International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing
This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

Special Issue on "Chinese Lexical Resources: Theories and Applications"
Guest Editor: Shu-Kai Hsieh

易繫辭曰上古結繩而治後世聖人易之以書契百官以治萬民以察說文敘曰蓋文字者經藝之本宣教明化之始前人所以垂後後人所以識古故曰本立而道生知天下之至賾而不可亂也教化既萌文心雕龍則謂人之立言因字而生句積句而成章積章而成篇篇之彪炳

Vol.18 No.2 June 2013 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 18, no. 2, June 2013

Contents

Special Issue Articles:

Chinese Lexical Resources: Theories and Applications

Papers

Foreword

The role of lexical resources in the field of NLP has been well recognized in recent years, great advances have been achieved in developing tools and databases, as well as techniques for the automatic acquisition, alignment and enrichment for lexical resources. However, comparing to the major European languages, the lack of available comprehensive lexical resources in Chinese, and the resulting under determination of lexical representation theory by empirical lexical data, have posed crucial theoretical issues and exacerbated many difficulties in Chinese processing application tasks.

The aim of this special issue is to solicit research papers addressing aforementioned issues. It is pleasing to note that we have gathered together a diverse range of papers in this issue, reflected in the titles of the papers. The first paper "Assessing Chinese Readability using Term Frequency and Lexical Chain" investigates the automatic assessment of Chinese readability by extracting information from E-HowNet lexical database. The second paper "Cross-Strait Lexical Differences: A Comparative Study based on Chinese Gigaword Corpus" conducts a contrastive study on Chinese Concept Dictionary (CCD) and Chinese Wordnet (CWN), with their lexical usage based on a large comparative corpus. The third paper "A Definition-based Shared-concept Extraction within Groups of Chinese Synonyms: A Study Utilizing the Extended Chinese Synonym Forest" proposes a multi-layered gloss association method to synonyms extraction by applying it to the CiLin Thesaurus. The last paper "Back to the Basic: Exploring Base Concepts from the Wordnet Glosses" conducts an empirical investigation of the glosses of the Chinese Wordnet as a resource for the task of base concepts identification.

I would like to thank all of the authors whose work features in this special issue, and all the reviewers for their valuable contributions.

Shu-Kai Hsieh

Guest Editor

Graduate Institute of Linguistics, National Taiwan University

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 1-18

Assessing Chinese Readability using Term Frequency and Lexical Chain

Yu-Ta Chen*, Yaw-Huei Chen*, and Yu-Chih Cheng*

Abstract

This paper investigates the appropriateness of using lexical cohesion analysis to assess Chinese readability. In addition to term frequency features, we derive features from the result of lexical chaining to capture the lexical cohesive information, where E-HowNet lexical database is used to compute semantic similarity between nouns with high word frequency. Classification models for assessing readability of Chinese text are learned from the features using support vector machines. We select articles from textbooks of elementary schools to train and test the classification models. The experiments compare the prediction results of different sets of features.

Keywords: Readability, Chinese Text, Lexical Chain, TF-IDF, SVM.

1. Introduction

Readability of an article indicates its level in terms of reading comprehension of children in general. Readability assessment is a process that measures the reading level of a piece of text, which can help in finding reading materials suitable for children. Automatic readability assessment can significantly facilitate this process. There are other applications of automatic readability assessment such as the support of building a web search engine that can distinguish the reading levels of web pages (Eickhoff, Serdyukov, & de Vries, 2010; Miltsakaki & Troutt, 2008) and the incorporation into a text simplification system (Aluisio, Specia, Gasperin, & Scarton, 2010). Traditional measures of text readability focus on vocabulary and syntactic aspects of text difficulty, but recent work tries to discover the connections between text readability and the semantic or discourse structure of texts (Feng, Elhadad, & Huenerfauth, 2009; Pitler & Nenkova, 2008).

Most of the existing work on automatic readability assessment is conducted for English

* Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, R.O.C.
E-mail: {s0960413, ychen, s0990413}@mail.ncyu.edu.tw

text. In contrast, research on readability assessment for Chinese text is still in its initial stage. This paper investigates the appropriateness of using lexical cohesion analysis to improve the performance of Chinese readability assessment. More specifically, we build lexical chains, which are sequences of semantically related terms, in an article to represent the lexical cohesive structure of texts, and then derive features from the result of lexical chaining to capture the lexical cohesive information. Consisting of term frequency features and lexical chain features, various combinations of features are evaluated for generating prediction models on Chinese readability using support vector machines (SVMs). The prediction models are trained and tested on articles selected from textbooks of elementary schools in Taiwan. The results are compared for different sets of features.

This paper is organized as follows. Section 2 introduces related work in readability assessment and lexical cohesion analysis. Section 3 discusses the research methodology of our analysis, including problem definition, text processing, feature deriving, and prediction model building. Section 4 presents the experiments and the experimental results. Section 5 gives conclusions and directions for future work.

2. Related Work

This section briefly surveys existing work in the areas of readability assessment and lexical cohesion analysis.

2.1 Readability Assessment

Traditional readability formulae for English are based on shallow features such as average sentence length and average number of syllables per word to approximate syntactic and vocabulary difficulty in text (Kincaid, Fishburne Jr., Rogers, & Chissom, 1975; McLaughlin, 1969). However, this kind of measure makes strong assumptions about text difficulty and may not be always reliable.

With the growth of computational power, researchers began to have the ability to use word frequency as a better measure of word difficulty (Chall & Dale, 1995; Stenner, 1996). Word frequency information can be used in two ways. One is to maintain lists of common and rare words and to use the percentage of words in the article that are present or absent in the lists as features to measure the reading difficulty of that article (Chall & Dale, 1995; Lin, Su, Lai, Yang, & Hsieh, 2009; Schwarm & Ostendorf, 2005). The other is to compute the numbers of occurrences of words from a corpus and to use the computed word frequencies as features to measure the reading difficulty (Stenner, 1996). The effects of both methods rely on careful choice of corpus used to generate the word lists and frequency information, however, the second method is more flexible in that it can be incorporated into other models such as the term frequency-inverse document frequency (TF-IDF) scheme.

Some researchers suggest that text readability can be measured by factors in semantic aspect in addition to vocabulary and syntactic ones. Aluisio *et al*. (2010) consider the ambiguity ratio of terms for each part-of-speech (POS) as a feature for assessing text readability in Portuguese. Feng, Jansche, Huenerfauth, & Elhadad (2010) use some features inspired by cognitive linguistics to measure text readability, such as the number of named entities and the distribution of lexical chains in an article.

Some Chinese-specific factors, such as radical familiarity, number of strokes, geometry or shape of characters, are also considered (Lau, 2006). However, it is unclear whether these character-level features can truly benefit the readability assessment on Chinese text. Recently, machine learning based approaches also have been proposed for accessing Chinese readability (Chen, Tsai, & Chen, 2011; Sung, Chang, Chen, Cha, Huang, Hu, & Hsu, 2011).

2.2 Lexical Cohesion Analysis

Two properties of texts are widely used to indicate the quality of a text, coherence and cohesion. According to Morris and Hirst (1991), coherence refers to the fact that there is sense in a text, while cohesion refers to the fact that elements in a text tend to hang together. The former is an implicit quality within the text, whereas the latter is an explicit quality that can be observed through the text itself. Observing the interaction between textual units in terms of these properties is a way of analyzing the discourse structure of texts (Stokes, 2004). Discourse structure of a text is sometimes subjective and may require knowledge from the real world in order to truly understand the text coherence. However, according to Hasan (1984), analyzing the degree of interaction between cohesive chains in a text can help the reader indirectly measure the coherence of a text. Such cohesion analysis is more objective and less computationally expensive.

Halliday and Hasan (1976) classify cohesion into five types: (1) conjunction, (2) reference, (3) lexical cohesion, (4) substitution, and (5) ellipsis. Among these types, lexical cohesion is the most useful one and is the easiest to identify automatically since it requires less implicit information behind the text to be discovered (Hasan, 1984). Lexical cohesion is defined as the cohesion that arises from semantic relationships between words (Morris & Hirst, 1991). Halliday and Hasan (1976) further define five types of lexical cohesive ties in text: (1) repetition, (2) repetition through synonymy, (3) word association through specialization/ generalization, (4) word association through part-whole relationships, and (5) word association through collocation. All of the semantic relationships mentioned above except for collocation can be obtained from lexicographic resources such as a thesaurus. The collocation information can be obtained by computing word co-occurrences from a corpus or be captured using an n-gram language model with $n > 1$.

Lexical chaining is a technique that is widely used as a method to represent lexical cohesive structure of a text (Stokes, 2004). A lexical chain is a sequence of semantically related words in a passage, where the semantic relatedness between words is determined by the above-mentioned lexical cohesive ties usually with the help of a lexicographic resource such as a thesaurus. Lexical chains have been used to support a wide range of natural language processing tasks including word sense disambiguation, text segmentation, text summarization, topic detection, and malapropism detection.

Different lexicographic resources capture different subset of the lexical cohesive ties in text. Morris and Hirst (1991) use Roget's thesaurus to find cohesive ties between words in order to build lexical chains. WordNet (Fellbaum, 1998) is an online lexical database and has predominant use in information retrieval and natural language processing tasks, including lexical chaining. The major relationship between words in WordNet is synonymy, and other types of relationships such as hypernymy and hyponymy are defined among synsets, sets of synonymous words, forming a semantic network of concepts.

HowNet is a lexical database for Chinese words developed by Dong (n.d.). The idea of HowNet is to use a finite set of primitives to express concepts or senses in the world. The whole set of primitives are defined in a hierarchical structure based on their hypernymy and hyponymy relationships. Each sense of a word is defined in a dictionary of HowNet using a subset of the primitives. HowNet so far has two major versions: the 2000 version and the 2002 version. The 2000 version defines a word sense by a flat set of primitives with some relational symbols that determine the relation between the primitive and the target word sense. On the other hand, the 2002 version of HowNet uses a nesting grammar to define a word sense. A definition consists of primitives and a framework. The framework organizes the primitives into a complete definition. Dai, Liu, Xia, & Wu (2008) propose a method to compute lexical semantic similarity between Chinese words using the 2002 version of HowNet. For traditional Chinese, E-HowNet (Extended HowNet) is a lexical semantic representation system developed by Academia Sinica in Taiwan (CKIP Group, 2009). It is similar to the 2002 version of HowNet with the following major differences: (1) Word senses (concepts) are defined by not only primitives but also any well-defined concepts and conceptual relations, (2) Content words, function words, and phrases are represented uniformly, and (3) The incorporation of functions as a new type of primitive. An example of word sense definition is shown in Figure 1. Due to the first major difference mentioned above, a word sense definition may contain another well-defined word sense, such as "大學" (university, college) in the example. A bottom level expansion of the definition can be obtained by expanding all well-defined concepts in the top level definition, as shown in Figure 2.

```
教授    N    {老師:
                  location={大學}}
```

Figure 1. Top level definition of a word sense in E-HowNet.

```
教授    N    {human|人:
                  domain={education|教育},
                  predication={teach|教:
                                    agent={~}
                                },
                  location={InstitutePlace|場所:
                                    domain={education|教育},
                                    telic={or({study|學習:
                                                    location={~}
                                              },
                                              {teach|教:
                                                    location={~}
                                              }
                                            )
                                          },
                                    qualification={HighRank|高等}
                            }
             }
```

Figure 2. Bottom level expansion of the definition of a word sense in E-HowNet.

It has been suggested that coherent texts are easier to read (Feng *et al.*, 2010), and some previous studies have used lexical-chain-based features to assist in readability assessment of English text (Feng *et al.*, 2009; Feng *et al.*, 2010). Some other ways of modeling text coherence are also used for readability assessment, such as the entity-grid representation of discourse structure and coreference chains (Barzilay & Lapata, 2008; Feng *et al.*, 2009; Pitler & Nenkova, 2008). However, none of these discourse-based factors are tested on Chinese text for estimating readability. In this paper, we evaluate a combination of term frequency features and lexical chain features for generating classification models on Chinese readability.

3. Assessing Readability using SVM

This section presents the methodology adopted for assessing readability of Chinese text using SVM. We first explain the problem of readability assessment, basic concepts of SVM classification, and the system design. Then we describe how we conduct the text processing step, followed by the features we use for representing each article in the corpus. Finally, we discuss the performance measures used in the experiments.

3.1 Problem Definition

Various types of prediction models have been tested on the task of readability assessment in previous research (Aluisio *et al.*, 2010; Heilman, Collins-Thompson, & Eskenazi, 2008), including classification and regression models. Since several studies obtain better results when using SVM classification than regression models (Feng *et al.*, 2010; Petersen & Ostendorf, 2009; Schwarm & Ostendorf, 2005), in this paper we treat the problem of Chinese readability assessment as a classification task where SVM is used to build classifiers that predict the reading levels of given texts.

