* 協調/協議/溝通/協商, etc.

d) Subframe 4_Qarrel

Definition: Interlocutors dispute actively or vehemently exchange different opinions on a certain Topic-Cause in a heated manner;

Foregrounded FE: Topic-Cause;

Semantic attributes: [event type: process], [register: underspecified], [manner: marked_disagreeable], [purpose: pertaining to the Cause];

Lemmas: *chaojia/zhenglun/zhengcha/zhengzhi* 吵架/爭論/爭吵/爭執, etc.

e) Subframe 5_Chat

Definition: Interlocutors engage in an informal verbal activity in a rather casual and entertaining manner;

Backgrounded FE: Topic;

Semantic attributes: [event type: process], [register: informal], [manner: casual], [purpose: underspecified];

Lemmas: *xianliao/liaotain/xiantan/tantian* 閒聊/聊天/閒談/談天, etc.

The five subframes as defined above may serve as semantic anchors for further exploring near-synonym sets, such as 交談 *jiaotan* vs. 交流 *jiaoliu* 'converse, exchange'; *taolun* 討論 vs. *shangliang* 商量 'discuss, talk about'; or *xietia*o 協調 vs. *goutong* 溝通 'negotiate', etc. It is at the level of subframes that we may find the most relevant information for fine-tuning lexical distinctions among near-synonyms (For further discussion, see [Liu *et al.* 2004]).

4.2 Semantic Inheritance

The layered structure of frame-based classification entails semantic inheritance from top to bottom in a hierarchical structure. At the subframe level, multiple inheritances may happen; i.e., a given subframe may inherit features from two or more different frames. For example, as mentioned above, verbs belonging to the 'Quarrel'-subframe highlight Topic-Cause for the marked Manner. This is a result of multiple inheritance. The 'quarrel' verbs inherit the element Topic from the Conversation frame, and the element Cause from the Hostile_encounter Frame, where a Cause is normally specified.

5. Conclusion: Frame-based Hierarchy of Verbal Information

To fully represent the semantic relations among verbs, a multi-layered classificational structure is needed, which helps to manifest different levels of semantic generalization. In this paper, the syntactically well-motivated level of subframes is proposed, rendering a 5-layered hierarchical structure of lexical semantic representation: *Domain* > *Frame* > *Subframe* > *Near-synonym sets* > *Lemma*. This model is also adopted in building the lexical network of Mandarin VerbNet [Liu *et al.* 2004]. With the proposal of subframes within the theoretical constructs of frame semantics, verb meanings may be defined with finer semantic distinctions that are syntactically motivated. The next task is to show that further fine-grained lexical distinctions are needed to differentiate near-synonyms within each subframe. The postulation of subframes is a necessary refinement of the frame-based approach to lexical semantics. With its detailed lexical information, subframes provide the most relevant semantic anchor for further disambiguating near-synonyms as well as individual lemmas.

References

Baker, C. F., C. J. Fillmore, and B. Cronin, "The Structure of the Framenet Database," *International Journal of Lexicography*, 16(3), 2003, pp. 281-296.

Fillmore, C. J., and B. T. S. Atkins, "Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors," *Frames, Fields, and Contrasts*, ed. by Lehrer, A. and E. Kittay, Hillsdale: Lawrence Erlbaum Associates, 1992, pp. 75-102.

Fillmore, C. J. *et al*., FrameNet, version 2.0, http://framenet.icsi.berkeley.edu/, 2005.

Huang, C. R., M. Yeh, and L.P. Chang, "A Corpus-based Study of Nominalization and Verbal Semantics: Two Light Verbs in Mandarin Chinese," In *Proceedings of the Sixth North American Conference on Chinese Linguistics*, 1995, Los Angeles: GSIL, USC.

Liu, M. C., T. Y. Chiang, and C. Chang, "The construction of Mandarin VerbNet: A frame-based corpus approach". Paper presented at the Symposium of Current trends and development of Chinese Lexical Semantics, 2004, Beijing University, Beijing, China.

Liu, M. C., and Y. C. Wu, "Beyond Frame Semantics: Insight from Mandarin Verbs of Communication," Paper presented at the 4th Chinese Lexical Semantics Workshop, 2003, City University of Hong Kong.

Liu, M. C., and C. R. Huang et al., "Sense and Sense-ability Series V: Mandarin VerbNet (I)--General Introduction and Communication Frames," Technical Report No. 04-01, Center for Knowledge-based Corpus Studies, Department of Foreign Languages and Literatures, National Chiao Tung University, Hsinchu, and Corpus Linguistics Research Group, Institute of Linguistics Academia Sinica, Taipei, 2004.

Miller, A., R. Beckwidth, C. Fellbaum, D. Gross, and K.J. Miller, "Introduction to WordNet: An on-line Lexical Database," *International Journal of Lexicography*, 3(4), 1990, pp. 235-244.

Niles, I., and A. Pease, "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology," In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering* (IKE'03), Las Vegas, Nevada, June 23-26, 2003.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 445-458

Feature Representations and Logical Compatibility between Temporal Adverbs and Aspects

Shih-Min Li*, Su-Chu Lin* and Keh-Jiann Chen*

Abstract

In this paper, we propose clear-cut definitions of distinct temporal adverbs and provide descriptive features for each class of temporal adverbs. By measuring time points in temporal axis, we revise and reclassify the temporal adverbs listed in [Lu and Ma 1999] into four classes of semantic roles, namely, time, frequency, duration, and time manner. The descriptive features enable us to distinguish temporal relations and predict logical compatibility between temporal adverbs and aspects.

Keywords: Temporal Adverbs, Aspects, Temporal Schema, Speaking Time, Reference Time, Event Time

1. Introduction

There are about 130 temporal adverbs in Mandarin Chinese. Lu and Ma [1999] classified temporal adverbs into two groups, speaking-time related adverbs (abbr: ST-related adverbs, 定時時間副詞) and reference-time related adverbs (abbr: RT-related adverbs, 不定時時間副詞). The ST-related adverbs consist of 27 temporal adverbs, which are subdivided into three subclasses. In the class of RT-related adverbs, 104 temporal adverbs are listed and subdivided into 18 subclasses. In Lu and Ma's definition, ST-related adverbs modify events that happen at specific time points; RT-related adverbs modify events that have happened in the past or will happen in the future. However, the subdivisions are vague, and the definition is incomprehensible. For example, *cengjing* 曾經, *ceng* 曾, *yeyi* 業已 and *yejing* 業經 are grouped into two different subclasses of ST-related adverbs. The former two are grouped into the same subclass, which includes actions or situations that happened or existed before the speaking time. The later two are grouped into the same subclass, which includes actions or situations have been completed or have occurred. In fact, it is difficult to differentiate between actions or situations that have 'happened' or have 'completed', especially when the situation type is an achievement situation with SHORTLY-PRECEDE(t_1,t_2) or NEARLY-EQUAL

* CKIP, Institute of Information Science, Academia Sinica, Taipei
E-mail: {shihmin, jess}@hp.iis.sinica.edu.tw; kchen@iis.sinica.edu.tw

[Received April 12, 2005; Revised June 13, 2005; Accepted August 15, 2005]

(t_1,t_2).[1] Moreover, temporal adverbs may not have the same syntactic behaviour even though they are classified into the same subclass. For instance, the ST-related adverbs *cong* 從, *conglai* 從來, *zhijin* 至今, *xianglai* 向來, *sulai* 素來, *lilai* 歷來, *su* 素, and *yixiang* 一向 are grouped into the same subclass. When they co-occur with the aspect markers *le* 了, *guo* 過, and *zhe* 著, however, *cong*, *conglai*, and *zhijin* are incompatible with *le* and *zhe*; in addition, *xianglai*, *sulai*, *lilai*, *su*, and *yixiang* are incompatible with *le* and *guo*. The reason for the difference in the compatibility of temporal adverbs with aspects will be discussed in the following.

In this paper, based on the application of Smith's proposal for temporal location, we suggest more clear-cut definitions and provide descriptive features for each subclass of temporal adverbs. The descriptive features help us to define temporal relations and to predict the compatibility between temporal adverbs and aspect markers.

2. Literature Review and Methodology

To make clear-cut differentiations, time points in the temporal axis and the notion of temporal location proposed by [Smith 1991] will be used here to redefine the temporal relations of the temporal adverbs treated in [Lu and Ma 1999]. We will use the data in the Academia Sinica Balanced Corpus (Sinica Corpus) to analyze Mandarin Chinese temporal adverbs.

Smith [1991] discussed aspectual systems in language. She mentioned that the temporal information in a sentence is located in the time when the situation occurs. Adverbials further specify temporal location. Time is represented as a straight line stretching in both directions from Speech Time. The linear representation is presented in (1):

(1) Time Line [Smith 1991]

---------- Speech Time ----------

Past Present Future

[1] The terms SHORTLY-PRECEDE(t_1,t_2) and NEARLY-EQUAL(t_1,t_2) were mentioned by Yang and Bateman [2002]. The predicate SHORTLY-PRECEDE(t_1,t_2) indicates that time point t_1 only shortly precedes time point t_2. The predicated NEARLY-EQUAL(t_1,t_2) indicates that time point t_1 lies close to time point t_2. Applying the terms to the explanation of achievements, we can treat t_1 as Yang and Bateman's t_i, the event initial time, and as Smith's [1991] I, initial points, and regard t_2 as Yang and Bateman's t_f, the event finishing time, and as Smith's F, final points. Yang and Bateman's expression for the predicates used to express temporal relations are revised of assertions in [Allen 1984].

Following Reichenbach [1947], to temporally locate all types of sentences, three times are required: Speech Time, Reference Time, and Situation Time. Speech Time (ST) is the center of the temporal system. Reference Time (RT) is the temporal standpoint of a sentence, and, in complex sentences, it is a secondary orientation point. Situation Time (SitT) is the time of the event or state, identified as the interval [I]. To show the temporal location of these three times, Smith [1991] gives example 2a and its time line 2b. In 2b, time 2 is Reference Time, while time 3 is Situation Time:

(2) a. Mary said last Tuesday that she was leaving in 3 days.

b. …..time 2……..….time 3…..…… time 1

last Tuesday + 3 days Speech Time

Mayr said Mary leave

Mandarin Chinese does not have the grammatical category of tense, so temporal location is expressed directly by adverbials and indirectly through the use of an aspectual viewpoint. Temporal adverbials function as locating adverbials, which are classified as deictic, anaphoric, or referential, according to the type of orientation they take [Smith 1991]. Furthermore, Smith illustrates each situation type and viewpoint type with temporal schemata. Below are the temporal schemata of the Mandarin Chinese aspectual markers *le*, *guo*, and *zhe*, which are represented by symbols I, F, F+1, …. and /:[2]

(3) Temporal schema for the –*le* perfective [Smith 1991][3]

I F

/ / (RVC)

(4) The Mandarin –*guo* perfective viewpoint [Smith 1991]

I ….F F+1

/ /

[2] I and F indicate initial and final points. F+1 indicates a stage distinct from the final stage. The dots indicate internal stages and the slashes indicate the interval process.

[3] RVC is an abbreviation of Resultative Verb Complements.

(5) The –zhe viewpoint [Smith 1991]

I…..…

////$_{State}$

Klein [1994] points out five temporal features and notes TT, TU and TSit. TT, topic time, is the time span to which the speaker's claim on the occasion is confined. TU is the time of utterance, which is the time at which the utterance is made. TSit is the time of the situation, which is the time at which the situation occurs.

In addition to Smith's and Klein's terminology, we can apply the event modules in the MARVS framework proposed by Huang *et al*. [2000] to analyze the temporal relations of temporal adverbs. Event modules are the basic building blocks of the event contour. Five event modules stand alone or in combination; they are Boundary, Punctuality, Process, State, and Stage. The event module Boundary is defined as an event module that can be identified with a temporal point and must be regarded as a whole (including a complete Event). It is adopted in this paper to define the Boundary Point.

Yang and Bateman further discussed the semantic temporal relations of an aspect system and proposed principled semantic conditions for aspect combination. In their opinion, the Chinese aspect system is actually composed of both aspect morphemes (-le, -zhe, -guo4, etc.) and aspect adverbials. The Chinese aspect system has, basically, seventeen simple primary aspect forms. These simple primary aspect forms belong to perfective, imperfective, or future-existing subsystem, according to the semantic properties of individual cases. Some simple primary aspect forms can combine to form an aspect of secondary type if their temporal attributes are in harmony. The temporal relation of the combination is represented graphically by time points t_i, t_f, t_r, and t_s[4].

In this paper, we adopt the terms proposed in the research described above to clarify the temporal relation of each subclass of temporal adverbs listed in [Lu and Ma 1999]. We use the notations ST, RT, ET, BP, Start, and End to define temporal relations. Each, respectively, denotes the speaking time, reference time, event time, boundary point, start point of the event, and end point of the event. The symbols $<$, $\leqq$, and $=$, respectively, indicates priority, inclusion and overlap in temporal location. For instances, the temporal features for *le*, *guo*, and *zhe* are defined as follows in our system and they are compatible with the temporal schemata proposed in [Smith 1991].

[4] As the second footnote mentions, t_i is the event initial time, and t_f is the event finishing time. The symbols t_r and t_s are the reference time and speaking time, respectively.

le: BP≦ST, which means that the prominent boundary point of the referred event precedes the speaking time;

guo: End<ST, which means that the end point of the referred event precedes the speaking time;

zhe: ET=RT, which means that the referred event time overlaps with the speaking time.

3. Temporal Relations and Compatibility between Temporal Adverbs and Aspects

Temporal adverbs were classified by Lu and Ma into two classes, ST-related adverbs and RT-related adverbs, according to the time point of the situation referred to. ST-related adverbs establish constraints between event time (ET) and speaking time (ST), but RT-related adverbs establish constraints between ET and reference time (RT). For example, the ST-related adverb *cengjing* only denotes events or situations that happen or exist before ST. To be precise, the end points of events occur prior to ST. or example, the sentence *wo shangci cengjing jieshao guo* 我上次曾經介紹過 'I have introduced it last time' is grammatical; however, **wo xiaci/mingtien cengjing jieshao guo* 我 *下次/明天 曾經介紹過 '*I have introduced it next time/tomorrow' is ungrammatical. The RT-related adverb *yijing* 已經 is used to refer to events that happen or exist before either a certain specific time, an event, or ST. For instance, the sentences *shang libaiwu wo yijing chuguo le* 上禮拜五我已經出國了 'I had gone abroad last Friday' and *xia libaiwu wo yijing chuguo le* 下禮拜五我已經出國了 'I have gone abroad next Friday' are both grammatical even though *xia libaiwu* is a future specific time.

Lu and Ma divided ST-related adverbs into three subclasses and RT-related adverbs into eighteen subclasses. Three subclasses of ST-related adverbs were subdivided into nine sub-subclasses. Eighteen subclasses of RT-related adverbs were subdivided into thirty sub-subclasses. Checking Sinica Corpus, we find that some of the RT-related adverbs in Lu and Ma's classification behave like ST-related adverbs. Some temporal adverbs denote the order of events and establish constraints only between two ETs. Some temporal adverbs refer to habitual situations, which are not related to temporal relations. Therefore, our classification of temporal adverbs is somewhat different from Lu and Ma's. In this paper, temporal adverbs are divided into four classes according to temporal relations. More detailed subclasses will be introduced in the following sections.

3.1 Temporal Relations and Compatibility between ST-related Adverbs and Aspects

In our classification, ST-related adverbs are classified into eleven subclasses. Each subclass is characterized and represented by its temporal features and is assigned one semantic type: Time, Duration, or Frequency. Table 1 shows the temporal features of each subclass.

Table 1. Temporal relation and semantic role of each subclass of ST-related adverbs

Temporal Adverbs	*Temporal Relation*	*Semantic Role*
*cengjing*曾經, *ceng*曾	End<ST	time
*yeyi*業已, *yejing*業經	End≦ST	time
*zhongyu*終於, *bijing*畢竟, *daodi*到底	End≦ST	time
*xianxing*先行	Start<ST	time
*zaori*早日, *jizao*及早, *chenzao*趁早	ST<Start	time
*zhongjiang*終將, *zhongjiu*終久, *zhonggui*終歸, *zonggu*總歸, *bijiang*必將, *chizao*遲早, *zaowan*早晚	ST<BP	time
*zhongjiu*終究	End≦ST	time
*cong*從, *conglai*從來, *zhijin*至今, *xianglai*向來, *sulai*素來, *lilai*歷來, *su*素, *yixiang*一向	Start<ST	duration
*zhan*暫, *zhanqie*暫且, *guqie*姑且, *quanqie*權且, *qie*且	ET=ST	duration
*yongyuan*永遠, *yong*永, *shizhong*始終, *zhi*直	ET=ST	duration
*yidu*一度	BP<ST	frequency

For instance, *cengjing* and *ceng* are grouped into the same subclass of ST-related adverbs. Since they denote actions or situations that happened or existed before ST, the end points of the situations precede ST. Thus, the temporal relation of *cengjing* and *ceng* is defined as End<ST.

Temporal features not only clearly indicate the relations among these ST-related adverbs but also predict and verify the correctness of the co-occurrences of temporal adverbs and aspectual markers, including *le*, *guo*, and *zhe*, by unifying their temporal features. The temporal features of *le*, *guo*, and *zhe* are BP≦ST, End<ST, and ET=RT, respectively. If the corpus data shows that a given sentence is grammatical after the combination step, its combination of temporal relation will be in harmony; otherwise, there may be something wrong with the temporal relations or some other essential factors may cause incompatibility. On the other hand, we assume that by detecting harmony in the combination of temporal relations between temporal adverbs and aspects, we can predicate whether a given temporal adverb is compatible with *le*, *guo*, and *zhe*. Here, we will use an example with *cengjing* and *ceng*. The corpus data in the Sinica Corpus show the following combinations of *cengjing* with *le*, *guo*, and *zhe*:

(6) *ceng jing* co-occurs with *le*, *guo*, and *zhe*

a. 我們的通信曾經給了她很大的快樂

women de tongxin cengjing gei ta henda de kuaile

She had been very happy while we wrote letters each other.

b. 我們曾經提過
women cengjing ti guo
we had mentioned before

c. 曾經提著小包袱
cengjing ti zhe xiao baofu
ever carried little backpack

The combination of *cengjing* with *le*, *guo*, and *zhe* is grammatical, so we conclude that the temporal relation of *cengjing* is correct and assign the semantic role of time to the subclass of *cengjing* and *ceng*. On the other hand, from the observation of the temporal relations of *le*, *guo*, *zhe*, and *cengjing*, the temporal relation of *cengjing* is compatible with that of *le*, *guo*, and *zhe*, respectively. Thus, before checking the data in the Sinica Corpus, we can predict the that co-occurrence of *cengjing* with *le*, *guo*, and *zhe*, respectively is grammatical. The grammatical sentences in (6) prove that our hypothesis and prediction are correct.

Temporal adverbs classified into the same subclass with the same temporal relation may not have the same syntactic behaviours. For instance, *cong*, *conglai*, *zhijin*, *xianglai*, *sulai*, *lilai*, *su*, and *yixiang* are grouped into the same subclass and assigned the semantic role of duration. The temporal relation of these temporal adverbs is defined as Start<ST. We find, however, that *cong*, *conglai*, and *zhijin* are incompatible with *le* and *zhe*; in addition, *xianglai*, *sulai*, *lilai*, *su*, and *yixiang* are incompatible with *le* and *guo*. Although this subclass of ST-related adverbs has the above temporal relation, their compatibility with aspectual markers is somewhat different. The Sinica Corpus sentences shown below reveal that *cong*, *conglai*, and *zhijin* are usually found in negative sentences. The feature [+NEG] brings about the different syntactic behaviours:

(7) 好像這件事從沒發生過
hoaxing zhejian shicong mei fasheng guo
It seems this thing hasn't happened before.

(8) 從來不會逃避
conglai buhui taobi
never evade

(9) 至今尚未出現在他生命中
zhijin shangwei chuxian zai ta shengming zhong
haven't appeared in his life

Temporal adverbs usually co-occur with aspects. The compatibility of temporal adverbs with aspect markers can verify the correctness of temporal relations. Therefore, we focus on not only temporal feature representations of temporal adverbs and aspects but also the co-relations between temporal adverbs and aspects. Table 2 is the reversion of Table 1 and it shows the compatibility of temporal adverbs with aspects. The asterisks indicate a lack of grammaticality.

Table 2. Temporal relations and compatibility between ST-related adverbs and aspects

Semantic Role	*Temporal Adverbs*	*Temporal Relation*	*Compatibility with le, guo, and zhe*
time	*cengjing*曾經, *ceng*曾	End<ST	*le* *guo* *zhe*
time	*yeyi*業已, *yejing*業經	End≦ST	*le* *guo* *zhe*
time	*zhongyu*終於, *bijing*畢竟, *daodi*到底	End≦ST	*le* *guo* *zhe*
time	*xianxing*先行	Start<ST	*le* *guo* **zhe*
time	*zaori*早日, *jizao*及早, *chenzao*趁早	ST<Start	**le* **guo* **zhe*
time	*zhongjiang*終將, *zhongjiu*終久, *zhonggui* 終歸, *zonggu* 總歸, *bijiang* 必將, *chizao* 遲早, *zaowan*早晚	ST<BP	**le* **guo* **zhe*
time	*zhongjiu* 終究	End≦ST	*le* *guo* *zhe*
duration	*cong* 從 (+NEG), *conglai* 從來 (+NEG), *zhijin* 至今 (+NEG), *xianglai*向來, *sulai*素來, *lilai*歷來, *su*素, *yixiang*一向	Start<ST	+NEG: **le* *guo* **zhe* others: **le* **guo* *zhe*
duration	*zhan*暫, *zhanqie*暫且, *guqie*姑且, *quanqie*權且, *qie*且	ET=ST	*le* **guo* *zhe*
duration	*yongyuan* 永遠, *yong* 永, *shizhong*始終, *zhi*直	ET=ST	*le* **guo* *zhe*
frequency	*yidu*一度	BP<ST	*le* *guo* **zhe*

3.2 Temporal Relations and Compatibility between RT-related Adverbs and Aspects

Having adopted the corpus-based approach and analyzed the time points along the temporal axis, we show in Table 3 the sub-classification of RT-related adverbs.

Table 3. Temporal relations and Compatibility between RT-related Adverbs and Aspects

Semantic Role	*Temporal Adverbs*	*Temporal Relation*	*Compatibility with le, guo, and zhe*
time	*yijing*已經, *yi*已, *zaoyi*早已, *zaojiu*早就, *dou*都	BP<RT	*le guo zhe*
time	*gang*剛, *ganggang*剛剛, *cai*才	BP≦RT	*le guo zhe*
time	*xian*先, *yuxian*預先, *shixian*事先	BP≦RT	*le guo zhe*
time	*jijiang*即將, *jiangyao*將要, *jiuyao*就要, *kuai*快, *xingjiang*行將	ST<RT<Start	**le* **guo* **zhe*
time	*like*立刻, *liji*立即, *jike*即刻, *mashang*馬上, *ganjin*趕緊, *gankuai*趕快, *ganmang*趕忙, *lianmang*連忙, *jimang*急忙, *jiu*就, *bian*便, *dangji*當即	RT≦Start	*le* **guo* *zhe*
time	*dunshi*頓時, *dengshi*登時, *shashi*霎時, *lishi*立時, *yixia(zi)*一下(子)	RT≦Start	*le* **guo* *zhe*
time	*turan(jian)*突然(間), *zouran*驟然, *mengran(jian)*猛然(間), *mengdi*猛地, *modi*驀地	RT≦Start	*le* **guo* *zhe*
duration	*zheng* 正, *zhengzai*正在, *zai*在	ET=RT	**le* **guo* *zhe*
duration	*hai*還, *haishi*還是, *reng*仍, *rengran*仍然, *rengjiu*仍舊, *yiran*依然, *yijiu*依舊, *zhaojiu*照舊, *zhaoyang*照樣, *zhaochang*照常	BP≦RT	*le guo zhe*

The RT-related adverb *yijing* represents actions or situations that finish, happen or exist before a certain specific time, another action, or a situation. The BP of a situation denoted by *yijing* must precede RT; therefore, its temporal relation is BP<RT. The Sinica Corpus data point *yijing* is compatible with aspect markers.

(10) 一定已經打破了紀錄

yiding yijing dapo le jilu

have certainly broken the record

(11) 已經吃過飯

yijing chiguo fan

have had a meal

(12) 蛇已經咬著他了

she yijing yao zhe ta le

A snake has bitten him.

We find that the combinations of temporal relations of some classes and those of aspect markers contradict the corpus data. The temporal adverbs classified in subclasses with *like*, *dunshi*, and *turan(jian)* indicate situations or events that happen immediately after ST or another situation. Their temporal relation is defined as RT≦Start. The corpus data show that these RT-related adverbs are incompatible with *guo*; however, their temporal relation RT≦Start is compatible with *guo*. Sentence (13) shows the temporal relation of *like* between RT and ST:

(13) 他一聽到消息，立刻回了電話

ta yi tingdao xiaoxi like hui le dianhua

He immediately called back when he got the news.

In sentence (13), RT is *ta yi tingdao xiaoxi*, and ST is the present. The temporal relation of *like* in sentence 13 is defined as RT≦Start<ST, which is compatible with the aspect markers *le*, *guo*, and *zhe*. However, the corpus data show that *guo* is not compatible with *like*. The temporal relation RT≦Start focuses on the start points of events, while the aspect marker *guo* focuses on subsequent stages following the end points of events. On one hand, an event can be viewed as a whole or as consisting of different parts. When it is viewed as a whole, the temporal relation is defined by BP. When it is viewed as consisting of different parts, its temporal relation is defined by either the start point or end point. Since Start is the focus of the temporal relation RT≦Start, the aspect marker *guo* indicating End is incompatible with it. Thus, RT-related adverbs that are assigned to the same subclasses as *like*, *dunshi*, and *turan(jian)* are actually incompatible with *guo* even though their temporal relations seems to be compatible. On the other hadn, the main feature of *–guo* is that it indicates a discontinuity with the present or other Reference Time [Smith 1991]. The subsequent stage cannot be considered part of the event and is separate from the situation, so the temporal relation of *guo* is not compatible with the temporal relation RT≦Start. Thus, *guo* is incompatible with the

temporal adverbs classified into subclasses with *like*, *dunshi*, and *turan(jian)*.

3.3 Temporal Relations and Compatibility between ET-related Adverbs and Aspects

In Lu and Ma's classification, some temporal adverbs are neither ST-related adverbs nor RT-related ones. These temporal adverbs establish constraints between two ETs, and denote the order, sequence, or succession of two situations. Thus, we classify these temporal adverbs into another class: ET-related adverbs. Table 4 lists the ET-related adverbs and their feature representations and co-relations with aspects.

Table 4. Temporal relations and compatibility between ET-related adverbs and aspects

Semantic Role	*Temporal Adverbs*	*Temporal Relation*	*Compatibility with le, guo, and zhe*
time	*xianhou*先後,xiang*jyi*相繼	$ET_1 \leqq ET_2$	*le guo zhe*
time	*tongshi*同時	$ET_1 = ET_2$	*le guo zhe*
duration	*ranhou*然後, *erhou*而後, *suihou*隨後, *suiji*隨即, *congci*從此	$ET_1 \leqq ET_2$	*le *guo zhe*

Since these temporal adverbs refer to the successive of events, their temporal relations have no relation with ST and RT. These temporal adverbs only concern the sequence of events in time.

3.4 Feature Representations and Compatibility between other Temporal Adverbs and Aspects

Some temporal adverbs mark habitual situations or the temporal manner of situations. Habitual situations represent that situations go on without focusing on any time points in the time axis. Temporal adverbs denoting the manner of situations are not related to temporal relations. Consequently, these temporal adverbs are classified into another subclass shown in Table 5.

Table 5. Feature representations and compatibility between other temporal adverbs and aspects

Semantic Role	*Temporal Adverbs*	*Features*	*Compatibility with le, guo, and zhe*
frequency	*changchang*常常,*chang*常, *shichang*時常, *wangwang*往往, *shishi*時時, *shike*時刻, *bushi*不時, *meimei*每每, *lao*老, *zong*總, *yizhi*一直	high frequency *experience	*le *guo zhe*
frequency	*ouer*偶爾, *ouer*偶而, *jianhuo*間或, *youshi*有時	low frequency *experience	*le *guo zhe*
time manner	*jian*漸, *jianjian*漸漸, *jianci*漸次, *zhujian*逐漸, *rijian*日漸, *rijian*日見, *zhubu*逐步	slow stage change *experience	*le *guo zhe*

time manner	*suishi*隨時	preparatory	**le* **guo* *zhe*
time manner	*anshi*按時, *anqi*按期	preparatory regular	**le* **guo* **zhe*

4. Conclusion

In sections 3-1 to 3-4, we classified the temporal adverbs listed by Lu and Ma into four main classes. The subclasses in each main class are somewhat different from Lu and Ma's although the majority are similar. Lu and Ma's two-class classification of temporal adverbs is vague, so we adopt four more detailed classes. The corpus data in the Sinica Corpus verify the correctness of our temporal relations. In addition, the co-relations of temporal relations between temporal adverbs and the aspect markers *le*, *guo*, and *zhe* help us to predict the co-occurrences of temporal adverbs with aspect markers. The feature representations of temporal relations are also helpful for assigning semantic roles to temporal adverbs in the Sinica Treebank.

In future work, if necessary or for other reasons, the subclasses of temporal adverbs may be subdivided into sub-subclasses. Furthermore, certain classes of temporal adverbs can co-occur grammatically, while certain classes cannot.

(14) 他曾先行離去

ta ceng xianxing liqu

He had left in advance.

(15) *他曾早日離去

ta ceng zaori liqu

He had left soon.

In sentence (14), the temporal adverb *xianxing* co-occurs grammatically with *ceng*; however, in sentence (15), *zaori* cannot co-occur grammatically with *ceng*. Temporal relations or other linguistic factors causing such differences in compatibility are worth investigating. In addition, the RT-related adverbs may be temporal points, temporal durations, or event times, as in '*ta yi tingdao xiaoxi*' in sentence (13). The reference of RT-related adverbs may be further investigated.

References

Allen, J. F., "Towards a General Theory of Action and Time," *Artificial Intelligence*, 23, 1984, pp.123-154.

Huang, C.-R., K. Athens, L.-L. Chang, K.-J. Chen, M.-C. Liu, and M.-C. Tsai, "The Module-Attribute Representation of Verbal Semantics: From Semantics to Argument Structure," *International Journal of Computational Linguistics and Chinese Language Processing*, 5(1), 2000, pp.19-46.

Klein, W., *Time in Language*, Routledge , London, 1994.

Lu, J., and Z. Ma (陸儉明&馬真), "關於時間副詞" *現代漢語虛詞散論*, 1999, pp.98-127.

Reichenbach, H., *Elements of Symbolic Logic*, Macmillan, London, 1947.

Smith, C. S., *The Parameter of Aspect*, Kluwer Academic Publishers, Dordrecht. 1991.

Yang, G.-W., and Bateman John A., "The Chinese Aspect System and its Semantic Interpretation," In *Proceedings of 2002 COLING II*, 2002, Academia Sinica, Taipei, pp.1128-1134.

Website Resources

Huang, C.-R., MARVS, http://corpus.ling.sinica.edu.tw/course/marvs/index.html,2001.

Academia Sinica, Sinica Corpus, version 4.0, http://www.sinica.edu.tw/SinicaCorpus/index.html, 2001.

Academia Sinica, Sinica Treebank, version 3.0, http://treebank.sinica.edu.tw/, 2005.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 459-472

信息處理用現代漢語虛詞義類詞典研究和工作單設計[1]

Study on the Machine Tractable Thesaurus Dictionary of Contemporary Chinese Functional Words for Information Processing and Design Information Terms for Dictionary Entries

陳群秀*

Qunxiu Chen

摘　要

目前，世界各國學者都十分重視語言信息處理的知識資源的建設，知識包括詞彙學知識、句法學知識、語義學知識、語用學知識乃至常識方面的知識，核心問題是語義學知識。在語義學知識中，詞彙的語義知識是最基本最重要的語義知識。在過去的 10 年中，我們清華大學和合作者對漢語實詞中主要的語類（例如：動詞、形容詞、名詞）的詞彙語義知識進行了系統、全面的研究，並且研製出“現代漢語述語動詞機器詞典”、“現代漢語述語形容詞機器詞典”、“現代漢語名詞槽關係系統”和“現代漢語語義分類詞典”。但是漢語虛詞的詞彙語義知識的研究特別是面向信息處理用的虛詞的詞彙語義知識的研究至今還是一個空白點。本文首先討論了漢語虛詞研究對漢語信息處理的意義。漢語是孤立語、詞根分析型語言，大多數漢語詞彙本身不能明顯地表達語法意義，句法手段主要靠虛詞和語序，況且虛詞對表示語態、語氣、時態體貌、能願情態、句子關係、程度、範圍、引入對象等句子語義和篇章結構有關鍵的作用，因此漢語虛詞詞彙語義研究對漢語信息處理有著特別重要的意義。其次，本文介紹了一個信息處理用現代漢語虛詞義類詞典的探索研究，是在漢語的語

1 本論文有關研究承 863 項目(項目號 2001AA114210)的資助。
* 智慧技術與系統國家重點實驗室，清華大學計算機科學與技術系，北京 100084，中國
The State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
E-mail: cqx@s1000e.cs.tsinghua.edu.cn

[Received April 12, 2005; Revised June 13, 2005; Accepted August 15, 2005]

義層面上研究漢語虛詞的分類的二十一世紀的表達類，亦即研究一個信息處理用現代漢語句子情態表示系統。然後，本文展示了我們初步研究的成果，我們將虛詞義類初步分爲語態範疇、時態體貌範疇、語氣範疇、能願情態範疇、肯定否定範疇、（句子）關係範疇、程度範疇、範圍範疇、對象範疇、其他範疇等十大範疇，每個大範疇根據虛詞表達的意義不同而再分爲若干中範疇和小範疇。再之後，本文提出信息處理用現代漢語虛詞義類詞典工作單的信息項的設計設想。最後，論文還倡議建立現代漢語語義知識庫平台。

關鍵字：信息處理用現代漢語虛詞義類詞典、信息處理用現代漢語句子情態表示系統、語態範疇、時態體貌範疇、語氣範疇、能願情態範疇、肯定否定範疇、（句子）關係範疇、程度範疇、範圍範疇、對象範疇、其他範疇、工作單、信息項、現代漢語語義知識庫平台

Abstract

At present, worldwide scholars from different countries attach great importance to the knowledge resource construction of language information processing. Knowledge includes the aspects of lexicology, syntax, semantics, pragmatics and even general knowledge. The knowledge of semantics is the core, of which the most basic and most important one is the semantic knowledge of vocabulary. During the last 10 years, scholars of Tsinghua University and Cooperators have conducted systematic and complete studies on lexical semantic knowledge of the main categories of Chinese content words (including verb, adjective, noun, etc), and developed and studied “the Machine Tractable Dictionary of Contemporary Chinese Predicate Verbs”, “the Machine Tractable Dictionary of Contemporary Chinese Predicate Adjectives”, “the System of Relations of Slots Centering on Nouns for Contemporary Chinese”, and “the Thesaurus Dictionary of Contemporary Chinese for Information Processing”. However, research on lexical semantic knowledge for information processing in particular of Chinese functional words has so far been limited. Firstly, this paper discusses the value of functional words in Chinese Information Processing. Chinese is a language whose typology involves isolated language and root-word-based analytical language. The vast majority of Chinese words can not themselves clearly express grammatical sense. Chinese syntactical methods mainly depend on the use of functional words and word order. In addition it is important that Chinese functional words express meaning of sentences and structures of texts, e.g. voice, tense & acpect, mood, modality, positive and negative, relation of sentences, degree, range, draw into object, etc. Therefore, the study of lexical semantics of Chinese functional words is great importance to Chinese Information Processing. Secondly, this paper

introduces our research on the Machine Tractable Thesaurus Dictionary of Contemporary Chinese Functional Words. The results of our initial research are presented. The Chinese functional words are divided into ten main-level categories, e.g. Voice Category, Tense & Aspect Category, Mood Category, Modality Category, Positive and Negative Category, Relation Category of Sentences, Degree Category, Range Category, Object Category, Other Category, etc. Each of main-level category is divided into a number of middle-level categories or minor-level category once and again according to expressing different meaning. This paper also presents some ideas of designing information terms for dictionary entries. Finally, we proposes the conception of constructing a base platform for the Contemporary Chinese Semantic Knowledge Bank.

Keywords: The Machine Tractable Thesaurus Dictionary of Contemporary Chinese Functional Words, The Modality Representation System of Chinese Sentences for Information Processing, Voice Category, Tense & Aspect Category, Mood Category, Modality Category, Positive and Negative Category, Relation Category of Sentences, Degree Category, Range Category, Object Category, Other Category, Dictionary Entries, Information Terms

1. 前言

中文信息處理的研究和進展，依賴于漢語計算語言學的詞彙學、句法學、語義學、語用學的研究和進展，核心問題是語義學。語義學是難度最大、起步較晚的一個薄弱環節。由于漢語缺乏屈折變化，是語義型語言，句法分析對句子的貢獻比英語等要小，因此語義分析對漢語機器理解尤爲重要。目前，自然語言理解、漢語信息處理處于一個關鍵時期，處在取得重大突破的前夜，最重要最困難的是語義學的研究和突破。

在計算語言學界，越來越多的專家把機器詞典的規模和質量看作是決定一個自然語言處理系統成敗的關鍵。對於漢語來說，由於缺乏形態變化，漢語的計算機自動分析和處理相對別的語言要困難得多，尤其需要重視語言知識庫特別是語義知識庫的建設。目前中文信息處理領域的語言知識庫有一些，主要是實詞的詞法詞典、實詞的語義詞典、句法規則庫和語料庫，但是至今還沒有一個系統的漢語虛詞詞典。國內外面向人的虛詞的研究也不少，但面向機器自動處理的虛詞研究卻不多，有的話也是零散的個別的，根本沒有系統研究。例如著名的 EDR、WordNet、FrameNet、MindNet 等都是概念詞典，都只有動詞、形容詞、名詞等實詞的概念，都沒有系統地研究虛詞的表達體系。漢語虛詞的語義知識的研究特別是面向信息處理用的虛詞的詞彙語義知識的研究也至今還是一個空白點。

清華大學一直重視和致力於中文信息處理領域的基礎研究，在中文信息處理基礎資源建設方面已經取得了一些成果。在過去的 10 年中，我們對漢語實詞中主要的語類（動詞、形容詞、名詞）的詞彙語義知識進行了系統、全面的研究，並且研製出現代漢語述

語動詞機器詞典、現代漢語述語形容詞機器詞典、現代漢語名詞槽關係系統和現代漢語語義分類詞典。但是對現代漢語虛詞的語義知識特別是面向信息處理用的漢語虛詞的詞彙語義知識一直是我們十分關注而還沒有完成的心願。

漢語是孤立語、詞根分析型語言，大多數漢語詞彙本身不能明顯地表達語法意義，句法手段主要靠虛詞和語序。漢語裏的虛詞往往可以顯示詞與語、片語與片語以及句子與句子之間的關係，成爲語句組織的脈絡。同時，虛詞對表達語態、時態體貌、語氣、能願情態、肯定否定、句子關係、程度、範圍、引入對象等句子語義和篇章結構有關鍵的作用，是篇章知識的主要來源。因此漢語虛詞詞彙語義研究對中文信息處理有著特別重要的意義。二十多年來，中文信息處理從字處理發展到詞處理進而發展到句處理這個層面，若要取得新的突破，必須進入篇章處理層面。因爲一則句子層面要想作深入處理，必須要依靠篇章的知識，不然的話，很多語法歧義、語義歧義、指稱、照應等問題無法解決。二則中文信息處理的一些應用領域（例如機器翻譯、自動文摘、自動問答系統的源文分析和篇章生成）需要對篇章做出有效的分析。因此對漢語虛詞作系統全面的研究不僅是有意義的而且是迫切需要的。但是漢語虛詞的個性很強，運用範圍很廣，運用頻度又高，有的一詞多類兼多義，而且漢語虛詞應用很靈活且缺省現象很嚴重，因此漢語的虛詞特別是信息處理用虛詞詞典研究具有很大難度。

2. 信息處理用現代漢語虛詞義類詞典（漢語情態表示系統）的初步研究

兩年多來，我們對現代漢語的虛詞做了大量的調研、分析和初步研究。我們研究的虛詞包括介詞、副詞、連詞、時態助詞、結構助詞、語氣助詞、助動詞、感歎詞、擬聲詞等詞類。我們對漢語虛詞的研究角度，不僅從虛詞的形式、語法作用角度分析，而且還從虛詞所表示的語義角度和語用角度分析，目的在于研究一個信息處理用現代漢語虛詞義類表達，是在漢語的語義層面上研究漢語虛詞的分類的二十世紀的表達類，亦即研究一個信息處理用現代漢語句子情態表示系統。經過初步研究，我們將漢語虛詞義類分爲語態範疇、時態體貌範疇、語氣範疇、能願情態範疇、肯定否定範疇、（句子）關係範疇、程度範疇、範圍範疇、對象範疇、其他範疇十大範疇，每個大範疇根據虛詞表達的意義不同再分爲若干中範疇和小範疇。下面將虛詞義類的大、中、小範疇分類例示如下：

1 語態範疇：指說話者選擇主體還是選擇客體作話題。

1.1 主動態

1.2 被動態

形態/准形態標誌：被、讓、由

1.3 使役態

形態/准形態標誌：動詞爲“讓、請、使、勸、叫、教…”等使役動詞；

2 時態體貌範疇：指因說話者不同的觀察點而表達的事件的時間進程軸上所處的特定階段或與時間無關而與動量有關的運動狀態和情貌

2.1 時態範疇

2.1.1 現在時

2.1.2 過去時

2.1.3 將來時

2.2 體貌範疇

2.2.1 預期體 框架標誌：將 V

2.2.2 即始體 框架標誌：即將 V

2.2.3 開始體 框架標誌：V 起來；V 起；V 開；V 上；

例如：戰爭打起來了；奏起軍歌；議論開了；議論上了；

2.2.4 剛始體 框架標誌：剛 V 起來；剛剛開始

例如：戰爭剛打起來；

2.2.5 進行體 框架標誌：V 著；正 V 著（呢）；正在 V 著；在 V 著；V_1 著 V_2；V_1 著 V_1 著 V_2；V 著 N；V 著點兒！V 著!

