Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 57-84

Back to the Basic: Exploring Base Concepts from the Wordnet Glosses

Chan-Chia Hsu[*] and Shu-Kai Hsieh[*]

Abstract

There has been no consensus as to what constitutes a set of base concepts in the mental landscape. With the aim of exploring base concepts in Chinese, this paper proposes that frequently-occurring words in the glosses of a lexical resource such as the Chinese Wordnet can be seen as a candidate set of base concepts because the glosses use basic words. The present study identified 130 base concepts in Chinese. The Base Concepts in EuroWordNet were adopted as a reference for comparison. While only 44.6% of the base concepts identified in the present study have an equivalent in the set of Base Concepts of EuroWordNet, the other base concepts extracted by our gloss-based approach also reflect a certain degree of basicness. It is hoped that both the overlap and the difference between different sets of base concepts identified in different languages and by different approaches can deepen our understanding of the basic core in the mind. Additionally, it is also hoped that the set of base concepts identified in the present study can have computational as well as pedagogical applications in the future.

Keywords: Chinese Wordnet, EuroWordNet, Base Concept, Gloss

1. Introduction

For the past few decades, a large body of research has been trying to touch on the basic core in the mind. Some studies (e.g., Wierzbicka, 1996) have aimed to figure out how a large number of concepts in the mind can be neatly organized with a basic set of concepts, leading us to the realm of human cognition. Furthermore, some studies have identified a set of base concepts that have had a wide range of computational applications.[1] WordNet (Miller *et al.*, 1990), for instance, is organized around a set of base concepts (i.e., *SuperSenses*), with which a large number of lexical items are associated through lexical relations. There have been many

[*] Graduate Institute of Linguistics, National Taiwan University, Taiwan
E-mail: chanchiah@gmail.com; shukaihsieh@ntu.edu.tw

[1] The term *base concept* should be distinguished from other terms related to the notion of basicness in the mind, such as *basic level concept*. See Section 2 for a more comprehensive review.

approaches to exploring what is basic in the mind, but there has been no consensus as to what constitutes a set of base concepts universal to all human languages.

This study aims at providing a new perspective to identify a candidate set of base concepts in Chinese. Our data consist of the glosses in the Chinese Wordnet. Since the glosses in the Chinese Wordnet use basic words, words that occur frequently in the glosses of the Chinese Wordnet can be assumed to be reflective of a candidate set of base concepts. After data extraction and introspection, the resulting set of base concepts in the present study is compared with the set of Base Concepts proposed in the EuroWordNet project (Vossen *et al.*, 1998). In selecting a set of base concepts, our method is based on the *frequencies* of words used in the glosses of the Chinese Wordnet, whereas the method adopted in the EuroWordNet project is based on the *relations* between synsets. It is thus noted that the set of Base Concepts in EuroWordNet is not seen as de facto, but as a reference. We use the Base Concepts in EuroWordNet as our reference because on the one hand, the Chinese Wordnet and EuroWordNet both derive from the WordNet framework, and on the other hand, the set of Base Concepts from EuroWordNet is based on many European languages. It is hoped that both the overlap and the difference between different sets of base concepts identified by different approaches can deepen our understanding of the basic core in the mind. Additionally, it is also hoped that the set of base concepts identified in the present study can have computational as well as pedagogical applications in the future.

This paper is organized as follows. Section 2 provides a comprehensive review of different approaches to the notion of basicness in the mind. Section 3 reviews the significance of glosses in different contexts. Section 4 introduces our experiment method and presents the set of base concepts identified in the present study. Section 5 discusses how our proposed set of base concepts in Chinese is different from that of EuroWordNet. Section 6 concludes the paper.

2. Defining the Core Lexicon in Language and the Mind

Over the past few decades, there have been various approaches to the notion of *basicness* in the mental landscape. Some have created lists of lexical items as basic words, mainly for pedagogical purposes. Some, from a cognitive perspective, have selected different sets of basic concepts at different levels of abstraction (e.g., semantic primitives, base concepts, basic-level categories, and basic domains).

The present study focuses on base concepts, which have contributed to the establishment of lexical resources (e.g., WordNet, EuroWordNet, and BalkaNet). Compared with basic words, base concepts have more computational applications than pedagogical ones. Compared with semantic primitives and basic domains, base concepts are selected in a more scientific procedure. Compared with basic-level categories, base concepts are hierarchically higher. A comprehensive review of different approaches to the notion of basicness in the mind will be given in the following.

2.1 Basic Words

One of the earliest efforts to address the notion of basicness in the lexicon is to identify a list of basic words, which is motivated by pedagogical needs.[2] Many basic vocabulary lists have been proposed, ranging from 300 words to more than 2,000 words (e.g., Dolch, 1936; Gates, 1926; Hindmarsh, 1980; Lee, 2001; McCarthy, 1999; McCarthy & O'Dell, 1999; Ogden, 1930; West, 1953; Wheeler & Howell, 1930). With the rapid development of computational analyses, such lists are mostly based on frequency counts. They can serve as useful references for pedagogical purposes, such as the design of a syllabus and the development of a language proficiency test. The main problem with most basic vocabulary lists is that the raw data on which the frequency counts are based may not be representative enough. Additionally, since what counts as a word is an issue in itself, an insight is needed when it comes to word forms and lexicalized phrases (McCarthy, 1999).

2.2 Semantic Primitives

In the discussion of basicness in the mind, more abstract than basic words are semantic primitives, or semantic primes, which are pursued mainly in the theory of Natural Semantic Metalanguage (Goddard, 2002; Wierzbicka, 1972, 1996).[3] A semantic primitive is basic in the sense that it is lexicalized in every language and that it cannot be defined or paraphrased in simpler terms. From a cognitive perspective, it is suggested that there is an innate set of semantic primitives representing "a universal set of fundamental human concepts" (Wierzbicka, 1996:13). Such a set is argued to be sufficient to define or paraphrase the entire vocabulary of a language. For example, the word *envy* can be defined as what follows (Wierzbicka, 1996:161):

[2] In previous studies, the terms "basic vocabulary", "sight vocabulary", "core vocabulary", and the like are sometimes interchangeable.

[3] For others who have adopted a similar approach in languages other than English, see Goddard (2002:12).