Readability can be classified according to grade levels, but the difference between adjacent grades may be insignificant, which makes the classification result less accurate. More importantly, grade-level readability is too fine for many applications and a broader range of readability level is more practical. For example, the U.S. government surveyed over 26,000 individuals aged 16 and older and reported data with only five levels of literacy skills (National Center for Education Statistics, 2002). Therefore, we divide reading skills of elementary school students into three levels: lower grade, middle grade, and higher grade, where lower grade corresponds to the first and second grade levels, middle grade corresponds to the third and fourth grade levels, and higher grade corresponds to the fifth and sixth grade levels.

In this paper, we try to evaluate different combinations of features for predicting the reading level of a text written in traditional Chinese as suitable for lower grade or middle grade. We will build one prediction model for lower grade level and another prediction model for middle grade level. These binary SVM classifiers can be combined to solve the multiclass problem of predicting the reading level of an article (Duan & Keerthi, 2005; Hsu & Lin, 2002).

While most studies on readability assessment view the reading levels as discrete classes, we think readability is continuous. That is, an article that is suitable for students of a certain level must also be comprehensible for students of higher levels. Similarly, if a student can understand an article of a certain reading level, he/she must also be able to understand any article of a lower reading level. Therefore, when building classifiers for lower grade, we use articles of grades 1 and 2 as positive data, while the others are negative data. When building classifiers for middle grade, articles of grade 1 through grade 4 are used altogether as positive data, while those of higher grade levels are used as negative data.

3.2 Text Processing

After the data set is collected, each article is undergone a word segmentation process as a pre-processing step before deriving features from the texts. Word segmentation is done using a word segmentation system provided by Academia Sinica (CKIP Group; n.d.). The segmentation result is stored in XML format, where POS-tags are attached to all words and sentence boundaries are marked.

It is reported by Yang and Petersen (1997) that chi-square test (χ^2) performs better than other feature selection methods such as mutual information and information gain in automatic text classification. Therefore, we use chi-square test to evaluate the importance of terms in the corpus with respect to their discriminative power among reading levels. The chi-square test is used to test the independence of two events, which, in feature selection, are the occurrence of the term and the occurrence of the class. Higher chi-square test value indicates higher discriminative power of the term to the classes. For each prediction model, we compute chi-square test value for each term in the corpus. Such information will benefit our feature derivation process described below. We do not perform stop word removal and stemming because Collins-Thompson and Callan (2005) report that these processes may harm the performance of classifier on lower grade levels.

3.3 Feature Deriving

The use of term frequencies as the primary information for assessing Chinese readability has been investigated (Chen, Tsai, & Chen, 2011), where TF-IDF values of the terms with high discriminative power are used as features for SVM classification. This paper investigates the appropriateness of using lexical cohesion analysis to improve the performance of Chinese readability assessment. Therefore, we build lexical chains for both the training and testing documents and deriving features from the lexical chains to capture the lexical cohesive aspect of the texts.

A general algorithm for generating lexical chains is shown in Figure 3, which is a simplified version of that proposed by Morris and Hirst (1991) as described in (Stokes, 2004). The chaining constraints in the algorithm are highly customizable and are the key to the quality of the generated lexical chains. The allowable word distance constraint is based on the assumption that relationships between words are best disambiguated with respect to the words that lie nearest to each other in the text. The semantic similarity is the most important factor that determines term relatedness and is generally based on any subset of the lexical cohesive ties mentioned above. Figure 4 shows an example of the lexical chaining result.

```
Choose a set of highly informative terms for chaining, $t_1$, $t_2$, …, $t_n$.
The first candidate term in the text, $t_1$, becomes the head of the first chain, $c_1$.
For each remaining term $t_i$ do
    For each chain $c_m$ do
        If the chain is most strongly related to $t_i$ with respect to allowable word
            distance and semantic similarity
        Then $t_i$ becomes a member of $c_m$,
        Else $t_i$ becomes the head of a new chain.
    End for
End for
```

Figure 3. A general lexical chaining algorithm.

Original Text:

在都會區房價飆高之時，銀行業整體壞帳率創下歷史新低，業界人士對此相當擔心，房價看來將持續疲軟到明年第1季，近2年承作的房貸物件將無上漲空間，尤其是泛公股行庫的整批房貸，多數是在「升緩、跌快」的市郊區，亦是銀拍屋的集中地，若明年房貸呆帳湧現，恐成爲銀行業系統性風險爆發的最大來源。

爲避免壞帳爆增牽連銀行的獲利，國銀和消金外銀間正默默建立共識，絕對不能以「衝業務」的理由，在房貸市場推銷低利產品，不單純是業者配合央行的特別監管，主要是台灣金融業再也經不起龐大虧損。

銀行業的「壞年」定義，意指容易發生貸款壞帳的條件氛圍，最常見的狀況即現階段的房價漲、投資氣氛濃；其相反即是「好年」，如全球金融海嘯期間或SARS期間，雖然房市冷、價格縮，有能力消費或擔保融貸的消費者卻都是「百中選一」的信用良好者，從銀行取得房貸的標的，對成數不奢望，還款時間卻往往超前計畫的50％以上，銀行鮮少因此發生壞帳。

根據金管會統計，目前本國銀行的壞帳持續改善，整體銀行平均逾放比在歷史低點的0.96％，完全擺脫多年前動輒4個百分點的可怕記錄。銀行業者認爲，融貸逾放的來源有2大項目，信用卡和房地產，前者經常維持在2-3％之間，後者則因貸出利率僅2％，銀行能夠獲利的空間很小，長期以來平均逾放率在1％，實在經不起任何房價超跌的折損衝擊。

房貸圈目前存有一種默認的共識，泛公股行庫的三商銀和民營銀行的中信銀，不該帶頭促銷房貸產品，而消金外銀則繼續強化個人徵信，從風險管控著手，多管道降低「壞年」留下的逾放壓力。

Derived Lexical Chains:

lexical chain 1: (1)銀行業-3 (2)銀行業-25 (3)業務-34 (4)銀行業-45 (5)金融-59

lexical chain 2: (1)整體-4 (2)物件-14 (3)期間-61 (4)期間-62 (5)時間-74 (6)目前-79 (7)整體-82 (8)前者-94 (9)後者-95 (10)目前-104

lexical chain 3: (1)成數-73 (2)百分點-86 (3)利率-96

lexical chain 4: (1)系統性-26 (2)來源-28 (3)條件-50 (4)氛圍-51 (5)狀況-52 (6)氣氛-56 (7)價格-64 (8)能力-65 (9)信用-68 (10)標的-72 (11)來源-90 (12)壓力-118

Figure 4. An example of lexical chaining result.

The algorithm is adopted in this paper for the construction of lexical chains. We select nouns in the balanced corpus created by Academia Sinica (CKIP Group, 2010) with word frequency higher than a given threshold as candidate terms for lexical chaining. We apply the method proposed by Dai *et al*. (2008) to compute semantic similarity between words using E-HowNet instead of HowNet as the lexical database. The difference is that the primitives of function type are treated as descriptors. Let *P* and *Q* be two word senses and the number of modifying primitives of *P* is less than that of *Q*. The semantic similarity between *P* and *Q* is computed by Equation 1,

$$\begin{aligned} Sim(P,Q) &= \alpha \times Sim(P', Q') \\ &+ \beta \times \frac{\sum_{0 \le i < |P|} \max_{0 \le j < |Q|} (Sim(P_i, Q_j))}{|P|} \\ &+ \gamma \times \frac{|S \cap T|}{|S| + |T|} \end{aligned} \tag{1}$$

where P' and Q' are the primary primitives of P and Q, respectively, $|P|$ and $|Q|$ are the numbers of modifying primitives in their respective word senses, S and T are the sets of descriptors of frameworks of P and Q, respectively, $|S \cap T|$ is the number of common descriptors of S and T, $|S|$ and $|T|$ are the numbers of descriptors in S and T, and α, β, and γ are the relative weights of the three parts.

After constructing lexical chains, we derive five features from the lexical chains for each article. The five features are the number of lexical chains, the average length of lexical chains, the average span of lexical chains, the number of lexical chains with span longer than the half length of the article, and the average number of active chains per word. The features are normalized by dividing the article length. Table 1 shows the lexical chain features and their representing codes used in this paper.

Table 1. List of lexical chain features.

Code	Feature
lc-1	Number of lexical chains
lc-2	Average length of lexical chains
lc-3	Average span of lexical chains
lc-4	Number of long lexical chains
lc-5	Average number of active chains per word

3.4 SVM Classification

We apply support vector machines (SVM) as the modeling technique for our classification problem. The goal of an SVM, which is a vector-space-based large margin classifier, is to find a decision surface that is maximally far away from any data point in the two classes. When data in the input space (X) cannot be linearly separated, we transform the data into a high-dimensional space called the feature space (*F*) using a function $\phi: X \rightarrow F$ so that the data are now linearly separable. Then in the feature space we find a linear decision function that best separates the data into two classes. An SVM toolkit, LIBSVM (Chang & Lin, n.d.), is used for building prediction models. When training the prediction model for each reading level, texts belonging to that reading level are used as positive data, while the rest of the texts are used as negative data. We follow the procedure suggested by Hsu, Chang, & Lin (2010) including the use of radial basis function kernel, scaling, and cross-validation.

3.5 Evaluation

In this paper, we use precision, recall, F-measure, and accuracy to evaluate the learned prediction models. For the test data, we use the same procedure for text processing and feature deriving. Correct prediction refers to the agreement between the predicted reading level and the original reading level. We compute the following quantities: true positive (*TP*) is the number of articles correctly classified as positive, false negative (*FN*) is the number of positive articles incorrectly classified as negative, true negative (*TN*) stands for the number of articles correctly classified as negative, and false positive (*FP*) refers to the number of negative articles incorrectly classified as positive. Precision, recall, F-measure, and accuracy are defined as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Re}\,\text{call} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F} - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

We will test on different sets of features to find the best feature combination for training the prediction models.

4. Experiments

In this section we present our experiment setup and the results of the experiments on the textbooks corpus using different feature combinations.

4.1 Experiment Environment

The program modules for the experiments are written in Java programming language running on a PC with Microsoft Windows environment, Intel Core 2 Quad CPU, and 2GB of RAM. The corpus used as empirical data is stored in a Microsoft Access database. The lexicographic resources used for lexical semantic similarity computation in the experiments are stored as pure-text files in CSV format. LIBSVM is used for learning and testing SVM prediction models.

4.2 Empirical Data

The corpus used as empirical data consists of articles selected from the textbooks of elementary schools in Taiwan. We collect the digital versions of the textbooks of three subjects, Mandarin, Social Studies, and Life Science, for all of the six grade levels from publishers Nan I and Han Lin, resulting in a total number of 740 articles. Table 2 shows details of the collected data set.

Table 2. Summary of the textbooks corpus.

Reading Level	**Grade Level**	**Mandarin**	**Social Studies**	**Life Science**	**No. of Articles**
lower	1st grade	42	0	73	115
	2nd grade	56	0	55	111
middle	3rd grade	61	53	0	114
	4th grade	67	50	0	117
higher	5th grade	83	58	0	141
	6th grade	88	54	0	142
Total		397	215	128	740

4.3 Experiment Design

In each experiment, we use one set of features with a fixed parameter setting and target a certain grade level. We equally divide the corpus into five data sets to support 5-fold cross validation, and we present the average precision, recall, F-measure, and accuracy of the five folds.

Since the textbooks corpus does not contain articles beyond elementary school levels, we only build prediction models for lower grade and middle grade. For convenience, we denote feature sets by a string with special syntax. Feature types are indicated in the string by the abbreviation of that feature type. For example, “lc” refers to the lexical chain feature type and “tf” refers to the TF-IDF feature type. Options of a feature type are indicated in the string by a dash followed by the code name for that option, attached to the end of the feature type indicator.

4.4 Experiments on Lexical Chain Features

To test the capability of lexical chain features on Chinese readability assessment, the lexical chain features listed in Table 1 are used and the results are shown in Table 3 and Table 4.