例如：他正寫著報告呢；門開著；她笑著說；說著說著哭了；牆上掛著軍用地圖；聽著點兒!你聽著!

2.2.6 繼續體 框架標誌：V 下去；V 下來；

例如：這種武器還要生產下去；戰士們堅持下來了。

2.2.7 完成體 框架標誌：V 了 $_1$；V 過 $_2$(了 $_1$)；

例如：吃過飯就去。

2.2.8 結果完成體 框架標誌：V 好；V 完；V 成；V 上；V 上來；V 到；V 著(zháo)；V 住；V 掉；V 下；V 下來；V 去：V 了 $_3$(lou)；

例如：穿好軍裝；穿上軍裝；穿完軍裝；抓住敵人；抓著敵人；抓到敵人；脫下軍裝；脫了 $_3$ 軍裝；脫去軍裝；脫掉軍裝；

2.2.9 剛歷體 框架標誌：V…來著；

例如：河對岸剛才打槍來著。

2.2.10 經歷體 框架標誌：V 過 $_1$；曾經 V 過 $_1$；曾 V 過 $_1$；

例如：執行過偵察任務；

2.2.11 試量體 框架標誌：VV；VV…看；V 了 $_1$V；

例如：把坦克修修；讓他來修修看；

2.2.12 小量體 框架標誌：V 一 V；V 一下；V 了 $_1$ 一 V；V 了 $_1$ 一下；

例如：稍微把準星調一調；

2.2.13 反復體 框架標誌：V 來 V 去；V 過來 V 過去；

例如：跑來跑去；在天上飛過來飛過去；

2.2.14 多發體 框架標誌：$V_1V_1V_2V_2$；

例如：說說笑笑；蹦蹦跳跳；

2.2.15 已然體 框架標誌：已經 V；已 V；早已 V；(是)N 處 V 的 N；

例如：是西單上的車；已經打響；

2.2.16 未然體 框架標誌：沒有 V；沒 V；尚未 V；

例如：他沒打槍；敵人尚未進攻；

2.2.17 經常體 框架標誌：經常 V；常 V；時常 V；

例如：經常騷擾敵人；

2.2.18 漸進體 框架標誌：漸漸（表示程度或數量的漸變增減）

2.1.19 瞬時體 框架標誌：忽然，倏然

2.2.20 一般體 框架標誌：(無)

例如：上前線；觀察敵情；

3 語氣範疇：指說話者表示交際意圖的語氣。

3.1 功能語氣

3.1.1 陳述語氣 句終標點：。

3.1.2 疑問語氣 句終標點：？

3.1.3 反詰語氣 形態/准形態標誌+句終標點：{豈|難道|何嘗|不成|不行}+？

3.1.4 祈使語氣 句終標點：（吧）+{。|！}

3.1.5 感歎語氣 形態/准形態標誌+句終標點：

{太|多|多麼|真|特|特別}+（{了|啊}）+!

3.1.6 確認語氣 框架標誌：({當然|自然})+(是)…+的；

3.1.7 非確認語氣 形態/准形態標誌：吧

3.2 情感語氣

3.2.1 詫異語氣 形態/准形態標誌：{竟|竟然|居然}

3.2.2 料定語氣 形態/准形態標誌：{果然|果真|敢情}

3.2.3 領悟語氣 形態/准形態標誌：{難怪|怪不得|敢情|噢|原來}

3.2.4 僥倖語氣 形態/准形態標誌：{幸虧|幸好|幸而|省得}

3.2.5 索性反正語氣 形態/准形態標誌：{索性|簡真|反正|偏偏|偏不}

3.2.6 不滿語氣 形態/准形態標誌：什麼；怎麼；

3.2.7 追究語氣 形態/准形態標誌：究竟；到底；

3.2.8 肯定承諾語氣 形態/准形態標誌：一定；肯定；

3.2.9 加強堅決語氣 形態/准形態標誌：就；

4 能願情態範疇：指因說話者固有的觀點而引發的評價、告誡等態度和強烈程度。

4.1 評價型能願情態

4.1.1 “能”類評價 形態/准形態標誌：能；能夠；只能；才能；

4.1.2 “要”類評價 形態/准形態標誌：要；想；

4.1.3 “會”類評價 形態/准形態標誌：會；必然；必定；不會；一定；

4.1.4 “敢”類評價 形態/准形態標誌：敢；

4.1.5 “可能”類評價 形態/准形態標誌：可能；不可能；

4.1.6 “也許”類評價 形態/准形態標誌：也許；或許；大概；大約；多半；不一定；

4.1.7 “沒法”類評價 形態/准形態標誌：沒法；無法；

4.1.8 “肯”類評價 形態/准形態標誌：肯；願意；情願；樂意；願；巴不得；想；

4.1.9 “可以”類評價 形態/准形態標誌：可以；可；

4.1.10 “準備”類評價 形態/准形態標誌：準備；預備；打算；計畫；

4.1.11 “得”類評價 形態/准形態標誌：得；不得；“一定得”；“准得”；

4.2 告誡型能願情態

4.2.1 “別”類告誡 形態/准形態標誌：別；勿；萬萬不要；

4.2.2 “要”類告誡 形態/准形態標誌：要；

4.2.3 “應該”類告誡 形態/准形態標誌：應；應該；應當；該；不該；不應；

4.2.4 “只有”類告誡 形態/准形態標誌：只有；

4.2.5 “必須”類告誡 形態/准形態標誌：必須；一定；務必；

4.2.6 “用不著”類告誡 形態/准形態標誌：用不著；不必；無庸；

4.2.7 “得”類告誡 形態/准形態標誌：得；

5 肯定否定範疇：指說話者對事件的肯定或否定態度。

5.1 全盤肯定 形態/准形態標誌：全；都；通通；全部；統統；

5.2 全盤否定　形態/准形態標誌：全不；全沒；都不；都沒；

5.3 部分肯定　形態/准形態標誌：並非…都；未必都；

5.4 部分否定　形態/准形態標誌：

5.5 堅決否定　形態/准形態標誌：統統不；通通不；

5.6 雙重否定　形態/准形態標誌：未嘗不可；未嘗沒有；未必不是；

5.7 不定　形態/准形態標誌：不 A；不 A（反義）；例：不上不下；不方不圓；不死不活；

6 關係範疇：指說話者在用複句表達事情時對事件間(分句間)的語義、時間、邏輯上的關系的認定。

6.1 並列關係：各分句在意義上是平排並列的。

常用的關聯詞語有：既…又…；又…又…；不但…同時…；…同時…；及；

6.2 承接關係：各分句間具有先後相繼關係。

有時用“首先…，其次…，然后…”，“一…就…”；“既而”；

6.3 遞進關係：后面分句的意思比前面分句更進一層。

常用的關聯詞語有：不但…而且…；不但…並且…；…並且…；不但…同時…；…同時…；不但…尤其…；…尤其…；不但…甚至…；…甚至…；不但…也…；…也…；不但…還…；不但…更…；不但…又…；不但…簡直…；不但…反而…；不但…竟然…；尚且…何況…；不但…連…也…；不但…既使…也…；不但…就是…也…；不但…哪怕…也…；不但…而且…甚至…；不但…而且…簡直…；不但…而且…尤其…；不但…哪怕…也…，甚至…也…；且不說…，也不說…，只說說…；要…還要…；

6.4 選擇關係：分句間有選擇關係。

常用的關聯詞語有：是…還是…；不是…就是…；或者…或者…；…，或者…；與其…不如…；與其…寧可…；不是…而是…；毋寧；

6.5 轉折關係：分句間意思上有所轉折。

常用的關聯詞語有：雖然…但是…；…但…；…可是…；雖然…可…；雖然…還…；…，雖然…；…卻…；…可(卻)…；…還…；…就…；…唯獨…；儘管…還是…；

6.6 因果關係：分句間具有原因和結果的關係。

常用的關聯詞語有：因爲…，所以…；…，因爲…；…，所以…；之所以…，是因爲…；因爲…，…；…，因此…；…，因而…；…，于是…；…，一則…，二則…；由于…，因此…；既然…，(就)…；…，反正…；…，結果…；到底；

例如：因難年還好辦，反正買不到菜，沒有什麼三盤四樣。

6.7 假設關係：分句間有假設和結果的關係。

常用的關聯詞語有：如果…，則…；…，如果…；要是…就…；…吧…吧；假如…，那未…；假如…就…；如果…，那未…；如果…，那麼…；若是…，那未…；如果…，就…；(如果)…的話，…就…；要不然的話，…；…就…；萬一…，那…；…一…，(那)…；

例如：冰一化，那魚是會發臭的。/不管吧，先進人物發展成道德敗壞，說明制藥廠的政治思想工作不得力；管吧，弄不好又是對先進人物的誹謗和妒忌。

6.8 條件關係：分句間具有條件和結果的關係。

常用的關聯詞語有：只有…才(能)…；不管…也要…；不管…都…；無論…一定要…；只要…就…；只要…(就)一定(能)…；除非…才…；除非…否則…；…一…就…；

例如：除非他去，否則這場球贏不了。/她一聞見煙味就咳嗽。

6.9 讓步關係：一個分句對另一個分句表示退讓一步。

常用的關聯詞語有：即使…也還要…；…，即使…，即使…也…；縱然…都…；…則…，但…；…是…，就是…；

例如：好是好，就是太貴了。/巧則巧矣，但恐不合于說漢話人的心理。

6.10 目的關係：前後分句分別表示行動和目的的關係。

常用的關聯詞語有：爲了…；…，爲的是…；…，也好…；…，以便…；…，免得…；…，以免…；

例如：爲啥不早點來，也好讓大媽把你當作女兒來疼你。/學好本領，以便更好地爲人民服務。/事先要準備好，免得臨時手忙腳亂。

6.11 發展關係：分句間表示程度隨著條件的發展而發展。一般第一個分句爲條件，后面的分句表示隨條件的發展而發展的程度。

常用的關聯詞語有：越…越…；愈…愈…；

例如：環境越是艱苦，他的革命意志越是堅定。

6.12 比擬關係：表示主要內容說完后意猶未盡，又比擬之，或是使未竟之意形象化，或是使未竟之意理性化。

常用的關聯詞語有：好象；就象；象；

例如：他嚇得從長椅上跳起來，就象書生幻想美人，美人來了卻是狐狸變的，嚇得他魂不附體。/徐麗莎恍恍惚惚地好幾天，象被狂風卷上了天，在空氣中翻著筋斗。

6.13 詮釋句：后段是直接解釋前段末某個成分的。

前一段末一般都可以加“：”號，或是後段加“說是”之類的詞語。

例如：當時的工作組也曾有過懷疑：這樣的人家能不能稱作城市貧民呢？/我的男朋友只愛我一點，說是跟我在一起時感到很快樂。/南京的新名勝，不用說首推中山陵。

6.14 時同關係(伴隨關係)：表示在做某件事同時或伴隨著發生某件事的關係。

常用的關聯詞語有：一邊…一邊…；邊…邊…；一面…一面…；連…帶…；隨…隨…；

例如：他一邊走一邊嗑著瓜子。/她一路上連哭帶罵。/兒子總是邊看電視邊吃飯。

7 程度範疇：指說話者對程度大小、程度高低、程度深淺的認定態度。

例如，表示悲痛的程度有：很；十分；十二分；相當；太；極其；萬分；無比，等等。又如，表示不滿程度的有：很；非常；大爲；極爲，等。

8 範圍範疇：指說話者對物體數量多少、時間和處所的範圍大小的認定態度。

例如，表示整體、全體或部分的有：統統；全都；無一；無一不；都；大都；大半；大體；多半；僅僅；只；才不過；惟獨；除了；除去；除開，等等。表示處所的範圍大小有：處處；到處；無處不；大多；只；太凡；僅，等等。表示時間的範圍大小有：時時；時時刻刻；時常；有時；偶爾，等等。

9 對象範疇：指說話者對涉及的個體、時間、地點、依據的引入。

9.1 引進對象範疇：指說話者對引進對象的表示。

例如，由“給”、“于”、“對”、“爲”、“由”、“替”、“向”、“沖”、“沖著”、“奔”、“奔著”、“當著”等引進的人或物對象。

9.2 來源去向經由範疇：表示時間、地點的來源、去向或經由的範疇。

例如，“打”、“從”、“來自”、“由”、“向”、“打從”、“從打”、“經”、“奔”、“朝”等。

9.3 依據憑藉範疇：表示說話、做事的依據和憑藉的範疇。

例如，“根據”、“據”、“憑”、“憑藉”等。

9.4 目的範疇：表示說話或做事的目的對象。

例如，“爲”、“爲了”等。

10 其他範疇：不屬于 1~9 範疇的其他範疇。

例如，象“啾啾”、“嘰哩呱啦”、“淅淅瀝瀝”、“哐當”、“嗷嗷”、“吧唧”、“吧嗒”這樣的象聲詞，還有表示並列名詞或名詞性片語的“和”、“與”、“跟”、“及”、“以及”等。

3. 信息處理用現代漢語虛詞義類詞典表示信息項目設計

對于機器詞典來說，所表示的信息項目的設立最爲重要。我們的信息處理用現代漢語虛詞義類詞典（現代漢語情態表示系統）信息表示項目設立原則是虛詞的語法形式、語法意義、語義意義、語用用法相結合，期望爲大家提供最精確完整和細緻的信息。我們的工作單中對每一個虛詞描述信息項目有：詞形、拼音、詞類、義項數目、義項序號、釋義、情態範疇（包括大範疇、中範疇、小範疇）、表示意義、表示意義的範疇值、形態/准形態標誌、框架標誌、常用的近義/同義關聯詞語、句例、備註。工作單樣式如下：

<table>
<tr><td>詞條序號</td><td colspan="3"></td><td>製作者</td><td></td><td>工作單號</td><td colspan="3"></td></tr>
<tr><td>詞形</td><td></td><td>拼音</td><td></td><td>詞類</td><td></td><td>義項數目</td><td></td><td>義項序號</td><td></td></tr>
<tr><td>釋義</td><td colspan="9"></td></tr>
<tr><td>情態範疇</td><td>大範疇</td><td colspan="2"></td><td>中範疇</td><td colspan="2"></td><td>小範疇</td><td colspan="2"></td></tr>
<tr><td>表示意義</td><td colspan="9"></td></tr>
<tr><td>表示意義的範疇值</td><td colspan="9"></td></tr>
<tr><td>形態/准形態標誌</td><td colspan="9"></td></tr>
<tr><td rowspan="2">框架標誌</td><td colspan="9"></td></tr>
<tr><td colspan="9"></td></tr>
<tr><td>常用的近義/同義關聯詞語</td><td colspan="9"></td></tr>
<tr><td rowspan="2">句例</td><td colspan="9"></td></tr>
<tr><td colspan="9"></td></tr>
<tr><td>備註</td><td colspan="9"></td></tr>
</table>

其中“拼音”填寫的是該虛詞在該範疇下的拼音讀寫，填寫的拼音必須帶聲調，用1、2、3、4、5 分別表示漢語的四聲和輕聲。“詞類”填寫的是該虛詞在該拼音和該範疇下的詞類。“義項數目”指的是該虛詞在該拼音該詞類下共有的義項數目，“義項序號”指的是該虛詞在該拼音該詞類該義項數目下的第幾個義項。“情態範疇”填寫的是該虛詞在該拼音該詞類該義項序號下屬于什麼大範疇、什麼中範疇、什麼小範疇。“表示意義”填寫的是該虛詞表示的語法意義、語義意義，包括語體。此項填寫有助于辨析該虛詞該義項與該虛詞其他義項的區別。“表示意義的範疇值”表示的是“程式範

疇”、“範圍範疇”等範疇的屬性值。例如，表示程度範疇的副詞有“很、十分、十分、十二分、萬分、極、相當、挺、無比”，填寫程度範圍值時“十分”一詞填寫“10”，“很”填寫值“10”。“十二分”填寫“12”、萬分填寫“10000”。“相當”填寫“8”、“挺”填寫“9”，“極”填寫“1000”，“無比”填寫“+∞”等。這樣將來機器理解時可以對它們進行比較和計算。“框架標誌”填寫的是該虛詞在使用環境中可以多少種模式框架，與什麼類詞性的詞搭配使用等等。“常用的近義/同義關聯詞語”顧名思義是填寫與之同義或相近意義的關聯詞語或與之成對使用、前後呼應的關聯詞語。“句例”填寫的是該拼音該詞類該義項序號該範疇下的句子各種用法的真實例子。“備註”中可填寫在以上各信息項中尚未包括的信息內容或作必要的說明。

4. 現代漢語語義知識庫平台建設的構想

目前我們正在編寫和修改“信息處理用現代漢語虛詞義類詞典”填寫規範，並且正在進行詞典工作單的試填寫實驗。信息處理用現代漢語虛詞義類詞典的管理軟件系統也初步設計和實現了，準備在試填寫實驗后進行工作單的正式填寫和錄入校對，以期儘快構建這個漢語虛詞義類詞典（亦即信息處理用現代漢語句子情態表示系統）。同時，我們也正在對已構建的“現代漢語述語動詞機器詞典”、“現代漢語述語形容詞機器詞典”、“現代漢語名詞槽關係系統”、“信息處理用現代漢語語義分類機器詞典”等四個實詞的語義詞典進行整合集成和機器學習功能的研究。待“信息處理用現代漢語虛詞義類詞典”構建後，我們準備將這五個實詞、虛詞的語義詞典整合集成一個現代漢語語義知識庫平台，此平台可爲中文信息處理提供豐富、全面、可靠的語義知識支持（包括詞彙語義知識和句子語義知識），可爲現代漢語語言學、語義學研究、對外漢語教學、中小學語文教學計算機輔助教學提供有力的工具和資源。目前我們正在進行現代漢語述語動詞機器詞典、現代漢語述語形容詞機器詞典在漢語信息處理方面應用的探索，也正在研究和設計這些語義詞典在對外漢語教學輔助教學方面的應用軟件。

參考文獻

林杏光，《詞彙語義和計算語言學》。北京：語文出版，1999。

林杏光，《複句與表達》。北京：中國物資出版社，1986。

陳群秀，”信息處理用現代漢語句型系統的初步研究”，第二十屆東方語言計算機處理國際學術會議（20th ICCPOL’2003）論文集《Advances in Computation of Oriental Language》，2003 年 8 月，205-212。北京：清華大學出版社。

申小龍，《漢語句型研究》。海南：海南人民出版社，1989。

賀陽，”漢語’語氣’(Modality)及其標誌簡表”，1991。

史有爲，羅建林，”漢語‘體’及其標誌簡表”，1991。

陳群秀，”漢語自然語言理解研究概況、前景及難點討論”，1990 年國際中文與東方語言計算機處理學術會議論文，長沙，1990。

陸儉明，馬真，《現代漢語虛詞散論》。北京：北京大學出版社，1985。

朱德熙，《語法問答》。北京：商務印書館，1985。

呂叔湘，《漢語語法分析問題》。北京：商務印書館,1979。

胡裕樹主編，《現代漢語（增訂本）》。上海：上海教育出版社，1981。

房玉清，《實用漢語語法》。北京：北京語言學院出版社，1984。

姚殿芳，潘兆明，《實用漢語修辭》。北京：北京大學出版社，1987。

黃伯榮，廖序東，《現代漢語（下冊）》。蘭州：甘肅人民出版社，1980。

陳群秀，"信息處理用現代漢語虛詞義類研究初步構想"，香港: 第四屆漢語辭彙語義學研討會 (4thCLSW)，2003 年 6 月。

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 473-482

An Unsupervised Approach to Chinese Word Sense Disambiguation Based on Hownet[1]

Hao Chen*, Tingting He*, Donghong Ji+ and Changqin Quan*

Abstract

The research on word sense disambiguation (WSD) has great theoretical and practical significance in many fields of natural language processing (NLP). This paper presents an unsupervised approach to Chinese word sense disambiguation based on Hownet (an electronic Chinese lexical resource). In our approach, 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. Lastly, the ambiguous words can be disambiguated after a similarity calculation process is completed. Our experiments involved extraction of terms, and an 82.62% average accuracy rate was achieved.

Keywords: Word Sense Disambiguation, Hownet,Second-Order Context, K-Means Clustering

1. Introduction

Ambiguous words are widely used in various kinds of natural languages and are distributed widely. Word sense disambiguation (WSD) has long been studied as an important problem in natural language processing (NLP), and the research on WSD has great theoretical and practical significance in many areas of NLP.

Much work has been done on WSD. In [Lu *et al*. 2002], the author presented an unsupervised approach to WSD based on a vector space model. This method defines the

[1] The study was supported by the National Natural Science Foundation of China (NSFC), (GrantNo10071028), the National Language Application Project of the Tenth five-year plan of China (GrantNo ZDI105-43B), the Ministry of Education of China, and a Research Project for Science and Technology (Grant No ZDI105-43B).

* Department of Computer Science, Huazhong Normal University, Wuhan, China 430079
E-mail: chenhaoccnu@163.com; tthe@mail.ccnu.edu.cn; quanchqin@163.com
The author for correspondence is Tingting He.

+ Institute for Infocomm Research, Heng Mui Keng Terrace 21, Singapore 119613
E-mail: dhji@i2r.a-star.edu.sg

[Received April 14; Revised June 15, 2005; Accepted August 15, 2005]

matrix of a word and calculates the importance of the context to depict the word, so that the precision position of the word in vector space can be located. However, a lot of information provided by the word sequence in the context is ignored by *tf.idf*. method. Li Juan-zi used *Cilin*(Chinese dictionary) to get semantic vectors, using a machine-readable method [Li 1999]. There are three drawbacks when using *Cilin* as a lexical knowledge base: (1) Semantic vectors are hard to determine accurately because of the coarse classification of word senses. (2) There are not enough words; *Cilin* only contains sixty thousand. (3) *Cilin* is based on the semantic frame of a hierarchical tree, with which it is hard to express the semantic relativity, especially the domain relativity between words. Yet such information plays rather important role in the WSD processing.

This paper presents an unsupervised approach to WSD based on Hownet. Compared with the above works, the main characteristics of ours are as follows:

(1) The synonymous set provided by Hownet can express word senses more concretely.

(2) Contexts that include ambiguous words are converted into vectors by means of second-order contexts, so that more information about word senses in contexts can be provided.

(3) Features are selected based on the extraction of terms.

Section 2 describes our method. In section 2.1, we introduce some related information about Hownet. Section 2.2 describes the method for establishing second-order contexts and the method for clustering these context vectors by means of the k-means algorithm. Section 2.3 describes the method for calculating the similarity between context vectors. In section 3, we describe experiments that we conducted using our method. We compare our method with other similar methods and find that we were able to obtain better results. In section 4, we conclude our work and offer some suggestions for further improvement.

2. An Unsupervised Approach to WSD Based on Hownet

With the rich lexical resources of Hownet, we firstly convert contexts that include ambiguous words into context vectors; then, the definitions (DEFs) of ambiguous words in Hownet can be determined by calculating the similarity between these context vectors. The whole process is completed automatically, so a sense-labeled corpus is not needed.

2.1 Hownet and WSD

Hownet [Dong and Dong 2000], which describes concepts in Chinese and English, is a knowledge database that determines the relationships between concepts or the attributions of concepts. In our method, we focus on the knowledge dictionary.

2.1.1 The Record Mode of the Knowledge Dictionary

The knowledge dictionary is the fundamental document in Hownet. In this document, a record contains a concept of a word and its description. Each record is made up of 4 parts .In each part, the part on the left side of "=" contains the name of the data and the part on the right side contains the value of the data. Their alignment is as follows:

W_C= Chinese Word;
W_E= English Word;
E _X= Example of the word;
G_C= Part of speech in Chinese;
DEF= Definition.

2.1.2 An Example of Ambiguous Word in Hownet:

One sense of "沽(Gu)":

No =032339;
W_C=沽(Gu);
G_C=v;
W_E= sell;
DEF=sell, commercial;

The other sense of "沽(Gu)":

No=032340;
W_C=沽(Gu);
G_C=v;
W_E= buy;
DEF=buy, commercial.

2.1.3 Word Senses of Ambiguous Words

Several different W_C in Hownet correspond to the same DEF; in other words, each word sense has a synonymous set. Table 1 gives examples that show the numbers of synonyms of

ambiguous words in Hownet.

Table 1. The numbers of synonyms of ambiguous words in Hownet

Ambiguous Word	DEF	The number of synonyms	Total number
健康	A value|属性值,kind|类型,ordinary|普,desired|良	1	47
	A value|属性值,physique|体格,strong|强,desired|良	38	
	attribute|属性,physique|体格,&Animal Human|动物	8	
造就	result|结果,#succeed|成功,desired|良	29	53
	cultivate|培养	24	

Perhaps the words in a set of synonyms are also ambiguous words. We can solve this problem using the following method: for each word in the set of synonyms, we collect its contexts and cluster these contexts by means of the k-means algorithm (section 2.2). Each word has several categories. We select one category with the smallest distance from the others, and the contexts in that category are the contexts that we want. Finally, we synthesize the contexts as the context set of the ambiguous word.

2.2 K-Means Clustering Based on Second-Order Contexts

The k-means clustering approach presented by [MacQueen 1967], is an important method for data mining and knowledge discovery, and it has the characteristics of simplicity and fast convergence.

When the k-means clustering algorithm is used, the first problem is translating the content of the document into a form that can be processed mathematically. Here, we convert contexts that include ambiguous words into vectors by means of second-order contexts. Our method is as follows:

(1) We Extract terms from the contexts that include ambiguous words by the method that offered in [Pantel and Lin 2001].

(2) A parameter called “m” represents the number of calculated occurrences. We then select terms with high “m” values as features. This step is based on our assumption that there is at least one feature in each context. If the context contains several features, we calculate the sum of the vectors of the features.

In the following, we will take 健康（health）as an example. We select the features whose numbers of occurrences are above 8 from 445 texts. Table 2 presents the co-occurrence frequency of two features in 445 texts; and Table 3 presents the vectors of 5 contexts in Table 2.

Table 2. The co-occurrence frequency of two features

feature/m $context_i$ |feature	持续 /11	快速 /11	发展 /13	老年人 /9	生活 方式 /9	精神 /8	心理 /10	教育 /16	身体 /8	教师 /9	事业 /10
C1 持续	\	7	7	1	1	1	2	1	1	0	5
C1 快速	7	\	7	0	2	1	2	0	1	0	5
C1 发展	7	7	\	1	3	1	0	2	3	2	6
C2 老年人	1	0	1	\	5	3	2	1	4	5	0
C2 生活方式	1	2	3	5	\	1	1	2	3	3	0
C3 精神	1	1	1	3	1	\	2	0	3	2	1
C3 心理	2	2	0	2	1	2	\	4	2	4	2
C3 教育	0	0	2	1	2	0	4	\	0	6	5
C4 身体	1	1	3	4	3	3	2	0	\	3	4
C4 教师	0	0	2	5	3	2	4	6	3	\	6
C5 事业	5	5	6	0	0	1	2	5	4	6	\

Table 3. Vectors of contexts

Feature $Context_i$	持续	快速	发展	老年人	生活方式	精神	心理	教育	身体	教师	事业
C_1	14	14	14	2	6	3	4	3	5	2	16
C_2	2	2	4	5	5	4	3	3	7	8	0
C_3	3	3	3	6	4	2	6	4	5	12	8
C_4	1	1	5	9	6	5	6	6	3	3	10
C_5	5	5	6	0	0	1	2	5	4	6	0

Then we perform clustering using the Novel K-means algorithm [Li *et al.* 2002]. K-means adopts the iterative way to renew. In each iteration, the value of the objective function decreases. When the smallest value of the objective function is related, the best clustering result can be obtained. During k-means clustering, determining k is always a key problem .If the ambiguous word has n word senses in Hownet, we can define k as being equal to n. But if the contexts we extract from the corpus do not contain n word senses, we can using the following method to get k: for each k, we cluster separately from 2 to n. Take k＝3 as an

example, to each category of the three, we can get the highest score between the category and the word sense in Hownet. Then we can take the average score of the 3 highest scores as the final score. Lastly, we can determine k when its corresponding score is the highest. In the example discussed in section 2.2, we find 5 contexts in which "health" occurs .We determine that k＝3 when the highest average score is obtained, in other words, we can get the best result by 3 categories .The results are as follows: （C_1），（C_2、C_4、C_5），（C_3）.

2.3 A Way to Calculate the Similarity between Two Vectors

The similarity between two context vectors can be calculated using the 'cosine' formula, which is show below:

$$sim(Q, D2) = \frac{\sum_{j=1}^{t} wqj \times wd_{ij}}{\sqrt{\sum_{j=1}^{t} \left(wqj\right)^2 \times \sum_{j=1}^{t} \left(wd_{ij}\right)^2}} \quad . \tag{1}$$

The formula makes good use of the vector space model, which can formalize the content of the context into a spot and convert it into a vector. This is because documents can be defined in real number space, models can be identified, many algorithms in other realms can be applied, and the calculability and maneuverability of documents in NLP has greatly improved.

3. Experiment and Results

3.1 Experiment Based on Hownet

The steps in the algorithm are as follows:

(1) Select some contexts containing ambiguous words (W) from a corpus;

(2) Cluster these contexts into several categories using the k-means clustering algorithm;

(3) Find the word sense (w_i) of W (w_1, w_2, w_n) in Hownet and compose the synonym set for each word sense (w_i). We then extract some contexts containing synonyms from the corpus and finally, convert the contexts into second-order contexts. Each word sense is replaced by a vector, which is constructed by means of second-order contexts based on its synonyms set.

(4) Calculate the similarity between the categories in (2) for word sense; the similarity of the word sense with the largest value is the correct one.

Experimental data are shown in Table 4. Concerning the experiment, several points must be noted:

(1) The training data in our experiment came from a modern Chinese corpus.

(2) The extraction of contexts takes the sentence as a unit, and a context that lacks features is removed.

(3) Hownet, 2000 version, by Dong Zheng-Dong was adopted in our experiment.

(4) Our experiment does not have the difference between open test and close test because of the unsupervised approach.

Table 4. Experimental data and results (Hownet)

Ambiguous words	Number of contexts		Number of experimental contexts		Accuracy (%)	
	Word sense	Number of synonyms	All contexts	Correct contexts		Average (%)
材料 (cai 'liao)	S1	3	65	51	78.46	82.62
	S2	61				
	S3	7				
改(gai)	S1	33	72	60	83.33	
	S2	44				
代表 (dai biao)	S1	20	84	68	80.95	
	S2	47				
	S3	0				
	S4	6				
健康 (jian'kang)	S1	8	50	40	80.00	
	S2	38				
	S3	1				
保持 (bao'chi)	S1	27	66	56	84.48	
	S2	51				
打气 (da qi)	S1	24	46	40	86.95	
	S2	13				
挂彩 (gua'cai)	S1	21	55	47	85.45	
	S2	22				
表现 (biao'xian)	S1	2	75	61	81.33	
	S2	37				
	S3	24				

3.2 Discussion

We have compared our results with results reported in the literature [Lu *et al*. 2002]; [Li 1999]. The results in [Lu *et al*. 2002] are listed in Table 5 (excerpted from reference [Lu *et al*. 2002] page 1086) along with the word senses of ambiguous words from the <Modern Chinese Lexicon>, 1996 edition. Figure1 shows the accuracy of 5 ambiguous words; the results [Li 1999] are listed in Table 6 (excerpted from [Li 1999] page 76), and the word sense of ambiguous words came from Cilin, a Chinese thesaurus.

Table 5. Results of experiments in reported in [Lu et al. 2002]

Polysemous words	Number of senses	Accuracy (%)
材料（cai'liao）	4	72.83
改(gai)	3	78.71
表现(biao'xian)	3	90.39
发表(fa'biao)	2	89.11
健康(jian'kang)	2	75.07

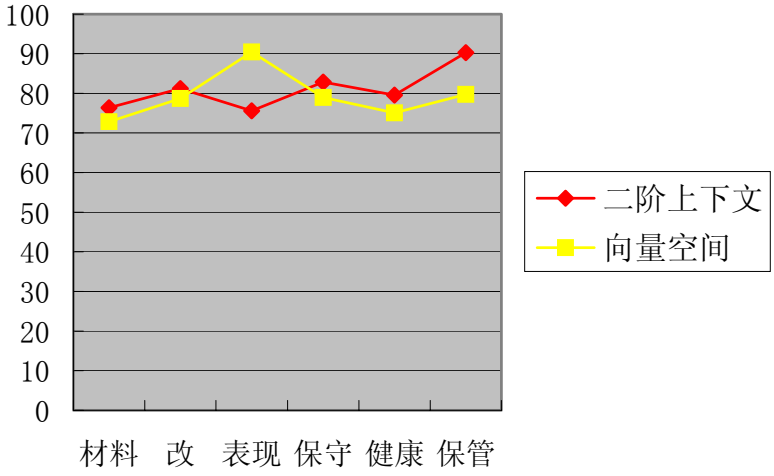


Figure1. Accuracy of 5 words based on Hownet

Table 6. Results of experiments reported in [Li 1999]

Polysemous words①	Sense ID②	Accuracy③ (%)
材料(cai'liao)	Dk17/ba06/al03	81.7
改(gai)	ih02/hg18/hj66	70.6
表现(biao'xian)	Jd06/di20/hj59	68.9
发表(fa'biao)	Hc11/hi14/jd03	73.4
健康(jian'kang)	ed43/eb37	70.1

①Ambiguous words, ②type of ambiguity, ③accuracy.

Notice: we did not select "Publish" as an ambiguous word in our experiment because several word senses of "Publish" in Hownet are very close to each other.

Compared with the results reported in [Lu *et al*. 2002], we obtained good results for some of the ambiguous words. What is more, for the selected 5 words, the average accuracy improved. Compared with the results reported in [Li 1999], the accuracy achieved was greatly improved.

4. Conclusion

Cognitive linguists believe that the similarity of contexts has a great influence on the process of differentiating diverse semantemes [Miller and Charles 1991]. This paper adopted this assumption that the similarity of contexts determines the similarity of semantemes of the words. The experiments proved that our approach is feasible. However, there are still some shortcomings:

(1) The selection of features. The features that we select must provide as much information for word sense disambiguation as possible and, at the same time, should contain as little noise as possible. This is important for clustering.

(2) The accuracy of clustering. The clustering of k-means clustering has some drawbacks in that it is sensitive to the center of the cluster and it is difficult to get a superior overall solution of overall situation. Thus the accuracy we achieved was influenced by the accuracy of clustering.

(3) Expanding the scope of ambiguous words. It is difficulty to evaluate the overall results of our approach, so many people are concerned about the issue. The applicability of our approach should be further tested in large-scale experiments and with other languages.

References

Dong, Z. D., and Q. Dong, "Hownet," http://keenage.com, 2000.

Li, F., B. Xue, and Y. L. Huang, "A Novel K-means Clustering Based on Initial Central Superior," *Computer Science*, 29(7), 2002, pp.94-96.

Li, J. Z., "The Research on Chinese Word Sense Disambiguation," Ph.D.Thesis, Beijing: Tsinghua University, 1999.

Lu, S., S. Bai, and X. Huang, "An Unsupervised Approach to Word Sense Disambiguation Based on Sense-Words in Vector Space Model," *Journal of Software,* 13(6), 2002, pp.1082-1089.

MacQueen, J., "Some methods for classification and analysis of multivariate observations," In *proceedings of 5th Symp. Math. Statist, Prob,* 1967, Berkeley, pp.218-297.

Miller, G. A., and W. Charles, "Contextual Correlates of Semantic Similarity " *Language and Cognitive Processes*, 6(1), 1991, pp.1-28.

Pantel, P., and D.K. Lin, “A Statistical Corpus-Based Term Extractor,”In *proceedings of AI*, 2001, Canadian, pp.36-46.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 483-494

以句式爲本的多義詞詞義辨識[1]

Word Sense Disambiguation Based on Syntactic Construction

蔡美智*

Mei-Chih Tsai

摘要

本文以動詞「送」的多義辨識爲例，通過語料庫句式分佈統計，研究詞義和句式間的對應關係，爲多義詞詞義辨識提供基礎資源。根據搭配句式的不同，多義詞「送」可切分爲遞送、贈送、送行、斷送四個義項，遣送、推送兩項則分別歸入送行義、遞送義，作義面處理。結果證明，句式搭配訊息確實能夠在詞義辨識過程中起一定作用。

關鍵詞：多義詞、詞義辨識、語料庫、義面

Abstract

This paper explores the correlation between lexical semantics and syntactic construction, with particular attention paid to delimiting the lexical semantic distinctions between the multiple senses of the polysemous verb *song4* 'to send'. Different types of syntactic constructions are first categorized according to the distribution of arguments, such as direct objects, indirect objects, and locatives. Four senses and two meaning facets are then identified and formulated.