X feels envy. =
sometimes a person thinks something like this:
    something good happened to this other person
    it didn't happen to me
    I want things like this to happen to me
because of this, this person feels something bad
X feels something like this

Specifically, Goddard (2002:14) has presented 58 "atoms of meaning", such as I, YOU, SOMEONE, PEOPLE, SOMETHING/THING, and BODY. Unfortunately, this line of research is open to valid criticisms due to a lack of a sound method of identifying semantic primitives (e.g., Riemer, 2006).

2.3 Base Concepts in WordNets

The notion of basicness has played a vital role in many lexical resources, such as English WordNet (Miller *et al.*, 1990),[4] EuroWordnet (Vossen *et al.*, 1998), and BalkaNet (Cristea *et al.*, 2002). In the architecture of English WordNet, synonyms are assembled in a set called *synset* (synonymous set). During the development of WordNet, synsets are organized into 45 lexicographical files based on the criteria of syntactic category and logical groupings. The 45 names of lexicographical files (e.g., noun.feeling and verb.cognition) are also called *SuperSenses*, which reveal the base concepts from the developer's perspectives.[5]

As an extension of the wordnet model, EuroWordnet further proposes a set of 1,024 core synsets - called **Base Concepts -** that are extracted from four wordnets and translated into the closest WordNet 1.5 synsets. To keep the set balanced and shared among these wordnets, 164 core base concepts of them were selected in terms of their (more) relations with other concepts and (higher) position in the hierarchy.[6] Based on the Base Concepts identified for EuroWordNet, the BalkaNet project adopts a similar approach and selects a set of Base Concepts by focusing on five Balkan languages, including Bulgarian, Greek, Romanian,

[4] WordNet is open to the general public at http://wordnet.princeton.edu.

[5] For the format of the lexicographical files, see *wninput(5WN)* at http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html.

[6] The 164 Base Concepts in EuroWordnet consist of 66 concrete synsets (nouns) and 98 abstract synsets (nouns and verbs). For more details, refer to http://www.globalwordnet.org/gwa/ewn_to_bc/ConcreteInfo.html and http://www.globalwordnet.org/gwa/ewn_to_bc/AbstractInfo.htm.

Serbian, and Turkish.[7]

2.4 Basic-level Concepts

In the context of cognitive linguistics, many experiments have shown that in taxonomies of *concrete* objects, there is one level of abstraction that is regarded as **basic** which distinguishes them from higher and lower-level categories (Cruse, 1977, 2000; Rosch *et al.*, 1976). For instance, in answering the question *what's that in the garden*, most speakers choose to say *a dog* rather than its hypernym *an animal* or its hyponym *an Alsatian* (Cruse, 1977:153-154). Compared with the ANIMAL concept and the ALSATIAN concept, the DOG concept is seen as a basic-level concept in that both its internal homogeneity and its distinctness from neighboring concepts are greater. The presumption of *basic-level concepts* has been also supported by language acquisition studies, which reveal a large percentage of children's early words are basic-level terms (Ungerer & Schmid, 2006).[8]

Some recent computational approaches have attempted to use algorithms to automatically extract the basic-level concepts. Izquierdo *et al.* (2008) automatically select basic-level concepts from WordNet based on the relations between synsets, while Lin (2010) proposes an algorithm that can automatically identify the cognitive level of a noun in WordNet based on the ability of the noun to form compounds and the position of the noun in a hierarchical chain.

A relevant discussion with regard to basic conceptualization in the study of language and the mind has been focused on *basic domains*, which derive directly from human *embodied experience* (e.g., sensory and subjective experience). Cognitive Grammar argues that a concept should be understood in terms of another more general, inclusive concept (Langacker, 1987:148). For example, the concept RADIUS makes sense only when it is viewed against the concept CIRCLE. Such a relationship can form a chain (i.e., the concept CIRCLE should be understood in terms of the concept SPACE), but the chain cannot be endless. Some concepts of a general nature, such as SPACE, TIME, and QUANTITY, are basic domains because they are characterized by a high degree of inclusiveness.

3. Definitions and Glosses in Different Contexts

Defining a word can be as easy as pointing to something the word refers to, but it can be as difficult as formulating "an ideal hypothetical norm which is a sort of compromise between

[7] For more information about the BalkaNet project, refer to http://www.dblab.upatras.gr/balkanet/ and http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=53 for similar works (e.g., Atserias *et al.*, 2003).

[8] Note that *basic-level concepts* should not be confused with *Base Concepts*. While a Base Concept occupies a high position in a hierarchy, a basic-level concept occurs in the middle of a hierarchy.

the generalization of inadequate experiential reality and a projected reality which is yet to be attained in its entirety" (Bernard, 1941:510). In different contexts, definitions and glosses play different roles, which will be reviewed in the following.

3.1 Definitions in Linguistic Semantics

When it comes to the meaning of a word, people may first think of looking up its definition in a dictionary. A good understanding of word meaning relies thus upon how the word can be defined. In the discussion of linguistic semantics, there are many ways to define the meaning of a word (Riemer, 2010:65-79). A definition can be ostensive, relational, or extensional, and it sometimes combines different approaches.

First, perhaps the most obvious, people often define a word in terms of ostension, i.e., by pointing out the objects a word denotes. Though an ostensive definition is useful for concrete nouns, it may cause many difficulties when used to define verbs, adjectives, adverbs, and function words (e.g., prepositions).

Second, a definition can place a word in relation to other words or events. For example, a word can be defined by its synonyms. However, since there are few absolute synonyms, the identity between a word and its synonyms can be challenged. A word can also be defined through an event, which is regarded as a typical context for the word. For instance, the verb *scratch* can be defined as "the type of thing you do when you are itchy" (Riemer, 2010:66). The weakness of such a definition is that it works only when the addressee of the definition can accurately infer the intended meaning on the basis of the given cue. That is, someone may not get the correct meaning of *scratch* if he or she does not scratch when feeling itchy.

Third, a definition can be extensional, and one of the commonest strategies is to define by a broad class (i.e., genus) and some distinguishing features (i.e., differentia). For example, *man* (in the sense of "human being") can be loosely defined as "rational animal" (Riemer, 2010:67). One of the main problems of a genus-differentia definition is that it can be too abstract to its addressee (Landau, 2001:167).