Table 3. Result of classifier for lower grade using lexical chain only.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5	0.76	0.57	0.65	0.81

Table 4. Result of classifier for middle grade using lexical chain only.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5	0.70	0.83	0.76	0.68

4.5 Comparison with TF-IDF Features

It is interesting to see whether incorporating a small number of TF-IDF features into lexical chain features can produce the same or even better results. We first use TF-IDF features generated from top 50 to top 500 terms to produce classifiers for lower grade. The precision, recall, F-measure, and accuracy of the classifiers using different number of TF-IDF features are shown in Table 5. Then, we add the five lexical chain features to the TF-IDF feature sets and repeat the same experiments. Their precision, recall, F-measure, and accuracy values are shown in Table 6. Figure 5 illustrates line graphs generated from F-measure values of the two tables, from which we find that the overall performance is improved for lower grade classifiers when using a combination of TF-IDF features and lexical chain features.

Table 5. Result of classifier for lower grade using TF-IDF features only.

Feature set	Precision	Recall	F-measure	Accuracy
tf-top50	0.78	0.87	0.82	0.88
tf-top100	0.81	0.86	0.83	0.89
tf-top200	0.80	0.89	0.84	0.90
tf-top300	0.82	0.89	0.85	0.90
tf-top400	0.86	0.89	0.87	0.92
tf-top500	0.84	0.89	0.87	0.92

Table 6. Result of classifier for lower grade using lexical chain and TF-IDF.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5 + tf-top50	0.85	0.85	0.85	0.91
lc-1-2-3-4-5 + tf-top100	0.83	0.87	0.85	0.91
lc-1-2-3-4-5 + tf-top200	0.90	0.83	0.86	0.92
lc-1-2-3-4-5 + tf-top300	0.95	0.91	0.93	0.95
lc-1-2-3-4-5 + tf-top400	0.93	0.93	0.93	0.96
lc-1-2-3-4-5 + tf-top500	0.93	0.89	0.91	0.95

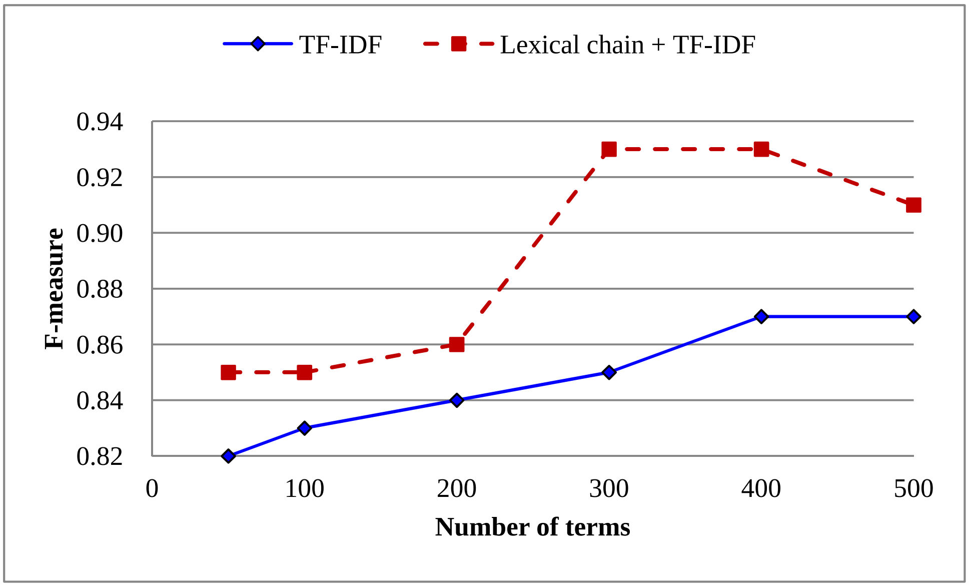


Figure 5. Result of classifier for lower grade.

The same set of experiments is conducted for the middle grade classifiers. Precision, recall, F-measure, and accuracy values of classifiers generated from TF-IDF features and the combination of TF-IDF and lexical chain features are shown in Table 7 and Table 8, respectively. The line graphs of F-measure values are shown in Figure 6, where the combined TF-IDF and lexical chain features generate the same or better F-measure in all cases. Therefore, incorporating a small number of TF-IDF features into lexical chain features is recommended for middle grade classifiers.

Table 7. Result of classifier for middle grade using TF-IDF features only.

Feature set	Precision	Recall	F-measure	Accuracy
tf-top50	0.81	0.88	0.84	0.79
tf-top100	0.81	0.90	0.85	0.81
tf-top200	0.83	0.92	0.87	0.83
tf-top300	0.86	0.90	0.88	0.84
tf-top400	0.82	0.92	0.87	0.83
tf-top500	0.82	0.95	0.88	0.84

Table 8. Result of classifier for middle grade using lexical chain and TF-IDF.

Feature set	Precision	Recall	F-measure	Accuracy
lc-1-2-3-4-5 + tf-top50	0.82	0.87	0.84	0.80
lc-1-2-3-4-5 + tf-top100	0.84	0.89	0.86	0.82
lc-1-2-3-4-5 + tf-top200	0.87	0.88	0.88	0.84
lc-1-2-3-4-5 + tf-top300	0.89	0.87	0.88	0.85
lc-1-2-3-4-5 + tf-top400	0.83	0.93	0.88	0.84
lc-1-2-3-4-5 + tf-top500	0.83	0.93	0.88	0.84

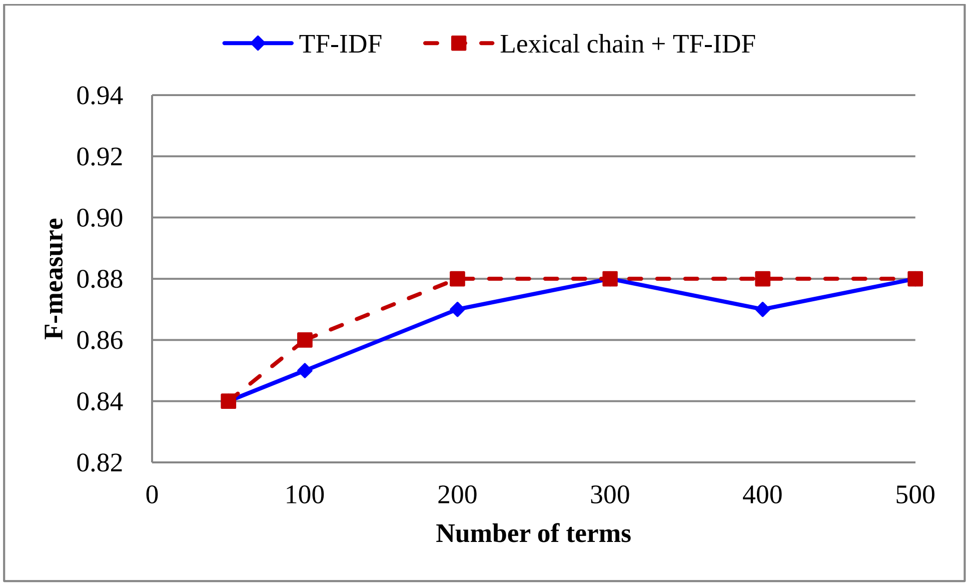


Figure 6. Result of classifier for middle grade.

5. Conclusions

This paper focuses on evaluating the effect of lexical cohesion analysis, more specifically, the effect of features based on lexical chains and term frequency, on the performance of readability assessment for Chinese text. The experiments produce satisfactory results on the textbooks corpus. Combining lexical chain and TF-IDF features usually produces better results, suggesting that both term frequency and lexical chain are useful features in Chinese readability assessment.

Future work can be done to have more articles annotated with reading levels or resort to other types of corpora where reading levels are inherent. On the other hand, lexical cohesion is only one of several aspects of text cohesion, and other aspects of text cohesion may also have some impact on the task of readability assessment. Several existing models of text cohesion, such as Coh-metrix and entity grid representation, try to model other aspect of text cohesion and have been extensively used in other natural language processing tasks such as writing quality assessment. Future work can be done to verify whether these models can benefit the task of readability assessment for Chinese text.

Acknowledgments

This research was supported in part by the National Science Council of Taiwan under Grant NSC99-2511-S-415-007-MY2.

Reference

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). Readability assessment for text simplification. In *NAACL-HLT 2010: The 5th Workshop on Innovative Use of NLP for Building Educational Applications*.

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1-34.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chang, C.-C., & Lin, C.-J. (n.d.). LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, Y.-H., Tsai, Y.-H., & Chen, Y.-T. (2011). Chinese readability assessment using tf-idf and svm. In *International Conference on Machine Learning and Cybernetics (ICMLC2011)*, Guilin, China.

CKIP Group. (n.d.). A Chinese word segmentation system, http://ckipsvr.iis.sinica.edu.tw/

CKIP Group. (2009). *Lexical semantic representation and semantic composition - An introduction to E-HowNet*. (Technical Report), Institute of Information Science, Academia Sinica.

CKIP Group. (2010). *Academia Sinica Balanced Corpus (Version 3.1)*. Institute of Information Science, Academia Sinica.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology (JASIST)*, *56*(13), 1448-1462.

Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring semantic similarity between words using HowNet. In *International Conference on Computer Science and Information Technology 2008*, 601-605.

Dong, Z. (n.d.). HowNet knowledge database. http://www.keenage.com/

Duan, K.-B., & Keerthi, S. S. (2005). Which is the best multiclass SVM method? An empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*.

Eickhoff, C., Serdyukov, P., & de Vries, A. P. (2010). Web page classification on child suitability. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database and some of its applications*. Cambridge, MA: MIT Press.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 229-237.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *The 23rd International Conference on Computational Linguistics (COLING 2010): Poster Volume*, 276-284.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 184-219), Newark, DE: International Reading Association.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 71-79.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). *A practical guide to support vector classification*. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Hsu, C.-W. & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415-425.

Kincaid, J. P., Fishburne R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel*. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.

Lau, T. P. (2006). *Chinese readability analysis and its applications on the Internet*. (Master Thesis), Computer Science and Engineering, The Chinese University of Hong Kong.

Lin, S.-Y., Su, C.-C., Lai, Y.-D., Yang, L.-C., & Hsieh, S.-K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet. *Computational Linguistics and Chinese Language Processing*, *14*(1), 45-84.

McLaughlin, G. H. (1969). SMOG grading - A new readability formula. *Journal of Reading*, *12*(8), 639-646.

Miltsakaki, E., & Troutt, A. (2008). Real time Web text classification and analysis of reading difficulty. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21-48.

National Center for Education Statistics. (2002). *Adult literacy in America* (3rd ed.). Washington, D. C.: U.S. Dept. of Education.

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, *23*, 89-106.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 186-195.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the ACL*, 523-530.

Stenner, A. J. (1996). Measuring reading comprehension with the Lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

Stokes, N. (2004). *Applications of lexical cohesion analysis in the topic detection and tracking domain*. (Ph.D. Thesis), Department of Computer Science, National University of Ireland, Dublin.

Sung, Y.-T., Chang, T.H., Chen, J.-L., Cha, J.-H., Huang, C.-H., Hu, M.-K., & Hsu, F.-Y. (2011). The construction of Chinese readability index explorer and the analysis of text readability. In *21th Annual Meeting of Society for Text and Discourse Process*, Poitiers, France.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 19-34

以中文十億詞語料庫爲基礎之兩岸詞彙對比研究

Cross-Strait Lexical Differences: A Comparative Study based on Chinese Gigaword Corpus

洪嘉馡[*]、黃居仁[+]

Jia-Fei Hong and Chu-Ren Huang

摘要

近幾年來，由於兩岸交流頻繁，兩岸使用的詞彙，也因此互相影響甚重，語言學界對於漢語詞彙的研究，不論在語音、語義或語用上的探討，發現兩岸對使用相同漢語時的詞彙差異有著微妙性的區別。而兩岸卻又的確是使用漢字體系的書寫系統，只有字形上有可預測的規律性對應。本文在以兩岸皆使用中文文字的原則上，在繁體中文與簡體中文的使用狀況來比對兩岸使用詞彙的特性與現象，以探究與語義對應與演變等相關的議題。

首先，在 Hong 和 Huang (2006) 的對應上，藉以英文 WordNet 爲比對標準，藉由比較北京大學的中文概念辭典(Chinese Concept Dictionary (CCD))與中央研究院語言所的中文詞網(Chinese Wordnet (CWN))兩個 WordNet 中文版所使用的詞彙，探討兩岸對於相同概念詞彙的使用狀況。本文進一步使用中文概念辭典與中文詞網所使用的詞彙，在 Gigaword Corpus 中繁體語料與簡體語料的相對使用率，探究兩岸對於使用相同詞彙，或使用不同詞彙的現象與分佈情形，並以 Google 網頁中所搜尋到的繁體資料與簡體資料進行比對、驗證。

關鍵詞：CCD, CWN, WordNet, Gigaword Corpus, Google, 兩岸詞彙, 詞義, 概念

[*] 國立臺灣師範大學 National Taiwan Normal University
E-mail: jiafeihong@gmail.com
[+] 香港理工大學 The Hong Kong Polytechnic University
E-mail: churenhuang@gmail.com

Abstract

Studies of cross-strait lexical differences in the use of Mandarin Chinese reveal that a divergence has become increasingly evident. This divergence is apparent in phonological, semantic, and pragmatic analyses and has become an obstacle to knowledge-sharing and information exchange. Given the wide range of divergences, it seems that Chinese character forms offer the most reliable regular mapping between cross-strait usage contrasts. In this study, we take general cross-strait lexical wordforms to discovery of cross-strait lexical differences and explore their contrasts and variations.