Keywords: Polysemy, Word Sense Disambiguation, Corpus, Meaning Facet

---

[1] 本研究蒙國科會補助，計畫編號 NSC91-2411-H-110-021，部分內容曾於第五屆漢語詞彙語意學研討會發表，全文承論文審查委員提供寶貴意見，謹此致謝。

* 國立中山大學中文系
E-mail: tsmei@mail.nsysu.edu.tw

[Received April 13, 2005; Revised June 15; Accepted August 15, 2005]

1. 詞義消歧與自然語言處理

自然語言中普遍存在多義現象，即單一詞形具有數個不同但彼此相關的意義，如華語動詞「送」，可以「送」信，「送」禮，「送」客，也可以「送」命。根據[Li and Huang 1999]的統計，《同義詞詞林》[2]收錄的 52716 目詞當中，多義詞約 7800 個，佔 14.8%；而《現代漢語多義詞詞典》[3]修訂版所收錄的更超過 8150。即使沒有數據統計，但多義詞中一定不乏高頻詞，上述多義動詞「送」便是[4]。

對語言使用者而言，多義詞詞義不難理解，例如華語人士可以毫不費力地分辨下列五句「送」的詞義：(1a)遞送，(1b)贈送，(1c)陪送，(1d)斷送，(1e)推送。

(1) a. 把手伸出來，他還會抓住你的手往嘴裡送。
b. 花也可以送。
c. 最後她哭了，也不許別人送，一個人回家去了。
d. 你把孟瑞的命送了。
e. 貝珍把東尼的小包往瑟勒思娣面前一送。

反觀自然語言處理，多義詞的詞義辨識，也就是所謂的詞義消歧(Word Sense Disambiguation)，卻是資訊擷取自動化極重要的環節，電腦遇到多義詞，必須根據上下文自動排除歧義，才能正確分析、解讀語料。

早期詞義消歧仰賴專人撰寫規則，耗費時日，能處理的詞項極爲有限。八０年代以降，詞典成爲詞義消歧主要的依據，電腦根據詞典中的詞義解釋自動取得辨識詞義的訊息。詞典解歧法落實了詞義自動抽取和判讀的理念，但如[Li and Huang 1999]所述，一旦上下文異於詞典的定義，消歧效果便大打折扣，[Lam *et al.* 1997]利用《現代漢語詞典》[5]的詞義解釋和《詞林》的義類代碼處理多義實詞，詞義辨識率僅 45.5%。鑑於詞典無法預測可能出現的語境，也無法提供所有的組合特徵，利用電腦自動取得大量組合知識的語料庫解歧法於焉產生，如[李涓子 1999]在詞典的基礎上結合語料庫，多義詞辨識率提高爲 52.13%；[Chen *et al.* 1999]同樣藉由語料庫解決自動翻譯所面臨的歧義和多義問題，效能也見提昇。

儘管一詞多義相當普遍，但在實際語句中起作用的大多只有一個。認知語言學家[Choueka *et al.* 1983]實驗證明，人類僅憑上下文中一個或少數幾個成分就能分辨多義詞

[2] 以下簡稱《詞林》。
[3] 以下簡稱《多義詞詞典》。
[4] 「送」在《漢語水平詞彙與漢字等級大綱》中列爲甲級詞，使用頻率在北京語言學院編《現代漢語頻率詞典》中高居 279，而在中央研究院編《中文書面語頻率詞典》中也高居 406。參見[Dew 1999]，153。
[5] 以下簡稱《現漢》。

的詞義，[Tabossi *et al.* 1993]稱之爲獨有特徵(characteristic feature)，如 *port* 的兩個義項「（安全的）港灣」和「（紅）葡萄酒」，可視語境中出現 *safe* 或 *red* 判定。因此除了詞典和大型語料庫之外，晚近的消歧策略進一步將搭配成份的語法訊息納入考量。如此一來，詞性相同的多義詞可以經由比較搭配成份的語法功能和語意類型，達成詞義辨識，如[王惠 2002]辨別華語名詞「辦公室」的兩種意義，[Chuang 2003]區隔英語動詞 *set* 和華語動詞「擺」的多重意義。[Ahrens *et al*. 2003] 進一步利用 MARVS[6]模組屬性動詞語意表達模式，提出多義詞詞義(sense)與義面(meaning facet)的辨識標準。

值得注意的是，具有詞義辨識功能的成份未必緊鄰關鍵詞出現，如例(1)中「送」多處句尾，搜尋關鍵詞前接成份，只有(1d)搭配的「命」和(1e)的「一」具有辨義作用，其他三句出現的「裡」、「可以」及「別人」不足以消歧。可見檢索範圍侷限在關鍵詞前後的緊鄰成份，不但無法全面勾勒使用語境，也無法呈現不同語境間的相關性，如下列兩則例句，一則動詞後接量詞「個」，一則後接動貌助詞「了」，這些訊息既無法區別詞義，也無從得知兩則句式相同。

(2) a. 送個 E-Mail 給我們。
b. 我的大學同學呀，就送了一個東西給我。

同理，下列例句經過語境檢索，只能顯示動詞可以後接介詞「給」和代詞「他」，無法揭露句式間的相關性，也無從比較不同句式的使用頻率。

(3) a. 有一天，一個朋友送給他一束鮮花。
b. 我想送他一樣東西。

[Vu *et al.* 1998] 經實驗發現，透過語境中的句構訊息，歧義辨識效果更佳。針對當今漢語語料庫的設計，[黃居仁 2003]指出兩項詞義研究的瓶頸：一、一般語料庫僅能就關鍵詞的語境進行羅列、排序和統計，分析歸納有賴人力逐筆檢視；二、大型語料庫有利於處理低頻詞義和用法，但高頻用例檢索起來動輒上萬，耗費可觀的研究成本。解決之道可參考英國不來頓大學建置的 WASPS 網站[7]所提供的詞彙素描(word sketch)功能，搭配成份不列詞類訊息，而是按句法功能分類。本文以華語動詞「送」的多義辨識爲例，展示現有消歧法的效能與侷限，再由句型出發，探究詞義與句式間的呼應關係，體現詞彙素描功能的重要性。

---

[6] 即 The Module-Attribute Representation of Verbal Semantics 縮寫。

[7] 網址：http:// wasps.itri.bton.ac.uk/

2. 現有消岐法的效能與侷限

2.1 詞典詞義項目不等

綜觀詞典釋義條例，「送」的詞義項列舉不一。《漢語大詞典》收錄十四條，《詞林》、《現代漢語八百詞》[8]和《現漢》收三條[9]，《多義詞詞典》收四條[10]。反觀使用實況，中央研究院現代漢語語料庫[11]羅列的 1334 則用例中，出現遞送、贈送、送行、斷送、遣送、推送六種詞義，只有《漢語大詞典》全數囊括。

2.2 詞典語法訊息不全

《八百詞》針對「送」的三個義項「遞送、贈送、送行」[12]逐項提供例句，先後介紹動詞後接直接賓語 DO、間接賓語 IO，介詞「給、往」，助詞「了」，結果補語 RC 及趨向補語 DC 七種用法，詳見下表：

表 1:《現代漢語八百詞》多義詞「送」後接成份

句式 \ 義項	遞送	贈送	送行
送＋DO	√	-	√
送＋IO	-	√	-
送＋給	√	√	-
送＋往	√	-	-
送＋了	√	√	√
送＋RC	√	√	
送＋DC	√		√

由動詞後搭配成分與詞義間的對應關係可以發現，具有辨義作用的成分只有間接賓語 IO 和介詞「往」兩種，爲贈送義和遞送義特有。一旦搭配成分句法功能相同，可以進一步分析語意屬性。上表顯示遞送義和贈送義搭配成分有兩處雷同：第一，都可以接「給」再帶 IO，但遞送義跟所有權轉移無關，搭配的 IO 是移位目標，如(4a)的「秘書室」，而贈送義涉及所有權轉移，所以搭配的是接受者，如(4b)的「小華」；第二，兩者都可以帶 RC，但遞送義搭配的是動詞，如(5a)的「走」，贈送義則搭配形容詞，如(5b)的「多」。遞送義和送行義後接成份也有兩處雷同：首先，兩者都可以帶 DO，但遞送義搭配無生賓語，送行義則搭配有生賓語，比較(6a-b)的「貨、他」；其次，兩者都可以帶 DC，但(7a-b)由「把」帶出的 DO「這封信、我」也呈現無生、有生對比。

---

[8] 以下簡稱《八百詞》。

[9] 即遞送義、贈送義、送行義。

[10] 除遞送、贈送、送行三則義項之外，增列斷送義。

[11] 以下簡稱中研院語料庫，網址：http://www.sinica.edu.tw/~tibe/2-words/modern-words/。

[12] 義項名稱各詞典互有出入，《八百詞》作「運送/傳送、贈送、送別」，《現漢》作「送行/送別、輸送/遞送、餽贈」，《詞林》作「遞送、贈送、送別」，爲求論述方便，本文統稱「遞送、贈送、送行」。

(4) a. 文件送給秘書室。 b. 鋼筆我送給小華了。

(5) a. 把小孩兒送走。 b. 送多了。

(6) a. 送貨上門。 b. 送他上火車。

(7) a. 把這封信送到他家裡。 b. 他把我一直送到了家。

詞典根據詞義提供例句，由於篇幅限制，一般無法收錄所有相關義項和用法，只能擇要列舉。如先前例(1d-e)呈現的斷送義和推送義，《八百詞》《現漢》《詞林》等當代詞典俱未編列[13]。而例(5a)(8a)顯示遞送義可以搭配屬人賓語「小孩兒、他」，(9b)則顯示贈送義可以後接動詞「完」作結果補語，與上述歸納「遞送義搭配無生賓語，贈送義搭配形容詞補語」不盡相符。

(8) a. 那一天，我替男孩叫了計程車送他回去。
b. 送完禮物後，接著進行的是壽星才藝表演。

電腦辨識詞義需反向操作，端賴用例匯集組合訊息，藉以指認詞義，如果詞典編列義項過繁或過簡，用法舉例不全，將影響辨識成效。語料庫廣泛提供真實用例，正可彌補詞典的不足。

3. 語料庫句式分佈統計

3.1 語料檢索

透過中央研究院現代漢語平衡語料庫所提供的線上檢索功能，輸入關鍵詞「送」可取得675則用例。但基於分詞標準，「送」後接介詞及後接補語多視爲合成詞，不予切分[14]，如「送給、送往、送上、送到」等用法需個別搜尋。[15]影響所及，以「送」搜尋只能取得「送了出去」這種用例，「送出去、送出」則需輸入不同關鍵詞才找得到。[16]

合詞處理是否有助於詞義辨識？檢視上述四個合成詞，「送往、送上」表單一義項

[13] 以台灣中央研究院現代漢語平衡語料庫的檢索結果觀之，斷送義和推送義兩者出現比例都不及百分之一，參見下節表二。

[14] 參見[詞庫小組 1996]，頁 24-26，18-20。

[15] 少數如「送」後接介詞「由」，後接補語「完、錯、重、一點兒」切分處理。

[16] 他如「送到、送至」，依分詞標準屬合成詞，但仍有部分用例切分處理，前者用例 163 則合，6 則分；後者用例 7 則合，23 則分。分詞人員語感分歧可見一斑。

遞送，參見例(9)，但其實是介詞「往」、趨向補語「上」只搭配遞送義。「送給、送到」則屬多義，前者如(10)所示，可兼表遞送、贈送，後者如(11)所示，可兼表遞送、送行。

(9) a. 老人家要被送往另一個地方。
b. 陳代縣長非常高興接受，並且送上一個小紅包。

(10) a. 把計畫送給政府審查。
b. 李冰曾把他的女兒送給江神作妻子。

(11) a. 我們連夜把他送到醫院。
b.父親把客人送到前廳。

由此可知合詞處理對多義詞詞義辨識幫助不大。

3.2 各義項出現率

經由多次輸入動介、動補合成詞檢索，[17]得 659 則用例，與單詞用例加總，計 1334 則，各義項分佈按比例高低表列如下：

表 2:「送」各義項出現頻率

頻率　義項	遞送	贈送	送行	遣送	斷送	推送
1334	637	522	124	37	9	5
100%	47.75%	39.13%	9.3%	2.78%	0.67%	0.37%

全部用例一共出現六個義項，分佈比例差異懸殊。前兩個義項「遞送、贈送」出現率合計 86.88%，加上排名第三的送行義已高達 96.18%，其他「遣送、斷送、推送」三個義項出現率加總不及 5%，無怪乎一般詞典只收前三個義項。

3.3 各義項句式分佈

本節根據動詞後接成分的組合關係，觀察多義詞「送」各義項的句式分佈，語料中一共出現二十五種結構形式。

A. 共有句式

二十五種後接結構中，十種爲不同義項所共有，依出現率高低列表如下：[18]

[17] 動介合成詞如「送給、送往」，動補合成詞如「送到、送至、送達、送抵、送來、送去、送進、送回、送上、送出、送入、送下、送返、送上來、送上去、送出去、送回去、送回來、送進去」。

[18] 表內左欄句式成分代號表補語，VP 表動詞組，L 表處所。

表 3:「送」各義項共同句式 71.74%

句式 \ 頻率	遞送	贈送	送行	遣送	斷送	推送	總計
送+C	33.96%	0.68%	0.68%	0.38%	0.22%	0.07%	35.99%
送+DO	2.2%	13.1%	0.82%	-	0.37%	-	16.49%
送+DO+VP	0.68%	0.9%	5.4%	1.43%	-	-	8.41%
送	1.35%	1.95%	0.68%	-	0.07%	0.3%	4.35%
送+DO+到 L	0.75%	-	1.13%	-	-	-	1.88%
送+給 IO+VP	0.3%	1.35%	-	-	-	-	1.65%
送+DO+給 IO	0.07%	1.43%	-	-	-	-	1.50%
送+VP	-	-	0.22%	0.52%	-	-	0.74%
送+DO+到 L+VP	-	-	0.07%	0.45%	-	-	0.51%
送+DO+給 IO+VP	0.15%	0.07%	-	-	-	-	0.22%
總計	39.46%	19.48%	9%	3%	0.66%	0.37%	71.74%

十種句式中僅動詞後接補語「送＋C」爲所有義項共有，參見例(12)；其次爲動詞居句末的句式，如例(1)所示，出現義項多達五個，但使用率不及「送＋DO」和「送＋DO＋VP」兩種出現四個義項的句式。

(12) a. 我很高興地拜讀了石川女士送來的資料。
b. 尤其我們認識不久，送重了太唐突，太輕了又沒意義。
c. 許多人都盼望在歲末能送走霉運。
d. 華僑們時興將妻小送回家鄉，讓孩子能接受完整的中國教育，
e. 要不是大白鵝趕來救他，可能送掉一條命。
f. 蘇普迷迷糊糊的送出一刀，正好刺中在狼肚腹上柔軟之處，

上述句式雖然可表多重義項，但出現頻繁高的只侷限於一個義項，如「送＋C」絕大多數表遞送，「送＋DO」表贈送，「送＋DO＋VP」表送行。其他六種句式都出現兩個義項，使用率俱不超過總量 2%。

B. 獨有句式

「送」後接成分二十五種組合當中，有十五種爲單一義項專屬句式，其中七種表遞送義，六種表贈送義，兩種表送行義，表列如下：

表 4:「送」各義項特殊句式 28.26%

遞送	用頻	贈送	用頻	送行	用頻
送+L+VP	3.45%	送+給 IO	10.49%	送+DO+C	0.22%
送+往 L+VP	2.5%	送+IO+DO	5.7%	送送	0.07%
送+往 L	1.35%	送+IO	1.8%		
送+L	0.9%	送+給 IO+DO	1.35%		
送+往 L+DO	0.07%	送+IO+VP	0.15%		

送+由 IO+VP	0.07%	送+予 IO	0.07%		
送+DO+到 L+給 IO	0.07%				
總計	8.41%		19.56%		0.29%

遞送義七種獨有句式中六種含處所 L，唯一接 IO 的句式則由「由」帶出；贈送義六種獨有句式皆含 IO。「遣送、斷送、推送」三個義項沒有專屬句式。

C. 句式用頻排序

表 5.1:「送」各義項句式排序

排序 句式	傳送 637	%	贈送 522	%	送行 124	%
一	送+C	70.9	送+DO	33.5	送+DO+VP	58
二	送+L+VP	7.2	送+給 IO	26.8	送+DO+到 L	12
三	送+往 L+VP	5.2	送+IO+DO	14.6	送+DO	9
四	送+DO	4.6	送	5	送	7.3
五	送	2.8	送+IO	4.6	送+C	7.3
六	送+往 L	2.8	送+DO+給 IO	3.6	送+DO+C	2.4
七	送+L	1.9	送+給 IO+DO	3.45	送+VP	2.4
八	送+DO+到 L	1.6	送+給 IO+VP	3.45	送送	0.8
九	送+DO+VP	1.4	送+DO+VP	2.3	送+DO+到 L+VP	0.8
十	送+給 IO+VP	0.6	送+C	1.7		
十一	送+DO+給 IO	0.2	送+DO+給 IO+VP	0.4		
十二	送+DO+給 IO+VP	0.2	送+IO+VP	0.4		
十三	送+DO+到 L+給 IO	0.2	送＋予 IO	0.2		
十四	送+由 IO+VP	0.2				
十五	送+往 L+DO	0.2				

表 5.2:「送」各義項句式排序

排序 句式	遣送 37	%	斷送 9	%	推送 5	%
一	送+DO+VP	51.4	送+DO	55.6	送	80
二	送+VP	18.9	送+C	33.3	送+C	20
三	送+DO+到 L+VP	16.2	送	11.1		
四	送+C	13.5				

4.「送」多重詞義辨識

4.1 各義項句式特點

A. 遞送義

「送」的多重義項當中，遞送義所搭配的句式最多樣，十五種句式當中又以後接補語用法最突出，使用率高達 70%。另兩種常見句式都是後接「（往）L＋VP」，差異在於處所是否由介詞「往」帶出。以上三種用法合計超過 80%。賓語的分佈值得注意，直賓使用率較間賓頻繁，可以單獨出現，也可以接「到 L、VP 或 IO，間賓則一定後接 VP，沒

有獨現用例。至於連動句式出現率近 15%，關鍵詞和第二動詞組之間可插入移動客體，如(13a)「便當」，目標處所，如(13b-c)「銀行、縣府民政局」，接受者，如(14a)「政府」，或決策者，如(14b)「警備總部」。例(15)顯示「給」無法換成「由」，證明兩者句式功能不同，第二個動詞「看」與決策無關，不適合「送由…處理」句式。

(13) a. 學校不准校外餐廳送便當進校園。
b. 在兩週後學校獲知結果，才能送銀行審核。
c. 名單早在十天前即送往縣府民政局審核

(14) a. 把計畫送給政府審查。
b. 當事人送由警備總部依法嚴辦。

(15) a. 農友蔡漢坤在田裡拔了一把蒜苗送給主持人看。
b.*農友蔡漢坤在田裡拔了一把蒜苗送由主持人看。

B. 贈送義

搭配贈送義的十三種句式當中，以後接直賓最常見，佔 33%，間賓 30.9%次之，兩者差距不大。雙賓同現居第三位，達 17.7%。前三種用法加總起來超過 80%。各類句式當中，「給」和間賓的搭配值得矚目：間賓單獨出現的情況下，帶「給」的比例 26.4%遠高於不帶「給」4.5%；但如果間賓後再接直賓，不帶的比例 14.3%反而比帶 3.4%更頻繁。至於連動用法不如遞送義發達，只有 6.5%。

(16) a. 熱情的民眾送兩箱草莓慰問。
b. 我帶回來一部佛經，就送給你做紀念吧。

C. 送行義

搭配送行義的九種句式當中，連動用法相當發達，出現率高達 61%，以後接「DO＋VP」最普遍(17a)，佔 58%。其次是「DO＋到 L」12%，後接 DO 佔 9%，名列第三。以上三種句式合計超過 80%。送行義可以緊接第二動詞組(17b)，但無法插入動作接受者，這一點與遞送義、贈送義不同，比較(14a)、(16b)。

(17) a. 有一個朋友送他上飛機。
b. 眼看媽媽向外走，邱子章無言一直送下樓。

D. 遣送義

搭配遣送義的四種句式當中，連動用法一枝獨秀，使用率超過 85%，用法以接「DO＋VP」爲大宗(18a)，出現率過半，也可以緊接動詞組(18b)。雖然遣送義的用法送行義都有，但送行義還可以獨自出現，或獨帶 DO。

(18) a. 目前送留學生出國留學還是需要。
b. 不但這樣，還有能力將子孫一個個送出國留學。

4.2 詞義分析

根據句式分佈，尤其是表四所呈現的特殊句式，可以確立「遞送、贈送、送行」爲三個義項，詞彙語意皆涉及動作主事者(Agent)，因動作而產生位移的客體(Theme)，以及移動目標(Goal)。三者間的差異在於遞送義的目標爲處所(Location)，贈送義的目標爲接受者(Receipient)，送行的客體爲屬人(Human)。

送 遞送義 <Agent, Theme, Goal>
|
[location]
贈送義 <Agent, Theme, Goal>
|
[receipient]
送行義 <Agent, Theme, Goal>
|
[human]

遞送義根據目標的屬性，諸如目的地(Destination)、機關(Organization)或方向(Direction)，參見下列用例，區分「傳送、送交、推送」三層義面。

(19) a. 吳德貴率先五指一抓一捏就往口裡送。
b. 有一農婦採集殞石後送莫斯科研究。
c. 左手向前一送，藏在衣袖中的匕首已刺了出去。

送 遞送義 <Agent, Theme, Goal>
|
[location]

Goal [destination] → 義面一：傳送
Goal [organization] → 義面二：送交
Goal [direction] → 義面三：推送

送行義根據目標的屬性，諸如出發地(Departure)、目的地或活動(Activity)，區分「陪送、護送、遣送」三層義面。其中陪送義無法進入被動句，護送義可以，參見下列例句，遣送義則如例(18)所示，一定後接 VP。

(20) a. 爸爸親自送我們到車站。
b.* 我們被爸爸親自送到車站。
c. 雙雙昏倒在地，被救護車送到醫院。

送 送行義 <Agent, Theme, Goal>
|
[human]
Goal [departure] → 義面一：陪送
Goal [destination] → 義面二：護送
Goal [activity] → 義面三：遣送

斷送義無法如遞送義和贈送義後接目標成分，如例所示需前置，也不似送行義僅搭配屬人賓語，視爲獨立義項。

(21) 我差點爲此送了命。

送 斷送義 <Agent, Theme, Goal >
|
[life]

以上根據詞項所搭配的句式劃分詞義，擁有特殊句式者列爲獨立義項，擁有共同句式者則列爲同一義項下的不同義面。經由相關句式統計，多義詞「送」可切分爲遞送、贈送、送行、斷送四個義項，遣送、推送兩項則分別歸入送行義、遞送義，作義面處理。結果證明，句式搭配訊息確實能夠在詞義辨識過程中起一定作用。至於這種以句式爲本位的詞義辨識法是否適用於其他多義動詞，是否對其他詞類也能有效應用，還需大規模地考察和驗證。其間語料庫如能提供句法功能訊息，必能大幅提高句式分析時效，加速建成詞義句式對應知識庫，根本強化自然語言處理中的詞義消歧資源。

參考文獻

中國社科院語言所詞典編輯室編，《現代漢語詞典》繁體字版，香港：商務印書館，2001。

王惠，"基於組合特徵的漢語名詞詞義消歧"，*CLCLP* 7(2), 2002, pp.77-88.

呂叔湘，《現代漢語八百詞》，北京：商務印書館，1980。

李涓子，《漢語詞義排歧方法研究》，博士學位論文，清華大學圖書館，1999。

袁暉主編，《現代漢語多義詞詞典》，太原：書海出版社，二版，2001。

黃居仁，〈語言的網路，知識的架構：由意義誘發的語言科技〉，台灣師範大學主辦第三屆語言與科技研討會邀請講席，2003 年 12 月 20 日。

梅家駒、竺一鳴、高蘊琦等，《同義詞詞林》，上海：上海辭海出版社，1983。

詞庫小組，《「搜」文解字—中文詞界研究與資訊用分詞標準》，台北：中央研究院資訊科學研究所，1996。

Ahrens, K., C.-R. Huang, and Y-H Chuang, "Sense and meaning facets in verbal semantics: A MARVS perspective," *Language and Linguistics,* 4(3) 2003, pp. 469-484.

Chen, H.-H., G.-W. Bian, and W.-C. Lin, "Resolving translation ambiguity and target

polysemy in cross-language information retrieval," *CLCLP,* 4(2) 1999, pp. 21-38.

Choueka, Y., and S. Lusignan, "A connectionist scheme for modeling word sense disambiguation," *Cognition and Brain Theory,* 6(1) 1983, pp. 89-120.

Chuang, Y.-H., "*Sense Distinction of Verbs in English and Mandarin Chinese: An Analysis of the Verbs 'Set' and 'Bai3'*,"MA thesis. National Taiwan Univ., 2003.

Dew, J. E., *6000 Chinese Words. A Vocabulary Frequency Handbook for Chinese Language Teachers and Students*, Taiepi: SMC Publishing, 1999.

Lam, S.-S., K.-F. Wong, and V. Lum, "LSD-C –A linguistics-based word sense disambiguation algorithm for Chinese," *Computer Processing of Oriental Languages,* 10(4) 1997, pp. 409-422.

Levin, B., and M. Hovεv, "What alternates in the dative alternation," *CSSP*, 2001.

Li, J., and C. Huang, "A model for word sense disambiguation," *CLCLP,* 4(2) 1999, pp. 1-20.

Tabossi, P., and F. Zardon, "Processing ambiguous words in context," *Journal of Memory and Language*, 32, 1993, pp. 359-372.

Vu, H., G. Kellas, and S. T. Paul, "Source of sentential constraint on lexical ambiguity resolution," *Memory and Cognition*, 26, 1998, pp. 979-1001.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 495-508

從構式語法理論看漢語詞義研究[1]

A Construction-Bsed Approach to Chinese Lexical Semantics

王 惠*

Hui Wang

摘 要

本文在構式語法理論背景下，結合漢語具體實例，探索以一種新視角與新思路來進行詞義研究。一方面，從上向下看（up-down），充分利用詞的不同結構中的表現來觀察詞義的細微差別，在整體結構中挖掘詞義特徵、描述詞義的分佈特點與動態變異，力圖通過大量的微觀分析、比較，得到規律性的認識，爲大規模的漢語詞義組合分析探索出一條路子，同時在詞義理論上有一些新發展。另一方面則從下向上看（down-up），借鑒詞義系統的分析方法，描述和解釋句法結構之間存在的各種語義關係，如多義、同義、反義、同形等現象。這在以前的漢語研究中也是沒有的，值得深入探索。

關鍵字：構式語法，形式-意義關係，結構義，詞義

Abstract

This paper, set within the theoretical framework of Construction Grammar, aims to pave the way for a comprehensive investigation of the behavior of Chinese lexical meaning with a focus on word sense and structure sense. Taking a top-down view, we discuss observed nuances in the meaning based on different syntax distributions, and describe the related semantic features, distributions and observed dynamic change. In addition, taking a bottom-up view, we use the analytical methods of the

[1]本研究受到新加坡國立大學學術研究基金 (No. R -102- 000 - 029 - 112) 資助。初稿曾在第五屆詞彙語義學研討會（CLSW5，新加坡，2004 年 6 月）上宣讀。先後承蒙李英哲教授、黃居仁教授、連金發教授等提出寶貴意見和建議，在此一併致謝。

* 新加坡國立大學中文系，Dept. of Chinese Studies, National University of Singapore
E-mail: chswh@nus.edu.sg

[Received April 10, 2005; Revised June 12, 2005; August 15, 2005]

semantic system and also describe and explain many kinds of semantic relations between different structures, such as polysemy, synonymy, antonym, and homonymy.

Keywords: Construction Grammar, Form-Meaning, Structure sense, Lexical Sense

1. 構式語法理論

最近十多年來，國際語言學界越來越多的學者開始採用構式語法（Construction Grammar，CG）理論進行語言分析。該理論的主要觀點可概括爲：

（1）語言的基本單位是形式-意義結合體——結構式（constructions）

結構式，本是傳統語法中的一個術語，自 Aristotle 時代以來，一直是語言研究的主要物件。現在，CG 理論對其重新作了定義，用來指稱語言中各種約定俗成的"形式-意義"結合體，如語素、合成詞、成語、習慣用語和半固定短語，也包括搭配格式與句式等等。任何語言表達式，只要它的形式、意義或用法不能從其組成成分或其他結構式中推知出來，就都屬于"結構式"的範圍。

（2）結構式本身具有獨立于其組成成分——詞彙之外的意義

構式語法認爲，整體意義大于部分之和，句子意義不能只根據組成句子的詞彙意義推知出來，句法結構本身也表示某種獨立的意義。比如，雙賓結構"Subj [V Obj1 Obj2]"(*He gave her a cake; He baked her a muffin.*) 表示"有意識地給予"(intended transfer or giving)。CG 理論還進一步指出，就像詞彙中存在多義詞和同義詞一樣，句法結構既有多義的（有基本義和隱喻引申義），也有同義的。

句法結構上的任何差異，都直接反映了所表達的意義不同。需要指出的是，"意義"在這裏既包括語義資訊，也同時包含焦點、話題、語體風格、方言差異等語用內容。因爲它們和語言形式的關係都是約定俗成的，是句法結構本身所具有的表達功能。

（3）構式語法的目的是全面描述語言事實（full range of facts）

在這一點上，構式語法與當前的主流理論——生成語法之間存在著很大的分歧，后者忽略不同形式之間的語義差異，認爲語言形式是獨立于語義或話語功能而獨立存在的自主系統，對"核心語法"中那一部分簡單、規則、普遍的語言形式的研究就可以揭示語言本質。構式語法則特別強調語義，非常重視所謂的"邊緣語法"。目前研究熱點更是集中在詞彙語義和不常用標記句式上，認爲對這些特殊格式所表達的豐富語義/語用信息的研究將有助于揭示語言普遍規律，從而全面解釋各種語言現象。

總的來說，構式語法既有一種開放的語言哲學觀，又具有一套嚴格的基于合一運算的形式表示方法，主張把語法與詞彙、語義、語用、韻律作爲一個整體來分析，這無疑是給當前的語言研究帶來了一種新的視野和新的方法。利用它，我們可以來解釋一些先前不好解釋、或先前想不到去解釋的語言現象[Goldberg 1995, 2002, 2003]。

近年來，構式語法已經開始引起漢語言學界的日益關注。Kathleen Ahrens [1995]、

Chao-ran Chen, Chu-Ren Huang, and Kathleen Ahrens [1995]、黃居仁等[1999]、張伯江[1999, 2000]、沈家煊[1999, 2000]、劉丹青[2001]、李淑靜[2001]、陸儉明[2002,2004]等將該理論方法與漢語實際緊密結合起來，對漢語一些特殊句式進行探索，不僅有益于漢語語言學的理論建設，也更深入地發掘和解釋了漢語裏固有的事實和規律。

但遺憾的是，運用構式語法理論進行漢語詞義分析的成果至今還很少見到，代表性作品只有黃居仁等[1999]。實際上，構式語法很大部分是從 Fillmore 的框架語義學(frame semantics)和認知語義學[Lakoff 1987]發展出來的，特別強調語義。Fillmore, Kay and Connor[1988]甚至明確提出地把構式語法的研究重心放在對詞彙語義的研究上。本文在這一方面做一些初步探索。

2. 結構義與詞義

按照現有的詞彙語義學理論，句法表現是主要動詞的投射結果。比如，give 是三價動詞，在句中就要求 3 個論元角色與之共現：給予者(或施事), 接受者, 給予的物體(或受事)。

（1）He gave her a cake.
他給她一塊蛋糕。

可是，我們知道，有時非三價動詞也可以出現在雙賓句子中，如：

（2）He sneezed the napkin off the table.
他打噴嚏把餐巾紙吹到桌子下麵了。

（3）She baked him a cake.
她爲他烘焙了一個蛋糕。

例（2）中的 sneeze 是個典型的不及物動詞，例（3）中的 bake 本來也不能帶間接賓語。爲了解釋這兩個句子的合法性，就不得不爲動詞增加額外的特殊義項：sneeze 表示“X CAUSE Y to MOVE Z by sneezing”（X 通過打噴嚏使 Y 移動到 Z）；bake 表示“X INTENDS TO CAUSE Y to HAVE Z by baking”（X 通過烘焙的行爲有意識地使 Y 擁有 Z）。

如此一來，sneeze，bake 都成了含有歧義的動詞：基本義及其在上述句式中產生的臨時動態義。這樣處理顯然難以令人信服，不僅與我們的語感相違，也難以解釋這樣的現象：這些用法只出現在特定的句式中。

相比之下，構式語法對這種現象的處理則更加合理。它認爲整體意義大于其組成成分意義的簡單相加，句法結構本身具有某種獨立于詞義之外的意義。比如，傳遞、致使移位、結果義都是雙賓結構本身具有的結構義，該結構式中要求有 3 個論元：施事、受事和接受者。這樣，就不必爲 sneeze、bake 專門設立一個僅僅適用于該結構的特殊詞義了。

這種分析給我們很大的啓發：應重視結構義與詞義的區分，不能把結構義強加到詞義上。

如果用這一標準來看現有的漢語詞典釋義，類似于 sneeze，bake 的釋義錯誤還真不少，下面我們結合具體實例來進行分析。

漢語中，“有+光杆名詞 N”具有一種特殊的結構義。對此，《現代漢語八百詞》（增訂本）作了明確的說明：

【有】① 表示領有，具有。b)有些名詞跟“有”結合，…有程度深的意思。
他可是～年紀了｜這個人～學問｜你比我～經驗。

譚景春[2000]進一步指出“程度深”的意思是：不僅領有名詞所表示的那種事物，所領有的那種事物數量大、品質好。例如：

這個人有學問＝這個人的學問大
有年紀了＝年紀大了
有經驗＝經驗多
有年頭兒＝年頭兒多
有人緣兒＝人緣兒好

對于這種現象，《現代漢語詞典》（2002 增補本）卻錯誤地把這種結構義當作動詞“有”的詞義，並爲此專門設立一個義項：

【有】⑤表示多，大：～學問｜～經驗｜～年紀了。

顯然，“有”的“表示大、多”的意義僅僅適用于“有＋光杆名詞 N”這個特殊結構式，離開了這個結構，“有”就沒有該義，如：

有一丁點兒學問
有兩套西服
有了一張桌子

有時，《現代漢語詞典》（修訂本）還把這種整體結構義歸入有關名詞的詞義中，如：

【年頭兒】②多年的時間：他幹這一行，有～了

“多年的時間”顯然不是“年頭兒”的意思，其中“多年的”是“有＋年頭兒”這個結構產生的。其實“年頭兒”就是“時間”的意思，因爲還可以說：

年頭兒不短。
年頭兒不長。

類似的例子還有不少，如：

【風度】美好的舉止姿態：有～｜～翩翩。

儘管“風度”現在多用于好的方面，但基本上還是個中性詞。因爲它可以和“氣質”或“舉止”對舉，還可以說“風度欠佳”。“風度”本身並不含有“美好的”意思，“美好的”是“有＋風度”這個結構賦予的。

【人緣兒】跟人相處的關係（有時指良好的關係）：沒～｜有～｜～不錯。

“人緣兒”只有在“有人緣兒”中才指“良好的關係”，“良好的”也是結構帶來的。因爲我們也還可以說“他的人緣兒很差”。

此外，因爲沒有自覺地區分結構義和詞義，還造成了《現漢》釋義體例不一。同樣的現象，有的增加了義項，有的又沒有，主觀隨意性較大。比如：可以出現在“有＋光杆名詞”這個結構中的有些名詞則沒有像設立一個專門的義項：

【日子】②時間（指天數）：他走了有些～了。

“有日子”也可以表示“有好長時間、時間長”的意思，比如“咱們有日子沒見面了”。詞典中並沒有給“日子”增加一個義項“多天的時間”。類似的情況還有：

【學問】②知識；學識：有～。
【年紀】（人的）年齡。
【經驗】①由實踐得來的知識或技能：他對嫁接果樹有豐富的～。
【運氣】①命運。
【錢】 ④錢財：有～有勢。
【眼光】②觀察鑒別事物的能力；眼力：這輛車挑得好，你真有～。

因此，我們在分析詞義時，要注意區分哪些是結構意義，哪些是詞彙意義。結構義是詞的臨時言語義的產生途徑之一，但只有那些擺脫了特定結構的意義，才能是詞義，在詞典中設立單獨的義項。這對提高詞典釋義的質量也是非常有益的。

3. 動態義與詞義

詞義對客觀事物的反映是抽象的概括。用概括的、數量有限的詞語去表達具體的、無限的事物，爲了解決這個矛盾，更準確地表達思想，人們就有意識地將詞義靈活變通使用，使之產生種種變異。一個詞本來沒有某項意義，使用到某一個特定的結構中卻可以表達某一項意義，這就是我們通常所說的“動態義”（也稱“臨時義”、“言語義”）。

古代注疏中的“隨文釋義”、“緣事生訓”、“因文詁義”，不少揭示的就是詞義中這種臨時產生的動態義，其中常見的一種詞義變異就是“詞類活用”引起的。如：

（4）夫子所謂生死而肉骨。（《左傳·襄公二十二年》）

（5）孟嘗君客我。（《戰國策·齊策》）

例（4）中的“肉”是名詞的使動用法，義爲“使……長肉”；例（5）中的“客”是名詞的是意動用法，義爲“以……爲客”。

這種詞類活用現象不僅在古代漢語中經常可以遇到，現代漢語中也仍存在。如：

（6）老拴，就是運氣了你。（魯迅《藥》）

（7）密密的人群混濁了空氣。（王安憶《本次列車終點》）

例（6）中的“運氣”、例（7）中的“混濁”由于處于“SVO”格式中的謂語位置上，臨時改變了詞性，同時引起意義的變化。

這種現象顯示了人們語言交際中的靈活性，但所產生的詞義必須在特定的結構式中才能應用，離開了上述語法位置，就不再存在。因此，不能把它們看作詞本身的意義，詞典裏也不應把這些由于臨時產生的動態義處理爲單獨的義項。下面，讓我們由此視角來看看漢語中的一些具體現象。

表示可量化的（measurable）屬性值的反義詞兩方在分佈上是不平衡的，表示程度高的或‘積極’意義的一方的詞義範圍可以包括消極意義一方的，相當于整個概念範圍。對于這種現象，很多學者都已經注意到了，黃國營、石毓智[1993]的描述更是非常細緻：“有反義關係的一對詞的其中一方承擔了另一方的任務，……比如，問句‘小張有多高’不論是多高多低都問到了，這時的‘高’就攝入了‘低’的語義。表示量大的、多的可以攝入表量小的、少的語義，反過來就不行了”。

上述描寫與觀察無疑是很準確的，但反義形容詞之間爲什麼會存在這種詞義差異？以前的研究沒有談到。我們認爲，其主要原因在于二者的分佈範圍不同，所謂表示程度高的或‘積極’意義一方的詞義實際上是詞義本身加上結構義。因爲，它們可以自由地進入表示估量或比較的結構式“有+數量 NP+ Adj”，而表示程度低的或‘消極’意義一方的形容詞則不可以，試比較：

（8）a.這張桌子有兩米長。　　b. *這張桌子有兩米短。

（9）a.這條魚有三斤重。　　b. *這條魚有三斤輕。

（10）a.這口井有十米深。　　b. *這口井有十米淺。

（11）a. 小李有他哥哥高。　　b. *小李有他哥哥矮。

（12）a. 這個禮堂有五個教室大。　b. *這個禮堂有五個教室小。

在這些特殊的結構中，正是因爲有了數量短語表示具體的屬性值，“長、重、深、高、大”等詞才能由屬性值提升爲上層的屬性，分別解釋爲“長度”、“重量”、“深度”、 “高度”、“大小”。如果一旦離開數量短語，這些詞也就只能表示其原有的屬性值意義，而不能代表屬性，從而失去了表示“整個概念範圍”的意義。如：

（13）這張桌子長。

（14）這條魚重。

（15）這口井深。

（16）小李高。

（17）這個禮堂大。

由此可見，程度高的或積極意義的形容詞所表示“整個概念範圍”的意義，實際上是它在“有+數量 NP+ Adj”結構中臨時獲得的動態義，不應看作形容詞本身的詞義，如：

【遠】①空間或時間的距離長（跟‘近’相對）：廣州離北京很～｜眼光要看得～。
【粗】①（條狀物）橫剖面較大（跟‘細’相對）：～紗｜這棵樹很～。
【熱】②溫度高；感覺溫度高（跟‘冷’相對）：～水｜趁～打鐵｜三伏天很～。
【濃】①液體或氣體中所含的某種成分多；稠密（跟‘淡’相對）：～墨｜～茶。
【硬】① 物體內部的組織緊密，受外力作用后不容易改變形狀（跟‘軟’相對）：堅～｜～木｜～煤。

但由于沒有自覺地區分動態義與詞義本身，不同的編纂者對這類形容詞的解釋就難免具有一定的隨意性，《現代漢語詞典》（修訂本）中有時就把動態義誤作詞義，設立一個專門的義項，如：

【長】②長度：南京長江大橋氣勢雄偉，鐵路橋全～６７７２米。
【寬】②寬度：我們國旗的～是長的三分之二｜這條河有一里～。
【高】②高度：那棵樹有兩丈～｜書桌長四尺，寬三尺，～二尺五。
【重】①重量；分量：這條魚有幾斤～？
【厚】②厚度：下了二寸～的雪。
【深】②深度：這裏的河水只有三尺～｜這間屋子寬一丈，～一丈四。
【大】②大小：那間屋子有這間兩個～｜你的孩子現在多～了？
【快】②速度：這種汽車在柏油路上能跑多～？

這種前後矛盾的做法，顯然不利于讀者掌握正確的詞義。

當然，詞義和動態義也並不是截然對立、毫無關係的。事實上，“許多言語義是在語言義的基礎上生發出來的。有一些言語義在一定條件下可以發展成語言義”[符淮青 1996]。例如，“后門”一詞原指“房子或院子后面的門”，后用來比喻舞弊的途徑，這個意義后來廣泛應用，爲社會所接受，能在不同的場合重複出現，成爲語言義。

【后門】 ①房子或院子等後面的門。

②比喻通融的、舞弊的途徑：走～｜開～。

在特定結構式中產生的臨時動態義，在什麼情況下可以進入詞義系統，發展成爲了新的詞義，目前還沒還有一個明確的判斷標準，需要進一步研究。

我們認爲，判斷一個動態義是否變成了新的詞義，要考慮到很多方面，如重複出現的頻率、被語言社團接受的廣泛性等等。但其中一個不可忽視的因素則是該義是否擺脫了某個特殊結構，而具有一個相對獨立的分佈空間。只有那些擺脫了特定結構的意義，才能是詞義。

4. 配價分析

依據目前的主流語言學理論（如 GB, LFG,語義角色或題元理論等等），謂詞在句中處于核心重要的地位，其內在的論元要求（配價）決定了補足語的數量及種類。一價動詞在句子中有一個補足語與之共現，如“走（小麗走了）”；二價動詞有兩個補足語，如“寫（小學生寫作文）”；三價動詞有三個補足語，如“給（小李給她一個蘋果）”。

但謂詞的配價是根據什麼得到呢？許多語言學家對這一問題進行了深入的研究，提出了不同的理論方法。俄羅斯學者阿普列相[1974]認爲，配價是一種語義現象，是從詞彙意義中直接引出的。比如，對“租”這一情景來說，必不可缺少的配價是五個：承租者，所租的東西，出租者，租金，租期。這五項缺一不可。如缺少其中某一配價，“租”這一情景會變爲另一情景。如，缺少“租期”這一配價，“租”就會變成“買”。但另一方面，他在界定語義配價概念時又說，一個詞的語義配價是把句法上相關的一些詞與這個詞結合起來的配價特點。在阿普列相的理論中，語義配價的實質是個沒有說清楚的問題[轉引自：郭聿楷 1999]。

我們覺得，配價顯然是與句法關係相聯繫的現象，動詞與其他詞語的組合，說到底必須在語句中實現。比如，下列例句中的“高”、“吃”是幾價呢？

（18）a 張三高。　　b 張三高李四一頭。

（19）a 他吃了一個蘋果。　　b 他吃了小李一個蘋果。

對此，張國憲[2002]認爲，（18a）裏的“高”只有一個論元（或說“配價成分”），但在（18b）裏則有三個論元，因此可以出現在雙賓結構中。陸儉明[2002]也指出，（19a）裏的“吃”，具有兩個論元：施事和受事；而在（19b）裏則具有 3 個論元：施事、受事和與事，因此可以帶兩個賓語。這也就是說，“高”由一價形容詞變成了三價形容詞；“吃”由二價動詞變成了三價動詞。

顯然，這裏存在一種嚴重的循環論證：首先，根據動詞在句子中與 n 個名詞成分共現說它具有 n 個論元，是 n 價動詞；然後又說，該動詞在句子中之所以要求與 n 個名詞成分共現，是因爲它的配價數爲 n。

對此，構式語法提出了另外一種解釋：b 例中的 3 個論元是直接由雙賓式的整體結構意義所帶來的，與謂詞本身的語義特徵和論元要求無關，因此，謂詞在上述句子中並未發生所謂的論元（“配價數”）變化。張伯江[2000]明確指出了這一點，“動詞的配價觀在揭示句式語義方面是力不從心的，論元結構更多的是句式的要求而不是動詞的要求”。

這種解釋不僅避免了循環論證和隨意性，說一個形容詞（如“高”）是一價形容詞，一忽兒又成了三價形容詞；而且進一步指出了動態論元的存在條件——只有在特殊的結構式（比如“動詞＋名1＋名2”雙賓句式）中，這種情況才有可能出現。

5. 詞義研究方法在語法分析中的應用

“語義和語法”不僅是構式語法的核心，也是其他現代語言學理論所共同關注的研究課題。這是在處理這些基本問題的方法上的不同，展示了構式語法與其他理論的本質差異。

構式語法首次明確提出，語素、詞、習慣用語、半能產的搭配以及句式都是約定俗成的“形式-意義”結合體。作爲連續統的兩端，詞彙和句法結構的本質上是相同的，它們之間沒有嚴格的界限，都是以某種形式表達了人類認知對現實的反映。詞反映了人類認識世界的基本概念，論元結構則反映了相關的動態場景：某人傳遞某物給某人，某物致使某物移動或改變狀態，某人經歷某事，某事經歷了狀態或位置的改變等等。

這種認識令人耳目一新，使我們進一步推論，既然結構式的意義與詞義的形成都跟一個民族的認知方式有關，那麼，句法結構與詞彙之間應該具有一定的共性，比如，都存在多義、同義、反義和同形等現象。這也就是說，詞義分析方法也可以用于語法研究。下面，我們結合一些多義和同形兩個方面的實例加以說明。

5.1 多義現象

與詞彙中存在多義詞一樣，句法結構也可以是多義的，即用相同的形式表示彼此不同但密切相關的一組意義。比如，英語雙賓結構（Subj V Obj1 Obj2）的基本義是“有意的給予性轉移（transfer or giving），如上文中的例（1）。由此，還可以引申出來一些其他相關的意義：

（20）a. Liza guaranteed Zach a book.