In summary, there are many strategies to define the meaning of a word, and all of them have their limitations. More generally, the difficulty of a definitional approach to semantics is that defining the meaning of a piece of language with more language in the same system will inevitably end up circular (Portner, 2005:4).

3.2 Definitions in Lexicography

Explaining what words mean (thus the concepts they encode) is the central function of a dictionary. While the mental lexicon is a "theoretical exercise", a dictionary can be seen as a "practical work" (Landau, 2001:153). On the one hand, a dictionary simulates the mental lexicon, offering the phonological, syntactic, and semantic information of a lexical item. On the other hand, a dictionary cannot be as detailed as the mental lexicon, and lexicographers need to decide what to include in a dictionary. Compiling a dictionary is seen as a craft, for lexicographers aim to make the most of their limited resources to cater for the communicative and pedagogical needs of dictionary users.

One of the most challenging and contentious aspects of the compilation of a dictionary is the creation of *definitions* for a dictionary entry. The term ‘definition’ would be a misnomer if it implies that word’s meaning can be precisely pinned down. There are many strategies to define a word in a dictionary (Lew & Dziemianko, 2006). The most traditional definition in a dictionary is the *analytical model*, i.e., the **genus-differentia definition**. A definition composed in this way typically consists of two elements: the *genus expression* that locates the definiendum in the proper semantic category, and the *differentia* (or plural form *differentiae*) that indicates the information which makes the word differ from other words of the same semantic category. For example, *appraisal* is defined as “a statement or opinion judging the worth, value or condition of something” (taken from Longman Dictionary of Contemporary English), where ‘a statement or opinion’ is the *genus expression* and the postmodifying expression ‘judging the worth, value or condition of something’ is the *differentia*. In many cases, it is not an easy task to produce a genus-differentia definition, and such a definition can be difficult for a dictionary user to understand. Another way to define a word in a dictionary is to adopt a *contextual* definition. A contextual definition of ‘appraisal’, for example, is stated as “if you make an **appraisal of** something, you consider it carefully and form an opinion about it” (taken from Collins COBUILD Advanced Dictionary of English).

Our concern here is not to deal with the issue of ‘what makes a good definition’, or search for the underlying necessary and sufficient conditions, but to evaluate the way the principle of *maximal economy* is reflected in a definition sentence. Zgusta (1971) proposed a list of criteria, one of which states that the lexical definition “should not contain words more difficult to understand than the word defined” (cited in Landau, 2001:157). In addition, the effectiveness of dictionary definitions can be evaluated from the user’s viewpoint (Cumming *et al.*, 1994; Lew & Dziemianko, 2006). For example, language learners have been found to prefer contextual definitions to analytical ones (Cumming *et al.*, 1994). An interim conclusion thus worth drawing is that a definition should contain no more words than necessary, consistent with the demands of intelligibility and information-transfer (Atkins & Rundell, 2008).

3.3 Glosses in Lexical Resources

The reviews so far naturally lead us to the glosses (definitions of word senses) in **lexical and ontological resources** developed in recent years. Glosses and example sentences are two essential components in the construction of lexical resources like WordNet, for they have been proved to be highly useful in discovering semantic relations and word sense disambiguation tasks (Kulkarni *et al.*, 2010). In the design of WordNet, word lemmas are grouped into *synsets* (synonymous sets), which are organized as a lexical network by a wide range of lexical relations (e.g., hyponymy and antonymy). The role of glosses is thus to explain explicitly the meaning of *synsets* which lexically encode the human concepts.

Most of the lexical relations that connect *synsets* are *conceptually inclusive relations*, such as *hypernymy-hyponymy* and *holonymy-meronymy*, which make the wordnet architecture a hierarchical conceptual structure, or a **lexicalized ontology**.[9] In connection with ontology studies, Jarrar (2006) suggests that glosses can be of great use in an ontology. For example, glosses are easier to understand than formal representations, so ontology developers from different fields can rely on glosses to a certain degree when they communicate. However, as Jarrar (2006) further suggests, a gloss in an ontology is not intended to provide some general comments about a concept, as a traditional definition in a dictionary does. Instead, a gloss in an ontology functions in an auxiliary manner, providing some factual knowledge that is critical to the understanding of a concept but can be difficult to formalize explicitly and logically. As a consequence, glosses in a wordnet as a lexical ontology are different from dictionary definitions.

Jarrar (2006) provides some guidelines for writing a gloss in an ontology. First, an ontology gloss should start with the upper type of the concept being defined. Second, an ontology gloss should be in the form of a proposition. Third, an ontology gloss should emphasize the distinguishing features of the concept being defined. Fourth, an ontology gloss can include some examples. Fifth, an ontology gloss should be consistent with the formal representation of the concept being defined. Sixth, an ontology gloss should be sufficient and clear. Generally, the glosses in the Chinese Wordnet fulfill the above criteria. Here is an example taken from the Chinese Wordnet:

(1)

書：有 文字 或 圖畫 的 出版品

shu you wenzi huo tuhua DE chubanpin

‘book: a publication with words or pictures’

[9] According to Gruber (1995:908), an ontology is “an explicit specification of a conceptualization”, and a wordnet can be thought of as a lexical ontology because of its lexical implementation of conceptualization, in comparison with other formal ontologies (e.g., SUMO) where the focus is put on logical constrains.

While the gloss looks like a *genus-differentia* definition in a dictionary, they are different in essence. The definition techniques used by lexicographers to indicate differentiation come from various conventions, while the ontology gloss aims to make a minimal commitment to conceptualization, which meets the need of logical conciseness. The study of the basic lexicon is crucially different from other tasks of lexical acquisition in that unlike the latter where the broad coverage is at issue, the former requires instead fine-grained data to be explored. In summary, we propose that glosses in lexical resources are the best source to study the core component of the basic lexicon.

4. Glosses in the Chinese Wordnet

In this section, we introduce the method of how we used gloss data from the Chinese Wordnet to touch on base concepts.[10] The glosses in the Chinese Wordnet can be seen as a sample corpus with fine-grained lexical information. Figure 1 shows the similar type frequency distribution of 46 part-of-speeches (proposed by the Sinica Corpus) in the Sinica Corpus and the Chinese Wordnet, respectively.