Based on Hong and Huang (2006), we discuss the same conceptual words between cross-strait usages by WordNet, Chinese Concept Dictionary (CCD) and Chinese Wordnet (CWN). In this study, we take all words which appear in CCD and CWN to check their lexical contrasts of traditional Chinese character data and simplified Chinese character data in Gigaword Corpus, explore their appearances and distributions, and compare and demonstrate them via Google website.

Keywords: CCD, CWN, WordNet, Gigaword Corpus, Google, Cross-Strait Lexical Wordforms, Semantics, Concepts

1. 緒論

兩岸使用詞彙的差異問題，在目前兩岸人民的各種交流中，早就已經呈現出許多無法溝通、理解困難，或是張冠李戴，表達不合宜的錯誤窘境。探討兩岸使用詞彙的差異性時，不僅讓大量使用詞彙的記者們，感受到兩岸的差異 (如：華夏經緯網，2004；南京語言文字網，2004；廈門日報，2004)，甚至，近年來未了因應對岸人民來台旅行，台灣交通部觀光局也整理了「兩岸地區常用詞彙對照表」(中華民國交通部觀光局，2011)，這些皆成爲漢語詞彙學與詞彙語義學上研究的重要課題(如：王鐵昆與李行健，1996；姚榮松，1997)。

以往對於這個議題的研究，不論語言學學者或文字工作者注意到這個問題時，僅能就所觀察到特定詞彙的局部對應，來提出分析與解釋而缺乏全面系統性的研究。本文的研究方法，第一是延續 Hong 和 Huang (2006)、洪 等人(洪嘉馡與黃居仁，2008)的研究方法，先以 WordNet 做詞義概念的判準，比對中文概念辭典與中文詞網裡，概念相同、語義相同的詞彙使用狀況；第二、是以有大量兩岸對比語料的 Gigaword Corpus 作爲實證研究的基礎，驗證中文概念辭典與中文詞網對於相同概念語義的詞彙，使用上，確實有其差異性的存在。這是一個以實際語料、實際數據進行比對，且具有完整性、全面性、概括性的研究。

又 Miller 等人 (Miller, Beckwith, Fellbaum, Gross & Miller, 1993)認爲他們可透過使用同義詞集來表現詞彙概念和描述詞彙的語義內容，所以他們建立了 WordNet，近年來，

也有不少研究團隊在處理以 WordNet 爲出發點的不同語言翻譯。

值得一提的是，同屬於漢語詞彙系統的繁體中文系統與簡體中文系統，在中央研究院語言所與北京大學計算語言所的研究團隊裡，也針對此議題，做了不少相關的研究，因此，本文想要探討的是，相同概念的漢語詞彙語義，在繁體中文與簡體中文的使用狀況。

另外，對於繁體中文系統與簡體中文系統的對應，我們以 WordNet 當作研究語料的基礎，是爲了可以建立一套符合詞彙知識原則並能運用於英中對譯的系統，如此一來，即可比對出兩個中文系統，在語言使用上的差異性。

2. 研究動機與目的

自兩岸交流日趨頻繁之後，本屬於同文同種的漢語系統，確有不少知識與信息交流的障礙，造成這樣的原因，莫過於兩岸詞彙使用的差異。相同的詞形，卻代表不同的詞義；或相同的語義，卻有兩種不同的表達詞彙。這種問題，已經讓許多文字工作者費盡心思，試圖來解決這樣的窘境；而語言學者對於這種現象，也試圖從語音、語義、語用等方面著手，希望從各種與語言相關的角度，來探究兩岸詞彙的差異。

在研究議題上，光是觀察到兩岸選擇以不同的詞形來代表相同的語義，如下述例子(1)、(2)，這樣是不夠的。以 Gigaword Corpus 的語料呈現台灣/ 大陸使用的狀況及其在 Gigaword Corpus 中的詞頻，如下：

(1) 台灣的「煞 (155/ 65)」、大陸的「非典 (354/ 33504)」
(「Sars (SEVERE ACUTE RESPIRATORY SYNDROME)」「嚴重急性呼吸道綜合症」的翻譯)

(2) 台灣的「計程車 (22670/ 68)」、大陸的「出租車 (422/ 5935)」

在詞彙語義學研究上，我們必須進一步追究，這些對比的動機，語言的詞彙與詞義演變的動力是否相關，對比有無系統性的解釋等。Chinese GigaWord Corpus 包含了來自兩岸的大量語料，其中，有約 5 億字新華社資料(XIN)、約 8 億字中央社資料(CNA)，可以看出台灣和大陸對於同一概念而使用不同詞彙的實際狀況與分佈。

如要追究動機與解釋等理論架構問題，當然不能只靠少數觀察到的例子，而必須建立在數量較大的語料庫上，以便做全面深入的分析。以上兩個對比爲例，其實在大陸與台灣的語料中，都有相當多的變例出現。

3. WordNet和中文詞網(CWN)

WordNet 是一個電子詞彙庫的資料庫，是重要語料來源的其中一個語料資料庫，WordNet 的設計靈感源自於近代心理語言學和人類詞彙記憶的計算理論，提供研究者在計算語言

學，文本分析和許許多多相關的研究(Miller *et al*., 1993; Fellaum, 1998)。在 WordNet 中，名詞、動詞、形容詞、副詞，這四個不同的詞類，分別設計、組合成同義詞集(synsets)的格式，呈現出最基本的詞彙概念，在這當中，以不同的語義關係連結各種不同的同義詞集，串成了 WordNet 的整個架構，也呈現了 WordNet 整個全貌。

自從 Miller 等人 (1993)、Fellaum (1998) 發展 WordNet 以來，WordNet 就持續不斷地更新版本，目前最新的版本是 WordNet 3.0 版，這些版本間的差異，包括了同義詞集的量和他們的詞彙定義。然而，對於拿 WordNet 來做研究語料的學者，多數還是以 WordNet 1.6 版爲最多，因爲這個版本是目前最多計算語言學學者使用的。在 WordNet 1.6 版裡，有將近 100,000 的同義詞集。

我們知道，雙語領域分類，可以增加我們各種領域詞彙庫的發展，同樣的，在上一段的內容，我們也提到關於以 WordNet 爲基礎，發展出繁體中文系統(Chinese Wordnet, CWN)與簡體中文系統(Chinese Concept Dictionary, CCD)的對譯，我們使用雙語詞網，作爲詞彙知識資料庫來實現、支持我們在詞彙概念上的研究。

在中英雙語詞網中，每一個英文的同義詞集，我們都會給予三個最適合且對等中文翻譯，而這些翻譯，如果不屬於真正的同義詞，我們也會標註他們的語義關係(Huang, Tseng, Tsai & Murphy, 2003)，又這些雙語詞網，也在中研院語言所詞網小組團隊的發展，將每一個同義詞集都與 SUMO 概念節點連結，進而開發出 Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) (Huang, Chang & Li, 2010)。當我們無法直接取得中英相對應的詞彙，我們在雙語詞網的資料庫裡，可以利用這些語義關係，進而發展並預測領域分類。

4. WordNet和中文概念辭典(CCD)

CCD，中文概念辭典(Chinese Concept Dictionary)，是一個中英雙語的詞網，整個架構發展也是來自於 WordNet (于江生與俞士汶，2004；于江生、劉揚與俞士汶，2003；劉揚、俞士汶與于江生，2003)。在 CCD 的發展手冊裡記載，研究團隊描述這些詞義的首要條件，是不可以破壞原本 WordNet 對於同義詞集定義概念與其語義關係的架構。另一方面，CCD 的研究團隊考量到可以存在許多在中文與英文的不同描述架構，所以，他們不止表現對中文詞彙內涵的表達，也發展了中文詞彙語義與概念的關係性，以利於強調中文的特質。

CCD 的研究團隊專注在整個 CCD 的架構，提出同一概念的同義詞集的定義，其所呈現的概念、定義和概念網的上下位語義關係，每一個同義詞集都有其基本關係，彼此之間亦有語義關係的存在。至於 CCD 的邏輯推演原則在語義網上的呈現，是運用到數學的形式而來的，是可以幫助研究者在中文語義分析上的使用。

自從 2000/09 開始，北京大學計算語言學研究所就已經開始著手以 WordNet 爲基準，研究 CCD，並建立一個中英雙語的詞網，一個可以提供各種不同研究的詞網，如機器翻譯(MT)，訊息擷取(IE)…等等。

基於 WordNet 英文概念與 CCD 中文概念是屬於兩個不同知識背景，也因此 CCD 中，他們兩者間的相互關係與概念，是非常複雜、繁瑣的。CCD 包括了大量且繁雜的成對、成組的小網絡，大致上，差不多有 10^5 的概念節點和 10^6 的成組小網絡的概念關係，他們的關係，呈現如下圖：

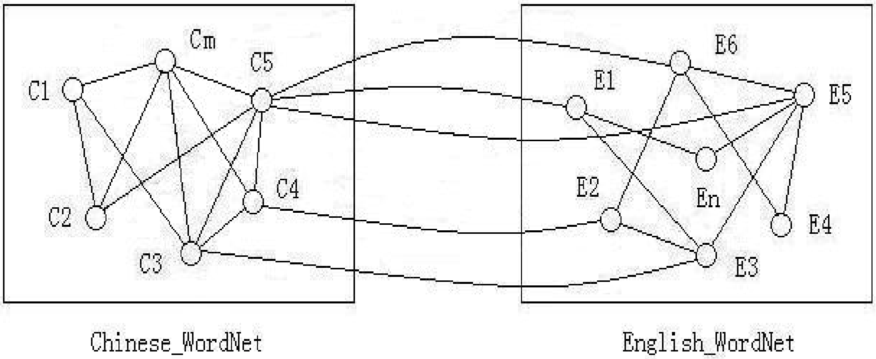


圖 1. WordNet 小網絡中複雜的關係結構

5. 文獻探討

對於兩岸詞彙對比的探討，過去的研究，多半著重在表面語言特徵的區別。如列舉語音方面、詞彙方面的對比(南京語言文字網，2004)；或以語音、詞彙、語法及表達方式等方面來分析語言差異的現象 (如：王鐵昆與李行健，1996；姚榮松，1997；許斐絢，1999；戴凱峰，1996)。

近年來，對於兩岸詞彙對比的研究，比較新的研究方法，是以 WordNet 爲基礎，取兩岸語料庫資料作比較，進而分析兩岸詞彙的對比(如：Hong & Huang, 2006)；或以 Chinese Gigaword Corpus (2005)爲基礎，探索兩岸對於漢語詞彙在使用上的差異現象，例如：相關共現詞彙(collocation)的差異、台灣或大陸獨用的差異、特定語境下的特殊用法的差異、語言使用習慣的差異等等(如：洪嘉馡與黃居仁，2008)。

6. 研究方法

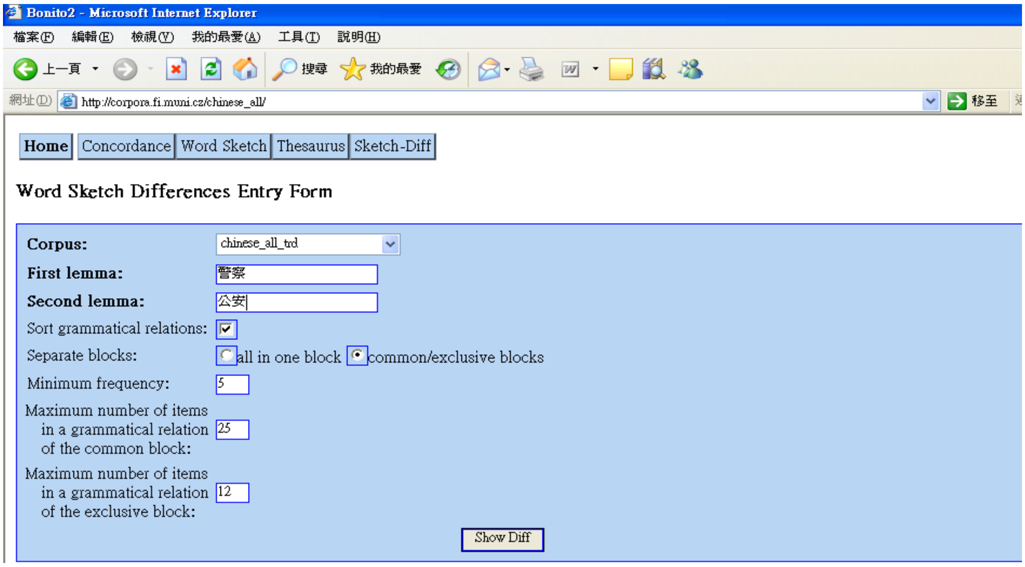
本研究以英文的 WordNet、繁體中文系統的中文詞網(CWN)、以及簡體中文系統的中文概念辭典(CCD)等三大資料庫爲主，對於繁體中文系統的英中對譯與簡體中文系統的英中對譯，我們先進行比對，試圖在比對中，尋找出兩者之間的差別與使用分佈。