Liza 保證給 Zach 一本書。

b. Liza refused Zach a book.

Liza 拒絕給 Zach 一本書。

c. Liza cost Zach his job.

Liza 使 Zach 失去了工作。

(20a) 表示一旦條件滿足，就會發生 transfer，Zach 將會得到一本書； (20b) 表示 transfer 將不再發生, Liza 使得 Zach 不能得到一本書； (20c) 表示給予（giving）的反義關係，即剝奪（taking away），Liza 使得 Zach 失去了工作。但即使是 *Cry me a river* 也可以通過隱喻表示 giving。

與英語的雙賓結構表達的“給予”義不同，漢語的雙賓結構雖然也是表示“有意識地使客體的所有權發生轉移”，但轉移的方向則是多樣的，既可以是“給予”義，也可以是“取得”義，甚至二者皆可[張國憲 2001; 陸儉明 2002]。如：

（21）我給他一本書。

（22）我拿他一本書。

（23）我借他一本書。

從這個角度來說，漢語的雙賓結構是多義的，句子的意義主要由謂語動詞本身的詞義特徵決定。比如，例（21）中的動詞是“給予”類的，整個句子的意義表示物體從“我”到“他”的轉移；例（22）中的動詞是“取得”類的，整個句子的意義表示物體從“他”到“我”的轉移；例（23）中的動詞“借”雙向的，就直接導致了句子的歧義，既可能是“我借給他一本書”，也可能是“我從他那兒借一本書”。

5.2 同形結構式

與多義現象相關的則是同形，即同一個結構式表示彼此無關的一組意義。

陸儉明[2004]提出一個問題：爲什麼層次構造和語法結構關係相同的同樣的詞類序列能形成各自表示不同“句式意義”的不同句式？

我們認爲，其實這正是同形現象在句法中的反映。比如說，“山上架著炮”既可以表示存在，表靜態（意思是“山上有炮”），又可以表示活動，表動態（意思是“山上正有人在架炮”），其原因就在于 “NPL+V+著+NP”結構是存現句的典型形式，同時

又可能是 SVO 語序中的謂語部分（狀中結構）。

類似的情況還有很多，如漢語中的“N+N”結構，既可以表示隸屬關係，也可以表示性質或類屬關係，因此，“大衣扣子”，有兩種意思，一是相當于“大衣上的扣子”，扣子是大衣有機的組成部分，個兒有的大，有的小——如袖口上的扣子就小；另一種則相當于“大衣上專用的扣子”，個兒都是大大的[陸儉明 2004]。

6. 結語

作爲近年來新興的一種語言學理論，構式語法已在國際語言學界越來越多學者的關注。它採用一種開放的語言哲學觀，首次明確提出把詞彙、語法、語義，甚至語用作爲一個整體來分析，既有一套嚴格的基于合一約束的形式描述系統，又對各種語言現象認知-語義基礎進行解釋。其突出貢獻是突破了單純結構分析的局限，使語言中的形式-意義關係得到很好的說明，並追求把描寫和解釋結合起來。

本文正是在這種理論背景下，探索著以新視角與新思路來進行詞義研究，一方面從上向下看，在整體結構中挖掘詞義特徵、論元（配價）要求，描述詞義的分佈規律與動態變異，另一方面則從下向上看，借鑒詞義系統的分析方法，描述和解釋句法結構之間存在的各種語義關係，如多義、同義、反義、同形等現象。這是以前的漢語詞義研究中所沒有的，需要進一步的深入探索。

参考文獻

符淮青，《詞義的分析和描寫》，北京：語文出版社，1996。

郭聿楷，“語義格與語義配價”，《外語與外語教學》，1999 年第 1 期，pp19-21。

黃國營，石毓智，“漢語形容詞的有標記和無標記現象”，《中國語文》1993 年第 6 期，pp 401-409。

黃居仁、張莉萍、安可思、陳超然，“詞彙語義和句式語義的互動關係”，《中國境內語言暨語言學》（臺灣）第 5 輯，1999 年，pp 413-438。

李淑靜，“英漢語雙及物結構式比較”，《外語與外語教學》，2001 年第 6 期。

劉丹青，“漢語給予類雙及物結構的類型學考察”，《中國語文》，2001 年第 5 期。

陸儉明，“再談“吃了他三個蘋果”一類結構的性質”，《中國語文》，2002 年第 4 期。

陸儉明，“詞語句法、語義的多功能性：對“構式語法”的解釋”，《外國語》，2004 年第 2 期，pp 15-20。

沈家煊，“ “在”字句和“給”字句”，《中國語文》，1999 年第 2 期。

沈家煊，“句式和配價”，《中國語文》，2000 年第 4 期。

譚景春，“名形詞類轉變的語義基礎及相關問題”，《中國語文》，1998 年第 5 期，pp 368-377。

譚景春，“詞的意義、結構的意義與詞典釋義”，《中國語文》，2000 年第 1 期，pp 69-78。

張伯江，“現代漢語的雙及物結構式”，《中國語文》，1999 年第 3 期。

張伯江，“論“把”字句的句式語義”，《語言研究》，2000 年第 1 期。

張國憲，“制約奪事成分句位實現的語義因素”，《中國語文》，2001 年第 6 期。

張國憲，“三價形容詞的配價分析與方法思考”，《世界漢語教學》，2002 年第 1 期。

Ahrens, K., “The meaning of the double object construction in Chinese,” In *Proceedings of the 6th North American conference on Chinese Linguistics (NACCL 6)*, 1995, GSIL, USA.

Chen, C.-R., C.-R. Huang, and K. Ahrens, “Construction as a theoretical entity: an argument based on Mandarin existential sentences,” In *Proceedings of the 10th PACLIC*. ed., by Benjamin K. Tsou. 1995, Hong Kong, pp 91-95.

Fillmore, C., P. Kay, and M. O’Connor, “Regularity and idiomaticity in grammatical conditions: The case of LET ALONE,” *Language,* 64, 1988, pp501-38.

Goldberg, A., *Construction: A Construction Grammar Approach to Argument Structure*, The University Chicago Press, 1995,

Goldberg, A., *Construction Grammar. Encyclopedia of Cognitive Science*. Macmillan Reference Limited Nature Publishing Group, 2002.

Goldberg, A., “Construction: A new theoretical Approach to language,” *Trends in Cognitive Science,* 2003.

Lakoff, G., *Women, fire, and dangerous things: what categories reveal about the mind* Chicago: University of Chicago Press. 1987.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 509-518

現代漢語中的形式動詞[1]

Dummy Verbs in Contemporary Chinese

俞士汶*、朱學鋒*、段慧明*

Shiwen Yu, Xuefeng Zhu and Huiming Duan

摘要

現代漢語中包括“加以”、“進行”等在內的一類動詞叫做“形式動詞”。本文在簡述了“加以”和“進行”的詞彙語義之后，比較詳細地討論了它們的語法屬性，還進一步探討了它們作爲句法成分和語義角色的標記的功能。

關鍵字: 形式動詞、詞彙語義、語法屬性、語義角色

Abstract

In contemporary Chinese, there is a subclass of verbs called Dummy Verbs. After briefly introducing the lexical meanings of two typical dummy verb, ‘Jiayi’ and ‘Jinxing’, this paper discusses the grammatical attributes of ‘Jiayi’ and ‘Jinxing’ in detail and further explores their functions as markers of syntactic constituents and semantic roles.

Keywords : Dummy Verb, Lexical Meaning, Grammatical Attribute, Semantic Role

1. 形式動詞的所指

北京大學計算語言學研究所開發的《現代漢語語法信息詞典》（以下簡稱《語法信息詞典》）收錄了若干“**形式動詞（Dummy Verb, DV）**”。在相關論著中，介紹了 DV 的概念及其語法屬性[俞士汶等 2003]。在《語法信息詞典》中忽略了 DV 的特殊性，只將

[1] 本文相關研究得到國家 973 項目(2004CB318102)和國家 863 計劃(2001AA114210，2002AA117010)的支持。

* 北京大學計算語言學研究所，100871 中國
Institute of Computational Linguistics, Peking University, 100871, China
E-mail: yusw@pku.edu.cn; duenhm@water.pku.edu.cn

[Received April 13, 2005; Revised June 13, 2005; Accepted August 13, 2005]

DV 和其他動詞等量齊觀，因而設計了同樣的語法屬性字段。本文力圖彙集語言學家的研究成果，突出形式動詞作爲句法成分和語義角色的標記的功能，這樣的知識對語言信息的自動處理是十分有價值的。

據筆者所知，最早提出漢語形式動詞概念的是呂叔湘先生[呂叔湘 1980]和朱德熙先生[朱德熙 1985]。朱德熙先生又稱“形式動詞”爲“虛化動詞”。其后又找到了有關形式動詞的論文 6 篇，除[苗傳江 1999]涉及信息處理外，[李峰等 1995; 毛宏願 1997; 閻仲笙 1998; 李晗蕾 2003;陳永莉 2003]都屬于本体研究或教學研究的範疇。

面向信息處理的《語法信息詞典》是在朱德熙先生的語法理論体系指導下研製的。本文的理論基礎也是朱德熙先生于 1985 年發表的論文《現代書面漢語裏的虛化動詞和名動詞》。該文列舉的形式動詞有“進行、加以、給予、給以、予以、作”。可以將朱先生列舉的 DV 分爲 3 組：①“加以、給予、給以、予以”，② “進行”，③ “作”。

本文除對形式動詞的共性進行一般性討論外，還重點討論兩個最典型的形式動詞：第①組中的“加以”和第②組的“進行”。本文不討論“作”。另外，《語法信息詞典》將“有”（取其“表示發生或出現”的義項，指“体重有增加”中的“有”）列入形式動詞是否得當，也不討論。

2. 形式動詞的詞彙語義

朱德熙先生對形式動詞的詞義有一個統一的說法：“這些動詞原來的詞彙意義已經明顯地弱化了，因此在某些句子裏把它們去掉并不影響原句的意思。”朱先生給出的例句有

關于矛盾的特殊性問題應當著重地（加以）研究。

他們花了整整一年時間（進行）調查。

對于這種損壞公物的行爲應當（給以）批評。

兩國政府將採取果斷措施與恐怖主義（作）鬥爭。

朱先生又指出這些句子中的 DV 還可以互換。改爲“進行研究”、“作調查”、“給予批評”、“予以批評”、“進行鬥爭”，原句意思不變。這也是 DV 詞彙意義弱化的表現。但“弱化”不等于沒有。深入了解 DV 的詞彙意義還是有用的。

“加以”是最典型的 DV，也是呂叔湘先生在《現代漢語八百詞》中唯一明確指出的 DV。記得在學算術時，老師強調了“乘”與“乘以”、“除”與“除以”的區別，“3*4”要讀作“3 乘以 4”，不能說“3 乘 4”；“8÷4” 要讀作“8 除以 4”，不能說“8 除 4”。當時記住了，并不理解。現在知道“3 乘以 4”可以解釋“用 4 乘 3”；“8 除以 4” 就是“用 4 除 8”。算術中不用“加以”和“減以”。“加以”在日常用語中有了約定俗成的特定意義和用法。《現代漢語詞典》對“加以”列了兩個義項。其一“表示進一步的原因或條件”，可以替換爲“加上”，查《現代漢英詞典》[外研社 1988]，

它的對譯的英語詞是 in addition 或 moreover，實際上它的詞性是連詞，不屬于 DV 範疇。其二“表示如何對待或處理前面所提到的事物”，與《現代漢語八百詞》的釋義“表示對某一事物施加某種動作”基本一樣，這是 DV 的意義和用法。需要注意到，這裏“施加的” 是動作，而不是事物（數也是一種抽象的事物）。

“進行”是另一個典型的 DV。《現代漢語詞典》的“進行”有兩個義項。其二是“前進”，不是 DV 的意思。其一是“從事（某種活動）”，正是 DV 的意思。但“從事”并不比“進行”更易懂、更常用。再查“從事”，從兩個義項中選出的一個相應的釋義是“投身到（事業中去）”，接著查“投身”，釋義是“獻身出力”，最后查“獻身”和“出力”。“獻身” 的釋義是“把自己的全部精力和生命獻給祖國”，“出力”的釋義是“拿出力量”，至此釋義只使用了常用的詞語，意思似乎也明白了，但同平時所理解和使用的“進行”又相距甚遠。如果從構詞角度考慮，“進行”顯然是由“進”和“行”構成的複合詞，“進”和“行”都有動詞意義，擇其相關的義項，“進”爲“向前移動”，“行”爲“表示進行某項活動”，無論動詞“進行”是聯合結構還是偏正結構，“行”總是處于中心（head）地位，于是出現了循環定義：又用“進行”解釋“行”。這樣看來，利用面向人的詞典中對詞義的解釋，計算機是不可能“理解”詞的語義的。大概人也不是通過查詞典理解“進行”這樣的詞的意義和學會它的用法的。

儘管形式動詞的意義已經弱化，但區分其不同義項（sense）仍然是值得探究的問題。在大規模基本標注的語料庫中調查形式動詞所帶賓語的情況以及作爲賓語的詞語的語義，可能會發現形式動詞的不同義項和用法。現在從事這樣的研究的條件已經具備，不過仍然需要投入精力和時間。本文缺乏這樣的研究結果，留下了遺憾。面向漢語信息處理的一些詞彙語義知識庫[董振東等 2001; 陳群秀 2001; 王惠等 2003; 于江生等 2003]也包含了“加以”、“進行”等形式動詞。由于缺乏應用的實踐，筆者現在還不能給出確切的評價。面向“自然語言理解”的形式動詞的詞義表達及其形式化還有很多工作要做。翻譯是區分義項、也是檢驗“理解”的手段之一，第 4 節涉及到“進行”的英語譯法。

3. 形式動詞的語法屬性

漢語語法學界很早就注意到句法和語義必須結合。陸儉明先生指出：“句法研究可以從形式入手，也可以從語義入手，但是如果從形式入手，所得結論需要找到意義上的依據；如果從意義入手，所得結果需要找到形式上的表現。”[陸儉明 2003]也就是說，從事漢語本体研究的學者是句法語義并重的。自然語言處理的終極目標是實現自然語言的機器理解，當然也必須句法語義并重。不過，在自然語言理解最終實現之前，要讓自然語言處理技術在信息處理應用領域發揮作用，筆者認爲，從句法形式入手不僅比較容易、比較現實，而且句法形式研究可以爲語義理解研究做好必要的鋪墊。因此，筆者與同事們對語義雖然也有所涉獵，但一直把漢語句法分析（包括方法與知識庫兩方面）作爲最重要的基礎研究而投入了熱情和精力。

《現代漢語語法信息詞典》是爲漢語自動分析和生成服務的電子詞典，它對于動詞

的語法屬性有相當詳細的描述[俞士汶等 2003]。可以很容易地從動詞庫中檢索出所有的 DV 及其全部屬性值。不過，該詞典按詞類設立屬性字段，動詞亦然，并沒有考慮 DV 的特殊情況。因此，對形式動詞的描述還是不夠充分、不夠縝密的。

以下分情況討論 DV 的句法功能及其各種語法屬性。呂叔湘先生只討論了“加以”的帶賓語的情況和受副詞修飾的情況。本文也從這裏開始。

3.1 形式動詞的賓語

形式動詞的最重要的特徵就是它帶賓語的情況。呂叔湘先生[呂叔湘 1980]指出，“（加以）必帶双音節動詞賓語”。朱德熙先生[朱德熙 1985]指出，“虛化動詞所帶的賓語只能是表示動作的双音節詞”。

呂叔湘先生是這樣區分漢語及物動詞和不及物動詞的：“及物動詞后邊可以帶一個表示承受動作的事物的名詞，稱爲賓語。不及物動詞不能帶這樣的名詞。”

《語法信息詞典》認爲“能帶真賓語的謂詞是及物動詞”，繼承了朱先生的定義，與呂先生的定義相容。需要注意：（1）漢語的及物動詞可以帶真賓語，不等于說在句子中一定帶賓語。第①組“加以”等 DV 必須帶賓語，這一點不同于一般的及物動詞，在語句中必須帶賓語的及物動詞可以叫做“粘賓動詞”。（2）賓語不一定是名詞，這一點表面上對呂先生的定義有所擴展，但并不與呂先生的觀念衝突，因爲呂先生已經指出“加以”必帶動詞賓語。真賓語可以分爲 3 類，即 ①体詞性的，②謂詞性的，③准謂詞性的。

形式動詞必須帶的賓語絕大部分都是准謂詞性的，少數是体詞性的。

作爲“加以”、“進行”賓語的双音節及物動詞失去了原有的一些特點，如不能再受副詞修飾，不能再帶賓語，如果這樣的賓語要擴充，也只能擴充爲体詞性的定中短語。這樣的賓語就是“准謂詞性賓語”。能夠擔任“准謂詞性賓語”的双音節動詞是動詞的一個子類，即“名動詞”。本文重點討論形式動詞的“准謂詞性賓語”，不宜爲“名動詞”多費筆墨。

并非只有形式動詞才能帶“准謂詞性賓語”。像“接受邀請”中的“邀請”也是“准謂詞性賓語”，屬于名動詞，但“接受”并不是形式動詞。另一方面，“拒絕邀請”中的“邀請”又不是“准謂詞性賓語”，因爲它除了可以擴充爲体詞性的定中短語，也可以擴充爲謂詞性的述賓結構，如“拒絕邀請那些不尊重當地風俗習慣的人”。

“加以”和“進行”帶賓語的情況也有所不同。“加以”只能帶准謂詞性賓語。“進行”的賓語的形式更爲豐富。多數是准謂詞性的，此時“進行”的詞彙意義弱化了；另一方面，也可以帶体詞賓語，如“進行戰爭”、“進行手術”、“進行面試”等，此時“進行”有更實在的詞彙意義，不能去掉，能夠替換“進行”的也應該是其他有實際意義的動詞，如“發動戰爭”、“施行手術”或“做手術”、“主持面試”。如果主張把帶体詞賓語的“進行”同帶准謂詞性賓語的“進行”分開，也未嘗不可。不過，當注意到充當“進行”賓語的那些名詞也是表示動作的，與作爲准謂詞性賓語的名動詞在語義範疇上是相似的，把這兩種情況處理爲同一個“進行”的兩種不同屬性也是可以的。

作準謂詞性賓語的名動詞基本上是及物的，但“進行”的准謂詞性賓語也可以是不及物動詞，如“進行合作”、“進行點名”。當“進行”有實在意義，其受事可以前置，這樣“進行”在句子中也可以不帶賓語，如“戰爭正在進行”、“手術進行得很順利”。

3.2 形式動詞受副詞修飾的情況

呂先生指出：“‘加以’前面如用副詞，必須是双音節的；單音節副詞后面不能用‘加以’，只能用‘加’。”例子有：“不加研究｜多加注意”。換句話說，如果要用單音節副詞修飾‘加以’，那麼双音節的‘加以’應該縮減爲‘加’。呂先生和朱先生在闡述詞的語法特性時，往往注意到詞語的音節數目。現在的《語法信息詞典》缺乏有關音節、韻律的信息，是個缺憾。

《語法信息詞典》設立了動詞是否可受“不”、“沒”、“很”等副詞修飾的屬性字段。查 DV 的這些屬性字段的值：“加以”和“進行”不能受“很”修飾，可以受“不”、“沒”修飾。

從 1998 年全年《人民日報》中檢索“不加以”、“沒加以”、“很加以”、“多加以”。找到了 10 例“不加以”：“不加以分析”、“對此美國不能不加以考慮”、“后來有識之士不得不加以勸導”、“不能不加以引介”、“不能不加以分析和考察”、“這種狀況如不加以改善”、“但如果不加以注意”、“如不加以整頓”、“許多人不但不加以制止”、“如不加以遏制整肅”。這些應該都是“加以”受副詞“不”修飾的可以接受的用法。不過，語言學家通常將受“不”修飾限定爲陳述性否定而排除假設性否定（或事態性否定）。呂叔湘先生說“加以”前不受“不”修飾，應該指的是陳述性否定。上述例子可能絕大部分都屬于假設性否定。由此可以領悟，副詞“不”的語義是十分複雜的。要準確判定“不”的語義，需要擴展到更大的語境，需要在更深入的層次上研究。

在 1998 年《人民日報》語料中確實沒找到“很加以”、“多加以”，也沒找到“沒加以”。但在更大的範圍內就找到了例句

越往下看越象小娃娃，可是老太太沒加以什麼批評。(出處:牛天賜傳)

像下面的句子

？由于長期對上游水土保持沒加以重視，所以造成近年下游頻發洪澇災害。

也應該是站得住的。因爲不是從真實文本中檢索出來的，故在前面加了問號。面向短語構成規律描述的《語法信息詞典》關于“加以”可以受“不”、“沒”修飾的描述只在較淺的層次上反映了詞語的使用情況。由于例句畢竟不多，現在對“加以”是否可受“不”或“沒”修飾，存在不同看法，是完全正常的。語言學界有“言有易，言無難”

和“例不十，法不立”的共識，是很有道理的。[俞士汶等 2003] 論及二選一型的語法屬性值的填寫時，曾記下了作者的心得：“在確定‘二選一’型屬性值時，一般說來，說‘可’較保險，不容易被駁倒；說‘否’則有一定風險，找到一個反例就使‘否’的立論站不住腳。但對于計算機處理來說，正是由于有剛性的值存在，才使得詞語的屬性值便于利用。”后來，在“詞的概率語法屬性描述”研究中，筆者提出了確定“可否”值的比較科學的、定量的判定準則[俞士汶等 2001]。不過，基于統計的結果也會受到語料規模的制約。

3.3 形式動詞的句法標誌功能

現代漢語的實詞劃分爲体詞和謂詞兩大類。相應的，短語也可以劃分爲体詞性的和謂詞性的兩大類。DV 是語句中謂語性成分的標誌之一。表示動作的名詞或体詞性短語，例如“戰爭、手術、詞彙語義的研究、行政處分”等等是不能作謂語的。但在它們前面加上 DV 構成述賓結構轉換成謂詞性成分，就可以作謂語了。

某些地區還在進行戰爭。

爲了提高機器翻譯的品質，我們必須進行詞彙語義的研究。

主管部門對違反紀律者給予行政處分。

双音節的名動詞雖然可以獨立作謂語，但是當它要同前面的含介詞或動詞的短語結構結合成更大的謂詞性短語時，爲了同前面的結構相匹配，名動詞先要複雜化，這時也必須加上 DV。

* 你們把這些資料整理。

你們把這些資料加以整理。

* 村長向群眾承認錯誤檢討。

村長向群眾承認錯誤進行檢討。

在這種情況下，“加以整理”是謂詞性的，是作句子謂語的狀中結構的中心成分。作句子謂語中心成分的連動結構“承認錯誤進行檢討”的一部分當然也是謂詞性的。DV 與所帶的体詞性成分構成的述賓結構在句子（或小句）中總是謂詞性的，通常作句子（或小句）的謂語（或其中心成分）。

更進一步，DV 還是其后的名動詞的受事的前置標誌（這已涉及語義、語用問題，將在 4.1 中論述）。

認識到 DV 是漢語句法的重要標誌，在缺乏形式標記的漢語中，對分析和理解句子

很有幫助，應當加以充分利用。

3.4 形式動詞的其他語法屬性

DV 的共性很多，當然不同組的 DV 也有不同的特性，每個 DV 都有自己的個性。研究詞的語法屬性就是要逐步深入地求同辨異。

⑴ 有 DV 都是粘著詞。

⑵ 所有 DV 不兼屬其他詞類。在動詞內部，也自成一個子類，與其他子類（不及物動詞、系動詞、助動詞、趨向動詞、兼語動詞、離合動詞、名動詞）都沒有交集。

⑶ 所有 DV 都不能單獨作主語、謂語、補語和狀語，而且第①組也不能單獨作賓語（“進行”可以作助動詞的賓語）。

⑷ 所有 DV 都不受程度副詞修飾，既包括單音節的“很”，也包括双音節的“非常”等。第①組 DV 都不能受“在”修飾，除“加以”外，也不能受“正在”修飾。

⑸ 第①組 DV 都不能帶由時量詞和動量詞構成的准賓語；除“給予”可帶“了、過”外，都不能帶助詞“著、了、過”。

⑹ 所有 DV 都沒有“AABB”、“ABAB”的形態變化。

綜上所述，DV 是動詞中頗具特色的一個子類。

4. 進一步研究的課題

4.1 形式動詞的語義和語用問題

前面已經談到，可以根據形式動詞所帶賓語的情況以及作爲賓語的詞語的語義，發現形式動詞的不同義項和用法。

再觀察以下例句：

⑴ 我們應當充分利用這批寶貴的資源。 / 這批寶貴的資源我們應當充分利用。

⑵ 我們應當把這批寶貴的資源加以充分（的）利用。

⑶ 對于這批寶貴的資源我們應當加以充分（的）利用。

⑷ 這批寶貴的資源我們應當加以充分（的）利用。

⑸ 這批資源很寶貴，我們應當加以充分（的）利用。

從句法語義角度考慮，(1)中的名詞短語“這批寶貴的資源”（NP）是動詞“利用”（v）的受事，(2)-(5)的句法形式雖然與(1)不同，但語義關係卻不變，NP 仍然是 v 的受事，不過這時 v 以名動詞 vn 的身份出現，由于作爲 DV 賓語的名動詞 vn 后面不能帶受事賓語，作爲受事的 NP 被前置了。在(2)中，已有介詞“把”作爲受事前置的標誌，DV“加以”

主要承擔句法功能。在(3)中，雖有介詞“對于”，但通常“對于”只是“與事”的標誌，但在有 DV 的句子中也承擔了受事的標誌。在(4)中，DV 成了 vn 的前置受事的僅有標誌。在(5)中，DV 的存在標誌了上文有一個語義上表示 vn 的受事的語言形式，不妨也把它看作廣義的前置受事。

總之，如果不用 DV，動詞 v（例子中的“利用”）是無標誌的，其受事可以前置，也可以后置。如果用了 DV，名動詞 vn 一定要有前置的受事，且是有標誌的。

[苗傳江 1999]論述“進行”句的語義結構，涉及到了句法語義範疇，但只討論了包含“進行”的幾種簡化的句型。實際上，真實文本中包含“進行”的句法形式要豐富得多、複雜得多。

至于寫文章或者說話什麼時候採用受事前置形式，則是受語用的制約：句子中哪些信息是已知的，作爲話題（topic），哪些成分又是關注的焦點（focus）。

4.2 語言信息處理中的形式動詞

在應用淺層分析的語言信息處理系統（如文獻檢索、信息提取與過濾等）中，也許可以把双音節的 DV 看作虛詞，列入停用詞表（stop list）。但在需要進行深層分析以得到句子的句法結構或語義框架的應用系統（如機器翻譯）中，對 DV 就要給予充分的關注。在詞性標注、詞義消歧（WSD）等工作中，如果能用上 DV 的句法語義信息，可以提高自動分析的精度。在北京大學現有的語言知識庫（現代漢語語法信息詞典，現代漢語語義詞典，中文概念詞典等）[俞士汶 2003]對 DV 的描述都是不夠充分的。爲了彌補這個不足，筆者倡議建設“廣義虛詞知識庫”，并且把 DV 也算作廣義虛詞[俞士汶等 2002]。

以漢語到英語的翻譯爲例，深入研究包含 DV 的句型的翻譯規律是有價值的。[苗傳江 1999]有一個例句：

我國正在對國營企業進行改革。

[苗傳江 1999]認爲，“這樣的句子，翻譯成英語時，‘進行’就沒有著落”。現在給出 5 種譯文：

1. We are carrying on reforms on the state-owned enterprises in our country.
2. We are making reforms on the state-owned enterprises in our country.
3. The reforms on the state-owned enterprises are being carried out in our country.
4. We are reforming the state-owned enterprises in our country.
5. The state-owned enterprises are being reformed in our country.

應該說，這 5 種譯文基本上都傳達了原文的意思。從前面的討論，可以瞭解到“我國正在對國營企業進行改革”同“我國正在改革國營企業”是有差別的。漢語之所以採用把“改革”的受事加以前置的句型，很可能是由于該受事是已知信息，而“改革”才是本句的焦點。漢語又用了“進行”，將名動詞“改革”作爲它的准謂詞性賓語，不僅顯得正式、嚴肅，而且反映了“改革”的過程性。在第 1,2,3 句譯文中，也使用 carry on 或 make 將英語動詞“reform”名詞化，與“進行”相當匹配。因此，通常在漢譯英時，“進行”并非沒有著落。實際上，相當于“進行”的英語詞并不少，至少有：carry on, carry out, undertake, undergo, conduct, engage, make, hold, commit, have, 等等。一是因爲多，二是名詞與這類動詞常有固定的搭配，如：hold discussion, make investigation ，這樣，英語中的這類動詞就不像漢語中的“進行”那麼引人注目。

5. 結語與致謝

筆者雖然長期在計算語言學領域耕作，但發表的論文絕大多數都是屬于語言工程或自然語言處理技術方面的。承蒙第五屆漢語詞彙語言學研討會（CLSW5, 2004 年 6 月 14 日至 16 日，新加坡）組織者賴金錠博士和姬東鴻博士的盛情邀請，要求我提供一個特邀報告。與以往不同，筆者這次選擇形式動詞這個題目嘗試進行微觀語法研究，雖然它也是爲綜合型語言知識庫的總体目標服務的，卻自知語言學功底甚淺，不太可能取得好的成果。不過又想，既弄斧，不妨就到魯班門前來弄，所以才壯膽將這篇文章先提交給程式委員會評審，還好沒有被拒絕。希望這篇報告能爲語言信息處理研究和語言本体研究的交流做出一點貢獻。

筆者曾就形式動詞的翻譯問題求教于王逢鑫教授，他在百忙中提供了例句“我國正在對國營企業進行改革”的 5 種譯文。劉雲博士、柏曉靜、王治敏、彭國珍等同學也協助作者收集參考文獻、提供翻譯例句。陸儉明教授、詹衛東博士、吳云芳博士、孫斌博士、呂學強博士、張化瑞老師、諶貽榮同學等對初稿提出了修改意見或補充資料。初稿的匿名評審者也給予了寶貴的建議。作者對以上各位的奉獻致以誠摯的謝意。

作者在將拙文提交給會議之后，收到刁宴斌教授的兩本論著。一本是他在馬慶株教授指導下完成的博士論文“虛義動詞論”，一本是 2004 年 3 月由遼寧師範大學出版社出版的專著《現代漢語虛義動詞研究》（37 萬字）[刁宴斌 2004]。如果在提交之前收到這兩本論著，筆者可能會另選題目。不過，在流覽了這兩本論著之后，我認識到刁宴斌教授的研究屬于現代漢語語法史框架下的本体研究，而本文畢竟是面向信息處理的，以尋求自動分析所需要的形式標記爲目標的探索還是有一定空間的，也爲本体研究的應用拓展了疆界。筆者始終認爲，爲了向機器理解的目標前進，面向信息處理的語言研究一定要從語言本体研究的成果中吸收營養。正是本著向語言學家學習的願望，筆者出席每年一度的 CLSW 系列會議，確實獲益匪淺。抱著同樣的願望，作者才有勇氣將拙文投給 IJCLCLP Special Issue of CLSW5 。

參考文獻

陳群秀，“現代漢語述語動詞機器詞典的擴充和槽關係研究”，《語言文字應用》，4，2001, pp. 98-104.

陳永莉，“DV 的範圍、次類及特徵”，晉陽學刊，3，2003, pp. 92-94.

刁宴斌，《現代漢語虛義動詞研究》，中國大連：遼寧師範大學出版社，2004 年 3 月.

董振東，董強，“面向信息處理的詞彙語義研究中的若干問題”，《語言文字應用》，3，2001，pp. 27-33.

李峰，柴耘，“‘V+以’類虛化動詞的賓語”，《新疆教育學院學報》，4，1995，pp. 68-71.

李晗蕾，“‘加以’的語用功能”蘇州教育學院學報，20(1)，2003, pp. 11-16.

陸儉明，《現代漢語語法研究教程》，北京：北京大學出版社，2003 年 8 月，pp. 161-162.

呂叔湘，《現代漢語八百詞》，北京：商務印書館，1，1980 年 5 月.

毛宏願，“話 DV 和形式化動詞”，《喀什師範學院學報》，4，1997.

苗傳江，“‘進行’句的語義結構”，見黃昌甯，董振東主編《計算語言學文集》，北京：清華大學出版社，1999, pp. 51-57.

外研社，《現代漢英詞典》，北京：外語教學與研究出版社，1，1988 年 11 月.

王惠，詹衛東，俞士汶，“現代漢語語義詞典規格說明書”，新加坡：《漢語語言與計算學報》，13(2), 2003，pp. 159-176.

閻仲笙，“說‘后續動詞性賓語動詞’”，河北師範大學學報（自然科學版），32(2)，1998, pp. 91-93.

于江生，劉揚，俞士汶，“中文概念詞典規格說明”，新加坡：《漢語語言與計算學報》，13(2), 2003，pp. 177-194.

俞士汶，“北京大學語言知識庫概況”，新加坡：《漢語語言與計算學報》，13(2)，2003 年 6 月，pp. 119-120.

俞士汶，段慧明，朱學鋒，“漢語詞的概率語法屬性描述”，《語言文字應用》，3，2001, pp. 21-26.

俞士汶，朱學鋒，劉雲，“現代漢語廣義虛詞知識庫的建設”，第二屆肯特崗漢語語言學圓桌會議（新加坡），2002 年 11 月 27 日，新加坡：《漢語語言與計算學報》，13(1)，2003 年 3 月，pp. 89-98.

俞士汶等，《現代漢語語法信息詞典詳解（第 2 版）》，北京：清華大學出版社，2003 年 2 月.

中國社會科學院語言研究所詞典編輯室編，《現代漢語詞典（修訂版）》，北京：商務印書館，1983 年 1 月.

朱德熙，“現代書面漢語裏的虛化動詞和名動詞——爲第一屆國際漢語教學討論會作”，《北京大學學報（哲社版）》，5，1985.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 519-532

A Synchronous Corpus-Based Study on the Usage and Perception of Judgement Terms in the Pan-Chinese Context

Oi Yee Kwong* and Benjamin K. Tsou*

Abstract

This paper reports on a synchronous corpus-based study of the *everyday* usage of a set of Chinese judgement terms. An earlier study on Hong Kong data found that these terms were more polysemous than their English counterparts within the legal domain, and were even more fuzzily used in general news reportage. The current study further compares their usage in general texts from other Chinese speech communities (Beijing, Taiwan, and Singapore) to explore the regional differences in lexicalisation and perception of the relevant legal concepts. Corpus data revealed the distinctiveness of the Singapore data, and that the contrasting frequency distributions of the terms and senses could be a result of the varied focus in reportage or the use of alternative expressions for the same concepts in individual communities. The analysis will contribute to the construction and enrichment of Pan-Chinese lexico-semantic resources, which will be useful for many natural language processing applications, such as machine translation.

Keywords: Synchronous Corpus-Based Study, Legal Concept and Terminology, Regional Differences, Pan-Chinese Lexico-Semantic Resources

1. Introduction

In this paper, we report on a synchronous corpus-based study on a set of semantically related legal terms, and propose a Pan-Chinese lexico-semantic resource for the legal domain, such as one in the form of a thesaurus, to differentiate the usage and perception of closely related legal concepts and terminology across various Chinese speech communities.

The situation with language and law is a very interesting one in Hong Kong. As pointed out by Tsou and Kwong [2003], the legal system in Hong Kong has operated through English

* Language Information Sciences Research Centre, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
E-mail: {rlolivia, rlbtsou}@cityu.edu.hk

[Received April 14, 2005; Revised June 15, 2005; Accepted August 14, 2005]

solely for more than 150 years. Following the implementation of legal bilingualism in the 90's, Hong Kong became the first community that follows the Common Law system which at the same time allows the use of both English and Chinese in court proceedings. In contrast to the many precisely lexicalised legal concepts in English, given its long and established tradition in Common Law, the use of their Chinese equivalents is apparently more fuzzy, as is evident from the lack of one-to-one correspondence of legal terms between English and Chinese. This difference in the cross-lingual lexicalisation of legal concepts between English and Chinese has thus become a substantial linguistic hurdle in the implementation of legal bilingualism, and is directly related to whether both languages could eventually be used in balanced and proper ways to the same effect in the legal domain. The apparent fuzziness with Chinese legal terms is nevertheless peculiar in the Hong Kong context, but not in other places where Chinese is also used as the official language in the legal domain, such as in Mainland China. A possible reason is that preciseness is somehow diluted upon translation. In English, for instance, a "verdict" and a "sentence" are sufficiently distinguished, despite their semantic relatedness (as both are related to the results of a trial). However, when expressed in, or more often translated into, Chinese, the preciseness is somehow weakened. On the one hand, the expression or translation in Chinese might have to take into account the corresponding syntactic constraints and stylistic differences in the two languages in order to sound natural, hence the variation in expressing the same concept in different contexts. On the other hand, the translation could sometimes be affected by usages in other places. For example, although "contract" is mostly expressed as "合約" in Hong Kong, it is sometimes expressed as "合同", which is the term used in Mainland China for this concept.