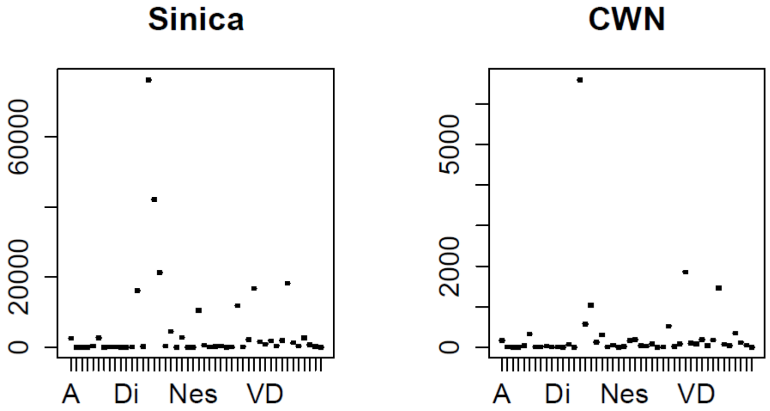

Figure 1. The POS distribution of the Sinica Corpus and the Chinese Wordnet

4.1 Extracting a Set of Frequently-occurring Words from the Glosses of the Chinese Wordnet

In our first experiment, we extracted a set of frequently-occurring words from the glosses of the Chinese Wordnet. Since a gloss in the Chinese Wordnet uses basic words instead of giving a scientific definition that can be incomprehensible to the user (Huang, 2008:22), the frequently-occurring words extracted from our experiment may reflect a certain degree of basicness in Chinese and even be considered to constitute a candidate set of base concepts in

[10] The Chinese Wordnet (CWN) has been released as an open-source project, and is freely available at http://lope.linguistics.ntu.edu.tw/cwn

Chinese. Our method and the results will be presented in the following.

Our first step was to extract all the glosses from the Chinese Wordnet. For glosses containing more than one period (i.e., the Chinese period 。), we discarded words preceding the first period because what precedes the first period in a gloss only provides grammatical properties. Next, what remained in the glosses was segmented by a segmentation system developed by Chinese Knowledge and Information Processing (CKIP). Consider the following example:

(2)

學生： 普通名詞。 在 學校 系統 內 讀書 學習 的 人。
xuesheng putongmingci zai xuexiao xitong nei dushu xuexi DE ren
‘student: someone who studies and learns in a school system’

In the example (2), *putong mingci* ‘common noun’ would be discarded, and then the remaining part of the definition would be segmented as shown in the example. With all the glosses segmented, a frequency wordlist with 19,852 words was created.

We manually checked the wordlist for meta-linguistic terms (e.g., *xingrong* ‘modify’) and mis-chunked words (e.g., **dedanwei* ‘DE + unit’). Only the first 1,000 words on the wordlist were checked both because our resources were limited and because it was assumed that core base concepts should be at the top of the frequency wordlist. For meta-linguistic terms, we chose to exclude them because it is obvious that they do not represent base concepts. For mis-chunked words, we either manually segmented them further (**dedanwei → de danwei*) or simply excluded them if they were not comprehensible (e.g., *dejian* ‘DE-simple’).[11] In such cases as *dedanwei*, the resulting words together with their frequencies were added to the wordlist if they had not been listed there, or the frequencies of the resulting words were revised. Take *de danwei* as an example. There were 328 *de danwei* in the data, and both *de* and *danwei* had been on the wordlist before *dedanwei* was further segmented. The frequencies of *de* and *danwei* were revised to be 15,653 and 1,178, respectively.[12]

To demonstrate how our new approach to identifying a set of base concepts is different from others, we decided to compare the resulting set in the present study with the set from EuroWordNet. Since all the Base Concepts in EuroWordNet are nouns and verbs, we focus on only nouns and verbs in the present study.[13] Therefore, words that were not tagged with V or

[11] The morpheme *jian* does not stand alone in Modern Chinese.

[12] Originally, there were 15,325 tokens of *de* and 850 tokens of *danwei* in the data.

[13] For which synsets in EuroWordNet were merged in the present study, see the appendix.

N were removed from our wordlist. In the end, the frequency wordlist based on the glosses of the Chinese Wordnet contained 17,018 words.

In EuroWordNet, there are 98 abstract Base Concepts and 66 concrete Base Concepts. However, as Vossen *et al.* (1998) have admitted, some synsets appear to represent almost the same concepts (e.g., {form 1; shape 1} and {form 6; pattern 5; shape 5}), so the number of the Base Concepts in EuroWordNet can be reduced. In such cases, we merged the two (or more) synsets into one. Finally, we retained 130 Base Concepts, i.e., 75 abstract concepts and 55 concrete concepts. Therefore, we also selected the top 130 words from our wordlist to be a candidate set of base concepts in Chinese.

When we examined the 130 words high on our wordlist, we found that some words needed to be replaced. First, two proper nouns were unsurprisingly high on the wordlist based on the Chinese Wordnet, i.e., *Zhongguo* ‘China’ (32th) and *Taiwan* ‘Taiwan’ (67th). The two words were excluded from the candidate set of base concepts. Second, since we focused on typical nouns and verbs, words typically not functioning as nouns or as verbs were excluded from our wordlist, regardless of their tags. Words discarded at this stage included:

(3)

負面	fumian	‘negative’
多	duo	‘numerous’
主要	zhuyao	‘primary’
大	da	‘big’
相同	xiangtong	‘the same’
小	xiao	‘small’
容易	rongyi	‘easy’
固定	guding	‘stable; fixed’
用來	yonglai	‘use…to…’
可以	keyi	‘can’
所在	suozai	‘a place where…’
受到	shoudao	a passivization marker in Chinese
沒有	meiyou	‘without’

In (3), words such as *da* and *xiao* usually function as adjectives, and *zhuyao* and *rongyi* can be adjectives or adverbs. The word *meiyou*, originally tagged as a noun, functions as a polarity operator rather than as a noun or as a verb.[14]

Another issue in the selection of the top 130 words from the glosses of the Chinese Wordnet was near-synonymy. For example, both *yong* 'use' and *shiyong* 'use' were high on our wordlist, and so were *wuti* 'object' and *wupin* 'object'. In deciding whether two words did represent the same concept, the present study counted on the Chinese Wordnet rather than on our own introspection or on further analyses. In the former case, *yong* 'use' and *shiyong* 'use' bear the relation of synonymy in the Chinese Wordnet. Therefore, the two words were considered to represent the same concept, and the frequencies of the two words were added together. In the latter case (i.e., *wuti* and *wupin*), the two words do not bear the relation of synonymy in the Chinese Wordnet. As a consequence, the two words were listed separately on our wordlist (cf. Table 1).