相同的概念，本歸屬於一個同義詞集，但因兩岸在詞彙使用上的差異，而有所不同，儘管如此，仍舊有一些兩岸使用相同的詞彙來表達相同的概念語義。本文中將從繁體中文系統與簡體中文系統的英中對譯資料裡，集中探究同一個同義詞集，在兩岸使用的詞彙是完全相同、完全不同的狀況。然後，再將這些完全相同、完全不同的詞彙，以 Gigaword Corpus 爲基礎，分析這些詞彙在這個語料庫裡，所呈現出兩岸使用的狀況。

接著，本文再以語料庫爲研究出發點，是以約十四億字的 Chinese Gigaword Corpus 爲主要語料來源，以中文詞彙速描爲搜尋語料工具 Chinese Gigaword Corpus (2005)、Chinese Word Sketch Engine、Kilgarriff *et al.* (2005)。Chinese Gigaword Corpus 包含了分別來自大陸、臺灣、新加坡的大量語料，包括約 5 億字新華社資料(XIN)、約 8 億字中央社資料(CNA)，及約 3 千萬字新加坡聯合早報資料(Zaobao)。本研究，僅就大陸新華社資料與臺灣中央社資料進行比對，因此，本文研究可以提供兩岸詞彙差異的大量詞彙證據。

最後，本文亦視 Google 爲一個擁有大量繁體中文、簡體中文的語料庫，試圖根據 Google 所搜尋到的繁體中文網頁與簡體中文網頁的資料，進行並驗證兩岸在詞彙使用差異上的實際使用證據。

爲了可以比較 Chinese Gigaword Corpus 的繁體中文與簡體中文，及兩者的使用差異性，我們採用中文詞彙速描系統(Chinese Word Sketch)進行檢驗。中文詞彙速描系統裡，有四大搜尋功能，分別爲：concordance、word sketch、Thesaurus、Sketch-Diff，其中「Sketch-Diff」這個功能就是比較詞彙差異的工具，可以看出兩岸對於同一概念而使用不同詞彙的實際狀況與分佈，也可以看出同一語義詞彙在兩岸的實際語料中，所呈現的相同點與差異性。我們主要利用中文詞彙速描中詞彙速描差異(word sketch difference)的功能。詞彙速描差異的實際操作介面，如圖 2：


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
移至
Home
Concordance
Word Sketch
Thesaurus
Sketch-Diff
Show Diff

http://corpora.fi.muni.cz/chinese_all/

Word Sketch Differences Entry Form

Corpus:	chinese_all_trd
First lemma:	警察
Second lemma:	公安
Sort grammatical relations:	☑
Separate blocks:	○ all in one block ● common/exclusive blocks
Minimum frequency:	5
Maximum number of items in a grammatical relation of the common block:	25
Maximum number of items in a grammatical relation of the exclusive block:	12

圖 2. 中文詞彙速描系統的詞彙描素對比

在此功能下，我們將已經比對過 CCD 與 CWN 對應不同對譯的詞彙，進一步探究兩詞彙的使用狀況與分佈。在本文中，主要是以比對兩岸詞彙詞頻爲主，倘若在 CCD 與 CWN 的對應中，確實是相同語義，卻在兩岸使用完全相同或完全不同的詞彙，那麼其各自使用的詞彙，在 Gigaword Corpus 裡繁體語料與簡體語料交叉比對後所得的詞頻，也應當會有近似的分佈現象，藉此數據，不但可以證明 CCD 和 CWN 在英中對譯上，繁體

中文系統與簡體中文系統，是有差別的，也可以證明，確實有兩岸使用不同詞彙來表達相同概念語義的用法，進而了解兩岸詞彙的實際現象，以進行本研究的分析。

7. CCD與CWN語料分析

繁體中文系統的英中對譯(CWN)與簡體中文系統的英中對譯(CCD)，依不同詞類，區分成：名詞、動詞、形容詞和副詞四大類來進行對比，以 WordNet 為主，檢測在同一個同義詞集中(Synset)，繁體中文系統的對譯詞彙和簡體中文系統的對譯詞彙，然後再進行比對。

在四大詞類中，我們可以清楚得知，在同一個同義詞集中(Synset)，繁體中文系統，可能有多個相對應的對譯詞彙，同樣地，簡體中文系統也可能有個相對應的對譯詞彙。在這些對譯詞彙裡，又有可能是兩邊使用的對譯詞彙完全一樣，稱之「完全相同」；如果，兩邊使用的對譯詞彙，沒有一個相同的，稱之「完全不同」，也就是「真正不同」；或者，只有使用其中一個或一個以上對譯詞彙，這個狀況，稱之「部份相同」，而在「部份相同」的對譯詞彙，如果兩邊的對譯詞彙使用的詞首相同，稱之「詞首相同」，如果只是使用到相同的字，則稱之「部份字元相同」，如：

表 1. CCD 和 CWN 對譯的各種分佈狀況

Synset	CCD 對譯詞彙	CWN 對譯詞彙	
bookshelf	書架、書櫃、書櫥	書架、書櫃、書櫥	完全相同
lay off	下崗	解雇	完全不同
immediately	立即	立刻	詞首相同
according	據報	根據	部分字元相同

對於 CWN 與 CCD 的對比，總共有 70744 個 Synset 是對譯相同的，分屬於形容詞、副詞、名詞和動詞這四個詞類當中，其中，以名詞在 CWN 與 CCD 的完全相同對譯中，所佔比例最高，有 66.79%；反之，動詞所佔比例最低，僅有 4.05%，其詳細的分佈情況，如下圖顯示：

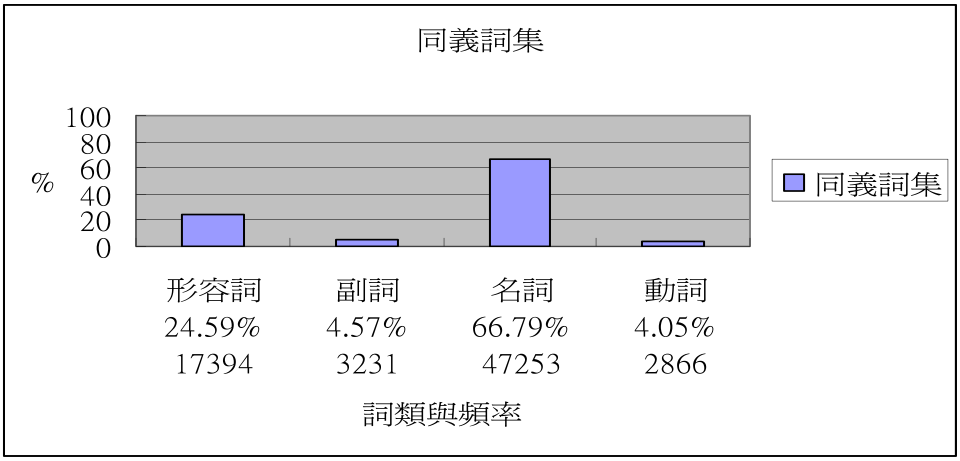


圖 3. CCD 和 CWN 依不同詞類呈現對譯相同的分佈

以 WordNet 爲基礎，CWN 所對譯的繁體中文與 CCD 所對譯的簡體中文，兩者使用完全不相同的情況，依各詞類的分佈情形，如下圖顯示：

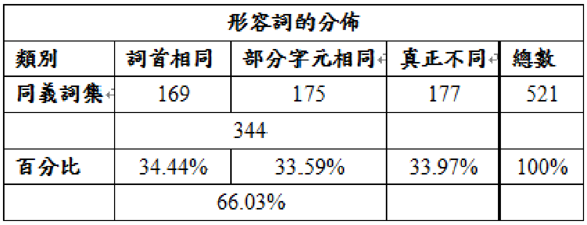
表 2. CCD 和 CWN 對譯不同的分佈狀況

	形容詞	副詞	名詞	動詞	總數
同義詞集數量	17915	3575	66025	12127	**99642**
不同對譯數量	521	344	18772	9261	**28898**
	2.91%	9.62%	28.43%	**76.37%**	**29.99%**
	最少			最多	

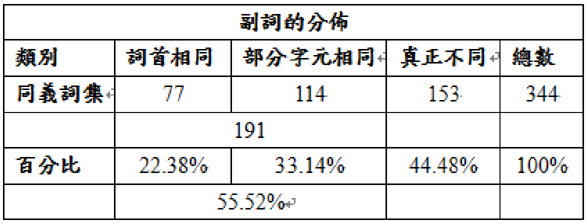
值得一提的是，在 CCD 和 CWN 翻譯不同的分佈狀況裡，很清楚得看到，「動詞」在兩岸的使用狀況，有極大的差異性，不過，由於在我們實際使用漢語時，常會以同類近義詞或語義相近相關詞來取代原本的詞彙，所以，我們又更進一步，更仔細地分析，希望將每一個詞類中，有這樣的使用情形分類出來，以得到真正兩岸使用不同詞彙的現象。

我們以詞彙的「詞首相同」、「部份字元相同」和「真正不同」這三大類爲主，分析 CCD 和 CWN 在形容詞、副詞、名詞和動詞這四個詞類當中，翻譯不同的分佈狀況，如下表顯示：

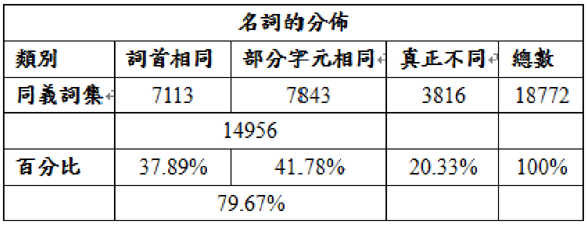
表 3. CCD 和 CWN 在各詞類中，對譯不同的分佈狀況



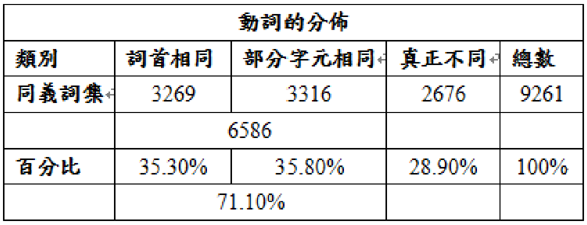
形容詞的分佈				
類別	詞首相同	部分字元相同	真正不同	總數
同義詞集	169	175	177	521
	344			
百分比	34.44%	33.59%	33.97%	100%
	66.03%			



副詞的分佈				
類別	詞首相同	部分字元相同	真正不同	總數
同義詞集	77	114	153	344
	191			
百分比	22.38%	33.14%	44.48%	100%
	55.52%			



名詞的分佈				
類別	詞首相同	部分字元相同	真正不同	總數
同義詞集	7113	7843	3816	18772
	14956			
百分比	37.89%	41.78%	20.33%	100%
	79.67%			



動詞的分佈				
類別	詞首相同	部分字元相同	真正不同	總數
同義詞集	3269	3316	2676	9261
	6586			
百分比	35.30%	35.80%	28.90%	100%
	71.10%			

從表 1 到表 3，我們可以清楚知道對於各詞類，CCD 和 CWN 在對譯不同的詞彙裡，仍然有些算是語義相近的相關詞彙，扣除這些相關詞彙後，兩岸詞彙在使用上的真正不同，就可清楚呈現。至於，上文中，所提及關於「動詞」是兩岸詞彙中，使用最多不同的狀況，我們從表 3 的分析得知，在動詞的使用上，因爲較常出現同類近義詞或語義相近相關詞來取代原本的詞彙的狀況，所以「詞首相同」和「部分字元相同」這兩類佔了