A set of semantically related legal terms was studied by Tsou and Kwong [2003] with respect to their usage in the legal domain and in general texts. The terms are "裁定" (hold, convicted), "裁決" (determine, verdict), "判決" (judgement, conviction), "裁斷" (find, finding), and "裁判" (Magistracy)[1], all of which have one or more senses referring to some aspects of "judgement". Their uses were studied via a corpus of bilingual court judgments[2] and a general corpus of news articles from Hong Kong. From the corpus of bilingual judgments, it was observed that these Chinese terms were considerably polysemous, such that most of them were found as the renditions for multiple English terms, which are distinct though closely related in meaning. Even more varied usages were found for the Chinese terms

[1] The English terms are the more common translations of the corresponding Chinese terms as observed from Hong Kong court judgments. They are included here for reference only and are not necessarily the absolute or correct translations *per se*.

[2] An explicit distinction between the use of "judgment" and "judgement" is drawn here, as the inclusion or omission of the "e" is not arbitrary. "Judgment" refers specifically to the concluding writing for a court trial, while "judgement" refers to the action of judging in general.

in the general corpus, and the sense boundaries appeared to be more ambiguous. Although the confusion might be insignificant to the perception of the general readers, the subtlety therein could bear significant conceptual difference in the stricter legal domain, which is far less tolerable of impreciseness and ambiguity. Attempts were made to identify near-synonymous senses from the corpus examples, and to arrange them in terms of their semantic relatedness into a verb hierarchy and a noun hierarchy, in a way similar to WordNet [Miller *et al.* 1990].

The current work aims at expanding on the above study, to further explore and analyse the different usages and finely grained senses of the aforementioned Chinese legal terms, among news texts from various Chinese speech communities including Hong Kong, Beijing, Taiwan, and Singapore. Based on the usages of these terms in the corpus texts from different communities, we look for local differences in the lexicalisation, if any, and thus the perception of the corresponding legal concepts, which might be a result of the differences in social structures or legal systems. On the one hand, the different societies and legal systems might share similar legal concepts which are expressed in the same ways. On the other hand, in the case where they do not share similar lexical items, it is important to see what alternatives are used in different places to express related concepts. This is a necessary though preliminary step in the development of a Pan-Chinese lexico-semantic resource for legal terminology.

Efforts have been made by researchers in lexical semantics on the study of semantic relations among Chinese lexical items, with a view toward organising the lexical items into semantic networks. Gao [2001], for example, proposed a quantitative measure for the closeness and differentiation of near-synonyms among verbs denoting physical actions from a range of lexical semantic features. Cheng [2001] discussed the differentiation of related words from their individual focus and orientation. Nevertheless, work on lexical semantics and corpus-based lexicography often only drew reference from one particular corpus. Huang *et al*. [2000] worked on verbal semantics and near-synonyms of Mandarin Chinese as used in Taiwan. Tongyici Cilin [梅等 1984] is based exclusively on Chinese as used in post-1949 Mainland. However, linguistic variation is significant and especially salient for Chinese language used in different communities [Tsou *et al.* 2004]. Our corpus-based, Pan-Chinese approach, beginning with a set of domain-specific lexical items, thus has additional advantages for its indigenousness, portability, and versatility. Such a Pan-Chinese lexical resource, when done in large scale, could contribute to natural language processing applications like machine translation and serve as a rich reference for legal and paralegal professionals. More importantly, the resource would capture the linguistic norms from more than one Chinese speech community.

In Section 2, we first briefly review the polysemy of legal terms and the complexity of translating legal terms from English to Chinese. Then in Section 3, we present the details of

the corpus analysis done in the current study. Results are discussed in Section 4, and we will conclude with future directions in Section 5.

2. Polysemy of Legal Terms

Tsou and Kwong [2003] started with a set of Chinese legal terms, all with senses related to "judgement" or "the action of judging" in the legal context, to study how the preciseness of legal concepts lexicalised in English is captured in their Chinese translations as shown in bilingual court judgments in Hong Kong, and how the preciseness of the latter is in turn preserved in a general corpus. The set of "judgement" terms includes "裁定" (hold, convicted), "裁決" (determine, verdict), "判決" (judgement, conviction), "裁斷" (find, finding), and "裁判" (Magistracy). They observed that in Hong Kong, despite the implementation of legal bilingualism for several years, legal concepts are not as precisely lexicalised in Chinese as in English. Thus while a particular legal concept could be relatively unambiguously expressed by one term in English, the same concept may not have a direct one-to-one corresponding term in Chinese. The fuzziness is carried over from legal contexts as in court judgments to informal contexts as in news reports. For instance, "裁決" has been identified as the translation equivalent for "decision", "verdict", and "award" in the bilingual corpus of court judgments. Similarly, the word "decision" has been rendered as "決定", "裁決", and "判決", among many possibilities. Such a complex correspondence (as further illustrated in Table 1 and Figure 1) between English and Chinese legal terms can be explained by the fact that the use of English is much more mature in the Common Law system in Hong Kong. Many legal concepts are thus lexicalised and can be precisely expressed in English, whereas this preciseness is greatly weakened when terms are translated into Chinese.

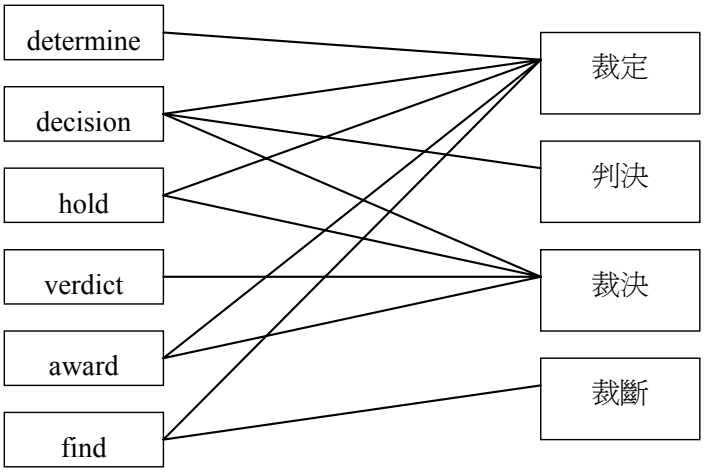


Figure 1. Interwoven mapping between English and Chinese legal terms

Table 1. Example of multiple renditions between English and Chinese legal terms

English	Chinese	Examples	
Decision	裁決	In Mayson v. Clouet [1924] AC980 the Privy Council approved the decision in Howe v. Smith (1884).	樞密院在 Mayson v. Clouet ([1924]AC980) 一案中對 Howe v. Smith([1884]27Ch.D.89) 一案的裁決表示贊同。
Verdict	裁決	The jury returned their verdicts on 2 September 1997.	1997 年 9 月 2 日，陪審團作出裁決。
Award	裁決	This was followed by a request that the tribunal should postpone making an award for two months.	這一段確認了賣方跟著請求仲裁庭延遲兩個月作出裁決。
Decision	決定	None of these decisions assist the appellant.	這些決定之中沒有一宗對上訴人有所幫助。
Decision	判決	The Court of Appeal, by a majority (Nazareth V-P and Liu JA, Rogers JA dissenting) reversed her decision, dismissing the claim to specific performance and holding that Douglas was entitled to forfeit the deposit.	上訴法庭以大比數（上訴庭副庭長黎守律和上訴庭法官廖子明同意，上訴庭法官羅傑志持異議）推翻她的判決，撤銷強制履行的申索和裁定第二與訟方有權沒收訂金。

Despite the complexity of multiple renditions, the morphemic structure of the individual Chinese terms might nevertheless indicate a core sense of the term, and thus suggest the focus of the relevant concept. For example, while the words “裁定”, “裁決” and “裁斷” share an identical morpheme “裁” (to judge), they could be differentiated by their second morphemes, which focus on “conclusion”, “decision”, and “inference” respectively. This distinction is similar to Cheng’s [2001] discussion of word families where similar and related words could be differentiated by their individual focus and orientation, or meaning facets.

In the current study, we are interested to see if the same kind of polysemy appears in the usage of the same legal terms in different Chinese speech communities. It is hypothesised that we may not find exactly the same usage of the same terms in the various communities, as their different social structures and legal systems might lead to different perception of the corresponding legal concepts, and the same concepts may not be equally salient for people in different communities. We will probe, using authentic corpus data, the perception of the various legal concepts in different communities, and see how the salience of “judgement” is reflected in the language used in different places; and if they do not use the terms in the same way, what alternative words are used to express similar concepts.

3. A Synchronous Corpus-Based Study

3.1 Materials

In this study, we further analyse three terms (called "target words" hereafter) which Tsou and Kwong [2003] studied, namely "裁定" (hold, convicted), "裁決" (determine, verdict), and "判決" (judgement, conviction). We leave out "裁斷" (find, finding), as it was only found in the bilingual court judgment corpus but not at all in the general corpus (LIVAC, as introduced below), and "裁判" (Magistracy), as it was found to mostly refer to the sense of "umpire" or "adjudication in a contest" when used in the general corpus.

Sentential contexts for the target words were extracted from a subset of the LIVAC corpus [Tsou *et al.* 2000]. LIVAC (http://www.livac.org) is a synchronous corpus developed by the Language Information Sciences Research Centre of the City University of Hong Kong. The corpus contains newspaper articles collected synchronously and regularly from six Chinese speech communities. The subset we used in the current study consists of texts from Hong Kong (HK), Beijing (BJ), Taiwan (TW) and Singapore (SG), covering local news, international news, sports news, entertainment news, and financial news, collected over the *same* period of time (for two years, 1997-98 and 2002-03). Each sub-corpus, that is, texts from each of the four places, contains about 5M Chinese characters, which yields about 3M words upon segmentation.

3.2 The Analysis

For each target word, 30 samples of their sentential contexts (where there was sufficient data) were randomly selected from each sub-corpus, and assigned a sense from the sense inventory as in Tsou and Kwong [2003] where appropriate. New senses were recorded when found. Upon sense tagging, the samples from the various sub-corpora were further analysed with respect to the sense distribution of each word in each community, and the similarities and differences of such distributions across the various communities. As in the previous study, the assignment of senses took into account the collocation patterns of the different senses and subcategorisation patterns of the verbal usages where applicable. In addition, bi-syllabic words containing the morphemes "裁", "定", "決" or "判" were retrieved from the sub-corpora. The retrieved words and their relative frequencies were studied, to disclose any perceptual difference of the related legal concepts and any alternative expressions of such concepts in the different communities.

4. Results and Discussion

4.1 Relative Frequency Distribution

The frequency of the target words from the various sub-corpora is shown in Table 2. The relatively low frequency of all the target words in BJ data is most notable. The small numbers readily indicate that court news does not receive as much attention in Beijing newspapers as in other places.

Table 2. Frequency of the target words in the sub-corpora

Word	HK	BJ	TW	SG
裁定	122	38	80	19
裁決	142	32	54	139
判決	160	66	210	341

Just comparing the absolute frequencies, "判決" ranked highest across the board. Its relative frequency is especially high for TW and SG. In the most dramatic case of SG data, there are 341 occurrences of "判決" but only 19 occurrences of "裁定". This is very different from, for example, HK data, where the relative frequency of "判決" and "裁定" differs by less than 10%. To a certain extent, this difference in relative frequency suggests a variation of focus in news reportage in the two communities, assuming that the use of these words in them is not arbitrary. According to many legal dictionaries [e.g. 《法學詞典》編輯委員會 1985; 劉清景 2001], "判決" and "裁定" refer to different aspects of the ruling of a court. In particular, "判決" is often associated with the final determination on the main issue in a trial, whereas "裁定" usually refers to the conclusions to other factual disputes during a trial.

4.2 Sense Distribution

Sample sentences of the target words from BJ, TW and SG data were examined and each occurrence of the words was assigned a sense with reference to the sense set defined by Tsou and Kwong [2003]. The sense distributions were compared to those reported for Hong Kong data in the same study. The results are tabulated in Tables 3 to 5 for "裁定", "裁決" and "判決" respectively (the number in brackets below each place refers to the number of samples checked and all figures reported are percentages). The second and third columns refer to data gathered from a bilingual corpus of Hong Kong court judgments and a subset of the current HK data from LIVAC respectively as reported in the earlier study. In this subsection we focus on the sense distribution with respect to individual words, and in the next we will further explore the regional differences observed.

4.2.1 裁定

"裁定" was earlier distinguished into four verb senses and two noun senses. No new sense was found from the data in the current study. The dominance of its verb usages is observed in all places except BJ. Over 70% of the samples for BJ were assigned sense 6 (as in "作出終審裁定", "阻礙判決、裁定的執行", "維持原判的裁定", etc.). This contrasts enormously not only with other places but also with data from the legal domain. It also contrasts dramatically with SG data where no nominal usages were found at all for "裁定". This may be a consequence of the small number of SG samples but is more likely a genuine difference in the usage of the word, as we will further discuss below. Another interesting observation is that in the legal corpus, "裁定" is seldom used to state the order given by the court (sense 3) and the BJ data are more or less in line with this. However, over 25% of the samples from HK, TW and SG fall under sense 3. This thus raises interesting questions regarding the saliency of the concepts in individual places.

Table 3. Sense distribution of "裁定"

Sense and Example	Legal (30)	HK (30)	BJ (30)	TW (30)	SG (19)
1. [v.] the court decides on the outcome of a case e.g. 法庭裁定…罪名成立。	43.33	43.33	3.33	16.67	31.60
2. [v.] the court resolves an issue in a case e.g. 法官裁定所提出的要求沒有得到滿意答覆。	36.67	20.00	6.67	30.00	26.30
3. [v.] the court gives an order e.g. 法官裁定港府要即時釋放他們。	**3.33**	26.67	**6.67**	26.67	42.10
4. [v.] to judge on some issue to resolve dispute e.g. 法庭需要裁定臨立會的合法性。	0.00	6.67	0.00	3.33	0.00
5. [n.] the resolution of an issue in dispute e.g. …裁定回覆不能令人滿意。本席認爲上述裁定正確無誤。	10.00	0.00	6.67	6.67	0.00
6. [n.] the decision on the outcome of a case e.g. 我認爲暫委法官的裁定是正確的。	6.67	3.33	**76.67**	16.67	**0.00**

4.2.2 裁決

Sense tagging is notorious for its difficulty as the meaning in the new occurrence of a word is not always so clear-cut that a pre-defined sense could be unambiguously assigned to it. In this regard, the tagging for "裁決" was most difficult and confusing. The difficulty may be largely attributed to its relatively general meaning. For example, according to the hierarchies suggested by Tsou and Kwong [2003], the verb sense 裁決/1 and the noun sense 裁決/5 are

the top nodes[3] in their respective hierarchies. Having the most general sense amongst others, it means that the word can be used in a relatively wide variety of contexts. When it comes to general news reports, they are not necessarily the correct and legitimate contexts. This is evident in two respects from Table 4.

Table 4. Sense distribution of "裁決"

Sense and Example	Legal (30)	HK (30)	BJ (25)	TW (30)	SG (30)
1. [v.] the court makes a decision based on evidence e.g. 若有法律觀點分歧，最終交由法庭裁決。	16.67	6.67	0.00	23.33†	33.33
2. [v.] the court decides on the outcome/sentence/etc. e.g. 法官裁決…把謀殺罪減爲誤殺罪…	0.00	3.33	0.00	23.33	20.00
3. [n.] the court's decision on the outcome of a case e.g. 陪審團達至誤殺的裁決。	30.00	6.67	20.00	13.33	0.00
4. [n.] the court's decision on monetary compensation e.g. …拒絕執行公約裁決…	23.33	0.00	0.00	0.00	3.33
5. [n.] the court's decision on a case and orders e.g. …會就法院上月底裁定港府要釋放十名越南人的裁決上訴。	16.67	53.33	40.00†	30.00	23.33
6. [n.] the resolution of an issue e.g. …關於證據接納性的裁決…	13.33	20.00	40.00†	0.00	0.00
7. [n.] religious orders, etc. e.g. …聽命於這類宗教裁決…	0.00	10.00	0.00	0.00	0.00
*8. [v.] the court decides on the outcome of a case e.g. 法官路易斯迪蘇沙裁決他的罪名成立。	--	--	--	3.33	10.00
*9. [v.] the court resolves an issue in a case e.g. …裁決搭客沒有起訴保險公司的權利。	--	--	--	--	6.67
*10. [v.] to judge on some issue to resolve dispute e.g. 三司必須裁決的只是關係到公眾利益的法律問題。	--	--	--	--	3.33
*11. [v.] the action of judging by referee in sports events e.g. …但二壘審竟判定外野手是接殺，然後訴請裁決…	--	--	--	6.67	--

[3] In the verb hierarchy, {裁決/1} subsumes {裁定/1, 判決/1}, {裁定/2, 判決/2}, {裁定/3, 裁決/2, 判決/3}, while {裁定/2, 判決/2} further subsumes {裁定/4}. In the noun hierarchy, {裁決/5, 判決/4} subsumes {裁定/6, 裁決/3}, {裁定/5, 裁決/6}, {裁決/4}, while {裁定/6, 裁決/3} further subsumes {判決/5}. Senses within curly brackets belong to the same synonym set [Tsou and Kwong 2003].

First, the use of "裁決" is almost abused in the SG data. Senses 8 to 10 in Table 4 were unexpectedly found from SG. They are essentially equivalent to individual senses identified earlier for "裁定" and "判決". In particular, 裁決/8 is apparently synonymous to {裁定/1, 判決/1}, 裁決/9 to {裁定/2, 判決/2}, and 裁決/10 to {裁定/4}. However, the low relative frequency for these senses of "裁決" and their absence from other regions suggest that these senses might more appropriately and specifically be replaced by the relevant senses of "裁定" and "判決" instead, as the examples for these senses in Table 4 sound slightly unnatural.

Second, the loose restriction on "裁決" is also reflected in the BJ and TW data (marked with † in Table 4) as the word is often used to refer to the decisions (or the action of making decisions) made by non-judiciary units (e.g. "完全由行政機關裁決" – TW, "世貿組織曾做出裁決" – BJ, etc.). Apart from this, an additional sense for "裁決" related to a referee's judgement (sense 11) was found from the TW data; while sense 4 (decision on monetary compensation) is so domain-specific and technical that it is rarely found outside legal documents.

4.2.3 判決

As seen from Table 5, all regions show a similar sense distribution for "判決", where senses 4 and 1 are the major senses. Sense 5 (i.e. conviction) is specific enough to appear only in the legal texts. Additional uses referring to a referee's judgement in a sports event were observed from TW and SG data.

Table 5. Sense distribution of "判決"

Sense and Example	Legal (30)	HK (30)	BJ (30)	TW (30)	SG (30)
1. [v.] the court decides on the outcome of a case e.g. 本席判決上訴得直。	13.33	20.00	20.00	36.67	16.67
2. [v.] the court resolves an issue in a case e.g. 上訴法院判決受託人有權提出呈請。	3.33	6.67	0.00	0.00	3.33
3. [v.] the court gives an order or sentence e.g. 國際法庭未被授權判決罪犯死刑。	0.00	6.67	3.33	3.33	0.00
4. [n.] the decisions made by court, and related orders e.g. …一宗上訴案的判決…	56.67	66.67	76.67	53.33	70.00
5. [n.] conviction, the judgment of being guilty e.g. 以上是二項罪名定罪判決的立場。	26.67	0.00	0.00	3.33	0.00
*6. [v.] the action of judging by referee in sports events e.g. 眾多裁判判決爭議的確曾阻擾比賽。	--	--	--	3.33	--
*7. [n.] the judgement of referee in sports events e.g. …因爲不滿裁判的一個判決而摔球拍…	--	--	--	--	10.00

4.3 Regional Variation

An obvious difference among the three target words is that verb uses tend to dominate for "裁定"[4] whereas "裁决" and "判决" are used more often as nouns. However, from Table 3, it can be seen that BJ has a lot more nominal usages (sense 6) of "裁定". This observation is nevertheless in line with the findings of Kwong and Tsou [2003] on verb-noun categorial fluidity in Chinese, where in texts from BJ, about 18% of the verbs were found to undergo the verb-noun shift, compared to about 15% of the verbs in the TW and HK data. Hence this linguistic phenomenon might account for the dominance of sense 6 of "裁定" in BJ data.

There remain some interesting questions regarding the differences in sense distribution across the various regions:

1. Table 3 shows that "裁定" is mostly used in SG to state the order given by the court, that is, sense 3. Does this tell us anything about the salience of court orders in SG reportage?
2. Table 3 also shows that "裁定" is least used in sense 3 in BJ, compared to other places. The frequency of its synonymous senses 裁决/2 and 判决/3 is also extremely low in BJ data. So is the concept missing or expressed in another manner?
3. 裁决/3 is absent from SG data, and so is its synonym 裁定/6. Do SG news reports pay little attention to verdicts? Otherwise where has the concept been absorbed into other expressions?

The difference in sense distribution across various regions is on the one hand a result of the different linguistic norms and styles of language use, as exhibited by the dominance of nominal usages in BJ. Hence even though BJ does not use 裁定/3 or its synonyms, the relevant concepts might have been expressed via nominal uses such as 判决/4. On the other hand, it could reflect the varied approaches and perception in different communities regarding the concepts of judgement. For instance, these concepts are apparently less salient in BJ contexts given the relatively low frequency of the target words in BJ data. Moreover, since SG is found to use 裁定/3 heavily but not 裁决/3, it suggests that SG news tends to treat the conclusion (verdict) and the consequence (sentence and order) as a whole.

To further understand this Pan-Chinese variation, we have simultaneously, though only briefly, surveyed the use of more similar and related lexical items from various resources. For instance, among the words mined from the seed morpheme "判" from the corpus materials used in this study, words like "判刑", "判監", "判罰", "判囚" and "判處" (all related to sentencing) are relatively more abundant in HK and TW than in SG, suggesting that HK and TW tend to distinguish between the verdict and the sentence more clearly. On the other hand,

[4] In particular, sense 1 and sense 6 of "裁定" both refer to the same meaning but differ in syntactic category.

a preliminary survey through some internet resources on legal documents (e.g. websites with PRC judgments[5] and Taiwan judgments[6]) shows that terms related to "decision" and "sentencing" like "判定", "判處", "判令", and "判決" are shared by judgments produced in Hong Kong, Mainland China, and Taiwan region. However, a related but probably more specific term, "判命" (which appears to combine decision, sentencing and order), is found only in Taiwan judgments. The variation and the subtle difference among such related terms would be useful and should be captured in a comprehensive Pan-Chinese lexico-semantic resource for the legal domain. Thus the use of these words, their relations with the target words, and their variation in the Pan-Chinese context all require further investigation.

5. Conclusion

Thus in this study we have further analysed the usage and sense distribution of a set of closely related legal terms pertaining to judgement in the Pan-Chinese context. Linguistic data reveal variation in the salience of these concepts in various Chinese speech communities and the distinctiveness of the SG data. Based on the subtleties among various uses, we have further probed the salience of these concepts in the various communities, and the differences therein might be a result of the difference in legal systems. For instance, the use of the target words in HK and SG might be more influenced by translation from English legal terms than in BJ and TW. Alternatives for expressing similar and related legal concepts should be further explored and the study should be expanded with other sets of closely related legal terms.

As mentioned in the beginning of this paper, the subtle differences among the target words may be insignificant to the general readers. However, when it comes to high quality translation, especially translations which bear legal implications, the preciseness therein will definitely be indispensable. Hence our analysis of the use of legal terms in various Chinese speech communities will provide useful information for the construction of a Pan-Chinese legal term lexical resource as we witness the growing maturity of the Chinese language in the legal domain. Such an enriched lexical resource would be useful to legal and paralegal professionals, and for legal document translation between English and Chinese, by machines or by humans, as well as for many other natural language processing tasks.

---

[5] http://www.chinaiprlaw.com/wsjx/wsjx.htm

[6] 法源法律網 http://db.lawbank.com.tw/FJUD/FJUDQRY01-1.asp

Acknowledgements

An earlier version of this paper was presented in the Fifth Chinese Lexical Semantics Workshop, Singapore, June 2004. This work was supported by Competitive Earmarked Research Grants (CERG) of the Research Grants Council of Hong Kong under grant nos. CityU1233/01H and CityU1317/03H.

References

Cheng, C.C., "論同義詞的語意輕重與側重", Presented at the *Second Workshop on Chinese Lexical Semantics*, 2001, Peking University, Beijing, China.

Gao, H., "A Specification System for Measuring Relationship among Near-synonyms of Physical Action Verbs," Presented at the *Second Workshop on Chinese Lexical Semantics*, 2001, Peking University, Beijing, China.

Huang, C-R., K. Ahrens, L-L. Chang, K-J. Chen, M-C. Liu, and M-C. Tsai, "The Module-Attribute Representation of Verbal Semantics: From Semantics to Argument Structure," *Computational Linguistics and Chinese Language Processing,* 5(1), 2000, pp.19-46.

Kwong, O.Y., and B.K. Tsou, "A Synchronous Corpus-Based Study of Verb-Noun Fluidity in Chinese," In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*, 2003, Singapore, pp. 194-203.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, "Introduction to WordNet: An on-line lexical database," *International Journal of Lexicography,* 3(4), 1990, pp. 235-244.

Tsou, B.K., and O.Y. Kwong, "When Laws Get Common: Comparing the Use of Legal Terms in Two Corpora," *Language and Linguistics,* 4(3), 2003, pp. 609-629.

Tsou, B.K., T.B.Y. Lai, and K. Chow, "Comparing Entropies within the Chinese Language," In K.Y. Su, J. Tsujii, J.H. Lee and O.Y. Kwong (Eds), *Natural Language Processing – IJCNLP 2004*, LNAI Vol. 3248, Springer-Verlag, pp.466-475.

Tsou, B.K., W.F. Tsoi, T.B.Y. Lai, J. Hu, and S.W.K. Chan, "LIVAC, A Chinese Synchronous Corpus, and Some Applications," In *Proceedings of the ICCLC International Conference on Chinese Language Computing,* 2000*,* Chicago, pp. 233-238.

《法學詞典》編輯委員會, 《法學詞典（增訂版）》, 上海辭書出版社, 1985.

劉清景, 《新編法律大詞典》, 學知出版事業股份有限公司, 2001.

梅家駒、竺一鳴、高蘊琦、殷鴻翔, 《同義詞詞林》, 商務印書館/上海辭書出版社, 1984.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 533-542

《人民日報》語料庫命名實体分類的研究

The Chinese Named Entity Categorization Based on the People's Daily Corpus

夏迎炬*、于浩*、西野文人*

YingJu Xia, Hao Yu and Fumihito Nishino

摘要

在信息檢索、信息抽取等應用中，命名實体的處理十分重要。本文在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体的細分類進行了深入的研究。提出了命名實体的多級分類并給出了每一級的詳細分類。爲了檢驗該分類体系的實際效果，我們在人民日報語料上進行了初步的標注。并使用常用的基于統計模型的命名實体識別算法在人民日報語料上做了一系列的對比實驗。實驗結果表明：面向機器處理的細分類能有效地提高識別系統的性能并最終有助于信息檢索和抽取。

關鍵字：命名實体、分類、語料庫、自然語言處理

Abstract

Named entity recognition is a very important part of information retrieval and information extraction. Classification is also very important. This paper investigates the sub-classification of named entities from the point of view of information retrieval and information extraction. This paper also presents multi-classification and gives detailed information about each sub-class. We have manually annotated people's daily corpus (1998) and conducted a serial of experiments using the statistical model of named entity recognition. The

* 富士通研究開發中心有限公司，100016 北京市朝陽區霄雲路 26 號鵬潤大廈 B306 室
Internet Application Laboratory, Fujitsu Research & Development Center Co., LTD.
Room B306, Eagle Run Plaza No. 26, Xiao Yun Road, Chao Yang District, Beijing, 100016, P. R. China
E-mail: {yjxia, yu, nisino}@frdc.fujitsu.com

[Received April 12, 2005; Revised June 15, 2005; Accepted August 14, 2005]

experimental results show that the sub-classes presented by this paper can enhance the recognition system's performance and aid information retrieval and information extraction.

Keywords: Named Entity, Classification, Corpus, Natural Language Processing

1. 前言

隨著因特網的飛速發展，網絡信息量呈指數增長。如何使用計算機幫助人們從海量的數據中有效地獲取所需信息是自然語言處理的熱點問題之一。在這一類信息檢索和抽取的應用中，命名實体的自動識別是極爲關鍵的步驟。其主要採用的方法有規則方法、統計方法和統計規則相結合的方法。由于規則的方法需要人工總結、編寫大量的規則并且可移植性差等缺點，人們開始將研究的重點轉移到機器學習方法上[Aberdeen *et.al* 1995; Sekine *et.al* 1998; Borthwick 1999; Sun *et.al* 2002; Bikel *el.al* 1997]。而對于基于機器學習的識別系統而言，語料庫起著至關重要的作用[黃昌寧 2002; 馮志偉 2001]。歷來受到學者們的高度重視，隨著基于統計的自然語言處理研究的不斷深入，對于語料庫的需求也日益強烈。

富士通研究開發中心有限公司與北京大學計算語言學研究所、人民日報信息中心合作，以 1998 年人民日報爲對象，製作了大規模漢語標注語料庫[俞士汶 2000; 段慧明 2002]，并已將上半年部分于 2001 年進行了公開，在很多研究單位得到了使用。在 1998 年全年的人民日報標注語料庫定義中，包括有命名實体的標記，語料庫中的命名實体分爲 4 類[俞士汶 1998]：機構團体名、地名、人名、其他專用名詞。與之對應的是國際上權威的評測會議 Message Understanding Conference 于 1995 年（MUC-6）第一次設立命名實体識別評測專項。MUC（MUC-6）定義了三大類（實体、時間和數位）、七小類（人名、機構名、地名、時間、日期、貨幣和百分比）命名實体。

經過多年的實踐證明，這樣的分類對于應用來說顯得過粗，不能起到很好的定義作用。本文的工作就是在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体細分類的顆粒度進行了深入的研究。提出了命名實体的多級分類体系并給出了每一級的詳細分類定義。力求既要適應語言信息處理與語料庫語言學研究的需要，又要能爲傳統的語言研究提供充足的素材；既要適合計算機的自動處理，又要便于人工校對。在本文的第 2 節將分別介紹對人名、機構名、地名子類分類体系。在第 3 節對命名實体簡稱的分類做了初步探索。最后在第 4 節中給出實驗結果及結論并展望未來的工作。

2. 命名實体細分類

2.1 人名分類

在 1998 年人民日報語料中，人名採用統一的標注模式，從標注上無法區分中國人名和外國人名。這樣的標注語料會給機器學習帶來噪聲。因爲不同國家的人名的內部特徵（主要是人名用字集）存在較大的差別。例如，日本人名的常用山本、太郎、大島、藤田等

字和詞；蘇俄人名常用斯、基、娃等字；歐美人名常用朗、魯、倫、曼、尼等字。所以，將他們分開標注和識別將會提高人名的識別性能。基于這樣的考慮，我們將人名的第一級分爲：中國人名、外國人名和不確定三類。其中不確定類別是爲了減少在機器學習的噪聲而引入的，這樣的分類能儘量保證同一類別中的特徵比較集中。當然在外國人名中，可以嘗試根據其人名用字集分爲不同的國別，比如：日本人名、歐美人名、其他等，這樣可以進一步使特徵集中。

在第一級分類的基礎上，我們注意到：即便是同一大類中的名字（比如中國人名），其構成及用字也有很大的不同，比如單詞型（只有姓）、双詞型（姓名）、三詞型（夫婦双姓等），可以通過這些特徵將人名進一步分類。由此，我們得到了第二級的分類。最后在此分類的基礎上，我們對語料庫中的人名進行了統計，發現有必要按照其構成特徵繼續進行細分類。比如單詞型可以細分爲單字型、二字型、三字型和多字型。需要指出的是分類越細，將導致機器學習的訓練數據越稀疏，這是我們不希望看到的情況，但是考慮到標注的語料庫既要滿足語言信息處理與語料庫語言學研究、又要爲傳統的語言研究提供充足的素材。我們還是決定進行第三級的分類。在機器學習的時候可以根據訓練語料的規模以及特徵的相似性將某些特徵合并在一起，比如將三字型和多字型合并等。

表 1 給出了人名的細分類以及每一分類在語料庫中的出現的頻數。

表 1: 人名分類

總類	一級子類	二級子類	三級子類
人名（48709）	中國人名（39372）	單詞型（2673）	單字型（396）
			二字型（2141）
			三字型（96）
			多字型（40）
		双詞型（36663）	單姓單名（6492）
			單姓双名（29983）
			双姓單名（57）
			双姓双名（105）
			其他（26）
		三詞型（36）	
	外國人名（9227）	單詞型（9208）	二字人名（495）
			三字人名（2824）
			四字人名（2083）
			多字人名（3806）
		双詞型（19）	
	不確定（110）	存疑待查（42）	
		語料錯誤（68）	

2.2 機構名分類

關于機構分類，目前有很多可以借鑒的分類体系。比如按照中華人民共和國民法通則的規定，將組織機構分爲企業、事業單位、社會團体、國家機關四個大類。在此四大類的基礎上又詳細的劃分了若干子類，比如國家機關又可以分爲：國家權力機關、國家權力機關分支機搆、國家行政機關、國家行政機關分支、派出機構、國家司法機關、人民法院、人民法院分支機搆、人民檢察院、人民檢察院分支機搆、政黨機關、政協組織等。但是這樣的分類顯然不符合信息檢索和抽取的需要。在綜合考慮現有的各種分類体系基礎上，我們對語料庫中的機構名進行了詞頻統計。制定了機構團体的二級分類体系。由于中外機構名在用詞等特徵上沒有明顯的差別。而且在下一級的子類的劃分上也基本相同。我們并沒有劃分中國和外國機構名這一級。只是在總類上標明了可以有中國機構和外國機構的區別并給出了其在語料庫中的頻數。

第一級的分類有 23 類，在第一級的基礎上，又對其中的經濟機構類、政治組織類、健康組織、媒体組織、体育機構、教育機構進行了細分類。

其具体的分類和頻數統計見表 2。

表 2: 機構名分類

總類（標注）	一級子類（標注）	二級子類（標注）
機構團体（41141） [中國：32450] [外國：8691]	文學藝術類(art_nt)(2000)	
	經濟機構類（economic_nt）（3410）	銀行（bank_economic_nt）(1020)
		基金會（fund_economic_nt）(296)
		股票名稱（stock_economic_nt）(623)
		其他經濟組織（org_economic_nt）(1477)
	政治組織類（political_nt）（1693）	政治聯盟（org_political_nt）(1465)
		政黨名稱（party_political_nt）(228)
	公司（company_nt）(7960)	
	健康組織（health_nt）（1477）	醫院（hospital_health_nt）(659)
		醫療健康組織，但不是醫院（org_health_nt）(818)
	軍事機構（military_nt）（2255）	
	飯店酒店（hotel_nt）(39)	
	女性組織（woman_nt）（142）	
	媒体組織（media_nt）（1696）	電視媒体（tv_media_nt）(304)
		廣播媒体（radio_media_nt）(183)

		出版社（publishhouse_media_nt）（967）
		電影媒体（movie_media_nt）（183）
	製造業（manufacturer_nt）（1977）	
	研究機構（research_nt）（2156）	
	宗教機構（religion_nt）（73）	
	工會（labourunion_nt）（156）	
	体育機構（sport_nt）（1645）	体育運動隊（team_sport_nt）（783）
		其他体育組織（org_sport_nt）（862）
	教育機構（edu_nt）（4400）	學校（school_edu_nt）（4035）
		其他教育組織（org_edu_nt）（365）
	能源部門（energy_nt）（1276）	
	政府部門（gov_nt）（11133）	
	公安部門（police_nt）（1016）	
	檢查部門（procuratorate_nt）（260）	
	法院（court_nt）（624）	
	律師事務所等（law_nt）（245）	
	海關（ciq_nt）（164）	
	其他組織（otherunion_nt）（2799）	

2.3 地名分類

地名分類的情況與機構名分類情況有些類似，也有很多可以借鑒的現有分類体系。比如全國地名委員會編的《地名信息系統規範》就把地名分爲：國名、首都、省級行政區域駐地、地級市、縣級市等共 76 類別，但這樣的分類更多是從行政區劃上考慮的。對于面向信息檢索和抽取的標注語料庫建設來說，需要綜合考慮信息檢索、機器學習方面的需求。我們最終將地名劃分爲 46 個一級子類。同樣沒有區分中國地名和外國地名，其具体的分類見表 3。

表 3：地名分類

總類	一級子類
地名（ns）（5226） [中國：3921] [外國：1305]	市場（market_ns）（418）
	賓館（hotel_ns）（266）
	劇院禮堂（theater_ns）（129）
	博物館紀念館（museum_ns）（374）
	機場（airport_ns）（255）
	車站（station_ns）（167）
	公園（park_ns）（298）
	草原（grassland_ns）（27）
	大廈寫字樓（mansion_ns）（126）
	地區（area_ns）（845）
	公路（road_ns）（62）
	山脈（mountain_ns）（31）
	街道（street_ns）（113）
	体育運動場所（sportplace_ns）（275）
	行政區（district_ns）（313）
	農場（farm_ns）（57）
	廣場（plaza_ns）（136）
	平原（plain_ns）（47）
	碼頭（dock_ns）（30）
	開發區（developarea_ns）（166）
	教堂寺廟（religion_ns）（51）
	住宅小區（uptown_ns）（51）
	墓地陵園（grave_ns）（55）
	沙漠（desert_ns）（20）
	書店（bookshop_ns）（17）
	電站（powerplant_ns）（5）
	示範區（demonstratearea_ns）（20）
	流域（drainagearea_ns）（35）
	水文站（waterinfostation_ns）（29）

	活動中心（center_ns）（141）
	基地（fundation_ns）（64）
	海峽海域（seaarea_ns）（27）
	餐飲場所（restaurant_ns）（39）
	藥店（drugstore_ns）（3）
	林場（forestcenter_ns）（17）
	茶場（teafield_ns）（4）
	油田（oilfield_ns）（5）
	礦區（mine_ns）（20）
	音樂廳（musichall_ns）（14）
	出版社（publishhouse_ns）（1）
	政府（gov_ns）（29）
	公司（company_ns）（2）
	植物園（arboretum_ns）（17）
	鐵路（railway_ns）（3）
	其他地名（other_ns）（349）
	會議場所（meetinghall_ns）（49）

3. 簡稱分類

在真實語料中，很多命名實体是以簡稱的形式出現。比如 “老張”、“小李”、“京九鐵路”、“京津高速”、“北大”、“政協”等。在 1998 年人民日報標注語料中對這些簡稱的標注方案是：對人名的簡稱標成“人名（nr）”，對其他的簡稱則同一標成“簡稱（j）”。比如：“老張/nr”、“政協/j”、“中/j 美/j 關係/n”。這樣的標注不但混淆了全稱和簡稱的區別（人名簡稱），而且模糊了簡稱之間的界限（地名簡稱和機構名簡稱）。另外，從語義角度、信息檢索和抽取等應用角度應該是將不同類型的簡稱區別開來并且在語料中給予標注。比如：從面向機器學習的角度，“中美關係”較好的標注結果應該是“中/aloc 美/aloc 關係/n”。（其中 aloc 表示簡稱地名）。

基于這樣的考慮，我們將簡稱首先分爲人名簡稱、地名簡稱、機構名簡稱和其他簡稱等幾類。在這幾類簡稱中，機構名簡稱的情況比較複雜，也最難識別。經在語料中統計分析發現，機構名簡稱大致有如下的幾種形式：連續型：“解放軍第 301 醫院（301 醫院）”，不連續型：“北京大學（北大）”，以及混合型：“東風汽車電子儀表股份有限公司（東風電儀）”。更多的具体例子見表 4。通過這樣的分析，我們將語料庫中的簡稱按表 5 所示進行了細分類。

表 4: 機構名簡稱示例

連續簡寫	上海華聯超市股份有限公司	上海華聯
	上海紫江企業集團股份有限公司	紫江企業
	解放軍第 301 醫院	301 醫院
	北京 25 中學	25 中
不連續簡寫	上海證券交易所	上證/上證所
	北京大學	北大
	電子工業部第六研究所	六所
	武漢鋼鐵集團公司	武鋼
連續簡寫與不連續簡寫混合	東風汽車電子儀表股份有限公司	東風電儀

表 5: 簡稱分類

人名簡稱（APER）（716）	
地名簡稱（ALOC）（24760）	
機構名簡稱（AORG）（6378）	連續型簡稱（AORG_SEQ）
	不連續型簡稱（AORG_DIS）
	混合型簡稱（AORG_OTH）

4. 實驗結果及結論

4.1 實驗結果

爲了驗證細分類帶來的效果，我們在人民日報語料上進行了一系列的實驗。實驗採用的[Wu 2003]所示的基于詞類和詞性類的統計模型來進行命名實体識別，訓練語料爲 1998 年 1~5 月份的人民日報標注語料；測試語料爲 1998 年 6 月份的人民日報語料。我們使用的評測指標有：精確率，召回率，F-值

$$准確率 = \frac{正確識別的實体數}{總的識別實体數} ，$$

$$召回率 = \frac{正確識別的實体數}{總的實体數} ，$$

$$F-值 = \frac{2*召回率*正確率}{召回率+正確率} 。$$

首先，我們比較了第 2、3、4 節介紹的人名、地名、機構名細分類和原分類体系之間的差別。然后我們比較了在新分類体系下，使用與不使用簡稱識別的系統性能。實驗

結果如表 6、表 7 和表 8 所示。表 6 是在原分類体系上進行訓練和測試得到的結果，表 7 是在本文介紹的細分類体系上得到的結果，但是不使用簡稱識別。表 8 是在新分類体系上使用簡稱識別的測試結果。從結果上看，對于人名識別來說，新舊分類体系的差別很大。在新的分類体系下系統的性能得到很大的提高。而對于地名和機構名來說，在新的分類体系下，系統性能也有提高但不如人名那樣顯著。其主要原因是人名的細分類更多的抓住了人名的構詞特徵，比如中外人名的區別、單字型和二字型之間的區別等。這樣的劃分是有利于機器學習的特徵抽取的。而地名和機構名的劃分在構詞規律上區別不大，更多的是語法和語義上的區別，比如地名分類中的“賓館”和“大廈寫字樓”類之間以及機構名分類中的“政府部門”和“公安部門”之間從用字規律上基本相同。其主要的區別是職能上的，這已經屬于語義的範疇了，在目前的水準下，處理語義信息是極其困難的。這也是在新的分類体系下，地名和機構名分類不如人名分類效果那麼明顯的原因。

但從表 7 和表 8 的對比，我們可以看出，簡稱的分類和識別對地名和機構名識別的性能提升很明顯。這是對簡稱進行細分類和識別帶來的效果。當然，由于其中加入了簡稱識別技術。簡稱識別的性能也將對系統產生影響。這樣的對比并不是很嚴格。

表 6: 原分類体系下的系統性能

	准確率	召回率	F 值
人名	92.87	89.33	91.06
地名	93.66	88.78	91.15
機構名	84.36	77.96	81.03

表 7: 新分類体系下的系統性能（無簡稱識別）

	准確率	召回率	F 值
人名	94.06	95.21	94.63
地名	93.87	90.78	92.29
機構名	83.76	82.42	83.08

表 8: 新分類体系下的系統性能（有簡稱識別）

	准確率	召回率	F 值
人名	93.96	96.18	95.06
地名	95.77	96.76	95.26
機構名	89.77	86.72	88.22

4.2 結論和未來的工作

本文的在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体的細分類進行了深入的研究。提出了命名實体的多級分類体系并給出了每一級的詳細分類定義并對命名實体簡稱的分類做了初步的探索。在人民日報 1998 年標注語料上根據新的分

類体系進行標注和測試。實驗結果表明：新的分類体系有助于面向信息檢索和抽取的機器學習。這樣的分類体系可以使自動識別系統的性能得到大幅的提高。

常規的命名實体類別包括三大類（實体類、時間類和數字類）、七小類（人名、機構名、地名、時間、日期、貨幣和百分比）命名實体。對于信息提取、文本挖掘、網絡內容管理等應用來說，這些類別已經不能滿足應用的需求，還有一些命名實体也非常必要。例如：“事件類實体”（第一屆中國網球公開賽、第五屆中國國際航空航天博覽會、首屆英國戲劇舞蹈節、中國首屆網絡相聲大賽、澳大利亞文化周）、“著作類實体”（斯德哥爾摩環抱公約、《中國人民銀行金融機構反洗錢規定》、《神雕俠侶》）、“股票名稱”（銀河創新、富龍熱力 、深萬山Ａ）等等。我們需要對新的實体類型進行大範圍的調查，看需要擴充哪些類型的命名實体，分別出現在哪些領域，從而建立起命名實体分類体系和與之配套的識別和標注工具，爲信息提取、文本挖掘、網絡內容管理等應用奠定基礎。

參考文獻

段慧明、松井久仁于、徐國偉、胡國昕、俞士汶， “大規模漢語標注語料庫的製作與使用”， *語言文字應用，2002 年，第 2 期，* pp. 72-77.