Finally, five words had two tags and were listed separately. They were *gaibian* 'change', *shiyong* 'use', *jisuan* 'calculate', *chansheng* 'produce, generate', and *fasheng* 'happen'. They are verbs in their literal sense, but they can be nominalized. For the five words, the frequencies of the verbal use and the nominal use were added together, and each word was listed only once in our wordlist since both the verbal use and the nominal use represent the same concept.

When words were excluded or merged with another word, another word immediately lower on the wordlist went up until we got 130 words. The final set of base concepts extracted from the glosses of the Chinese Wordnet on the basis of the frequencies will be presented and discussed in the following section.

[14] In the glosses of the Chinese Wordnet, a typical context where *guding* occurs is as follows:

職業	婦女：	有	固定	工作	的	女子。
zhiye	funu	you	guding	dongzuo	de	nuzi
career	woman	have	stable	job	DE	female

career woman: a female who has a stable job

In this example, *guding* is used to modify *gongzuo* 'job'. We decided to exclude *guding* because it functions neither as a typical noun nor as a typical verb, but typically functions as a modifier in the glosses of the Chinese Wordnet. Additionally, the tag automatically assigned to *guding* (i.e., Nv) is problematic.

4.2 Results

By extracting words that occur frequently in the glosses of the Chinese Wordnet, we obtained a candidate set of words representing base concepts in Chinese. We attempted to map each word extracted in the present study to a Base Concept in EuroWordNet, either concrete or abstract. Note that if a word has no equivalent in the set of Base Concepts in EuroWordNet, we simply translated the word into English. Moreover, those without an equivalent in the set of Base Concepts in EuroWordNet were classified on the basis of their semantic characteristics. Table 1 summarizes the results. Following Table 1, each category will be presented.

Table 1. The distribution of base concepts extracted in the present study

CATEGORY		#	%
match	abstract	34	26.2%
	concrete	24	18.5%
non-match	positions	7	5.4%
	people	6	4.6%
	organizations	6	4.6%
	measurement	5	3.8%
	other (abstract) nouns	28	21.5%
	other abstract verbs	20	15.4%
TOTAL		130	100.0%

- ***Abstract concepts mapped to the Base Concepts of EuroWordNet***

Word	Freq.	Type	Synset in EuroWordNet
事件 shijian	2837	abstract	{event 1}
有 you；具有 juyou；擁有 yongyou	1930	abstract	{have 12; have got 1; hold 19}
使 shi	1293	abstract	{cause 6; get 9; have 7; induce 2; make 12; stimulate 3}
爲 wei	1276	abstract	{be 4; have the quality of being 1}
單位 danwei	1178	abstract	{unit 6; unit of measurement 1}
狀態 zhuangtai	736	abstract	{situation 4; state of affairs 1}
時間 shijian	722	abstract	{time 1}
方式 fangshi	511	abstract	{method 2}
動作 dongzuo	442	abstract	{action 1}

活動 huodong	382	abstract	{activity 1}
關係 guanxi	359	abstract	{relation 1}
空間 kongjian	357	abstract	{space 1}
方向 fangxiang	331	abstract	{direction 7; way 8}
內容 neirong	317	abstract	{cognitive content 1; content 2; mental object 1}
改變 gaibian	316	abstract	{change 11}
結果 jieguo	314	abstract	{consequence 3; effect 4; outcome 2; result 3; upshot 1}
做 zuo	314	abstract	{act 12; do something 1; move 19; perform an action 1; take a step 2; take action 1; take measures 1; take steps 1)
知識 zhishi	291	abstract	{cognition 1; knowledge 1}
訊息 xunxi	277	abstract	{message 2; content 3; subject matter 1; substance 4}
發展 fazhan	258	abstract	{development 1}
特質 tezhi	219	abstract	{quality 1}
運動 yundong	219	abstract	{motion 5; movement 6}
情況 qingkuang	205	abstract	{situation 4; state of affairs 1}
形狀 xingzhuang	204	abstract	{form 1; shape 1}
能力 nengli	186	abstract	{ability 2; power 3}
給 gei	179	abstract	{furnish 1; provide 3; render 12; supply 6}
做出 zuochu	171	abstract	{act 12; do something 1; move 19; perform an action 1; take a step 2; take action 1; take measures 1; take steps 1}
態度 taidu	160	abstract	{attitude 3; mental attitude 1}
顏色 yanse	155	abstract	{color 2; coloring 2}
方法 fangfa	153	abstract	{method 2}
變化 bianhua	151	abstract	{alter 2; change 12; vary 1}
時段 shiduan	146	abstract	{amount of time 1; period 3; period of time 1; time period 1}
從事 congshi	143	abstract	{act 12; do something 1; move 19; perform an action 1; take a step 2; take action 1; take measures 1; take steps 1}
感覺 ganjue	139	abstract	{feeling 1}