很大的因素，在 9261 個詞彙裡，就有 6586 個詞彙，大約是 71.10%，其真正兩岸對於動詞的不同使用，則有 2676 個詞彙，大約是 28.90%。

我們將以圖 3 和表 3 中，四種詞類裡，使用完全相同的詞彙與真正不同的詞彙，藉由 Gigaword Corpus 來分析兩岸人民對於詞彙使用的實際狀況。

8. 實驗設計與詞彙差異分析

以 WordNet 爲中心所對譯出 CCD 的簡體中文和 CWN 的繁體中文，比較兩者的對譯，有完全相同、完全不同與部份相同等三大類，在此，本研究僅就前兩類的資料，再以 Gigaword Corpus 爲依據，檢測實際語料中所呈現的狀況，同時，也以目前在網路上搜尋功能相當強大的 Google 作爲驗證的對象，比對利用在 Google 所搜尋的資料來驗證兩岸詞彙的對比。

8.1 Gigaword Corpus

首先，我們先取兩岸使用詞彙「完全相同」的資料，檢測這些資料在 Gigaword Corpus 中，分屬在繁體中文與簡體中文的使用頻率，再計算每個詞彙的頻率在繁體中文資料與簡體中文資料裡所佔的比例，如此，即可知道每一個詞彙，在繁體中文與簡體中文裡，出現和使用的情形。理想的想法，如果一個詞彙在兩岸使用的情況是非常接近的，其兩者詞頻比例的差距，應該是非常小的。我們試著將同一詞彙在兩岸使用的詞頻比例相減，以便檢測這些使用上完全相同的詞彙，又因其差距的數值過小，所以我們以放大 100000 倍後的數值來呈現，其分佈情形，如下圖所示：

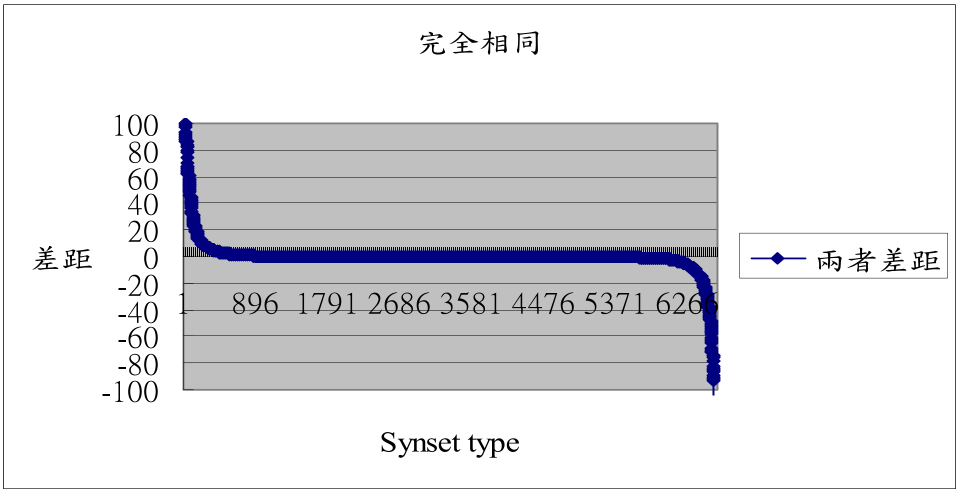


圖 4. 兩岸使用完全相同詞彙的分佈情況

從圖 4 來看，曲線彎曲的前後兩端，代表兩者的差距較大，靠左邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠右邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。在使用完全相同詞彙中，在分析數據呈現上仍有些使用差異的現象，這是值得我們深入

探討的議題。在 CCD 的簡體中文和 CWN 的繁體中文裡，有 6637 筆兩岸使用完全相同詞彙，我們以 Gigaword Corpus 的語料進行檢測，發現中央社/新華社語料所使用分佈差異的平均值爲 0.0143%。Gigaword Corpus 對於兩岸詞彙使用差異分佈在這個平均值內的詞彙，共計有 5880 筆。換句話說，兩岸使用完全相同的詞彙裡，在 Gigaword Corpus 使用狀況較爲相近的有 5880 筆，使用狀況較爲不相同的仍有 757 筆。在中央社與新華社語料中分別有 354 筆和 403 筆。這 757 筆資料是「同中有異」的詞語，値得我們將來進一步分析。

在 6637 個兩岸使用完全相同詞彙中，圖 4 雖然顯示其頻率差距幾乎是零。但是，如果我們由差距最小的第 3076 個詞，依前後各取 30%的(就是第 2153 個詞彙取到第 4144 個詞彙），將差距再放大呈現如圖 5。從圖 5 的差距數値顯示，是非常非常小的。一方面證明兩岸使用這些詞彙的情形，是非常非常接近的。另一方面，當間距放大後，我們看到差異分佈呈平滑的 S 字型，這也與預期中自然語料分佈的狀況相符。

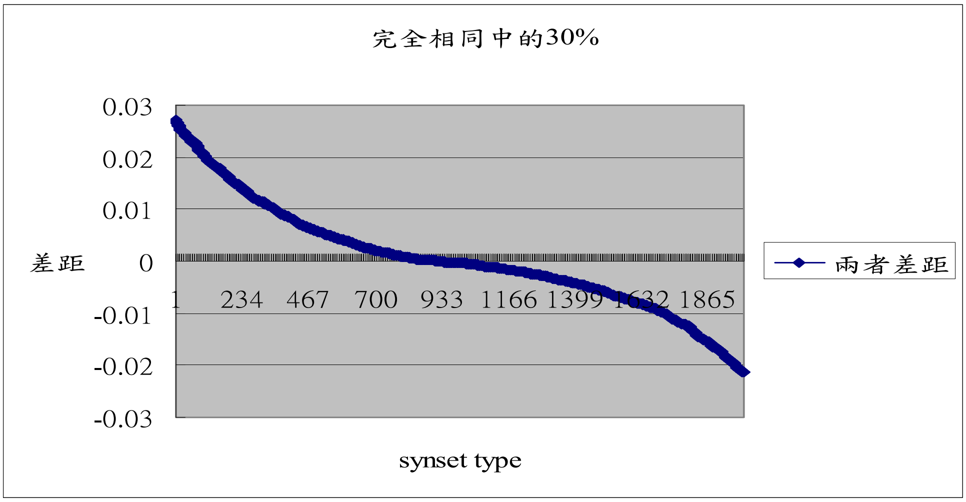


圖 5.兩岸使用完全相同詞彙中，差距最小的 30%的分佈情況

下面表 4，說明兩岸對於相同詞彙，在 Gigaword Corpus 中 CAN 的繁體中文與 XIN 的簡體中文，兩者使用狀況是非常接近的。

表 4. 兩岸使用完全相同詞彙的分佈狀況示例

詞彙	詞頻		附註
	CNA (繁體中文)	XIN (簡體中文)	
酒桶	32 (0.157μ)	20 (0.155μ)	使用狀況非常接近
絲瓜	1380 (6.78μ)	96 (0.748μ)	使用狀況有差異
柳葉刀 雙刃小刀	2 (0.00982μ)	273 (2.13μ)	使用狀況有顯著差異

接著，我們以相同的實驗方法與步驟來檢測兩岸使用完全不同的詞彙，檢測這些資料在 Gigaword Corpus 中呈現的分佈，在此，本文僅取數量較大的名詞和動詞來做比對，並且擷取語料的原則是出現在 CCD 所使用的詞彙，是 XIN 的詞頻大於 CNA 的詞頻；出現在 CWN 所使用的詞彙，是 CNA 的詞頻大於 XIN 的詞頻，其分佈情形，如下圖所示：

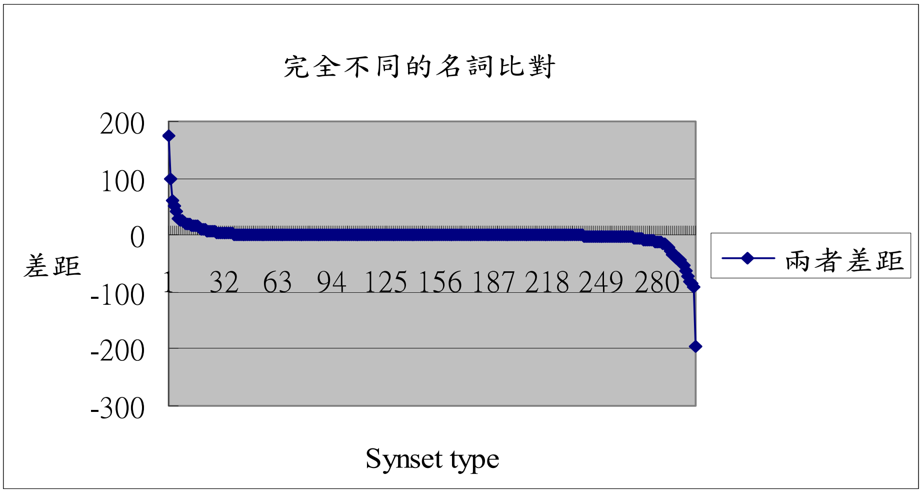


圖 6. 兩岸使用完全不同的名詞詞彙的分佈情況

在兩岸使用完全不同的名詞詞彙裡，共計有 302 筆資料，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們一樣採取兩者差距最小的 30%來檢測，其計有 91 筆資料，從圖 7 的差距數值來看，可以證明，這些不同的詞彙，在所屬的語言系統裡，其使用狀況的獨特性，換言之，同一個詞彙，在繁體中文系統裡，使用的頻率較高，在簡體中文系統裡，使用的頻率較低，反之亦然，而呈現相對之分佈狀態，這樣的情形，在圖 7 的差距數值和表 5 例子中得到驗證。

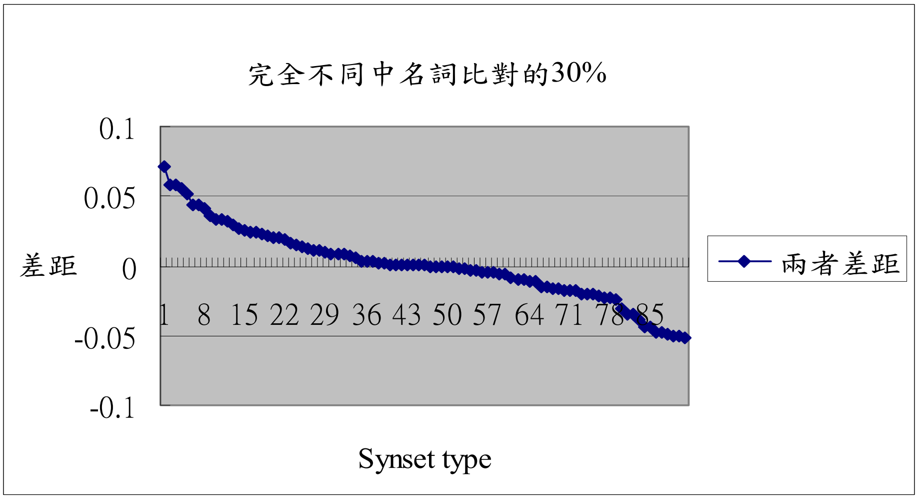


圖 7. 兩岸使用完全不同的名詞詞彙中，差距最小的 30%的分佈情況

下面表 5，說明兩岸對於完全不同的名詞詞彙，在 Gigaword Corpus 中 CAN 的繁體中文與 XIN 的簡體中文，兩者使用的分佈狀況。

表 5. 兩岸使用完全不同的名詞詞彙的分佈狀況示例

詞彙		詞頻				附註
CCD	CWN	CCD		CWN		
		XIN	CNA	CNA	XIN	
風帽	頭罩	10 (0.0779μ)	2 (0.0098μ)	101 (0.4963μ)	37 (0.2882μ)	使用狀況對比明確
雙休日	週末	1383 (10.7736μ)	25 (0.1228μ)	17194 (84.4908μ)	6105 (47.558μ)	使用對比較不明確
屏幕 CRT 屏幕	映像管	3086 (24.04μ)	118 (0.5798μ)	427 (2.0983μ)	1 (0.0078μ)	使用對比較不明確

至於在兩岸使用完全不同的動詞詞彙裡，共計有 461 筆資料(如圖 8 所示)，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們採取一樣的方式來進行檢測，其 30%的資料，共計有 140 筆，從圖 8、圖 9 的差距數值來看，確實可以證明這些使用不同的動詞詞彙，在繁體中文系統與簡體中文系統，有其使用狀況的對比性。