馮志偉， “中國語料庫研究的歷史和現狀”， *Proceedings of ICCC2001,* pp.1-24.

黃昌寧、李涓子， “語料庫語言學”， *商務印書館， 2002,* pp. 244-256.

俞士汶、朱學鋒、段慧明， “大規模現代漢語標注語料庫的加工規範”， *多語言信息處理國際會議，2000* , pp. 19-24.

俞士汶等， “現代漢語語法信息詞典詳解”，*清華大學出版社， 1998*, pp. 546.

Aberdeen.J, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain, “MITRE: Description of the ALEMBIC System Used for MUC-6,” In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995, pp. 141-155.

Bikel, D.M., S. Miller, R. Schwartz, and R. Weischedel, “Nymble: a high-performance learning name-finder,” *In Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington D.C., 1997, pp. 194-201.

Borthwick, A., “A Maximum Entropy Approach to Named Entity Recognition,” PhD Dissertation*,* New York University, 1999.

Sekine, S., R. Grishman, and H. Shinou, “A decision tree method for finding and classifying names in Japanese texts,” In *Proceedings of the Sixth Workshop on Very Large Corpora,* Canada, 1998, pp.171-178

Sun, J., J. Gao, L. Zhang, M. Zhou, and C. Huang, “Chinese Named Entity Identification Using Class-based Language Model,” In *Proceedings of the 19th International Conference on Computational Linguistics,* Taipei, 2002, pp. 967-973.

Wu, Y., J. Zhao, and B. Xu, “Chinese Named Entity Recognition Combining Statistical Model wih Human Knowledge,” *ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition,* 2003, Japan.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 543-552

双向考察和驗證：

并列成分中心語的語義關係和 CCD 的

名詞語義分類体系[1]

Bidirectional Investigation:

The Semantic Relations between the Conjuncts and the Noun Taxonomy in CCD

吳云芳*、李素建*、李芸*、俞士汶*

Yunfang Wu, Sujian Li, Yun Li and Shiwen Yu

摘要

并列結構是語言信息處理中的難點。本文一方面基于中文概念詞典 CCD 的語義分類体系來考察名詞性并列結構并列成分中心語的語義關係，大部分并列結構呈現出語義相似的特性，少部分并列結構呈現出語義相關、語義相對的特性；另一方面透過并列成分中心語的語義關係來審視 CCD 的語義分類体系，對語義分類体系中的一些語義類、以及語義類之間的關係進行了深入思考。這是語言形式（并列成分中心語）和語言意義（語義類和語義關係）的双向考察和驗證。

關鍵字: 并列結構、語義關係、語義相似、分類体系

---

[1] 本文研究得到了中國國家 973 項目(2004CB318102)、中國博士后科學基金(2004035029)和 863 項目(2001AA114210)的支持。

* 北京大學計算語言學研究所，北京 100871
Institute of Computational Linguistics, Peking University, Beijing, 100871
E-mail: {wuyf, yusw}@pku.edu.cn

[Received April 10, 2005; Revised June 16, 2005; Accepted August 15, 2005]

Abstract

This paper presents a bidirectional investigation on linguistic form and meaning. On the one hand, based on the Chinese Concept Dictionary (CCD), this paper examines the semantic relations between the heads of the conjuncts of nominal coordination. Most of the conjuncts show semantic similarity, a few of them show semantic association, and a few of others show semantic opposition. On the other hand, the semantic relations between the conjuncts provide a new perspective on the noun taxonomy and suggest ways to improve the taxonomy.

Keywords: Coordinate Structure, Semantic Relation, Semantic Similarity, Taxonomy

1. 引言

并列結構（coordinate structure）是語言信息處理中的難點。一般認爲并列成分是相似的，并列結構的自動識別研究幾乎全是圍繞并列成分的相似性來進行。[Okumura and Muraki 1994] 和 [Agarwal and Boggess 1992] 對英語并列結構的研究，[Kurohashi and Nagao 1994] 對日語并列結構的研究，[周強 1996] 和 [孫宏林 2001]對漢語并列結構的研究，都是基于“并列成分相似”這樣的語言學假設，在此前提下設計規則和演算法。漢語語言研究同樣認爲并列成分是相似的，[吳競存，梁伯樞 1992] 指出，詞性相同、結構相同、語義類相同、音節數相同的項并列是最理想、最嚴格的并列。

中心語（head）是當代句法理論中的一個核心概念，擴展的短語結構文法（GPSG）、中心語驅動的短語結構文法（HPSG）都把中心語擺在了重要的位置。中心語是其父親節點句法語義特徵的集中体現者，那麼，并列成分的相似也應該集中体現在各并列成分的中心語上。CCD（Chinese Concept Dictionary，中文概念詞典）是北京大學計算語言學研究所研製開發的漢語語義詞典[于江生、俞士汶 2002]，基本上沿襲了 WordNet 的語義分類体系。

本文一方面基于 CCD 的語義分類体系來考察名詞性并列結構并列成分中心語的語義關係，一方面透過并列成分中心語的語義關係來審視 CCD 的語義分類体系，這是一個双向考察和驗證的過程。[Resnik 1993] 基于早期版本的 WordNet 的名詞語義分類体系研究表明，動詞 drink 的直接賓語是 beverage 的下位詞語；在新版本的 WordNet 的語義分類体系中，[Miller 1999] 引用 Resnik 的研究成果來證明上下位語義關係存在的合理性，這是語言現象和語義分類体系的双向驗證過程。

2. 從 CCD 看并列成分中心語的語義關係

沿襲 WordNet 的分類体系，CCD 的名詞分爲了 25 個基本語義類[2]。在經過了詞語切分和

[2] 這 25 個語類是：1) 動物(animal), 2) 人(person), 3) 植物(plant), 4) 人工物(artifact), 5) 自然物(natural object), 6) 身体(body), 7) 物質(substance), 8) 食物(food), 9) 屬性(attribute), 10) 數量

詞性標注的《人民日報》1998 年 1 月 1－10 日語料基礎上，作者手工標注了語料中出現的有標記的短語層面的并列結構[3]，從中抽取了 2101 個名詞性并列結構，基于 CCD 對并列成分中心語語義關係進行了定量考察。本文的例句均取自于此。待試驗的并列結構都是兩項的，多項并列結構可看作是多個兩項并列結構的疊加，和兩項并列結構應該具有相同的語義約束。待試驗的并列結構僅包括并列成分中心語是名詞的并列結構，如“被[習慣勢力和陳舊觀念]所束縛”，“全部是由[國家、集体]投資”。對名詞性并列結構，各并列成分的最右端一個詞默認爲是中心語；當并列成分是光杆詞語時，其自身也就是中心語。考察兩個并列成分中心語的語義關係，其計算機操作過程可概要地敍述爲：1）提取兩個并列成分的中心語，并列標記之前一個詞是前并列成分的中心語，并列結構結尾處最后一個詞是后并列成分的中心語；2）在 CCD 名詞語義知識庫中對應各中心語的語義類，當詞語是多義詞有多個語義類歸屬時，由人工甄別選擇正確的語義類；3）生成并列成分中心語語義類同現列表。考察結果如表 1 所示。

表 1：名詞性并列結構并列成分中心語語義關係考察

有共同祖先節點（屬于同一語義類）:	1639	78%
無共同祖先節點（不屬于同一語義類）:	462	22%
總計：	2101	100%

表 1 顯示，78%的名詞性并列結構其并列成分的中心語屬于同一語義類，是“同類并列”，呈現出語義上的相似性（semantic similarity）；而有 22%的名詞性并列結構其并列成分的中心語不屬于同一語義類，是“非同類并列”。這麼大比例的“非同類并列”與我們的先驗期待不相符合，可能存在兩個原因：1）或是 CCD 的 25 個名詞基本語義類的設定不合適，至少從并列結構的角度來看不合適；2）或是并列結構本來就不是我們所想像的那樣完全遵從“同類并列”的原則。前一種可能爲我們提供了一個新的視角來審視 CCD 的語義分類体系；后一種可能要求我們重新分析并列成分中心語的語義關係。

2.1 并列成分中心語語義相似

大多數名詞性并列結構并列成分中心語呈現出語義相似的特性。CCD 是用標記樹來表示語義關係的，我們就用“樹”的術語來描述這些語義關係：同一初始語義類下并列的概念稱爲兄弟節點；同一初始語義類下的上下位概念，不論距離遠近稱爲祖孫節點；同一初始語義類下的其他概念，如果不屬于同一同義詞集合（synset）、不形成兄弟節點和祖孫節點，則稱爲遠距離節點。根據并列成分中心語在語義分類樹上的相互位置，語義相似又可分爲 5 種情況。1) 詞形相同。如“競爭機制和激勵機制”，這約占并列結構總數

(quantity), 11) 關系(relation), 12) 通信(communication), 13) 時間(time), 14) 認知(cognition), 15) 情感(feeling), 16) 動機(motivation), 17) 自然現象(natural phenomenon), 18) 過程(process), 19) 行爲(activity), 20) 事件(event), 21) 群体(group), 22) 處所位置(location), 23) 所有物(possession), 24) 形狀(shape), 25) 狀態(state)。

[3] 這個標注了并列結構的語料從 www.icl.pku.edu.cn 網址上可自由下載，供研究之用。

的 7%。2) 同一個同義詞集合，如“產品有[20 個大類、1000 多個品種]”，這約占并列結構總數的 3%。3) 兄弟節點。如“促進整個[長江、黃河]流域生態環境的好轉”，這約占并列結構總數的 22%。4) 祖孫節點。如“中國願意加強同[聯合國和其他國際組織]的協調”，這約占并列結構總數的 3%。5) 遠距離節點。如“[專家、各界觀眾]也提出許多修改意見”，這約占并列結構總數的 43%。

2.2 并列成分中心語語義相關

語義相關（semantic association）是另一種重要的詞語之間的語義聯繫。人腦思維中容易同時激活同一情境（situation）或同一框架（frame）下的不同概念，不同語義類并不能妨礙這種激活，情境或框架足可以成爲激活因子（trigger）。例如，面對“商業事件”這一情境時，人們很容易聯想到買主、賣主、商品、錢，以及買、賣的行爲[Fillmore 1982]。又如，面對“醫療行爲”這一情境時，人們很容易聯想到醫生、護士、醫院、疾病、費用等等相關概念[董振東、董強 2000]。同一情境下的相關概念在一定的語境中就可形成并列。例如：

（1） a 有利于提高[企業和資金]運作效率。

（企業[＋社會團体]，資金[＋所有物]，情境：商）

b 從這裏出發的[車輛和人群]如洪水般流向麥加。

（車輛[＋人工物]，人群[＋人們]，情境：道路交通）

c 促進更多的[中文、中國]信息上網際網路。

（中文[＋通信]，中國[＋處所位置]，情境：中國）

d 環境教育的[師資、教材]都非常缺乏。

（師資[＋人]，教材[＋人工物]，情境：教育）

e 造就出一批批自強不息、直面挑戰的[企業和企業家]。

（企業[＋社會團体]，企業家[＋人]，情境：企業）

這些不同語義類的詞語因在同一個情境下共存而可形成并列，并列的詞語通過不同的方式“指引了（index）”或是“喚起了（evoke）”相同的普遍情境。HowNet 致力于反映概念之間和概念的屬性之間的各種關係，同一情境下的不同概念之間存在著相關聯的描述，并列成分中心語語義相關在 HowNet 的描述中也可得到部分驗證。例如，對（1）a、d 并列成分中心語，HowNet 的描述是[4]：

[4] 此處參考的是 HowNet 2000 版本，向董先生表示謝意。

（2） a 企業：InstitutePlace|場所,*produce|製造,*sell|賣,commercial|商
資金：$spend|花費，#money|貨幣，commercial|商
d 師資：human|人,*teach|教,education|教育,mass|眾
教材：readings|讀物,*teach|教,education|教育

2.3 并列成分中心語語義相對

有時并列成分中心語呈現出語義相對的特性。例如：

（3） a 接收河西醫院全部[人員和資產]。 （人員[＋人們]，資產[＋所有物]）
b 漫漫史河的[許多實事、眾多人物]。 （實事[＋事件]，人物[＋人們]
c [社會心理、人們情感]變化是值得抒寫的。（心理[＋認知]，情感[＋情感]
d 贏得了寶貴的[時間和空間]。 （時間[＋時間]，空間[＋處所位置]）
e 典型本身的[真實事蹟和先進思想]， （事蹟[＋行爲]，思想[＋認知])

（3）中并列的概念在某種意義上是對立的，表示兩個互補的集合，這兩個集合相并就形成一個對語言交際而言完整的集合，對這個完整的集合我們還無法用一個更爲抽象的語詞來指稱。人們常說“人財兩空”，“人”和“財”在漢語言人們的認知世界中是對立的和互補的，因此有了 a 的并列。同樣“人”和“事”、“心理”和“情感”、“時間”和“空間”也是對立互補的。e 是“認知”類名詞和“行爲”類名詞并列。哲學上強調理論和實踐的統一，在人們的思維中同樣注重“認知”和“行爲”的辨正統一，“認知”和“行爲”類名詞在語言中經常形成并列結構：

（4） a 提出了近 15 年內的基本[措施和政策]。
（措施[＋行爲]，政策[＋認知]）
b 以自己的[聰明才智和實際行動]，譜寫青春之歌。
（聰明才智[＋認知]，行動[＋行爲]）

2.4 并列成分中心語語義既不相似也不相關也不相對

語言中存在少數的并列結構，其并列成分中心語的語義既不相似也不相關也不相對。例如：

（5） a 草案確定了反恐怖活動戰略計畫的執行[機構和辦法]。
b 領導幹部受到[人民和法律]的監督。

對此我們還無法進行有效的描述和解釋。[儲澤祥等 2002]從“語用需要、經濟原則”的角度描述兩個名詞的非常規聯合，但沒有涉及句法語義的解釋。

3. 從并列成分中心語語義相似看 CCD 的名詞語義分類体系

把名詞性并列結構并列成分中心語語義相似看作是一種客觀存在的語言特性，那麼并列結構爲我們提供了一個很好的視角來審視 CCD 的語義分類体系，這種審視對其他的語義分類体系也很具參考價值。

3.1 “人們”、“社會團体”語義類名詞和“人”語義類名詞可形成并列—考慮移動併入

CCD 在“群体（group）”語義類下設有“人們（people）”一小類（記作[＋人們]），表示“任何一群人（any group of human beings）”，而 25 個基本語義類中又設有“人（person）”一類。現代漢語中，[＋人們]和[＋人]名詞經常形成并列，例如：

（6） a 日益被[各層領導和社會公眾]所認識。 （領導[＋人]，公眾[＋人們]）
b 雅俗共賞，極受[專家和人民群眾]喜愛。 （專家[＋人]，群眾[＋人們]）

[＋人們]、[＋人]名詞能自由形成并列，而可以不論是否是“群体”，即數的多少，這是由于漢語沒有數的形態變化而造成的。b 中并列結構若翻譯成英語必得是“experts and common people”。“人們”、“人”這兩個語義類在漢語中是相近的，應該合併在一起。事實上，董振東先生的 HowNet、北京大學的《語義詞典》[王惠等 2003]都是將“公眾”、“群眾”這樣的詞置于“人”語義類下。因此，在 CCD 中可以將“人們”小類從“群体”類中移出併入“人”語義類。

CCD 在“群体（group）”語義類下設有“社會團体（social_group）”一小類（記作[＋社會團体]），[＋社會團体]名詞經常和[＋人]名詞形成并列，例如：

（7） a [求職者和用人單位]反映最爲強烈的，（求職者[＋人]，單位[＋社會團体]）
b [旅客和航空公司]都受到損失， （旅客[＋人]，公司[＋社會團体]）

[＋人]名詞和[＋社會團体]名詞都具有“施事”功能，很多動詞對它們具有相同的選擇限制（selectional restriction）。例如，“反映”的主体可以是“求職者”，也可以是“單位”，因此可形成 a 的并列；“受到損失”的主体可以是“旅客”，也可以是“公司”，因此可形成 b 的并列。除了人所特有的一些生理動作[＋社會團体]名詞不能勝任，例如不能說“**單位吃”，“**公司跑”，[＋社會團体]名詞可以充當大多數動詞的施動者，如“單位贈送錦旗”，“公司轉讓債權”，“航空公司請求延期”等等。雖然表面上[＋

社會團体]名詞不具有生命，但它由具有生命的人所組成，并且由其中的代表法人來實施某種行爲，因此句法功能上[＋社會團体]名詞和[＋人]名詞有很多相似之處。《語義詞典》將“社會團体”置于“人”語義類下作爲一個次類[5]，這是比較合適的。由此，在 CCD 中可以將“社會團体”小類從“群体”類中移出併入“人”語義類。

同爲“群体”語義類的[＋人們]名詞和[＋社會團体]名詞可形成并列。例如：

（8） a 要求[各國政府和全人類]採取緊急行動，（政府[＋社會團体]，人類[＋人們]）
b [地方政府和人民群眾]積極支持部隊，（政府[＋社會團体]，群眾[＋人們]）

因此，從并列結構形成的角度考慮，“人們”、“社會團体”可適當併入“人”語義類。

3.2 “社會團体”語義類名詞和“行政區”語義類名詞可形成并列——考慮移動靠近

“群体”語義類下的“社會團体”名詞和“處所位置”語義類下的“行政區（district）”名詞（記作[＋行政區]）可以形成并列。[＋社會團体]名詞既可以指稱共用某些社會關係的人們，也可以表示這些人們所在的處所位置，例如“銀行”，在（9）a 中表示“銀行的領導者或人們”，而在（9）b 中表示“處所位置”之義。反之，[＋行政區]名詞既可以表示占有一定空間的處所位置，又可以指稱相關的社會團体，例如“北京”，在（10）a 中表示“處所位置”之義，而在（10）b 中表示“北京的領導者或人們”。表現在并列結構上，[＋社會團体]名詞和[＋行政區]名詞可以自由形成并列，如（11）中的例子。

（9） a 這 13 家銀行也作出了積極反應。
b 他走進了那家銀行。

（10）a 1921 年 9 月 26 日生于北京。
b 北京按照國際奧會的要求，如期將申辦報告文本送交國際奧會審閱。

（11）[俄羅斯和北約]建立戰略夥伴關係。（俄羅斯[＋行政區]，北約[＋社會團体]）

由此可見，“社會團体”和“行政區”這兩個語義類在漢語中是相近的，語義類設置中應使兩者靠近。

[＋社會團体]名詞可以和[＋人]名詞形成并列，[＋社會團体]名詞也可以和[＋行政區]

[5] 《語義詞典》將此語義記作“團体(group)”。

名詞形成并列，但[＋人]名詞卻很少能夠和[＋行政區]名詞形成并列。可見，詞語之間的并列關係是不可傳遞的。

3.3 “抽象物”類語義類更易形成并列—分類宜粗

“實体（entity）”類語義類形成并列結構時，其語義類相同要求更細，而“抽象物（abstraction）”類語義類形成并列結構時，其語義類相同要求略粗。事實上，同屬于“抽象物”類的“屬性”、“關係”、“通信”、“認知”、“群体”、“狀態”6 個語義類之下的詞語可以相當自由地彼此形成并列。例如：

（12）a 要堅持不懈改善[生態環境和生產條件]。（環境[＋狀態]，條件[＋屬性]）
b 可持續發展的[定義和內容]。（定義[＋通信]，內容[＋認知]）
c 被[習慣勢力和陳舊觀念]所束縛。（勢力[＋屬性]，觀念[＋認知]）
d 有著[悠久的文明和豐富的文獻傳統]。（文明[＋群体]，傳統[＋認知]）
e 一國兩制事業的[可行性和輝煌前景]。（可行性[＋屬性]，前景[＋狀態]）

人們對抽象事物的認識其實并沒有那麼清晰的分類意識。假如問一個人“環境”、“傳統”、“可行性”的語義類分別是什麼，他或許會回答“環境”的語義類是“認知”，“傳統”的語義類是“通信”，“可行性”的語義類是“狀態”。各家語義分類体系對抽象詞語的歸類也存在諸多的不一致性。《語義詞典》將“環境”和“傳統”籠統地歸入“抽象事物”，將“可行性”歸入“屬性”。HowNet 將“環境”和“可行性”歸入“屬性”，將“傳統”歸入“規矩”（相當于 CCD 中的“認知”類）。而對具体事物（實体）就是另一番光景了，沒有人會認爲“桌子”是“食品”，或者“狗”是“人工物”。表現在語言上，具体事物（實体）的語義類之間不能隨意并列，偶爾并列也依賴于一定語境的支撐。現代漢語并列結構的形成啓示我們，對路徑長度相同的兩對節點，“具体事物”類下的兩個節點語義距離較大，而“抽象事物”類下的兩個節點語義距離較小。由此可知，一方面在并列結構的自動識別過程中，對抽象名詞的語義相似性要求可適當放寬，而對具体名詞的語義相似性要求需適當加嚴；另一方面在語義類設置過程中，對抽象類名詞的分類宜粗而不宜過細。

4. 從并列成分中心語語義相對看 CCD 的名詞語義分類体系

4.1 高度抽象的詞語容易形成并列—宜從一個新的角度進行語義歸類

上文 2.3 談到，表示相對意義的詞語經常形成并列。例如：

（13） a [社會心理、人們情感]變化是值得抒寫的。（心理[＋認知]，情感[＋情感]）
b 贏得了寶貴的[時間和空間]。 （時間[＋時間]，空間[＋處所位置]）

需要注意的是，（13）中各并列成分多是概括的、抽象的詞語，它們不指稱具体的概念，一旦將其中一個抽象概念換作具体概念，并列就不能成立。這種抽象概念往往就是某個語義類的“標籤”，其本身的意義和用法與語義類內部具体詞語差別很大。例如“時間”的語義類是[＋時間]，但和具体的時間詞“今天”、“明年”用法語義差別很大，可以說“今天學習”，但不能說“**時間學習”。“時間”能和別的語義類的詞語形成并列，如“時間和精力”，“時間和空間”，但具体的時間詞卻只能跟時間詞自己并列，不能跟別的語義類的詞語形成并列，“**今天和精力”，“**明年和空間”這樣的并列在語言中是不存在的。又例如“情感”的語義類是[＋情感]，但它和具体表示情感的詞語（例如“溫情”、“恐慌”）在用法語義上差別很大，“情感”和“溫情”、“恐慌”在形成并列結構時鮮有共同點。像語義類標籤的這些高度抽象的詞語可看作是廣義的屬性（attribute），語義類之下的具体實例 （instances）可看作是屬性值（attribute values），我們有 VALUE(時間) = 今天|明年，VALUE(情感)=溫情|恐慌。我們懷疑，類似語義類標籤的這些高度抽象的詞語，是否應該從一個新的角度進行語義歸類。

5. 結語

本文一方面基于 CCD 的語義分類体系，考察了現代漢語名詞性并列結構并列成分中心語的語義關係，呈現出四種：語義相似、語義相關、語義相對、語義既不相似也不相關也不相對。語義相似是并列成分之間最主要的語義關係，但并不是所有的并列成分都呈現出語義相似的特性。另一方面，透過并列成分之間的語義關係其中主要是語義相似的關係，我們重新審視了 CCD 的語義分類体系，對詞語之間的語義關係進行了深入的思考，爲語義類的設置提供了一些有價值的參照座標。本文對并列成分中心語的語義關係和 CCD 的名詞語義分類体系進行了双向考察和驗證，這其實也就是形式和意義的相互驗證。

參考文獻

儲澤祥等，《漢語聯合短語研究》，長沙，湖南大學出版社，2002。

董振東、董強，知網. 見：http://www.keenage.com. 2000

孫宏林，《現代漢語非受限文本的實語塊分析》，北京大學計算機系博士學位論文，2001。

王惠、詹衛東、俞士汶，“現代漢語語義詞典規格說明書”，《漢語語言與計算學報》,13(2), 2003, pp.159-174.

吳競存、梁伯樞，《現代漢語句法結構與分析》， 北京，語文出版社，1992。

于江生、俞士汶， “中文概念詞典的結構”，《中文信息學報》，16(4), 2002, pp.12-20 轉 44 頁.

周強，《漢語語料庫的短語自動劃分和標注研究》，北京大學計算機系博士學位論文，1996。

Agarwal, R., and L. Boggess, "A simple but useful approach to conjunct identification," In *Proceedings of 30th Annual Meeting of Association for Computational Linguistics,* 1992, Newark, Delaware, pp. 15-21.

Kurohashi, S., and M. Nagao,1994. "A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures," *Computational Linguistics* 20,(4), 1994, pp. 507-34.

Miller, A., "Nouns in WordNet", In Fellbaum, C., (ed.), *Wordnet: An Electronic Lexical Database*, Cambridge: MIT Press, pp. 23-46, 1999.

Okumura, A., and K. Muraki, "Symmetric pattern matching analysis for English coordinate structures," In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994, University of Stuttgart, pp. 41-46.

Resnik, P., *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. Dissertation, University of Pennsylvania, 1993.

Fillmore, C. J., "Frame semantics," In *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Corporation, 1982, pp.111-137. 中譯本：詹衛東譯，2003，框架語義學。《語言學論從》第 27 輯，北京：商務印書館。382-412 頁。

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 553-570

Source Domains as Concept Domains in Metaphorical Expressions

Siaw-Fong Chung*, Kathleen Ahrens* and Chu-Ren Huang+

Abstract

The use of lexical resources in linguistic analysis has expanded rapidly in recent years. However, most lexical resources, such as WordNet or online dictionaries, at this point do not usually indicate figurative meanings, such as conceptual metaphors, as part of a lexical entry. Studies that attempt to establish the relationships between literal and figurative language by detecting the connectivity between WordNet relations usually do not deal with linguistic data directly. However, the present study demonstrates that SUMO definitions can be used to identify the source domains used in conceptual metaphors. This is achieved by identifying the relationships between metaphorical expressions and their corresponding ontological nodes. Such links are important because they show which lexical items are mapped under which concepts. This, in turn, helps specify which lexical items in electronic resources involve conceptual mappings. Looking specifically at the concept of PERSON, this work also establishes connectivity between lexical items which are related to "Organism." Therefore, the methodology reported herein not only aids the categorizing of lexical items according to their conceptual domains but also can establish links between these items. Such bottom-up and top-down analyses of lexical items may provide a means of representing metaphorical entries in lexical resources.

Keywords: Concept, Ontology, Conceptual Metaphor, Source Domain, WordNet, SUMO

* Graduate Institute of Linguistics, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei, Taiwan R.O.C. 106. Phone: +886-2-3366-4104 Fax: +886-2-2363-5358
E-mail: claricefong6376@hotmail.com; kathleenahrens@yahoo.com
+ Academia Sinica, No.128, Sec. 2, Academia Road, Nangang, Taipei, Taiwan R.O.C. 115
Phone: +886-2-2352-3108 Fax: +886-2-2785-6622
E-mail: churen@gate.sinica.edu.tw

[Received April 15, 2005; Revised June 15, 2005; Accepted August 15, 2005]

1. Introduction

Many studies have found that metaphors are not represented separately in lexical resources [Alonge and Castelli 2002ab, 2003; Peter and Wilks 2003; Lönneker 2003]. When metaphors are found in lexical resources, they are most often represented using meaning entries in addition to non-metaphorical ones. WordNet [http://www.cogsci.princeton.edu/~wn/; Fellbaum 1998] is one of the lexical resources that sometimes lists metaphorical meanings as different senses along with other non-metaphorical ones.

However, where conceptual metaphors [Lakoff 1993; Lakoff and Johnson 1980] are concerned, there is no uniform representation of source (concrete) and target (abstract) domains in WordNet. Due to this fact, studies have been carried out which have attempted to represent figurative meanings by indicating the source-target domain pairing in addition to the literal meaning [Lönneker 2003; Alonge and Lönneker 2004; Peters and Wilks 2003]. In order to do so, these studies first established the relationship between the literal and figurative entry in the lexical resource. As a result, the mapping between the source and target domains could then be extracted.

Researchers who have attempted to incorporate metaphors into WordNet include Eilts and Lönneker [2002], who created the Hamburg Metaphor Database, a database which provides French and German metaphors by creating links between metaphorical expressions and their related synsets in WordNet. Another possible way to establish the relationships between literal and figurative entries is to determine the semantic relations between lexical entries. For instance, Lönneker [2003] and Peters and Wilks [2003] suggested that the meronymic relations in WordNet can be used to identify the links between lexical items by establishing a connection between the *event* and its *participants* or the *action* carried out by the participants. However, neither of these studies used ontologies to link the concepts in the source or target domain. Alonge and Castelli [2003] were the first to suggest that the EuroWordNet Top Ontology needs to be extended with more concepts in order to deal with figurative language. The advantages of ontologies have been outlined by Navigli and Velardi [2004]; they include the fact that a) an ontology has a wide "coverage" of domain concepts; b) that it is a result of "consensus" reached by a group of people and c) that most importantly, it is easily accessible through electronic resources.

Most of the above mentioned works, which tried to generate the underlying connectivity between synsets, have based their mechanisms on distributed models of processing [Rumelhart and McClelland 1986] or the coarse-coding of lexicons [Harris 1994]. Harris stated that the "representational units" in a coarse-coding mechanism "do not match the information represented…in a one-to-one fashion" [Harris 1994]. Rather, they are connected to one another, and when one unit is activated, the others will also be activated. It is through

these units that clusters of information (which contain one or more concepts) are built. This activation of concepts governs the generation of underlying relations between WordNet relations.

Working within this framework, our aim in this work is to establish links between metaphorical items by identifying the shared concept carried by a cluster of lexical items. For instance, when one active concept (from a lexical item) such as "Growth" is activated, the related concept, such as "Organism" is also activated. This link between lexical items in the same domain is necessary to show which lexical items are mapped under which concepts; i.e., lexical items with the same shared concepts will be sorted into the same source domain of a metaphor. In this paper, the building of a "concept" involves the use of knowledge nodes from an ontology, which is a shared understanding of some domain of interest [Uschold and Gruninger 1996]. Keil [1979] stated that the knowledge representation in an ontology "has unique properties and is highly structured. Moreover, it constrains the nature of semantic and conceptual knowledge." The particular ontology we use in this paper, SUMO, was developed by the IEEE Standard Upper Ontology Working Group. Huang, Chung and Ahrens (In press) applied SUMO to explore how an ontology can be used to predict metaphorical mappings. Our work herein extends that of Huang *et al*. (In press) by focusing on how the source domain of a metaphor can be determined via SUMO definitions.

All metaphorical instances in this work are identified using the Conceptual Mapping (CM) model [Ahrens 2002], and data in this study come from both English and Chinese corpora. Searching for ontology nodes is facilitated by the Academia Sinica Bilingual Ontological Wordnet [Sinica BOW, Huang, Chang and Lee (2004) http://BOW.sinica.edu.tw], a system that integrates WordNet, the English-Chinese Translation Equivalents Database (ECTED), and SUMO.

2. Conceptual Metaphor, Concept Domain and Ontology

Ahrens [2002] suggested that the metaphorical expressions can be analyzed in terms of the entities, qualities and functions that can map between a source and a target domain. When these conventionalized metaphorical expressions have been examined, an underlying reason for these mappings can then be postulated. This particular study collected data from native speaker intuitions to determine the mappings from the source to the target domain. For example, in the three examples from the metaphor LOVE IS PLANT, given below, the Mapping Principle (MP) of "Love is understood as plant because plants involve physical growth and love involves emotional growth" was extracted based on the fact that all the examples in some way had to do with growth.

1. (a) 兩 人 的 愛 苗 最近 才 剛 萌芽
liang ren de ai miao zuijin cai gang mengya
two people MOD love seedling lately just recently sprout
'Their love just begins to sprout lately.'

(b) 我 對 他的 愛意 漸漸 滋長
wo dui tade ai-yi jianjian zizhang
I for his love gradually grow
'My love for him has grown gradually.'

(c) 愛情 需要 辛勤 的 灌溉
aiqing xuyao xinqin de quanqai
love need industriously water
'Love needs to be watered industriously.'
[Ahrens 2002]

Ahrens, Chung and Huang [2003] extended this study by proposing a corpus-based approach to establish the systematicity between source and target domain pairings (i.e. Mapping Principles (MPs)). They suggest that each source-target domain pairing will have a prototypical instance of mapping determined by the most frequent mapping, as compared with other mappings. In a later paper, Ahrens 2004, they use Suggested-Upper-Merged-Ontology (SUMO) in combination with WordNet to determine Mapping Principles when there is no highly frequent mapping. SUMO can be used to infer knowledge through automatic reasoning as well as to constrain the falsifiability of the MP.

This study extends the previous works of Ahrens, Chung and Huang [2003, 2004] by using SUMO to define the concepts involved in source domains through the use of two major databases -- WordNet (1.6), and SUMO nodes along with their definitions. The integration of WordNet and SUMO by Niles and Pease [2003] enables us to examine the ontological nodes in SUMO that have hyperlinks to WordNet semantic definitions.

3. Metaphor Analysis: CAREER IS A PERSON as a Sample

In order to find conceptual metaphors in corpora, a single target word was used for the target domain. Four target domains were chosen, namely, CAREER, CULTURE, STOCK MARKET and ECONOMY, of which the latter two are composed Chinese-English data (see Table 1).

The target domains of STOCK MARKET and ECONOMY have been discussed by Ahrens *et al*. [2003] and Chung, Ahrens and Sung [2003], and in this paper, we further refine the previous findings. In this paper, also, the Chinese only CAREER and CULTURE target domains will be added to strengthen the methodology discussed herein.

Table 1. The sources and frequencies of the corpora instances found

Target domains	**Sources**	**Number of hits**	**Number of metaphorical expressions**
(Chinese) *Shiye* 事業 CAREER	Sinica Corpus (http://www.sinica.edu.tw/SinicaCorpus/)	1062	84
(Chinese) *Wenhua* 文化 CULTURE	Sinica Corpus	2000	335
(Chinese-English) *Gushi* 股市 STOCK MARKET	1997 *Huashishingwen* 華視新聞焦點 1997 *Gungshangshibao* 工商時報 [Chung, Ahrens and Sung, 2003]	NA	135
	1994 Wall Street Journal (Available at the Language Data Consortium http://www.ldc.upenn.edu/ldc/online/index.html)	500	130
(Chinese-English) *Jingji* 經濟 ECONOMY	Sinica Corpus	2000	311
	1994 Wall Street Journal	500	215

The four target domains were used to extract instances from the corpora. After all instances were extracted, they were analyzed manually for the instances of metaphors. A metaphor was identified when there was a source-target domain mapping. The following sentence shows a metaphorical instance for the target domain CAREER: 他事業的生命力日趨旺盛 "the life-force of his career is becoming exuberant day-by-day," the concrete meaning of "life force" is mapped onto CAREER. In another example,事業創傷 "the wound of career," the more concrete source "wound" is mapped onto the abstract target CAREER. Through similar analysis, all metaphorical instances were marked and extracted. Once all the metaphorical instances had been identified, the next step was to define the source domains of these instances. In order to do so, the corresponding WordNet and SUMO nodes for each lexical item were searched for in Sinica Bow. Through this system, each metaphorical instance was keyed in at the WordNet page. This was done to extract the WordNet explanations which were linked to the corresponding SUMO nodes in the system. In order to obtain the WordNet explanation, a prior step was performed, i.e., the most appropriate meaning was selected from the list of senses available. This selection was done manually, but most often, the most appropriate meaning was found to be the most concrete meaning in the list. For example, when the Chinese keyword *chuangshang* 創傷 "wound" from the target domain CAREER was searched for in Sinica Bow, the senses listed in Table 2 were extracted.

Table 2. The search result for chuangshang 創傷 "wound" in Sinica Bow

WordNet (1.6)	WordNet Explanations	Corresponding SUMO Nodes
Sense 1:trauma	an emotional wound or shock often having long-lasting effects	EmotionalState(情緒狀態)
Sense 2:wound	any break in the skin or an organ caused by violence or surgical incision	Injuring(傷害)

Among these senses, the more concrete sense (i.e., the more concrete meaning that was mapped from the source domain) was selected. In this case, sense 2 was selected and then the corresponding node (the rightmost column in Table 2) was found. In this case, it is "Injuring." The SUMO definition for "Injuring" is "The process of creating a traumatic wound or injury. Since injuring is not possible without some biologic function of the organism being injured, it is a subclass of biological process." The keywords in the SUMO definition (shaded in the previous sentence) are the terms that helped us categorize the metaphorical instances. Through the collection of SUMO definitions for all the metaphorical instances, all the similar metaphorical expressions with the same related nodes are grouped into categories. For instance, all the metaphorical expressions related to "Organism" were grouped together. Then, from these instances, the source domains were decided. The expressions that were grouped under the same source domain formed a cluster of lexical items under a domain. In the following discussion, we will provide a detailed example using the CAREER domain.