- ***Concrete concepts mapped to the Base Concepts of EuroWordNet***

Word	Freq.	Type	Synset in EuroWordNet
物體 wuti	1382	concrete	{inanimate object 1; object 1; physical object 1}
人 ren	1353	concrete	{human 1; individual 1; mortal 1; person 1; someone 1; soul 1}
位置 weizhi	598	concrete	{location 1}
物品 wupin	521	concrete	{inanimate object 1; object 1; physical object 1}
動物 dongwu	518	concrete	{animal 1; animate being 1; beast 1; brute 1; creature 1; fauna 1}
建築物 jianzhuwu	511	concrete	{building 3; edifice 1}
身體 shenti	413	concrete	{body 3; organic structure 1; physical structure 1}
部份 bufen	369	concrete	{part 3; portion 2}
數量 shuliang	369	concrete	{amount 1; measure 1; quantity 1; quantum 1}
地方 difang	336	concrete	{place 13; spot 10; topographic point 1}
表面 biaomian	329	concrete	{surface 1}
地點 didian	315	concrete	{location 1}
團體 tuanti	256	concrete	{group 1; grouping 1}
植物 zhiwu	235	concrete	{flora 1; plant 1; plant life 1}
金錢 jingqian	232	concrete	{money 2}
文字 wunzi	226	concrete	{word 1}
食物 shiwu	213	concrete	{food 1; nutrient 1}
部位 bufen	206	concrete	{part 3; portion 2}
物質 wuzhi	197	concrete	{matter 1; substance 1}
作品 zuopin	170	concrete	{creation 3}
液體 yiti	158	concrete	{liquid 4}
區 qu	158	concrete	{part 9; region 2}
物 wu	153	concrete	{inanimate object 1; object 1; physical object 1}
裝置 zhuangzhi	146	concrete	{device 2}

- ***Positions***

Word	Freq.	Translation
中 zhong	1764	middle
上 shang	573	up
後 hou	277	back; behind
內 nei	277	inside
以上 yishang	243	above
下 xia	192	down
正面 zhengmian	155	front, facade

The seven words do not have an equivalent in the set of Base Concepts of EuroWordNet though their potential hypernyms such as *weizhi* and *difang* can be mapped to synsets such as {location 1} and {place 13; spot 10; topographic point 1}. We suggest that the seven concepts may be regarded as a set of basic locative concepts in Chinese. Generally, the set exhibits a degree of symmetry in the sense that some words (i.e., *shang* and *xia*; *zhengmian* and *hou*) form pairs.

It is noted that the word *yishang* is ambiguous. It can mean 'above' or 'more than', and the latter sense is not locative. However, since we assume that the 'more than' sense might metaphorically derive from the 'above' sense, *yishang* is assigned to the present category.

- ***People***

Word	Freq.	Translation
姓 xing	1025	name
他人 taren	685	others
自己 ziji	386	self
女子 nuzi	174	woman
對方 duifang	170	the other party
男子 nanzi	164	man

Though *ren* 'human' can be mapped to the synset {human 1; individual 1; mortal 1; person 1; someone 1; soul 1}, in the candidate set of base concepts in Chinese are still some other words that denote people. As in the set of locative words, this set also exhibits a degree of symmetry (i.e., the self/other distinction: *taren/duifang* and *ziji*; the gender distinction: *nanzi* and *nuzi*). Such distinctions appear to be basic, and that is captured in our experiment.

- ***Organizations***

Word	Freq.	Translation
機構 jigou	314	institute
國家 guojia	264	country
政府 zhengfu	261	government
組織 zuzhi	256	organization
大學 daxue	221	university
學校 xuexiao	140	school

Our method extracted more words denoting organizations and institutes than the EuroWordNet project. However, some words extracted in our experiment are not hierarchically high. For example, *daxue* is just a subcategory of the educational institute.

- ***Measurement***

Word	Freq.	Translation
一 yi	1264	one
計算 jisuan	747	calculate
兩 liang	541	two
個 ge	918	a measure word
種 zhong	404	kind, type

Measurement is an important dimension of semantic primitives. Wierzbicka (1996:44-47) has identified a few quantifiers as semantic primitives. Our experiment identified five words that are not included in the Base Concepts of EuroWordNet: *yi* and *liang* are quantifiers, and both are also identified in Wierzbicka (1996) (i.e., ONE and TWO); *ge* and *zhong* are common classifiers in Chinese; *jisuan* is a typical verb in the measurement domain.

We could further categorize the remaining 28 nouns that are not in the set of Base Concepts of EuroWordNet but were extracted in our design. However, that would be of no more significance than creating a miscellaneous category like this, for the remaining subcategories might contain as few as one or two members. For instance, we could create a category for perception, which is intuitively an important dimension. However, in the present study, a category for perception may include no more than *shengyin* and *wundu*.

- ***Other (abstract) nouns***

Word	Freq.	Translation
對象 duixiang	5322	object; target
事物 shiwu	797	event; object
範圍 fanwei	525	range
程度 chengdu	481	extent, degree
其他 qita	454	other
行爲 xingwei	393	behavior
者 zhe	334	someone; something
事 shi；事情 shiqing	330	thing; job; business
聲音 shengyin	329	sound; voice
工具 gongju	280	tool
條件 tiaojian	252	condition
不同 butong	233	difference
標準 biaozhun	228	standard
文化 wenhua	209	culture
功能 gongneng	184	function
目標 mubiao	177	goal
古代 gudai	177	ancient times
系統 xitong	170	system
參考點 cankaodian	169	reference point
目的 mudi	163	purpose
領域 lingyu	161	field, domain
西元 xiyuan	154	A.D.
情緒 qingxu	152	emotion
生物 shengwu	149	creature
心理 xinli	145	mentality
地位 diwei	143	status
溫度 wendu	140	temperature
過程 guocheng	138	process

Almost all of the members in this category are abstract concepts. The only exception is *shengwu*. Its literal translation would be “creature”, so *shengwu* can seemingly be mapped to the synset {animal 1; animate being 1; beast 1; brute 1; creature 1; fauna 1}. Actually, the two concepts are not the same. In English, *creature* refers to a living organism that can move voluntarily, as the gloss in WordNet states. On the other hand, *shengwu* in Chinese refers to any living organism, whether it can move voluntarily or not. Therefore, we decided not to map the two concepts together.

- ***Other (abstract) verbs***

Word	Freq.	Translation
進行 jinxing	940	proceed
用 yong；使用 shiyong	723	use
發生 fasheng	539	happen, occur
產生 chansheng	458	produce
位於 weiyu	333	be located
達到 dadao	311	achieve
製成 zhicheng	308	be made into
得到 dedao	294	get
感到 gandao	244	feel
影響 yingxiang	234	influence
移動 yidong	234	move
預期 yuqi	225	expect
接受 jieshou	200	accept
開始 kaishi	187	start
取得 qude	184	gain
超過 chaoguo	167	exceed
失去 shiqu	161	lose
發出 fachu	155	give off
作爲 zuowei	150	serve as
工作 gongzuo	147	work (v.)