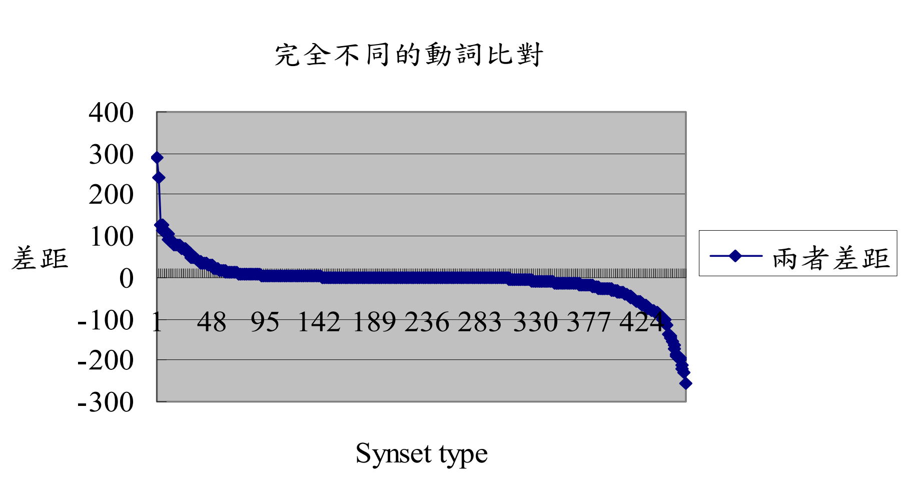


圖 8. 兩岸使用完全不同的動詞詞彙的分佈情況

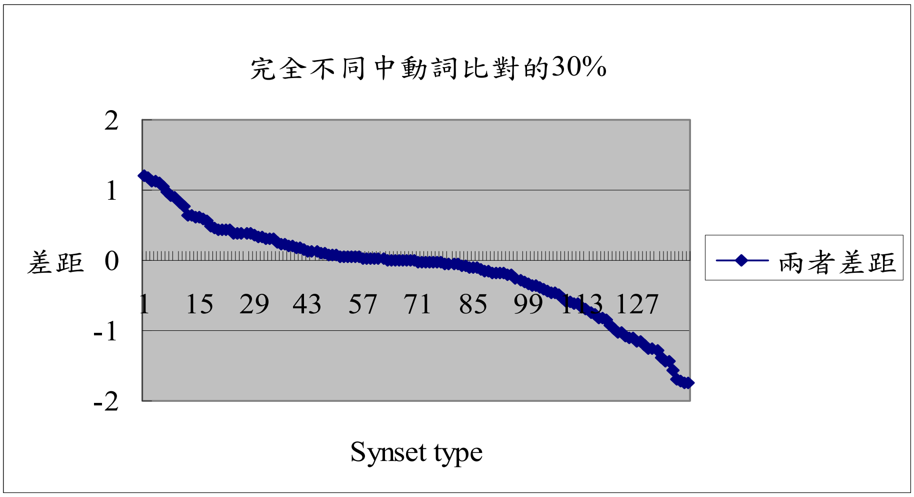


圖 9. 兩岸使用完全不同的動詞詞彙中，差距最小的 30%的分佈情況

兩岸使用完全相同詞彙的平均值是 0.0143%，那麼，理論上，兩岸使用不同詞彙的比例，應該大於這個平均值，倘若小於這個平均值，則有可能是兩岸使用相同概念詞彙時，產生混用的現象。在使用不同的名詞詞彙中，以大陸獨有詞的比例來排序，發現有 174 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 168 筆資料小於這個平均值；在使用不同的動詞詞彙中，以大陸獨有詞的比例來排序，發現有 320 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 348 筆資料小於這個平均值。這個數據證實了一個直覺的觀察，就是說兩岸詞彙互相影響滲透的現象日益顯著。以目前的數據看來，台灣的用法影響大陸略強於於大陸的用法影響台灣。

8.2 Google搜尋引擎

對於兩岸詞彙對比研究而言，除了根據具有學術性質的語料庫的資料來進行對比之外，我們也利用一般民眾每天都會使用的網路資料來進行對比，試圖了解民眾在日常生活中對於兩岸詞彙的使用狀況。因此，我們選定以 Google 搜尋引擎所找到的資料做爲對於兩岸詞彙對比研究的對象，進行搜尋後所得到的結果，即可觀察到「所有中文網頁」與「繁體中文網頁」的訊息，雖然沒有直接顯示「簡體中文網頁」的資訊，但在「所有中文網頁」的筆數與結果，可以看到包含「繁體中文網頁」與「簡體中文網頁」的訊息，換句話說，「簡體中文網頁」的筆數與結果，就是「所有中文網頁」的筆數扣掉「繁體中文網頁」的筆數，例如：

(3) 出租車

所有中文網頁：約有 141,000,000 項結果
繁體中文網頁：約有 4,330,000 項結果
簡體中文網頁：約有 136,670,000 項結果

由(3)的查詢結果顯示，「出租車」在簡體中文網頁的使用頻率多於繁體中文網頁的使用，表示「出租車」一詞，一般民眾在大陸地區是比較常使用的；相反地，在台灣地區則是比較少使用的。

有些是關於音譯的詞彙，在兩岸的使用上也有所不同，例如：美國總統 Obama，台灣的音譯名是「歐巴馬」，大陸的音譯名是「奧巴馬」，從 Google 搜尋的網頁資料，如(4)所示，台灣使用「歐巴馬」的筆數多於大陸；反之，大陸使用「奧巴馬」的筆數多於台灣，如此一來，顯示音譯詞彙方面在台灣與大陸皆有獨特使用的對應詞彙，也可以看出兩岸對於音譯詞的差異性及使用的頻率。

(4) 台灣的「歐巴馬 (4,670,000/ 1,230,000)」、
大陸的「奧巴馬 (10,100,000/ 115,900,000)」

再者，在兩岸人民的生活中，也有因爲一些制度、環境、日常生活、習慣而產生出的特殊用語，例如：台灣的「學測」、「免洗筷子」與大陸的「維穩」、「一次性筷子」…等，皆可從 Google 搜尋網頁的資料顯示出兩岸對於某些詞彙的獨用，或者對於相同的概念卻以不同的詞彙來呈現。

儘管 Google 搜尋網頁的資料可以顯示繁體中文網頁的偏用或簡體中文網頁的偏用，以呈現台灣、大陸的兩岸詞彙使用差異性，然而，兩岸人民在各方面的交流、接觸日益頻繁狀況下，彼此使用對方詞彙的狀況也日趨頻繁，以致於漸漸失去所謂台灣繁體中文系統獨用的詞彙或大陸簡體中文系統獨用的詞彙，例如：警察與公安。根據洪 等(洪嘉馡與黃居仁，2008)的研究結果，「警察」一詞應屬於較常被使用在台灣繁體中文系統，但是，在 Google 搜尋網頁的資料卻發現，「警察」一詞亦已在大陸地區廣泛被使用了。

(5) 警察
所有中文網頁：約有 335,000,000 項結果
繁體中文網頁：約有 24,500,000 項結果
簡體中文網頁：約有 310,500,000 項結果

藉由 Google 搜尋網頁的資料，雖然可以呈現出台灣、大陸的兩岸詞彙對比使用狀況，但是，畢竟網路上的資源是比較多元化、也比較具有複雜性，而且，我們無法從網頁訊息得知兩岸網頁總數各若干，所以，按常理推斷，大陸的網頁應該比台灣多很多。此外，兩岸目前交流較頻繁，常有互相引用，無法排除，目前，即可看出大陸用「警察」用法是愈來愈多。這也說明，Google 所搜尋到的資料，僅可以當作目前台灣與大陸一般民眾對於某些詞彙的使用狀況，而無法真正提供兩岸詞彙或世界華語對比的研究。

9. 結論

兩岸詞彙在使用上的相同、不同或些許的差異，甚或混雜使用，在交流頻繁的情形下，已經日趨明顯，如何區分並釐清兩岸詞彙的個別語義架構，又能在其架構下，增加我們對於漢語詞彙語義系統性演變脈絡的理解，是我們從事語言研究者不容忽視的議題。本文藉由 WordNet 所發展出的繁體中文系統 CWN 與簡體中文系統 CCD，進行兩岸詞彙的比對，再將比對過後的詞彙，以收集實際大量語料的 Gigaword Corpus 爲基礎，檢測兩岸在詞彙上使用的現象與分佈狀況；亦可由 Gigaword Corpus 所呈現的狀況，證明繁體中文系統 CWN 與簡體中文系統 CCD 在比對上的正確度與可靠性；也證實了 CCD 和 CWN 將兩岸詞彙對比的使用狀況質化呈現，而 Gigaword Corpus 則是以實際語料來驗證兩岸詞彙對比的使用狀況量化呈現。我們更進一步發現了兩岸共用詞彙有「同中有異」的現象，而對比詞彙也產生了互相滲透影響的現象。值得更深入探討研究。同時，應用具有大量繁體中文、簡體中文的 Google 搜尋網頁的資料進行兩岸人民使用詞彙的對比與差異分析，在此，發現具有學術性質的語料庫，如本文所使用的 Gigaword Corpus 在作爲兩岸詞彙對比研究或世界華語對比研究時，其研究成果與學術價值是比 Google 所提供的資料高很多的。

參考文獻

Chinese Word Sketch Engine: http://wordsketch.ling.sinica.edu.tw/.

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

Hong, J.-F., & Huang, C.-R. (2006). WordNet Based Comparison of Language Variation - A study based on CCD and CWN. Presented at *Global WordNet (GWC-06)*. 61-68. January 22-26. Jeju Island, Korea.

Huang, C.-R., Chang, R.-Y., & Li, S.-b. (2010). *Sinica BOW: A bilingual ontological wordnet*. In: Chu-Ren Huang *et al*. Eds. Ontology and the Lexicon. Cambridge Studies in Natural Language Processing. Cambridge: Cambridge University Press.

Huang, C.-R., Tseng, E. I. J., Tsai, D. B. S., & Murphy, B. (2003). Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Languages and Linguistics*. *4*(3), 509-532.

Kilgarriff, A., Huang, C.-R., Rychly, P., Smith, S., & Tugwell, D. (2005). *Chinese Word Sketches*. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.

Lexical Data Consortium. 2005. Chinese Gigaword Corpus 2.5.: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993). Introduction to WordNet: An On-line Lexical Database. In *Proceedings of the fifteenth International Joint Conference on Artificial Intelligence.*

于江生、俞士汶(2004)。中文概念詞典的結構。*中文信息學報(Journal of Chinese Information Processing)*, *16*(4), 12-21。

于江生、劉揚、俞士汶(2003)。中文概念詞典規格說明。*Journal of Chinese language and Computing*, *13*(2), 177-194。

王鐵昆、李行健(1996)。兩岸詞彙比較研究管見。*華文世界*，第 81 期，台北。

中華民國交通部觀光局(2011)。*兩岸地區常用詞彙對照表*，http://taiwan.net.tw/m1.aspx?sNo=0016891.

姚榮松(1997)。論兩岸詞彙差異中的反向拉力。*第五屆世界華語文教學研討會*，世界華語文教育協進會主辦，1997 年 12 月 27-30 日，台北劍潭。

南京語言文字網(2004)。*兩岸普通話大同中有小異*，http://njyw.njenet.net.cn/news/shownews.asp?newsid=367.

洪嘉馡、黃居仁(2008)。語料庫爲本的兩岸對應詞彙發掘(A Corpus-Based Approach to the Discovery of Cross-Strait Lexical Contrasts). *Language and Linguistics*, *9*(2), 221-238, 2008. Taipei, Nankang: Institute of Linguistics, Academia Sinica。

許斐絢(1999)。*台灣當代國語新詞探微*。國立台灣師範大學華與文教學研究所碩士論文，台北。

華夏經緯網(2004)。*趣談海峽兩岸詞彙差異*，http://www.huaxia.com/wh/zsc/00162895.html.

廈門日報(2004)。*趣談兩岸詞彙差異*，http://www.csnn.com.cn/csnn0401/ca213433.htm.