For the target domain CAREER, all 84 metaphorical instances are listed in Table 3. Each of the items in Table 3 was looked up using the Sinica Bow system to find their corresponding WordNet senses and, later, the SUMO nodes.

Table 3. Metaphorical expressions related to shiye 'career' (tokens are in brackets)

新創(1)	紮實(1)	溶進...之中(1)	軌道(1)	意識(1)	異軍(1)	幕後功臣(1)	前途(2)
創造(5)	起步(1)	收起來(1)	轟轟烈烈(2)	搖身一變(1)	改革(1)	玩掉(1)	投入(1)
開創(1)	走向(2)	投身...中(1)	走上(1)	創傷(1)	兵符(1)	投向(1)	壯大(1)
共創(1)	第一步(1)	基礎(5)	火車頭(1)	打拼(1)	前程(1)	開發(1)	成長(1)
再創(2)	闖(2)	追求(4)	退出(2)	掙(1)	競爭(1)	登上...位子(1)	角色(1)
挑戰(1)	關(5)	風險(1)	躍進(1)	拼(1)	抗爭(1)	包袱(1)	
策略(2)	關卡(1)	供輸(1)	放手(1)	大舞台(2)	投(1)		
趨勢(3)	過程(1)	賭(1)	生命力(1)	階梯(1)	衝刺(1)		

Table 4 shows how the SUMO definitions help us differentiate between source domains. However, due to limited space, only selected instances from Table 3 will be discussed.

Table 4. Defining source domains through WordNet and SUMO

Expressions	WordNet senses	WordNet explanations	SUMO nodes	SUMO definitions
策略 'tactic'	6: ambush	the act of concealing yourself and lying in wait to attack by surprise	ViolentContest (暴力性的競爭)	A Contest where one participant attempts to physically injure another participant.
軌道 'track'	2: track	a pair of parallel rails providing a runway for wheels	Transportation Device (運輸工具)	A TransportationDevice is a Device which serves as the instrument in a Transportation Process which carries the patient of the Process from one point to another.

For the expressions listed in Table 4, their SUMO definitions (the rightmost column) provide the keywords referring to the source domain to which these expressions might belong[1]. For instance, the keyword "Contest" for 策略 "tactic" might refer to WAR or COMPETITION, whereas "TransportationDevice" and "Transportation" for 軌道 "track" might refer to VEHICLE. When all the lexical items in Table 3 are looked up, the similarities of these items at the upper ontological levels can be established. In this paper, we will use CAREER IS A PERSON as an example. All items in (2) are found to have "Organism" as the shared concept[2].

(2) 意識 "consciousness" 生命力 "the force to live" 成長 "grow"
創傷 "wound" 第一步 "first step" 起步 "start a step"
放手 "let go" 走向 "walk towards"

Table 5 below shows the SUMO definitions for the items in (2). The original method produced the list of all instances in Table 3 in the form shown in Table 5; however, due to space limitations, this paper does not include the full list of all 84 items in Table 3. Only the ones related to "Organism" are shown. Among the SUMO nodes listed in Table 5, "Awake" and "Emotional State" are related to the upper node "State of mind" or a psychological process. The other nodes are related to a physical aspect of the organism, such as "Body

[1] There are also underlined keywords in the "WordNet Explanations" column. Although these keywords might also contribute to determining the source domains of metaphors, what we wish to discuss is connectivity at the ontological level.

[2] Note that though an item like *tousheng* 投身 "to throw oneself to" can be intuitively linked to PERSON, it is not listed in (2). This is because this lexical item is not found in the Sinica Bow. These items are not categorized in this paper. The limitation of the Sinica Bow will be discussed later.

motion" for "Walking."

Table 5. CAREER IS A PERSON: SUMO definitions

Expressions	WordNet senses	SUMO nodes	SUMO definitions
意識 "consciousness"	1: consciousness	Awake (清醒)	Attribute applies to Organisms that are neither Unconscious nor Asleep.
創傷 "wound"	2:wound	Injuring(傷害)	The process of creating a traumatic wound or injury. Since Injuring is not possible without some biologic function of the organism being injured, it is a subclass of BiologicalProcess.
放手 "let go"	1: let_go	EmotionalState (情緒狀態)	The Class of Attributes that denote emotional states of Organisms.
生命力 "life force"	1: animation	Living (活的)	This Attribute applies to Organisms that are alive.
成長 "grow"	3: mature	Growth (生長)	The Process of biological development in which an Organism or part of an Organism changes its form or its size.
起(步) "start a step"	1: pace	Walking (行走)	Any BodyMotion which is accomplished by means of the legs of an Organism on land for the purpose of moving from one point to another.
(走)向 "walk toward"	1: foot		
第一(步) "first step"	1: pace		

Although the linking concept is found to be "Organism," the concept of "Organism" is too broad, because it comprises all living things, including all plants and animals. For conceptual metaphors, preference is given to source domains that are in contact with human conceptualization, i.e., more concrete concepts which human can easily recall when describing an abstract idea. Furthermore, as indicated in Table 5, the type of organism should not only have the abilities to grow and to walk, but should also to have an emotional state. Based on these criteria, PLANT and ANIMAL are ruled out as possibilities, and HUMAN or PERSON is suggested.

Table 5 also lists the keywords in SUMO that are related to "Organism" (shaded). These keywords are important because, as explained in the parallel distributed model, one active unit will lead to the activation of other representation units. If a semi-automatized program is to be generated for the purpose of metaphor extraction, one first needs to establish the link between active units. In this case, the active units can be these keywords, and in the future research, the links between these keywords can be established using a computerized program. However, at the present stage, the analysis of metaphors, the selection of WordNet senses, and the selection of keywords has to be carried out manually.

In this paper, we incorporate more target domains to test the reliability of using SUMO nodes. If the same repeated SUMO nodes are found, then there is a systematic pattern that

links the lexical items within a similar source domain. In the following sections, we will skip the steps for obtaining SUMO nodes and focus more on comparing the different target domains which share the same source domain (of PERSON).

3.1 CULTURE IS A PERSON

The example CULTURE IS A PERSON is shown in (3) below.

(3) 使 客家 文化 走入 現代 現實 之中 啊！
shi kejia wenhua zouru xiandai xianshi zhizhong Exclam
cause hakka culture walk into modern reality inside Exclaim
"To make Hakka culture walk into the modern world!"

In this example, CULTURE is seen as something that can "walk." Comparing it with CAREER IS A PERSON, we find that the similarities and differences are those summarized in Figure 1.

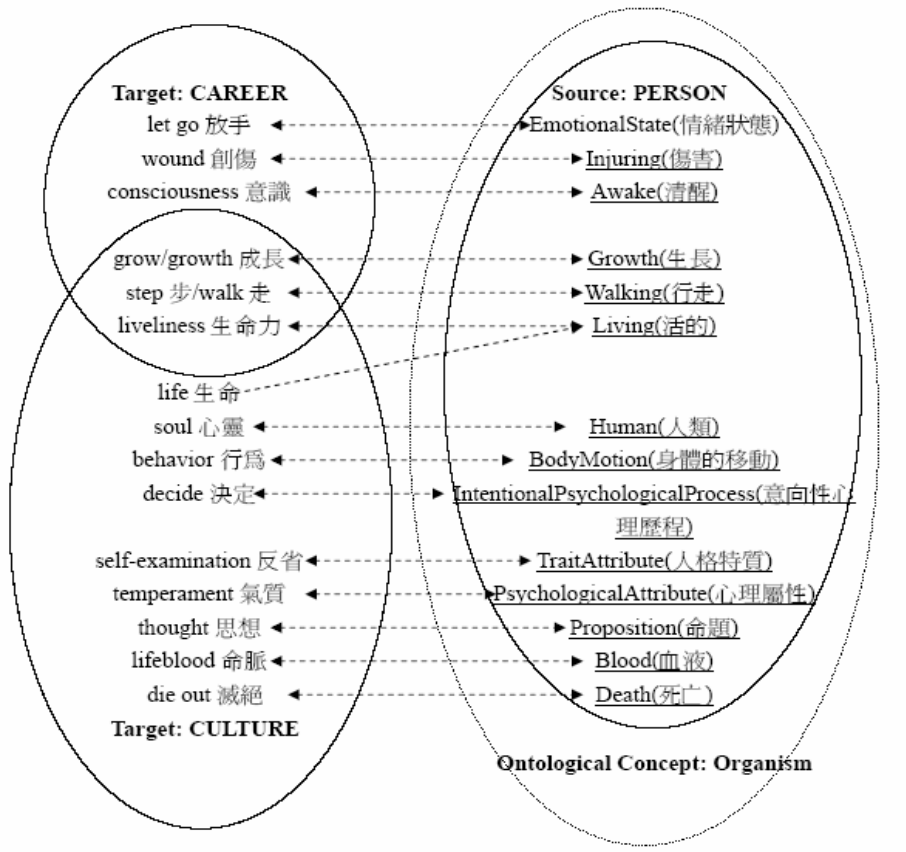


Figure 1. Categorizing the ontological nodes for CAREER and CULTURE

In this figure, the target domains CAREER and CULTURE are placed on the left side and the source domain PERSON on the right. All metaphorical instances in the target domains are related to their ontological nodes on the right side inside the circle of PERSON. Since all the SUMO nodes are related to "Organism" in one way or another, they can be subsumed under the ontological concept of "Organism," which forms the outer circle of PERSON. As mentioned previously, the metaphor CAREER IS AN ORGANISM is not selected because some nodes in "Person" (especially those related to psychological ones) cannot be accounted for by the other subsets of "Organism" such as PLANT and ANIMAL.

From Figure 1, we can see that a concept (i.e., PERSON) can be generated by looking at a cluster of lexical items (i.e., metaphorical expressions) and by looking at their SUMO nodes. The more general concept of "Organism" can be seen as the linking concept of all these lexical items. The overlapping area between CULTURE and CAREER (concepts such as "Growth" and "Walking") are linked to lexical items that are more lexicalized, as these items can apply to more target domains than the other items can. In other words, lexical items within several overlapping source domains (such as 成長 "growth" and (走)向 "walk towards") tend to be lexicalized faster than the other lexical items.

3.2 STOCK MARKET IS A PERSON in Chinese and English

The application of SUMO to two knowledge systems (Chinese and English) was discussed by Huang, Chung and Ahrens (in press). The focus of this paper is to investigate which conceptual nodes will be similar or different when a similar source domain (PERSON) is created in two different languages. For example (4), the Chinese metaphor STOCK MARKET IS A PERSON is found.

(4) 紐約 股市 復甦 的 活力 驚人
neuyue *gushi* *fusu* *de* *huoli* *jingren*
New York stock market recovery DE vitality shock people
"The vitality of the recovery of the New York stock market is surprising."

An English example is *the nervous stock market tumbled 67.85 points yesterday*. A comparison of the Chinese and English data for STOCK MARKET is shown in Figure 2.

Previous analysis of Chinese and English STOCK MARKET was reported by Chung, Ahrens and Sung [2003], who also re-evaluated Charteris-Black's [2001] English and Spanish data. However, neither study incorporated ontologies in their analyses.

In Figure 2, all the metaphorical instances of STOCK MARKET are linked to their

ontological nodes under PERSON. No overlapping metaphorical items that have the same meaning are found. However, the ontological nodes "BiologicalAttribute" and "OrganismProcess" overlap in the Chinese and English data. If one compares Figure 1 and 2, one finds that there are no overlapping metaphorical expressions in Chinese and that only three ontological nodes are repeated in both figures. These nodes are "IntentionalPsychologicalProcess," "TraitAttribute" and "PsychologicalAttribute."

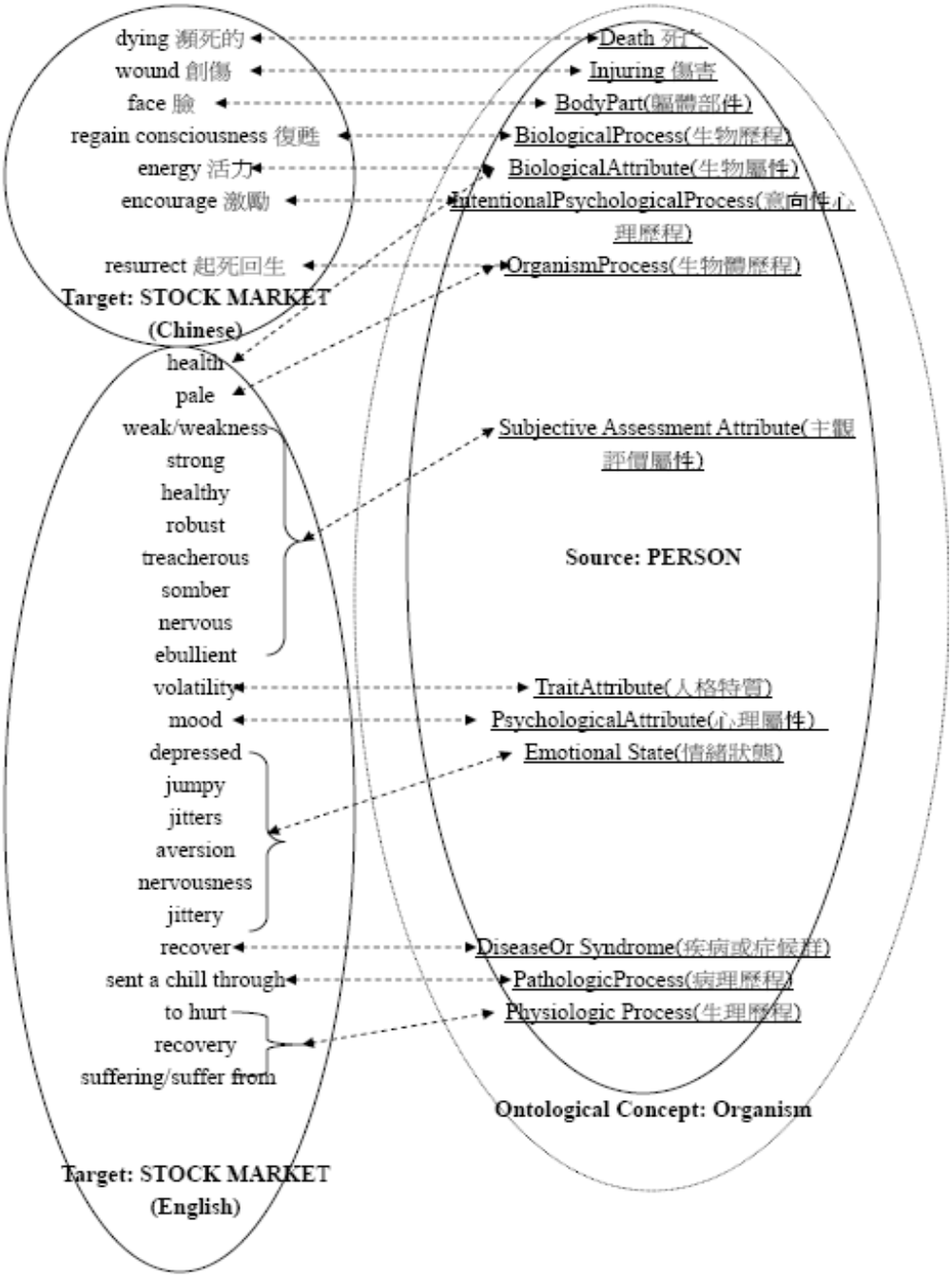


Figure 2. Categorizing the ontological nodes for STOCK MARKET

The few overlapping ontological concepts in Figures 1 and 2 suggest that the metaphorical items in CAREER and CULTURE are different from those in STOCK MARKET, even though all of them are linked to the concept PERSON. In the next section, the ontological nodes of ECONOMY will be examined.

3.3 ECONOMY IS A PERSON in Chinese and English

The example ECONOMY IS A PERSON in Chinese is shown in (5) below.

(5) 只 想 暫時 維持 不 讓 經濟 衰退
zhi *xiang* *zhanshi* *weichi* *bu* *rang* *jingji* *shuaitui*
only think temporary maintain NEG. let economy degenerate
"only want to temporarily maintain and not let the economy degenerate"

An English example is *the economy remains anemic in that period*. A comparison of the SUMO nodes for Chinese and English ECONOMY IS A PERSON is shown in Figure 3.

From Figure 3, the Chinese and English data of ECONOMY overlap with the concept "Growth" in both languages. In fact, previous works [Ahrens *et al.* 2003; Chung *et al*. 2003ab] suggested that the meaning of "growth" is mapped most frequently in both Chinese and English ECONOMY metaphors[3]. As shown in Figure 1 earlier, the concept of "growth" is also found in the target domains CAREER and CULTURE, indicating that this concept can be used to describe not only ECONOMY but also CAREER and CULTURE. This repetition of the same lexical items in several target domains may mean that this lexical item is a common concept and that its frequency of occurrence as a metaphor is high, indicating that its metaphorical meaning is strongly conventionalized.

In order to understand which aspects of PERSON are involved in all the target domains CAREER, CULTURE, STOCK MARKET and ECONOMY, a comparison of ontological nodes will be discussed in Section 4 below.

[3] This paper only considers the types of lexical items and their related SUMO nodes (regardless of whether the items that appear many times will affect the number of times an ontological node will appear). This is so because the aim here is to extract the ontological nodes that form a concept, not their frequency.

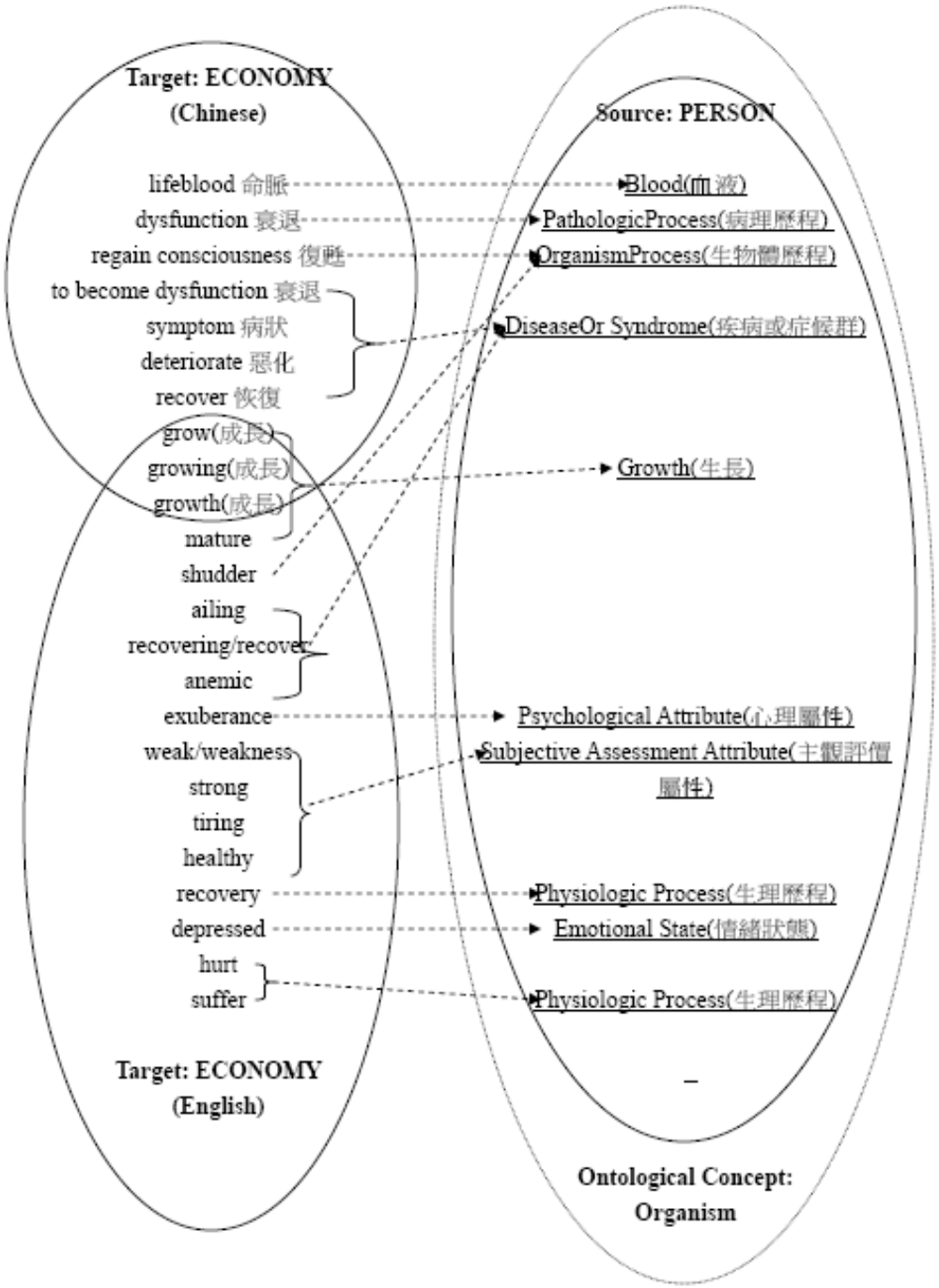


Figure 3. Categorizing the ontological nodes for ECONOMY

4. Discussion

When ontological nodes are contrasted, they are found to fall into two main types, namely, physical and psychological. This means that both the physical and psychological aspects of a person are selected when a metaphor is created.

(6) **Physical**		**Psychological**
DiseaseOr Syndrome	Injuring	Subjective Assessment Attribute
Growth	PathologicProcess	Emotional State
Physiologic Process	TraitAttribute	Psychological Attribute
OrganismProcess	Awake	IntentionalPsychological Process
Walking	BiologicalProcess	Proposition
Living	BodyMotion	
BiologicalAttribute	BodyPart	
Blood	Damaging	
Death	Human	

The distribution in (6) shows that many aspects of the physical concept of human are mapped when constructing metaphors. Comparing Figures 1, 2 and 3, one finds that the node "DiseaseOrSyndrome" comes from both Chinese and English ECONOMY and STOCK MARKET only (where more lexical items in ECONOMY have this node). A similar situation is found with the ontological node of "Growth," where more types of lexical items come from the target domain ECONOMY (rather than the other three target domains). The preference for (both English and Chinese) ECONOMY for the physical aspect of PERSON shows that these two languages do not differ greatly in terms of the types of lexical items found.

On the other hand, most of the lexical items that denote the psychological aspect of a person come from STOCK MARKET (refer to "Subjective Assessment Attribute" and "Emotional State" in Figure 2). Compared with English ECONOMY, which maps onto the physical aspect of a person, STOCK MARKET seems to be described more in terms of a psychological change of a person. This finding is interesting because as Ahrens [2002] noted in her discussion of the Mapping Principle Constraint, that "a single target domain must use different source domains for different reasons." In this case, the selection of the single source domain by two different target domains in the same language occurs for different reasons. With this in mind, one can see the underlying motivation for the formulation of different

conceptual metaphors. That is, the motivation governing the selection of a source domain by a target domain is not *ad hoc*; rather, it is governed by principles. This is only reflected when the ontological nodes involved in the source domain concepts are extracted.

5. Conclusion

The results obtained from the WordNet lexical representation and SUMO searches prove two major points: First, the manual analysis performed using the CM Model in previous studies can be further automatized. Second, there is conceptual connectivity between the metaphorical expressions found within a source domain. As the philosopher Kamppinen pointed out, "representations do not operate in isolation," they are "clustered into systems of representations, cognitive schemata" [1993]. In this case, this cluster of representations is connected through the overlapping concept or "Organism" in the lexical items found. This connectivity between lexical items can be established by using computational tools such as SUMO definitions, along with a linguistic model (i.e., the CM Model) to identify conceptual metaphors in corpora.

Despite the usefulness of WordNet and SUMO for determining source domains, there are two limitations to this study. First is the process of selecting senses from WordNet, which has to be carried out manually. Second is the missing entries that are found in the Sinica Bow look-up (especially Chinese adjectives). Even though the few missing items do not affect the results of categorization as a whole, their inclusion would have made this study more complete.

Nevertheless, the incorporation of an ontology helps pinpoint at which level of knowledge conceptual metaphors occur. The discovery concerning the motivation for using conceptual metaphors contributes to identifying cognitive differences across speech communities. From the perspective of anthropology and language processing, this study has provided linguistic evidence about how humans represent concepts mentally when using metaphors. It is hoped that the line of research discussed herein will stimulate more research on how computational approaches can help set parameters for determining metaphorical senses and point the way to creating a systematic relationship between literal and figurative synsets in WordNet. In addition, it may also be possible to conduct psycholinguistic experiments to verify the metaphorical expressions that have been extracted from corpora.

Acknowledgements

We would like to thank the CLSW-5 reviewers for their comments on this paper and National Science Council, Taiwan, for support under grant #NSC91-2411-H-002-082-ME, #NSC92-2411-H-002-076- ME, #NCS92-2411-H-001-010-ME and #NCS 93-2411-H-001-004-ME. Any remaining errors are the sole responsibility of the authors.

References

Ahrens, K., S. F. Chung, and C. R. Huang, "From Lexical Semantics to Conceptual Metaphors: Mapping Principle Verification with WordNet and SUMO," In D-h Ji, K. T. Lua and H. Wang, (Eds) *Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5)*, Singapore: COLIPS, 2004, pp. 99-106.

Ahrens, K., S. F. Chung, and C. R. Huang, "Conceptual Metaphors: Ontology-based Representation and Corpora Driven Mapping Principles," In the *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*. 2003, pp. 35-41.

Ahrens, K., "When Love is Not Digested: Underlying Reasons for Source to Target Domain Pairings in the Contemporary Theory of Metaphor," In Y.C. E. Hsiao, (Ed.) *Proceedings of the First Cognitive Linguistics Conference*, 2002, pp. 273-302.

Alonge, A., and M. Castelli. "Encoding Information on Metaphoric Expressions in WordNet-like Resources," In *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*. 2003, pp. 10-17.

Alonge, A., and M. Castelli, "Metaphoric Expressions: An analysis of data from a corpus and the ItalWordNet database," In *Proceedings of the First Global WordNet Conference,* 2002a, pp. 342-350.

Alonge, A., and M. Castelli, "Which Way Should We Go? Metaphoric Expressions in Lexical Resources," In *Proceedings of the third Language Resources and Evaluation Confernece (LREC-02)*, 2002b (11), pp. 1948-1952.

Alonge, A., and B. Lönneker, "Metaphors in Wordnets: from Theory to Practice," In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004),* 2004, pp. 165-168.

Charteris-Black, J., and T. Ennis, "A Comparative Study of Metaphor in Spanish and English Financial Reporting," *English for Specific Purposes*, 20 (3), 2001, pp. 249-266.

Chung, S. F., K. Ahrens, and C. R. Huang, "Using WordNet and SUMO to Determine Source Domains of Conceptual Metaphors," In D-h Ji, K. T. Lua and H.Wang, (Eds). *Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5)*. Singapore: COLIPS, 2004a, pp. 91-98.

Chung, S. F., K. Ahrens, and C. R. Huang, "RECESSION: Defining Source Domains through WordNet and SUMO," In *Proceedings of the Linguistic Society of Korea (LSK) International Conference*, Volume 2: General Sessions, Seoul: The Linguistic Society of Korea (LSK) and Yonsei Institute of Language and Information Science (ILIS), 2004b, Seoul, Korea, pp. 43-52.

Chung, S. F., K. Ahrens, and C. R. Huang, "ECONOMY IS A PERSON: A Chinese-English Corpora and Ontological-based Comparison Using the Conceptual Mapping Model," In *Proceedings of the 15th ROCLING Conference*, 2003a, Taipei, Taiwan, pp. 87-110.

Chung, S. F., K. Ahrens, and C. R. Huang, "ECONOMY IS A TRANSPORTATION_DEVICE: Contrastive Representation of Source Domain

Knowledge in English and Chinese," In *Proceedings of the special session of UONLP, 2003 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2003b, Beijing, pp. 790-796.

Chung, S. F., K. Ahrens, and Y. H. Sung, "STOCK MARKETS AS OCEAN WATER: A Corpus-based, Comparative Study in Mandarin Chinese, English and Spanish," In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2003c, Singapore, Singapore, pp. 124-133.

Eilts, C., and B. Lönneker, "The Hamburg Metaphor Database," *Metaphorik.de*, 2002(3), pp. 100-110.

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

Harris, C.L., "Coarse Coding and the Lexicon," In C. Fuchs and B. Victorri, (eds) *Continuity in Linguistic Semantics*, Amsterdam and Philadelphia: John Benjamins, 1994. pp. 205-229.

Huang, C. R., S. F. Chung, and K. Ahrens, "An Ontology-based Exploration of Knowledge Systems for Metaphor." In R. Kishore, R. Ramesh, and R Sharman, (eds.) *Ontologies in the Context of Information Systems: An AIS SIG-ODIS Initiative*, Springer, In Press.

Huang, C.R., R. Y. Chang, and S. B. Lee, "Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO." In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC2004), Lisbon, Portugal, 2004.

Kamppinen, M., "Cognitive Schemata," In M. Kamppinen, (Ed.). *Consciousness Cognitive Schemata, and Relativism*, 1993, pp. 133-168.

Keil, F.C., *Semantic and Conceptual Development: An Ontological Perspective*, Cambridge, Mass. and London, England: Harvard University Press, 1979.

Lakoff, G., and M. Johnson, *Metaphors We Live By*, Chicago: The University of Chicago Press. 1980.

Lakoff, G., "The Contemporary Theory of Metaphor." In Andrew Ortony (ed.). *Metaphor and Thought*, Second Edition. Cambridge: Cambridge University Press. 1993, pp. 202-251.

Lönneker, B., "Is There a Way to Represent Metaphors in WordNets? Insights from the Hamburg Metaphor Database," In the *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*, 2003, pp. 18-26.

Navigli, R., and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites," *Computational Linguistics*, 30 (2). 2004, pp. 151-179.

Niles, I., and Pease, A, "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology." In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering,* Las Vegas, Nevada. 2003.

Peters, W., and Wilks, Y, "Data-Driven Detection of Figurative Language Use in Electronic Language Resources," *Metaphor and Symbol*, 18(3). 2003, pp. 161-173.

Rumelhart, D.E., and J.L. McClelland, (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, Cambridge, Mass.: MIT. 1986.

Uschold, M., and M. Gruninger, "Ontologies: Principle, Methods and Applications," *Knowledge Engineering Review*, 11(2). 1996, pp. 93-155.

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 571-580

隱喻性成語的語義映射[1]

Semantic Mapping in Chinese Metaphorical Idioms

李芸*+、李素建+、王治敏+、吳云芳+

Yun Li, Sujian Li, Zhimin Wang and Yunfang Wu

摘　要

成語中的隱喻是一種非常普遍的現象。從本義到其隱喻義，可以看作是一種此類事物語義映射到彼類事物的認知活動。本文對 2,347 條隱喻性成語進行了統計，分析了它們的語言特點、語法結構、語法功能、語法結構與語法功能的關係以及語義類別。分析了隱喻性成語中的語義映射的類型。大多數隱喻性成語都是由關涉性語義成分和描述性語義成分共同構成，其中關涉性語義成分是實現顯性事物到隱性事物語義映射的基礎，而它們的關聯則依賴于描述性語義成分。

關鍵字：隱喻，隱喻性成語，語義映射

Abstract

Metaphors in idioms are universal in languages. The progression from conceptual meaning to metaphorical meaning is a cognitive activity of mapping from one thing to another. This paper, based on the analysis of 2,347 metaphorical idioms, their linguistic features, grammatical structures, grammatical functions, and semantic categories, tries to identify categories of metaphorical idioms. Most of the metaphorical idioms in Chinese are found to be composed of projective semantic

[1] 本文相關研究得到中國國家 863 計畫(2001AA114210，2002AA117010)和國家 973 計畫(2004CB318102)的支持。

* 北京大學計算語言學研究所
Institute of Computational Linguistics, Peking University, Beijing, China
E-mail: {liyun2003, lisujian, wangzm, wuyf}@pku.edu.cn

+ 中國社會科學院語言研究所
Institute of Linguistics, Chinese Academy of Social Science, Beijing, China
E-mail: liyun@cass.org.cn.

[Received April 20, 2005; Revised June 17, 2005; Accepted August 21, 2005]

elements and descriptive semantic elements, among which the former are the basis for mapping from explicit to the implicit meaning, while the latter are the key to understanding the relevance of the two parts.

Keywords: Idiom, Metaphorical Idiom, Semantic Mapping

1. 引言

傳統的隱喻理論將隱喻看作是一種語言現象，是一種用于修飾話語的修辭現象。然而，隱喻不僅僅是一種語言現象，它更重要的是一種人類認知現象。它是人類將其某一領域的經驗用來說明或理解另一類領域經驗的一種認知活動。在人類其他的文化和藝術活動過程中，我們到處都能看到隱喻的存在[束定芳 1998]。

漢語成語濃縮了中華民族博大精深的文化智慧，是中國的文化瑰寶。它的語言精煉，結構嚴謹，含義深刻，富有表現力。成語從形式上看，結構緊密，渾然一体。從意義上看，脫離字面的意思，產生了統一的概念，或者有了比喻義。產生比喻義的成語不在少數，研究這些成語的構成特點，以及從字面義到引伸義或到比喻義之間的語義映射，借以豐富當前隱喻研究的成果。

爲此，我們選取了研究的語料，一個是《分類漢語成語大詞典》，王勤、馬國凡、許正元、孫玉溱編著，山東教育出版社，1988 年 11 月，有 9,507 條成語。另一個是《三知成語詞典》，有 13,768 條成語，是從網上下載的共享軟件。

兩個詞典描述成語的屬性字段有多有少，爲了使兩個詞典信息互補，就把它們的交集部分作爲統計的基礎，兩個詞典共同的部分有 7,105 條成語。從中找出含有比喻義的成語 2,347 條。

2. 隱喻性成語

我們把含有隱喻意義的成語叫做隱喻性成語。具体到從數據庫中如何挑出這些隱喻性成語來，我們採取的方法是，如果釋義中出現“喻”，比如“比喻…”、“借喻…”“舊常喻…”“后以喻…”“喻”等等，就挑出來待進一步考察。當然，釋義中含有“喻”字的也有少量不屬于隱喻意義的，比如解釋“喻”字的“喻：…”就不是，還有“喻示”。去除這些成語“不可理喻”、“家喻戶曉”、“不言而喻”“罕譬而喻”等。如此挑選下來，得到 2347 條隱喻性成語。

2.1 來源和語言特點

隱喻性成語主要來源于典籍，產生于古代和近代，現代較少。有些脫胎于俗語、諺語、歇后語。比如來源于歷史的有：得隴望蜀，樂不思蜀、垂簾聽政。來源于典籍的有：鄭人買履，刻舟求劍。來源于寓言故事的有：狐假虎威。來源于俗語的有：無米之炊、依樣（畫）葫蘆、平地一聲雷、一鼻孔出氣、坐山觀虎鬥、一蟹不如一蟹、驢唇不對馬嘴、井水不犯河水、前怕狼，后怕虎等。

隱喻性成語中按是否含有比喻詞分兩類：

1）含有“如”、“似”、“若”、“象”、“比”、“同”、“猶”等喻詞的。

本体喻源均出現：執法如山，壽比南山，情同骨肉，情同手足，勢如破竹，心如鐵石，從善如登，從惡如崩。

本体不出現：如魚得水，如饑似渴，如履薄冰。

2）不含有“如”、“似”、“若”、“象”、“比”、“同”、“猶”等喻詞的。

本体喻源均出現：羊腸小徑。

本体不出現：月下老人，尚方寶劍，靡靡之音。

2.2 語法結構

從隱喻性成語內部的語法結構來統計，主要有聯合式、主謂式、偏正式、連動式、動賓式、複句式、緊縮式、兼語式、補充式、複雜式 10 種情況。如下表：

表 1: 隱喻性成語的語法結構分佈

結構	數量	比率(%)	示例
聯合式	616	31.21	坐言起行
主謂式	437	22.14	左右開弓
偏正式	381	19.30	鑽頭覓縫
連動式	182	9.22	做賊心虛
動賓式	145	7.35	坐于塗炭
複句式	107	5.42	坐山觀虎鬥
緊縮式	73	3.70	著手成春
兼語式	16	0.81	炙手可熱
補充式	12	0.61	運斤成風
複雜式	5	0.25	尾大不掉

從上表可以看出，隱喻性成語內部構成主要是聯合式、主謂式和偏正式，其中聯合式占到了 31.21%，并且比主謂式高出將近 10%。聯合式主要是 NP＋NP、VP＋VP 形式。

2.3 語法功能

把隱喻性成語作爲一個整体看，可以做句子的任何成分，如主語、謂語、賓語、定語、狀語、補語等，也可單獨成句。

2.4 語法結構和語法功能的關係

在成語知識庫中，詳細描述了每個成語的用法，包括成語內部的語法構成、可以充當的句法功能，以及褒貶義色彩等。所有這些我們都可以通過統計得出結果，進而分析它們

之間的各種關係。下面以聯合式隱喻性成語爲例，統計出來的語法結構與語法功能之間的關係如下表所示。其他結構略去。

表 2: 聯合式隱喻性成語的語法功能

語法結構	數量	比率(%)	語法功能	示例
聯合式	146	23.70	謂語、定語	趑趄不前
聯合式	73	11.85	謂語	坐薪懸膽
聯合式	70	11.36	賓語、定語	墜茵落溷
聯合式	54	8.77	賓語	坐言起行
聯合式	37	6.01	謂語、賓語	左輔右弼
聯合式	36	5.84	謂語、賓語、定語	裝神弄鬼
聯合式	35	5.68	謂語、定語、狀語	自吹自擂
聯合式	25	4.06	主語、賓語	陽春白雪
聯合式	20	3.25	主語、賓語、定語	左道旁門
聯合式	17	2.76	定語	鐘靈毓秀
聯合式	11	1.79	謂語、定語、補語	珠圓玉潤
聯合式	9	1.46	謂語、賓語、定語、狀語	舍本逐末
聯合式	8	1.30	謂語、賓語、狀語	破釜沉舟
聯合式	7	1.14	謂語、狀語	張冠李戴
聯合式	7	1.14	定語、狀語	再接再厲
聯合式	6	0.97	狀語	無聲無臭
聯合式	5	0.81	謂語、賓語、分句	兔死狐悲
聯合式	4	0.65	謂語、定語、分句	吐故納新
聯合式	4	0.65	謂語、補語	食玉炊桂
聯合式	4	0.65	定語、補語	佛口蛇心
聯合式	3	0.49	謂語、賓語、補語	天昏地暗
聯合式	3	0.49	定語、分句	載舟覆舟
聯合式	3	0.49	賓語、狀語	披肝瀝膽
聯合式	2	0.32	狀語、補語	一五一十
聯合式	2	0.32	主語、謂語	懷瑾握瑜
聯合式	2	0.32	主語、定語	民脂民膏
聯合式	2	0.32	謂語、狀語、補語	提綱挈領
聯合式	2	0.32	謂語、分句	換湯不換藥
聯合式	2	0.32	定語、狀語、補語	生龍活虎
聯合式	2	0.32	補語	落花流水
聯合式	2	0.32	賓語、定語、狀語	盲人瞎馬

聯合式	1	0.16	狀語、分句	七擒七縱
聯合式	1	0.16	主語、謂語、賓語、定語	天經地義
聯合式	1	0.16	主語、謂語、賓語	管窺蠡測
聯合式	1	0.16	主語、定語、狀語	單槍匹馬
聯合式	1	0.16	主語、補語、分句	綱舉目張
聯合式	1	0.16	主語、補語	別鶴孤鸞
聯合式	1	0.16	主語、賓語、狀語	蛛絲馬跡
聯合式	1	0.16	主語、賓語、兼語	蝦兵蟹將
聯合式	1	0.16	主語、賓語、分句	繁文縟節
聯合式	1	0.16	主語	白雲親舍
聯合式	1	0.16	分句	伯歌季舞
聯合式	1	0.16	賓語、定語、補語	花殘月缺
聯合式	1	0.16	賓語、補語、分句	歪打正著

上表表明，聯合式隱喻性成語既可以作謂語又可以作定語的數量最多，有 146 條，占到了 23.70%。其他依次爲作謂語，73 條，占 11.85%；作賓語和定語，70 條，占 11.36%。

對以上聯合式隱喻性成語所能充當的語法功能，按照各語法成分分組匯總，從多到少排列，得到下表。

表 3: 聯合式隱喻性成語的語法功能分佈

語法功能	謂語	定語	賓語	狀語	主語	補語	分句
成語數量	386	371	279	85	58	35	19
比率(%)	31.31	30.09	22.63	6.89	4.70	2.84	1.54