For a similar reason as in the case of nouns, a miscellaneous category is also created for the remaining 20 verbs. Additionally, as in the case of nouns, all the verbs here represent abstract concepts.

5. Discussion

Generally, as Table 1 shows, 72 words (55.4%) extracted from the glosses of the Chinese Wordnet have no equivalent in the set of Base Concepts in EuroWordNet. This suggests that our gloss-based approach can yield a very different set of base concepts from the set in EuroWordNet.

On the one hand, the 58 words that were identified in our experiment and could be mapped to an equivalent in EuroWordNet may be considered to represent concepts at the core of the mental landscape. These concepts can be singled out by different approaches, and they are prominent not only in the languages in EuroWordNet but also in Chinese. Therefore, the concepts represented by the 58 words may be regarded as basic in the mind.

On the other hand, words that were identified in our experiment but could not be mapped to any equivalent in EuroWordNet also reflect a certain degree of basicness in the mind. Like the Base Concepts in EuroWordNet, most of them are abstract and represent concepts hierarchically higher than basic level categories (cf. Section 2). Additionally, many of them (e.g., *chengdu* 'extent', *fanwei* 'range') are like basic domains (cf. Section 2), exhibiting a high degree of inclusiveness. Nevertheless, our gloss-based approach did obtain a few words representing sister concepts that are hierarchically lower, such as *shang*/*xia* 'up/down' and *nanzi*/*nuzi* 'male/female'.

In effect, it is natural that base concept sets vary from approach to approach. The number of the concepts in the lexicon is considerably larger than the number of base concepts. Take the present study for example. There are 17,018 candidate words in our frequency wordlist, and we only identify 130 words as potential base concepts in Chinese. The potential base concepts scatter around the mental lexicon; when we take a different perspective, adopt a different method, and have a different focus, we are very likely to extract a completely different set of concepts. That is why a study like the present one is of great significance. To really touch on the basic core of the mental landscape, we need to try a wide variety of approaches. Concepts surviving in different approaches can be seen as basic in the mind. On the other hand, since the pool is always much larger than the target set, concepts identified only by a certain approach are still significant rather than random and can reflect a certain extent of basicness from a certain perspective.

The limitations of this study are as follows. First, the design of EuroWordNet and the Chinese Wordnet is a key concern in the present study. As Vossen *et al.* (1998) admit, the data of some local wordnets were not well-structured when the base concepts were selected from each of the local wordnets. Also, the coverage of the Chinese Wordnet may not be comprehensive enough, for the project starts with words with a mid frequency. When EuroWordNet and the Chinese Wordnet are further updated, the resulting sets of base concepts and their comparison may give a different picture accordingly. Second, the gloss language is an issue in a gloss-based study like the present one. As a matter of fact, many words in our frequency wordlist have a low frequency, and many of such words can be replaced by other words with a higher frequency (An, 2009:172-182). If that is done, there will be fewer words in our wordlist, and the frequencies of some words will become higher. Therefore, a different set of base concepts in Chinese could be yielded.

Intriguingly, our method identified 58 words that could be mapped to an equivalent in EuroWordNet. This number is exactly the same as that of Goddard's (2002:14) "atoms of meaning". Additionally, this number is not far from that of the *SuperSenses* in WordNet (i.e., 48). Though the contents of the sets vary from approach to approach and need further examination, there appears to exist a certain range regarding the number of base concepts in the lexicon.

Alternatively, in previous research, the most commonly used words are determined by word occurrence frequency, but frequency is heavily dependent on the corpus selected. If the corpus is not large enough, or not balanced, the result will not be accurate enough. Recent developments of *distributional models* in semantics have shown success in this aspect. For example, Zhang *et al.* (2004) propose a metric for the distribution of words in a corpus. This will be left for future research.

6. Conclusion

Identifying the basic words that represent the core concepts is a crucial issue in lexicography, psycholinguistics, and language pedagogy. Recent NLP applications as well as ontologies also recognize the urgent need for the methodology for extracting and measuring the core concepts. In this paper, we have illustrated how glosses in a wordnet can be used to extract base concepts and provide evidence for basic conceptual underpinnings.

There is scope for the research to be extended in the direction of empirically-grounded evaluation of the results. We are also interested in putting the analysis in the contexts of multilingual wordnets. These are left as items for our future studies.

Acknowledgments

We would like to express our gratitude to Chih-Yao Lee for his technical support in the data analysis. Our gratitude is also extended to the reviewers for their insightful suggestions. Our appreciation also goes out to Harvey Hsin-chang Ho for his meticulous reading of the entire manuscript and his suggestions for its improvement. However, any weaknesses that remain are our own.

References

An, H.-L. (2009). *Studies of Chinese Gloss Language: Theories and Applications*. Shanghai: Xuelin. (安華林。《漢語釋義元語言：理論與應用研究》。上海：學林出版社。)

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Atserias, J., Villarejo, L., Rigau, G., Salgado, J. G., & Unibertsitatea, E. H. (2003). Integrating and porting knowledge across languages. In *Proceeding of Recent Advances in Natural Language Processing*, 31-37.

Bernard, L. L. (1941). The definition of definition. *Social Forces*, 19, 500-510.

Cristea, D., Puscasu, G., Postolache, O., Galiotou, E., Grigoriadou, M., Charcharidou, A., Papakitsos, E., Selimis, S., Stamou, S., Krstev, C., Pavlovic-Lazetic, G., Obradovic, I., Vitas, D., Cetinoglu, O., Tufis, D., Pala, K., Pavelek, T., Smrz, P., Koeva, S., & Totkov, G. (2002). *Definition of the Local Base Concepts and Their Mapping with the ILI Records*. Deliverable D.4.1, WP4, BalkaNet, IST-2000-29388.

Cruse, A. (1977). The pragmatics of lexical specificity. *Journal of Linguistics*, 13, 153-164.

Cruse, A. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Cumming, G., Cropp, S., & Sussex, R. (1994). On-line lexical resources for language learners: Assessment of some approaches to word formation. *System*, 22, 369-377.

Dolch, E. W. (1936). A basic sight vocabulary. *The Elementary School Journal*, 36, 456-460.

Gates, A. I. (1926). *A Reading Vocabulary for the Primary Grades*. New York: Teachers College, Columbia University.

Goddard, C. (2002). The search for the shared semantic core of all languages. *Meaning and Universal Grammar: Theory and Empirical Findings* (Vol. I), ed. by C. Goddard & A. Wierzbicka, 5-40. Amsterdam: John Benjamins. 5-40.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43, 907-928.

Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

Huang, C.-R. (2008). *Principles of Distinguishing and Describing Word Senses in Chinese* (5th ed.). Taipei: Academia Sinica. (黃居仁。《意義與詞義》系列《中文詞彙意義的區辨與描述原則》第五版。臺北：中央研究院。)

Izquierdo, R., Suárez, A., & Rigau, G. (2008). Exploring the automatic selection of basic level concepts. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing*.

Jarrar, M. (2006). Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *Proceedings of the 15th International World Wide Web Conference*, 497-503.

Kulkarni, M., Kulkarni, I., Dangarikar, C., & Bhattacharyya, P. (2010). Gloss in sanskrit wordnet. *Sanskrit Computational Linguistics*, 6465, 190-197.

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography* (2^{nd} ed.). Cambridge: Cambridge University Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar* (Vol. I). Stanford: Stanford University Press.

Lee, D. Y. W. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics*, 29, 250-278.

Lew, R., & Dziemianko, A. (2006). A new type of folk-inspired definition in English monolingual learners' dictionaries and its usefulness for conveying syntactic information. *International Journal of Lexicography*, 19, 225-242.

Lin, S.-Y. (2010). *A Computational Study of the Basic Level Nouns in English*. Unpublished doctoral dissertation, National Taiwan Normal University.

McCarthy, M. J. (1999). What constitutes a basic vocabulary for spoken communication? *Studies in English Language and Literature*, 1, 233-249.

McCarthy, M. J., & O'Dell, F. (1999). *English Vocabulary in Use: Elementary*. Cambridge: Cambridge University Press.

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

Ogden, C. K. (1930). *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber.

Portner, P. H. (2005). *What Is Meaning? Fundamentals of Formal Semantics*. Malden: Blackwell.

Riemer, N. (2006). Reductive paraphrase and meaning: A critique of Wierzbickian semantics. *Linguistics and Philosophy*, 29, 347-379.

Riemer, N. (2010). *Introducing Semantics*. Cambridge: Cambridge University Press.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.

Ungerer, F., & Schmid, H. J. (2006). *An Introduction to Cognitive Linguistics*. New York: Longman.

Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., & Peters, W. (1998). *The EuroWordNet Base Concepts and Top-Ontology*. Deliverable D017D034D036 EuroWordNet LE2-4003.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Wheeler, H. E., & Howell, E. A. (1930). A first-grade vocabulary study. *The Elementary School Journal*, 31, 52-60

Wierzbicka, A. (1972). *Semantic Primitives*. (Translated by A. Wierzbicka & J. Besemeres) Frankfurt: Athenäum Verlag.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Zgusta, L. (1971). *Manual of Lexicography*. The Hague: Mouton.

Zhang, H., Huang, C., & Yu, S. (2004). Distributional consistency: A general method for defining a core lexicon. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1119-1222.

Appendix

The appendix provides the Base Concepts in EuroWordNet.

I. Concrete Synsets

amount 1

animal 1

apparel 1

artifact 1

furniture 1

asset 2

being 1

beverage 1

body 3

bound 2

building 3

causal agent 1

compound 4

chemical element 1

cloth 1

commodity 1

structure 4

consumer goods 1 (= commodity 1)

covering 4

creation 3

decoration 2

device 2

document 2

land 6

entity 1

extremity 3

plant 1

fluid 2

food 1

furnishings 2 (= furniture 1)

garment 1 (= apparel 1)

group 1

human 1

object 1

instrument 2

instrumentality 1 (= instrument 2)

language unit 1

line 21

line 26 (= line 21)

liquid 4 (= fluid 2)

location 1

material 5

substance 1

monetary system 1

mixture 5

money 2

natural object 1

opening 4

part 3

region 2

part 12 (= part 3)

passage 6

work 4 (= creation 3)

place 13 (= location 1)

point 12

possession 1

product 2

representation 3

surface 1

surface 4 (= surface 1)

symbol 2

way 4

word 1

worker 2

writing 4

writing communication 1 (= writing 4)

II. Abstract Synsets

ability 2

abstraction 1

act 1

act 12 (= act 1)

interact 1

action 1 (= act 1)

activity 1

aim 4

allow 6

change 12

period 3

attitude 3

attribute 1

attribute 2 (= attribute 1)

be 4

be 9

cause 6

cause 7 (= cause 6)

cease 3

think 4

change 1

change 11 (= change 1)

change size 1

move 4

move 5 (= move 4)

change of state 1

quality 4 (= attribute 1)

knowledge 1

cognitive content 1

color 2

communicate 1

communication 1 (= communicate 1)

concept 1

condition 5

result 3

consume 2

convey 1

course 7

cover 16

create 2

decrease 5

definite quantity 1

development 1

direction 7

disorder 1

distance 1

utter 3

event 1

express 6 (= utter 3)

experience 7

express 5 (= utter 3)

feeling 1

form 1

form 6 (= form 1)

provide 3

take 17

give 16 (= provide 3)

move 15 (= move 4)

happening 1

have 12

idea 2

improvement 1

increase 7

information 1

kill 5

knowhow 1

travel 1

magnitude relation 1

message 2

method 2

movement 6

need 5

need 6 (= need 5)

path 3 (= course 7)

phenomenon 1

production 1

property 2 (= attribute 1)

psychological feature 1

quality 1 (= attribute 1)

ratio 1

relation 1

relationship 1 (= relation 1)

relationship 3 (= relation 1)

remember 2

remove 2

represent 3

say 8

sign 3

situation 4 (= condition 5)

social relation 1

space 1

spacing 1 (= space 1)

spatiality 1 (= space 1)

state 1 (= condition 5)

structure 4

time 1

unit 6

visual property