從表 3 中看出，聯合式隱喻性成語可以作句子的任何成分，還可以做分句。主要作謂語、定語和賓語，三者占到了 84.02%。

2.5 語義類別

這些成語參考《分類漢語成語大詞典》，按照成語所表達的意義分成了 48 個大類，290 個小類。這 48 個大類，按原詞典順序排列，分別是：01 國家類、02 政治類、03 法律類、04 軍事類、05 工作類、06 生產類、07 學習類、08 學問類、09 教育類、10 言語類、11 文章類、12 藝術類、13 地理類、14 行旅類、15 風光類、16 花木類、17 時令類、18 時間類、19 計謀類、20 得失類、21 功罪類、22 勞逸類、23 錢財類、24 富裕類、25 貧窮類、26 數量類、27 程度類、28 建築類、29 人才類、30 友誼類、31 修身類、32 品德類、33 情感類、34 意志類、35 智愚類、36 容貌類、37 体態類、38 儀表類、39 禮慶類、40 婦女類、41 愛情類、43 家庭類、44 飲食類、45 生老類、46 疾病

類、47 福禍類、49 死亡類、50 其他類。其中，42 婚姻類、48 迷信類，在我們的統計表中未出現，因此不作考慮。

3. 隱喻的語義映射

3.1 隱喻的本質

隱喻從本質上講，是一種認知活動，是人類理解周圍世界的感知和形成概念的工具。語言中的隱喻是人類認知活動的結果與工具[束定芳 1998]。

按照理查茲的互動(interaction)理論，隱喻是兩個主要部分“本体”(Tenor)和“喻源[2]” (Vehicle)之間的互動，隱喻本身指包含兩個部分的整体[胡壯麟 2004]。因此，隱喻涉及到兩個領域，用其中一個領域 A 來說明另一領域 B 。兩個領域中，被說明的領域（即 B）稱爲目標領域(target domain)，另一個（即 A）稱爲源領域(source domain)。兩個領域有不同的語義網絡，語義結構特徵具有相異性。但是，本体和喻源之間之所以能夠互動，是因爲它們之間有“共同點”(ground)或者“相似性”。相似，使隱喻成爲可能；相異，使隱喻更加鮮明，使認知更深入。

3.2 隱喻的語義映射

隱喻意義是兩個語義領域之間的語義映射。隱喻是以喻源和本体之間的相似性作爲意義轉移的基礎的。以“A 是 B ”結構出現的映射是將有關喻源域 B 的適當的知識結構的一部分映射到目標域結構 A 上。爲了以喻源域的詞語來理解目標域，人們必須具有對喻源域的適當的知識。

例如：“愛情是一次旅行(Love is a journey)”，分析如下：

愛情	是	一次	旅行。
A	是		B。
Target domain			Source domain
目標領域			源領域
Tenor			Vehicle
本体			喻源
映射： 本体	←（映射）		喻源
F: A	←	B	
F: B→A			

[2] 或者譯爲“喻体”。

于是，“愛情是一次旅行”隱喻是將“旅行”結構映射到“愛情”域。

語義映射的原則是：從具体到抽象、從熟悉到陌生、從簡單到複雜等。語義映射具有方向性。

3.3 隱喻性成語中的語義映射

語言中的隱喻是一種以詞語或句子爲焦點，以語境爲框架的語用現象。孤立的詞嚴格意義上講不能稱爲隱喻。只有在具体的語境中才能判斷一個詞是否用作隱喻[束定芳 1998]。但是，成語則不然了。成語是人們長期以來慣用的定型詞組或短語，脫離了字面的意思，獲得了統一的概念或者比喻義。因此，隱喻性成語從隱喻程度性上來說，成語應該算作死寂隱喻(dead and buried)或者是非活躍隱喻(inactive metaphor)[3]。

隱喻涉及到語義的轉移，這一轉移帶有方向性。一般情況下，喻源的有關特徵被轉移到本体上[束定芳 1998]。對于隱喻性成語來說，要轉移到本体上的意義已經固定下來，即成語的隱喻義。

隱喻性成語的語義映射可以分爲以下兩類：

1) 從“物”映射到“物”；

例如：[大風大浪] 原指自然界的大的風浪。現用來比喻激烈的鬥爭和艱險的歷程。

[清規戒律] 原指佛教徒所遵守的規則和戒條。現比喻束縛人的繁瑣不合理的規章制度。

[無名小卒] 卒：古時指士兵。不出名的小兵。比喻沒有名望或地位的人。

2) 從“事”映射到“事”。

例如：[風雲變幻] 風雲：比喻變化動盪的局勢；變幻：變化不定。像風雲那樣變化不定。比喻局勢複雜，變化迅速，難以預料。

[載舟覆舟] 民眾猶如水，可以承載船，也可以傾覆船。比喻人民是決定國家興亡的主要力量。

[明鏡高懸] 傳說秦始皇有一面鏡子，能照人心膽。比喻官員判案公正廉明。

[插翅難飛] 插上翅膀也難飛走。比喻陷入困境，怎麼也逃不了。

[百足之蟲，死而不僵] 百足：蟲名，又名馬陸或馬蚿，有十二環節，切斷后仍能蠕動。比喻勢家豪族，雖已衰敗，但因勢力大，基礎厚，還不致完全破產。

[東風壓倒西風] 原指封建大家庭裏對立的兩方，一方壓倒另一方。現比喻革命力量對于反動勢力占壓倒的優勢。

---

[3] 參見束定芳，論隱喻的本質及語義特徵，載于《外國語》1998 年第 6 期。文中提到隱喻程度性，引用 Goatly 的隱喻分類： 1)死喻(dead metaphor)； 2)死寂隱喻(dead and buried)； 3)非活躍隱喻(inactive metaphor)；其中又分爲沉睡式隱喻(sleeping metaphor)和陳舊隱喻(tired metaphor)兩種。 4)活躍隱喻(active metaphor)。

[天無二日] 日：太陽，比喻君王。天上沒有兩個太陽。舊喻一國不能同時有兩個國君。比喻凡事應統于一，不能兩大并存。

另外，一些成語典故屬于這一類。如：負薪救火，畫蛇添足，削足適履，刻舟求劍，盲人摸象，坐井觀天，守株待兔，南轅北轍，揠苗助長，緣木求魚，亡羊補牢，囫圇吞棗等，通過故事來說道理。

3.4 語義結構成分＝描述性語義成分＋關涉性語義成分

我們從另外一個角度來看隱喻性成語的構成及語義映射過程。

在語義結構中，有兩個組成部分，一是描述性成分，一是關涉性成分。即：語義結構成分＝描述性語義成分＋關涉性語義成分[施春宏 2003]。

隱喻性成語的語義結構大致有三種形式：

1)兩種語義成分都有，表現爲：描述性語義成分＋關涉性語義成分，如“同床異夢”，“井底之蛙”，“青出于藍”“弦外之音”“添枝加葉”，“千錘百煉”，“精雕細刻”。其中的 “床”、“夢”，“井”、“蛙”，“青”、“藍”，“弦”、音”，“枝”、“葉”，“錘”、“煉”，“精”、“刻”是關涉性語義成分，而其他爲描述性成分。

2) 沒有明確的關涉性語義成分，表現爲：描述性語義成分（＋關涉性語義成分），如“蠢蠢欲動”，“顛撲不破”，“無可救藥”。

3) 沒有描述性語義成分，表現爲：（描述性語義成分＋）關涉性語義成分，如“酒囊飯袋”，“陽春白雪”，“吳下阿蒙”（吳下：現江蘇長江以南；阿蒙：指呂蒙。居處吳下一隅的呂蒙。比喻人學識尚淺）“晨鐘暮鼓”，“美人香草”，“風花雪月”（本來泛指四時景色。也指浮泛的寫景言情的詩文題材。又指男女之間的愛情。后來多指反映沒落思想情調，堆砌辭藻而內容空泛的詩文）。

關涉性語義成分是實現顯性事物（喻源）到隱性事物（本体）語義映射的基礎，而它們的關聯則依賴于描述性語義成分。以友誼類隱喻性成語爲例說明。

例 1・抃風舞潤：抃：鼓掌；潤：雨水。如燕在風中飛翔，象商羊（鳥名）在雨中起舞。原指同類事物相互感應。后比喻意氣相合。

關涉性語義成分：風，潤 → 同類事物

描述性語義成分：抃，舞

例 2・車笠之盟：笠：斗笠。比喻深厚的友誼。參見“乘車戴笠”。

關涉性語義成分：車笠→同類事物

描述性語義成分：盟

例 3・乘車戴笠：乘：坐；笠：斗笠。乘車時，戴斗笠的。比喻友誼深厚，不因貧富不同而有所改變。

關涉性語義成分：車、笠→ 同類事物

描述性語義成分：乘，戴

例 4・唇齒相依：依：依靠。象牙齒和嘴唇一樣相依靠。比喻相互關係密切，互相依存。

關涉性語義成分：唇，齒→ 同類事物

描述性語義成分：相依

例 5・唇亡齒寒：嘴唇沒有了，牙齒也暴露在外邊感到寒冷。比喻關係密切，利害相同。

關涉性語義成分：唇，齒→ 同類事物

描述性語義成分：亡，寒

例 6・心有靈犀一點通：靈犀：傳說犀牛是一種靈獸，角中有條白紋通向頭腦，感應靈敏。比喻戀愛著的男女双方心心相印。也泛指彼此心心相印。

關涉性語義成分：心，靈犀→男人女人的心

描述性語義成分：有，一點通

如果成語中沒有明確的關涉性語義成分，而主要顯現描述性語義成分，象第 2 種情況，實際上，其關涉性語義成分實際蘊含其中而爲交際所默認。如：

[蠢蠢欲動] 蠢蠢：爬蟲蠕動的樣子。比喻敵人策劃進攻或壞人準備搗亂破壞。

[顛撲不破] 顛：跌倒；撲：敲、拍打。無論怎樣摔打都不會破。比喻言論或學說符合客觀規律，永遠不會被推翻。

[無可救藥] 藥：治療。病勢嚴重，無法醫治。也比喻壞到極點，無法挽救。

“蠢蠢欲動”的關涉性語義成分雖不在成語中，但所指是很明確的，指爬蟲，轉移到人，而且是敵人或壞人，便形成了隱喻義，比喻敵人策劃進攻或壞人準備搗亂破壞。“顛撲不破”的關涉物件是物件，意思是無論怎樣摔打這個物件都不會破；借指言論或學說，意爲言論或學說永遠不會被推翻。“無可救藥”的關涉對象是人身体上的病，無法醫治，借指人思想上的錯誤或事態的惡劣程度已經嚴重到了極點，無法挽救。

3.5 隱喻的多樣性

隱喻具有多樣性，也就是說，用一個隱喻的喻源與不同的本体相結合可以產生不同的意義，即所謂的“比喻之兩柄(handles)”現象（引自錢鐘書《管錐編》）。反過來，一個事物也可以有多種喻源，即一物多喻。

從成語知識庫中，可以挖掘到“喻源域”與“本体域”之間多對多的關係。進一步擴充到真實文本中，自動發現詞語的隱喻用法或隱喻義，爲計算機自動處理語義打下基礎。

4. 結語

以上對成語中的隱喻現象從多個方面作了細緻地統計和分析，但是還有待進一步深入研究。對隱喻的計算研究將會推動自然語言理解的深入，相信我們對隱喻性成語的研究能對隱喻理論的研究以及對中文信息處理研究作出貢獻。

致謝

感謝諶貽容同學、段慧明老師和俞士汶老師對建設成語知識庫的大力支持和幫助，感謝匿名評審老師提出的寶貴意見。

參考文獻

王勤、馬國凡、許正元、孫玉溱編著，《分類漢語成語大詞典》，山東教育出版社，山東，1988 年。

三知成語詞典，網上軟件“天空軟体站” http://www.skycn.com，2004 年。

束定芳，“論隱喻的本質以語義特徵”，《外國語》第 6 期，1998 年，pp.10-19。

趙升奎，“比喻——跨範疇的語義映射過程”，《雲南師範大學學報》第 35 卷第 2 期，2003，pp.119-121。

施春宏，“比喻義的生成基礎及理解策略”，《語文研究》第 4 期，2003 年，pp.19-24，載于人大複印資料，《語言文字學》2004 年第 2 期。

胡壯麟，《認知隱喻學》，北京大學出版社，北京，2004 年。

嚴世清，《隱喻論》，蘇州大學出版社，蘇州，2000 年。

季廣茂，《隱喻理論與文學傳統》，北京師範大學出版社，北京，2002 年。

劉潔修，《成語》，商務印書館，北京，1985 年。

Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 581-592

基于現代漢語語法信息詞典的詞語情感評價研究[1]

Research on Lexical Emotional Evaluation Based on the Grammatical Knowledge-Base of Contemporary Chinese

王治敏*、朱學鋒*、俞士汶*

Zhimin Wang, Xuefeng Zhu and Shiwen Yu

摘要

本文將情感評價屬性特徵納入現代漢語語法信息詞典的詞語屬性描述體系，基于人民日報基本標注語料庫，探討以定性和定量相結合的方式對漢語詞語的情感標注進行研究。根據真實文本實例的統計、歸納，對詞典中詞語的情感傾向加以描述，然后在詞典中形式化。詞語的情感評價屬性的計算處理對文本過濾、信息抽取、網頁評價等有重要的參考價值。

關鍵字: 詞語情感評價、語義韻律、搭配規律

Abstract

This paper introduces the attributes of emotional evaluation in the Grammatical Knowledge-base of Contemporary Chinese. Lexical emotion tagging is studied by means of both qualitative and quantitative approaches. Based on the statistical results from the People's Daily tagging corpus, lexical emotional trends are described and formulated in our Knowledge-base. Lastly, we also discuss potential applications of emotion tagging in related tasks such as text filtering, information retrieval and web page evaluation.

---

[1] 相關研究得到中國國家 973 項目(2004CB318102)和國家 863 計劃(2001AA114210，2002AA117010)的支持。

* 北京大學計算語言學研究所，100871 中國
Institute of Computational Linguistics, Peking University, 100871 China
E-mail: {wangzm, yusw}@pku.edu.cn

[Received April 11, 2005; Revised June 13, 2005; Accepted August 15, 2005]

Keywords: Evaluation of Lexical Emotion, Semantic Prosody, Collocation Regulation

1. 引言

隨著信息處理領域“信息檢索、文本過濾、自動文摘、網頁評價”等技術的不斷發展，研究者開始利用詞語所表現出來的情感屬性來提高實用系統的智能化水準，出現了以網頁評價、產品評價爲目標的情感分析研究。例如: [T'sou *et al*. 2005] 運用詞語的情感屬性來衡量名人的公眾看法: [Sanjiv *et al*. 2001] 從股票信息板塊獲取股票市場的評價；[咎紅英 2003]關于名人網頁的相關度評價研究等等。因此詞語表現出來的正面、負面的情感評價屬性特徵越來越受到學者們的關注。

經研究，我們發現詞語蘊涵的情感屬性對其句子中共現的詞語有很大的限制，其共現詞語往往也要求具有統一的情感傾向。例如：以“潰逃”爲例，“潰逃”是個貶義詞，當它進入句子中與其共現的主語成分大都是含有貶義的壞人。例如：敵軍～、匪軍～、反動派～、土匪～、壞蛋～、罪犯～、走私犯～。與其共現的狀語成分也表示貶義的含義。例如：倉惶～、 狼狽～。也有互爲共現的詞語表現出不一致的情感傾向。例如：“擺脫”不是貶義詞，但通常與表示消極情感傾向的詞語共現，如：～困難、～困境、～貧困、～不發達狀態、～羞恥和孤獨、～危機、～老套套、～束縛、～危險、～制裁、～困擾。雖然后面所帶的詞語都是表示消極、負面的，但是整個句子卻表現一種積極、肯定的情感傾向。由此“擺脫”也帶上了積極、正面的色彩。

母語是漢語的中國人也許在毫無察覺的情況下下意識地運用詞語的情感色彩，而對外國留學生或計算機則需要學習才會理解。這些規律如果能夠從真實語料庫中提取，然后對這些規律進行定量的分析，形式化到知識庫中，無論對中文信息處理還是對外漢語教學都是很有價值的。

詞語情感評價在受限領域已經有了一些研究和探索。例如：[蘇玉梅 2004]提出了汽車領域網頁褒貶相關度評價模型。該模型對網頁實體進行褒貶態度評價，其中包含了一系列評價要素，如褒貶結構、領域標準等關鍵方法。該模組被應用到天網知名度系統，得到了很好的試驗效果。

對于漢語通用領域的詞語情感色彩評價研究，目前還沒有人利用大規模文本來做這樣的事。北京大學計算語言所長期致力于中文信息基礎資源的研究和開發。其重要的研究成果《現代漢語語法信息詞典》（簡稱《語法信息詞典》）和人民日報標注語料庫爲詞語情感評價研究提供了非常好的基礎資源。語法信息詞典共計收詞 73000 多條[2]。該詞典在中文信息處理的自動分詞、詞語標注、機器翻譯、信息提取、信息檢索、概念詞典建設等方面發揮了較大的作用[俞士汶 2003] 。

這兩年北大計算語言所又得到了 863 項目的支援。預期達到的目標是在突破關鍵技術基礎上，研製符合奧運多語言信息服務需求的大規模通用基礎資源，並建設綜合型語

[2] 這是《語法信息词典》2003 年的詞條數量，截止到 2004 年 12 月，詞條數量已經增至到 8 萬條。

言知識庫。語法信息詞典的擴充作爲其中一個子任務，要求在 2004 年年底增至到 8 萬詞條，利用這次詞典擴充的機會，擬將情感評價特徵納入語法信息詞典的屬性描述體系。

2. 情感評價屬性的界定

關于詞語的情感評價，中國語言學界稱之爲詞語的感情色彩（詞的褒貶）。相關的詞彙學著作有過論述[符淮青 1985; 劉叔新 1990]，但他們只是簡單地說明詞語感情色彩的定義和枚舉相應的例子，語法信息詞典的情感評價屬性描述和傳統語言學的詞語感情色彩研究有一定區別。爲了能夠和語言學的感情色彩相區別，我們在以后的描述中使用“情感評價”這一術語。

北大現代漢語教材[1997]指出：感情色彩指詞義所附帶的表示褒貶態度的色彩。詞的感情色彩同詞的意義關係密切，詞義對客觀事物有肯定評價的，一般有褒揚色彩。如：英雄、勇士、悅耳、富饒…，詞義對客觀事物有否定評價的，一般有貶斥的感情色彩。如：奸賊、賭徒、陰險、平庸…。從上述的定義可以看出詞語的褒貶色彩是一種表示程度很高的情感評價，這部分詞語在實際語言當中比例並不是很多，大多數詞語雖然無法說出它們的褒貶，但在語言環境中可以表現出積極或消極（正面或負面）的情感傾向。這些詞語在句子中表現出來的情感傾向和詞語共現所表現出來的搭配規律很少有人關注，而這部分信息對于文本過濾，信息安全、網頁評價有很重要的應用價值。我們以“陷入”爲例，“陷入”在感情色彩方面並無貶義色彩，但是在具體的語言環境中往往要求其共現詞語含負面、消極的因素。由此“陷入”也表現出消極、負面的情感傾向，這點我們在下面的例句中得到了驗證。陷入(～惡性循環狀態、～苦悶邊緣、～困境、～金融危機、～金融麻煩、～被動、～僵局、～困擾、～危險、～窘境、～被壓迫民族的地位、～惡性循環、～一連串的追殺之中、～彷徨、～庸人的狹小圈子裏、～了一個雷區、～衰退、～癱瘓、～混亂、～了地獄、～低潮、～嚴重的分裂、～孤立狀態、～茫然～、一個怪圈，一個誤區、～低谷、～恐慌、～重重包圍、～沉思、～凝重的沉思)。

“陷入”的使用頻率很高，僅僅兩個月的人民日報就出現了 33 次，表現出負面信息的句子有 32 例可以得到確認，只有一例負面評價不好確認，如：～沉思，“沉思”在現漢中是“深思”之義，表示一種中性含義，例句原文如下：“吳書記眉頭漸漸鎖緊，陷入了沉思。”回到原文就可發現“沉思”所在的語境所表現出來也是一種愁苦的樣子，也應該是負面含義。同時還找到了一句“凝重的沉思”，也是表示一種不好的心緒。以此我們可以得出一個結論，“陷入”通常表示消極、負面的評價。與“陷入”同義的還有“跌入 、陷沒、陷落”,它們在句子中的共現詞語也同樣表示消極、負面的感情色彩。動詞短語的中心語動詞“陷入”和其后面所共現的名詞的情感標注趨于一致，給定了“陷入”和其后面共現詞語的情感屬性，機器就可以判定關鍵詞所在句子所描述事件的好壞。因此語法信息詞典對詞語的描述並不限于褒貶的評價，同時還要對詞語進行正面、負面等不同程度的評價。

語言是動態的，靜態詞庫裏的詞語在實際的語言應用中會受到其搭配詞語的影響，一個本來沒有多少褒貶意義的詞語在進入句子框架后可能會表現出強烈的情感傾向。這

方面的研究，國外已經開始得到重視。他們稱之爲語義韻律。[Partington 1996]將語義韻律定義爲“超越單個詞語界限的色彩意義的擴展”(Semantic prosody refers to the spreading of connotational coloring beyond single word boundaries) ，語義韻律由音韻學上的“韻律”概念轉化而來，音韻學上的韻律現象是用來概括語音研究中切分成分在語流中具有超切分特徵。音韻學中的同化、異化、連讀都屬于韻律的研究範疇。受這種現象的啓發，[Sinclair 1987，1991]把韻律加以推廣，認爲這種現象同樣存在語言的詞彙層面，同時利用大規模的語料庫資源對詞語的韻律進行了搭配意義方面的研究。

綜合國內外的研究成果，詞語的情感評價應該從兩個層面來考慮。一方面應該是靜態詞彙層面的研究，即詞語在靜態詞庫中所表現出來的褒義、正面、負面、貶義等情感屬性，這些屬性可以直接在語法信息詞典中描述。另一方面詞語的情感信息在進入句子框架下會發現情感偏移現象，即詞語評價屬性的動態句法研究。詞語情感評價的動態研究將會有助于發現詞語之間的動態搭配規律，同時對于描述語言規律，探求語言認知心理有很好的啓示。

3. 語法信息詞典的評價類別

漢語中的詞語具有情感評價的詞語並不限于名詞、形容詞，其他詞類的詞語也有類似的情感評價傾向。語法信息詞典中可以加上情感色彩描述的有 12 類。具體詞類如下：

名詞(n)

(1) 春光 慈父 牛市 英雄 勇士 勞模

(2) 癌細胞 愛滋病毒 悲歌 弊端 殘骸 慘禍 熊市 黑窩 痞子

動詞(v)

(1) 鍛煉 發明 發揚 防止 奉獻 改善 感謝 鼓勵 灌溉 激發 獎勵

(2) 暴虐 爆發 爆炸 貶低 瀕臨 殘殺 挫傷 顛覆 妒忌 訛詐 妨礙

形容詞(a)

(1) 美麗 聰穎 恭敬 燦爛 光滑 單純

(2) 傲慢 暴虐 悲淒 憋悶 沉痛 惆悵

狀態詞(z)

(1) 碧油油 光燦燦 甜絲絲 水汪汪 美滋滋 熱熱鬧鬧

(2) 悲慘慘 癡呆呆 病歪歪 惡狠狠 瘋顛顛 邋裏邋遢

區別詞(b)

(1) 錦繡(～中華～風光～前程) 稀世(～珍寶～珍品～之寶～杰作)

(2) 填鴨式(～教學)劣質(～食品) 僞劣(～產品) 違禁(～物品)冒牌(～貨～開發商)

副詞(d)

(1) 乘興(～追擊) 穩步(～發展～推進～提高～反彈～增長～上揚～升值)豁然(～開

朗) 竭誠（～服務）闊步(～前進) 銳意(～進取) 捨身(～拼搏～救人～爲國)

(2) 大肆(～燒殺～翻供～索要收取錢物～進行分裂祖國的活動～進行盜竊)

公然(～對鄰國巴基斯坦進行威脅～在巴大肆進行國家恐怖主義活動～敲詐開出天價～拒絕～縱容其進行製造“兩個中國”，～造假，僞造歷史文獻)

乘機(～搗亂～鑽洞而入～發洩～甩包袱) 遲遲(～未察覺～不表態)

不巧(～中國選手來日后多人患感冒～東京的櫻花開得早)

不慎(～掉到地上摔成重傷～買了僞劣產品)成心(～把泡菜醃老了)

嘆詞(e)

(1) 哈哈(～，我鯊膽有救了，膽子還可大一點) 好傢伙(～，當時我差點暈過去)

呵呵(～，就來就來)

(2) 嗚呼(～！她不死于盜匪之手而死于親人之口，真是天下第一等大悲了！)

(～，如果人心與人心之間都掛著防賊的鎖，那單位還有生氣嗎？)

成語(i)

(1) 百廢俱興 比翼雙飛 別具匠心 彬彬有禮 博古通今 赤膽忠心 寵辱不驚

(2) 暗箭傷人 黯然失色 稗官野史 笨嘴拙舌 遍體鱗傷 不擇手段 稱王稱霸

慣用語(l)

(1) 飽眼福 爆冷門 賣力氣 鳴不平 鬧新房 上檔次

(2) 不入流 吃獨食 吃啞巴虧 穿小鞋 發善心 鼓倒掌 悶葫蘆 上賊船 背包袱

簡稱(j)

(1) 五講四美 雙擁 兩彈一星

(2) 死緩 老弱病殘 假冒僞劣 危舊房

代詞(r)

(1) 您 足下

(2) 鄙人 吾輩

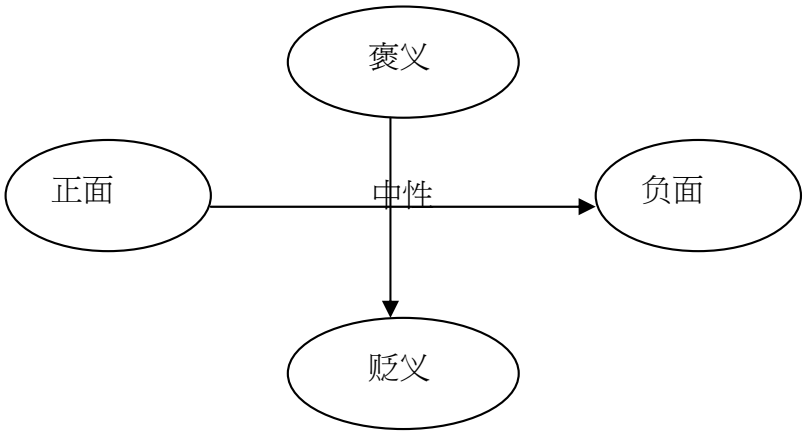
擬聲詞(o)

(1) 嘿嘿 哈哈 喳喳

(2) 喀嚓 嗡嗡

(1)表示“正面或褒義”的詞語，(2)表示“負面或貶義”的詞語。

語法信息詞典的情感屬性設定爲【褒義|正面|中性|負面|貶義】五個級別。



正負面評價是詞彙內容表現的客觀評價，不包含個人的主觀態度。例如：“慘禍”指嚴重的災禍。雖然災禍是指人爲不可抗拒的各種災害，具有消極、負面的傾向，但沒有任何人爲褒揚或貶損的主觀情緒。正面/負面判斷的方式一方面來自詞義本身，另一方面也來自詞義的上下文環境。詞語的【正面、負面】在表面上不容易判別，需要通過和搭配詞語的共現獲得某種情感傾向，例如 “單純”，如果說孩子單純，一般指孩子純潔，表現正面評價；如果說計算機單純，指的就是計算機思想單一，頭腦簡單，四肢發達，表示的是一種負面評價。

【褒義、貶義】和語言學的褒貶基本相當，詞語的褒貶評價從詞彙本身也可以看出，和正負面評價不同的是，【褒義、貶義】的詞語往往帶有人爲的主觀因素。例如：“敵人和強盜”就加入了人們對壞人的一種評價態度，因爲從詞義本身我們就可以斷定這兩個詞是貶義。除此之外，【褒義、貶義】也傾向于在一組近義詞或反義詞中判定。如“固執和頑固”都表示負面色彩，但兩者有不同程度的差別，“頑固”相對于“固執”貶義的程度更深一些。因此可以設定： 固執【＋負面】；頑固【＋負面 ＋貶義】。

根據上述標準，我們可以把名詞(1)(2)可以細分爲：

【正面】春光

【褒義】慈父 牛市 英雄 勇士 勞模

【負面】癌細胞 愛滋病毒 悲歌 弊端 殘骸 慘禍

【貶義】熊市 黑窩 痞子

詞語的【褒義、貶義】表現出來的情感傾向比較明顯，實際上可以算作【正面、負面】的特殊形式。通常情況下，從詞語的【褒義、貶義】可以推導出詞語的【正面、負面】評價，但是【正面、負面】評價並一定能推導出詞語的【褒義、貶義】評價。對于【褒義、貶義】這兩種特殊形式，往往並不適合描述上面列舉的全部詞語類別。因此在實際填寫過程中詞類的屬性設定也略有差別。例如：擬聲詞(o)、嘆詞(e)只設定【正面、負面、中性】三個屬性選擇，而且個別還要有所調整，比如代詞(r) “鄙人，吾輩”按照

上述標準不好做出判定，因此評價屬性定義爲【敬語|謙語】。

4. 基于語法信息詞典的詞語評價調查

詞語情感評價不僅僅涉及到特定詞或與其搭配的短語本身，還涉及到兩者之間相互作用而產生微妙的情感意義。語義表現出來的正面和負面的感情色彩往往通過直覺發現，但是這種憑直覺發現詞語所具有的感情色彩無法驗證，而基于千萬字量級的漢語基本標注語料庫和語法信息詞典提供的知識足夠使我們把這些現象通過統計揭示出來。例如：語法信息詞典提供的常用詞彙“有”和“味”都是中性詞語，但是它們進入句子框架后卻表現出具有明顯的情感評價傾向。它們在真實語料中有什麼樣的情感表現？它所表現的情感傾向的概率是多少？這些問題我們都可以通過千萬字量級的人民日報語料庫加以驗證。統計發現有以下兩種情況。

第一種情況，“有味”獨立出現，有時后面也加兒化，一共有 12 例

1、在街道上唱，公園裏演，蠻有味兒。

2、人安靜下來，越看越愛看，越看越有味。

3、即使是拈得雞肋者，也會叫道“食之有味，棄之無禮”。

4、聽上了廣播，日子是越過越有味了

5、做到真實可信，親切有味，引人入勝

6、人間有味是清歡———大型畫冊《中國竹工藝》賞析

7、可以化腐朽爲神奇，變無味爲有味。

8、先期讀到《雷達散文》的人士，都對雷達的散文評價甚高，認爲雷達的散文自然而有味。

9、同裏人自己則說，同裏的橋有韻、有味，橋是同裏的性格，

10、她們洋溢著自然天成的性情美：自信，進取，活力四射，光彩照人，有爲更有味。

11、讀書一日，就有一日之益。“讀書有味身忘老。”這是讀書從“苦讀”進到“樂學”境界的表現。

12、乘警巡船時在船艙內聞見氣味異常，找到有味車的車主，車主先謊稱是食品添加劑，后又稱是化工品。

第二種情況又包含兩個小類。第一小類“有味”作爲固定短語的一部分出現。出現在“津津有味”的例子有 36 例，和“津津有味”共現的動詞一般都是“聽、看、閱讀、喝、講、談起、品嘗、吃、嚼、介紹”等。例如：

1、沒想到所有家長都聽得津津有味。

2、斯圖拉普一直捧著一本名爲《我心依舊》的小說讀得津津有味。

3、一邊拿著一本豎排的中文書在津津有味地閱讀。

4、正當我津津有味地瞧著時，忽然傳來“當”的一聲，

5、便情不自禁地坐在攤前，匯入津津有味的食客行列。

第二小類，“有味”出現在固定短語“有滋有味”的例子有 28 例。“有滋有味”一般在句子中做補語。與其共現的動詞有“看、喝、吃、啃、說、品味、生活、過，覺，唱、當得、干得、打發”。

1、別看不是專業演出，可鄉親們依然看得有滋有味。

2、她都堅強地挺過來了，而且生活得有滋有味。

3、擺碟花生米，就可以有滋有味地喝起來。

4、生活也必然會有滋有味，精精神神，充滿無窮樂趣。

5、“都都”生活良好，有滋有味地吃了生日蛋糕。

還有一句出現在“有情有味”中。例如：

7、他的畫是有情有味有看頭和經得起琢磨的。

以上的所有例句來自于一年半的人民日報語料。“有味”這個短語在出現的所有例句中，76 句有 75 句帶有褒義的感情色彩，約占所有出現例句的 98.7%。整個句子表現出正面的情感傾向，只有一句是表示負面地含義。即：第一種情況中的例 12“…有味車…”。“有味車”中的“有味”是指不好的氣味，應該是“有氣味”的簡寫，在所有的語料當中只有一例說的是車上有味的含義，表現出消極的負面評價。而表示正面、褒義情感傾向的“有味”是指“有味道”的含義。我們在語料中找到了 15 個這樣的例子。例如：

1、也不論是講述生動有趣的、與飲食有關的故事，總之得有味道，品出多味的生活。

2、但該劇的成功，或說最大的特色，即是臺詞很有味道，語言精彩。

它們雖然字形相同，但是所表示的意思是不一樣的。如何在文本中自動判別“有味”屬于哪種類型值得進一步研究。

5. 計算機的形式化研究

語法信息詞典採用成熟的關係資料庫技術，詞語的評價屬性作爲原詞典的一個擴充項也沿用此種結構描述。目前語法信息詞典已經把“津津有味”、“有滋有味”收入詞典，而“有味”不作爲一個詞條。人民日報基本標注語料庫統一看作一個切分單位。在這方面兩者是不完全等同的。

1、越/d 看/v 越/d 愛/v 看/v ，/w 越/d 看/v 越/d 有味/a 。/w

2、沒/d 想到/v 所有/b 家長/n 都/d 聽/v 得/u 津津有味/i 。/w

3、可/c 鄉親/n 們/k 依然/z 看/v 得/u 有滋有味/l 。

根據這種情況，首先把詞典含有“有味”的四字短語標注上評價屬性。對于“有”和“味|味兒”，我們採用分開處理，類似的情況還有很多，例如：“有思想、有氣質、有頭腦、有意見”。不同的名詞和“有”組合所表現的情感傾向是有差別的，因此我們

對“有＋N”的形式化描述重點放在后面的N上,根據N和“有”在語料中的真實表現給出他們的情感屬性概率值。

語法信息詞典形式化的描述信息如下：（下表中頻率數值根據一年半人民日報統計得到）

詞語	拼音	詞性	褒貶評價	正負面評價	搭配詞語	概率
津津有味	Jin1jin1you3wei4	i	褒義	正面	聽、看、閱讀、喝、講、談起、品嘗、吃、嚼、介紹	1.0
有滋有味	You3zi1you3wei4	l	褒義	正面	看，喝、吃、啃、說、品味、生活、過，覺，唱、當得、幹得、打發	1.0
有	You3	v	中性	正面	味、味兒	0.917
味	Wei4	n	中性	正面	有	0.917

當然“有味”也可以當作一個詞條收入語法信息詞典，這時對“有味”的描述就和“津津有味、有滋有味”相類似。不過現在將“有味”分開處理，就是要從搭配的角度考察詞語進入句法環境之后所表現出來的情感變化，從方法論的角度爲詞語的情感搭配研究提供一種新思路。

6. 詞語情感評價的應用價值

詞語的情感評價研究是機器理解漢語的新拓展，讓機器真正理解自然語言一直是研究者一個遙遠的夢想。很多學者對此都覺得希望渺茫，其中一個十分重要的原因是，機器在自動分析時只是簡單的模式匹配，根本談不上理解，而目前可利用的語義資源也十分有限，詞語的情感評價資源將是一個重要的補充。比如當我們聽到“經濟衰退，股市下滑”的新聞報導時，我們就會由此判斷最近經濟不景氣，而機器無法判斷，但是如果給定 “衰退”和“下滑”具有負面情感傾向的屬性特徵，我們就可以利用機器預測股市的發展。詞語的情感評價研究如果能夠利用現有語料庫的豐富資源，給出量化和定性分析，不僅對文本過濾，信息安全、網頁評價的智能化研究有潛在的應用價值，而且對語言學習者也是一個很重要的信息，它能夠幫助學習者在實際語言使用中選擇正確、恰當的詞語。特別是對于母語非漢語的外國留學生而言，如果不理解詞語的情感評價屬性信息，就會在實際交際當中出現錯誤，而這些詞語所表現出來的情感傾向一方面通過實際的交流獲得，另一方面來源于工具書上的信息。但是影響最大的《現代漢語詞典》沒有提供這方

面的屬性。因此語法信息詞典的詞語情感描述研究也會爲外國留學生更深刻地理解漢語詞語提供一個重要的語言參考資源。

上面從計算機實現角度對詞語表現出來的情感傾向進行了簡單分類，而且依據大規模語料的證據給予了計量的表示，這詞語的上下文環境對詞語的情感判定起到關鍵作用。因此結合上下文的情感評價研究值得進一步探索。

當然知識庫手工情感標注往往造價很高。專家們試圖尋找各種自動方式的情感分類技術。例如[Pang 2005]提出一種從文本中自動抽取詞語情感多級分類的研究方法。[Takamura *et al.* 2005] 利用電極螺旋模型（spin model）（相當于詞語情感的正負極）來抽取詞語的情感極的研究。情感評價分類是自動分析的基礎，在研究初期手工標注工作還是必要的，未來的工作可以考慮利用機器學習等方法實現基于大規模語料的情感詞語的自動抽取。

致謝

筆者在研究過程中，北京大學計算語言所胡景賀同學、諶貽榮同學、呂學強博士、吳云芳博士提出了很好的建議，在此向他們表示衷心的感謝，同時也要特別感謝黃居仁教授提供的重要參考文獻以及對文章的修改建議。

參考文獻

大衛.克裏斯特爾[英]，沈家煊譯，《現代語言學詞典》，*商務印書館* 第 4 版, 2002，pp：290。

俞士汶等，《現代漢語語法信息詞典詳解》，*清華大學出版社*（第 2 版），2003，pp:40-41。

應英等，“漢語情感意義的機器標注研究初探，” *中文信息學報* （第 2 期），2002。

昝紅英, “名人網頁的相關度評價，” *中文信息學報*, 2003.5，pp:5。

蘇玉梅, “中文網頁褒貶態度的機器評價，” 碩士論文，2004，pp:7。

符淮青，《現代漢語詞彙》，*北京大學出版社* 1985，pp:28。

劉叔新，《漢語描述詞彙學》, *商務印書館* 1990，pp:11。

北京大學中文系現代漢語教研室，《現代漢語》,*商務印書館*，1997，pp:207-208。

呂叔湘，《漢語語法分析問題》，*商務印書館*，1979。

呂叔湘，《現代漢語八百詞》，*商務印書館*，1980。

Kleinberg, J., and $\dot{E}$. Tardos, “Approximation algorithms for classi_cation problems with pairwise relationships: Metric labeling and Markov random fields,” *Journal of the ACM*, 49(5): 2002, pp. 616-639.

Kushal, D., S. Lawrence, and D.M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classi_cation of product reviews,” *In Proceedings of WWW,* 2003, pp. 519-528.

Moshe K., and J. Schler, “The importance of neutral examples for learning sentiment,” *In Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN)*. 2005.

Palmer, F., ed. “Selected Papers of J.R. Firth, ” *Landon: Longman*, 1952-1959, 1968.

Pang, B., and L. Lee, “Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales,” *Proceedings of ACL 2005*, pp. 115-124.

Parington, A., “Patterns and Meanings :Using Corpus for English Language Research and Teaching”, 1996, pp. 68

Sanjiv, D., and M. Chen, “Yahoo! for Amazon: Extracting market sentiment from stock message boards,” *In Proceedings of the Asia Paci_c Finance Association Annual Conference (APFA)*. 2001.

Sinclair, J.M., “Looking Up,” *London and Glasgow: William Collins*, 1987.

Sinclair, J.M., “Corpus. Concordance, Collocation,” *Oxford University Press*, 1991.

Takamura, H., T. Inui, and M. Okumura, “Extracting Semantic Orientation of Words Using Spin Model,” *Proceedings of ACL* 2005, pp. 133-140.

T'sou, B. K.T., O.Y. K. wong, W.L. Wong, and T. B.Y. Lai, “Sentiment and Content Analysis of Chinese News Coverage,” *In International Journal of Computer Processing of Oriental Languages*, 18.2. 2005, pp. 171-183.