劉揚、俞士汶、于江生(2003)。CCD 語義知識庫的構造研究。*中國計算機大會(CNCC’2003)*。

戴凱峰(1996)。*從語言學的觀點探討台灣與北京國語間之差異*[A Linguistic Study of Taiwan and Beijing Mandarin]。政治作戰學校外國與文學系碩士論文，台北。

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 35-56

基於字典釋義關聯方法的同義詞概念擷取：以《同義詞詞林（擴展版）》爲例[1]

A Definition-based Shared-concept Extraction within Groups of Chinese Synonyms: A Study Utilizing the Extended Chinese Synonym Forest

趙逢毅*、鍾曉芳+

F. Y. August Chao and Siaw-Fong Chung

摘要

同義詞在資訊擷取與語義分類上是很重要的語料資訊，但將兩詞歸納爲同義其原由則值得令人探討。從語義(sense)的觀點來說，多義詞組歸到特定同義組合中，其語義中應有與該類字詞同義集合。此類型的代表爲《同義詞詞林》(梅家駒、竺一鳴、高蘊琦與殷鴻翔，1983)，將漢語同義字詞區分成具結構類別。而從計算語言學方法來說，同義詞關聯需要參考語料庫中詞組的出現頻率，輔以機器學習方法來計算同義詞相似度。然而前者專家分類原則是透過語感進行，若沒有對同義詞的類別原則加以定義，則後人便會產生對同義詞的混淆。後者機器學習方法使用統計方法來辨別相似詞彙，則會缺乏語義的辨別。爲了瞭解同義詞組的概念內涵，本研究提出基於辭典釋義文字的關聯計算原則，試透過計算共同擁有的釋義文字出現比率，以解析兩詞彙間所包涵之釋義概念。並且以《同義詞詞林（擴展版）》爲例，從釋義義涵的角度列舉出適合詮釋該詞組

[1] 本研究第二作者蒙國科會計畫專案補助編號 101-2410-H-004-176-MY2，初稿亦獲匿名審查委員指正，謹此誌謝。

* 國立政治大學資訊管理學系
Department of Management Information Systems, National Chengchi University
E-mail: fychao.tw@gmail.com

+ 國立政治大學英國語文學系
Department of English, National Chengchi University
E-mail: sfchung@nccu.edu.tw

的詞彙，突顯該類別所包涵的語義。最後，比較 SketchEngine (Kilgarriff *et al.*, 2004)中所取得的同義詞(similar words)之間的差異。本研究計算結果雖然會受辭典釋義內容影響，但辭典釋義內容相較於人工分類原則與統計語料庫所得的數值資料，較能從詞義上詮釋詞彙之間的共有概念。我們希望能透過釋義關聯方法更瞭解詞彙間的交集概念，亦希望能在同義詞的語義計算上，提供辭典釋義與詞條編寫上的思考。

關鍵字:同義概念, 同義詞詞林, 釋義, 辭典

Abstract

Synonym groups can serve as resourceful linguistic metadata for information extraction and word sense disambiguation. Nevertheless, the reasons two words can be categorized into a particular synonym group need further study, especially when no explanation is available as to why any two words are synonymous. Lexical resources, such as the Chinese Synonym Forest (or Tongyici Cilin) (Mei *et al*. 1983), assemble lexical items into hierarchical categories via manual categorization. Other than this, statistical measures, such as co-existing probability, have been adopted widely to verify synonymous relationships. Nevertheless, a purely statistical method does not provide description that can help interpret why such a synonymous relationship occurs. We propose a novel method for the study of shared concepts within any synonym group by comparing co-existing words in the dictionary definition of each member in the group. The co-existing words are seen as the representatives of shared concepts that can be used for interpretating any hidden meaning among members of a synonym group. We also compare our results with the thesaurus function in the Sketch Engine (Kilgarriff *et al*. 2004), which uses statistical data in the form of Sketch scores. The results show that our method can produce concept words according to dictionary definitions, but this method also has its limitations, as it works only with a finite number of synonyms and under limited computing resources.

Keywords: Shared Concept, Synonym, Chinese Synonym Forest, Dictionary Definition

1. 簡介

同義詞語料在自然語言處理與資訊擷取技術領域上是很重要的參考資源。透過同義詞語料不僅能讓機器學習方法更瞭解使用者所闡述的文字內容概念，亦可以對提問文字舉一反三。在詞彙同義的歸納上，大概可區分為兩種進行方向：詞義訓詁與機器學習。第一，“義訓者，觀念相同，界說相同，特不說兩字之製造及其發音”(黃侃述與黃悼，1983)。

「同義相訓」是在詞義訓詁上的主要工作。這項工作往往需要花費許多人力與時間，才能清楚地分析出詞彙之間的同義內涵。此類型的成果眾多，如梅家駒等(1983)所編撰的中文《同義詞詞林》(以下稱《詞林》)，收錄來自詞素、詞組、成語、方言詞與古語等詞，共五萬三千多詞彙，並且依照同義詞分類義涵有系統地分成不同的類別。其後經由哈爾濱工業大學信息檢索研究室(HIT IR Lab)刪除舊詞與罕用詞，並依新聞語料加入常用新詞，使擴展版詞彙量增加到七萬多。《詞林》中，詞語分類原則是「相對、比較」(梅家駒等，1983)，詞語所屬類別與列舉位置哲學有作者們不可言喻巧思。第二，機器學習方法進行同義詞辨析近年來不斷發展，但著重在使用同義字詞進行上下文之中詞義的消歧。在進行同義消歧的處理過程中，又可區分爲監督式與非監督式兩種學習方法，來進行辨別是否可歸爲同義字群。上述的兩類機器學習方法都需要參考語料庫的詞類頻率計算後，才能得到字詞之間的相似程度。在進行此類方法時，常見的問題是缺乏大型語料庫與同義詞資料稀疏的通則(劉挺與車萬翔，網頁資料擷取於 2012)。

本研究使用辭典釋義內容，首先對詞彙之間的共有概念計算原則進行討論，再對《同義詞詞林（擴展版）》(下稱《擴展版》)之分類進行釋義關聯計算。本研究試以共同使用的釋義用詞，擷取能表達該分類的共有概念詞組。除了計算《擴展版》中的同義類別義涵，並透過釋義涵蓋數與最大平均釋義關聯詞值比較同義類別中的詞彙，標記較適合用以表達該類別的詞彙。最後，我們將所使用的釋義關聯比較原則與 Sketch Engine(Kilgarriff *et al.*, 2004)的語料庫統計方法進行比較，對比兩種方法在詞彙共同義涵計算上異同。Chinese Sketch Engine 語料統計方法基於大量中文語料進行語法統計與共同出現詞頻計算取出同近義詞(Huang *et al*. 2004)。雖然此語料庫平台可以呈現語言行爲相似的詞彙，但近義詞語義關係無法透過此類統計方法的結果明確說明；相反的，釋義關聯能計算兩多義詞彙之間較合適的同義釋義，需依賴辭典釋義進行關聯計算。另外，由於《教育部重編國語辭典修訂本》(下稱《國語辭典》)的釋義屬非專業領域辭典，因此仍有涵蓋面欠缺或不全，此乃是本方法的限制。

2. 釋義關聯

詞彙是「語言中表達意義的最小獨立單位」(黃居仁，2005)，辭典的釋義文字則是組合查詢者瞭解的詞彙，以表達詞條的概念涵義。釋義說明所用的單一詞語包括被解釋字(詞)的部份或完全的涵義，而且在釋義的詞語都屬於知識層級上較爲通俗的語句或概念。例如辭典之中，在解釋“*鱗波*”詞條時，會使用“*魚鱗*”、“*波紋*”等詞彙來說明詞彙的義涵。基於釋義字彙的共同出現頻率原則，提出基於釋義詞彙共同出現的涵蓋比例(Percentage of co-appearance of relations, CoAP) 的語義關聯程度 (Semantic Relation Degree, SRD) 的比較方法 (趙逢毅與鍾曉芳，2011)，比較詞彙間在釋義文字的涵蓋關聯。計算方式如下：

$$CoAP(x) = \frac{X \bigcap Y}{X} \qquad (1)$$

$$SRD_{xy}=2\frac{CoAP(y)CoAP(x)}{CoAP(y)+CoAP(x)},\ 0\le SRD_{xy}\le 1 \tag{2}$$

其中 x, y 是兩個待計算的詞條字/詞彙， X 與 Y 分別是 x, y 兩字的釋義字彙組合。*CoAP* 是釋義詞彙共同出現的涵蓋比例，即是計算共同出現的釋義字彙在原有字彙中所佔之比例。語義關聯程度 (SRD) 即是將兩詞語 *CoAP*(*x*), *CoAP*(*y*)相互之間概念比例的方向性消除，所以使用平均數以表示兩詞彙之間共同釋義詞彙語義關聯程度。在進行釋義關聯計算中，會使用釋義詞彙之中僅以包涵概念義涵較多的動詞與名詞詞類進行計算。計算過程在此以「*漣漪*」與「*鱗波*」兩詞爲例：

在《教育部重編國語辭典修訂本》釋義中

「漣漪」爲“水面上細微的波紋。”

「鱗波」爲“水面似魚鱗狀的波紋。”

計算釋義詞彙經中研院斷詞系統[2]取得釋義文字的詞性之後，僅使用動詞與名詞進行計算。計算過程如下：

*「漣漪」爲“**水面(Nc)** 上(Ncd)細微(VH)**波紋(Na)**”*

*「鱗波」爲“**水面(Nc)**似(VG) 魚鱗狀(Na)**波紋(Na)**”*

因此，

$$CoAP(\text{漣漪})=\frac{(\text{水面 波紋})}{(\text{水面 上 細微 波紋})}=\frac{2}{4}$$ ，同理

$$CoAP(\text{鱗波})=\frac{(\text{水面 波紋})}{(\text{水面 似 魚鱗狀 波紋})}=\frac{2}{4}$$

$$SRD_{\text{漣漪,鱗波}}=2\frac{CoAP(\text{漣漪})CoAP(\text{鱗波})}{CoAP(\text{漣漪})+CoAP(\text{鱗波})}=2\frac{\frac{2}{4}*\frac{2}{4}}{\frac{2}{4}+\frac{2}{4}}=0.5$$

如果使用的釋義詞彙完全相同，則透過上述得到的 SRD 值會較高，即爲釋義相同義涵詞彙；反之，若共有釋義詞彙佔有的比例較低，則因沒有共同的釋義內容，而使兩個詞彙在直接的釋義內容上，無法找到共同詞彙交集。當使用這釋義詞彙進行計算時，有三項主要的缺點：(1)共有釋義詞彙數目的計算是以斷詞後的字詞組爲基準，並以字型比較原則進行計算，無關詞彙本身意義。如：「*魚鱗*」與「*龍鱗*」，雖然兩者都是以鱗片進行比喻，但在以字型爲基礎的字/詞組比較時，卻無法將兩者進行關聯；(2)多數詞彙的完整釋義句子都不長，從而使共有釋義詞彙數目對 SRD 值的影響十分敏感，如前述「*鱗波*」一詞。若將釋義「*鱗波*」改爲 *“水面*$\boxed{\text{上}}$*似魚鱗狀的波紋。”*，則經斷詞之後爲 *“水面 (Nc) 上 (Ncd) 似 (VG) 魚鱗狀 (Na) 波紋 (Na)。”*，則共有釋義詞彙由原本的兩詞

[2] 中研院斷詞系統，Chinese Knowledge and Information Processing (CKIP) Word Segmentation System, http://ckipsvr.iis.sinica.edu.tw/

變成三詞 *“水面 (Nc)上 (Ncd)波紋 (Na)。”* 並使 $CoAP(\text{漣漪}) = \frac{(\text{水面 上 波紋})}{(\text{水面 上 細微 波紋})} = \frac{3}{4}$ 且 $CoAP(\text{鱗波}) = \frac{(\text{水面 上 波紋})}{(\text{水面 似 魚鱗狀 波紋})} = \frac{3}{4}$ 最後兩個的 SRD 值則會是 $SRD_{\text{漣漪},\text{鱗波}} = 2\frac{\frac{3}{4} * \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = 0.75$ 而提高 25%； (3)所有共同釋義詞彙的權重皆相同，不會區別字詞在釋義句中的語法角色，因此釋義各字與詞彙的詮釋比重皆相同。如經斷詞之後的「*鱗波*」爲 *“水面(Nc) 上(Ncd) 似(VG) 魚鱗狀(Na) 波紋(Na)。”*，依句型文法與詮釋的角度解讀，其概念表達僅需要 *“魚鱗狀”*、*“波紋”*、*“水面”* 三詞彙，但上述釋義關聯方法則無法把功能詞去除。基於釋義文字的字詞組爲表達辭條概念意涵的原則之下，我們則將釋義文字的釋義也加入兩詞彙之間的關聯比較，因此用來計算的共用釋義詞彙能夠不僅局限於單一釋義階層的詞彙(字型相同)比較上。以使用相同的辭典內容，再次解釋出現的共有釋義詞彙意義，並納入關聯計算函式中，除了以增加共有釋義詞彙廣度，同時也依重覆出現的釋義詞彙計次來決定參與計算權重，以改善前述的三項缺點。計算方式如下：令 X^1 爲條目 x 所有符合過濾條件的辭典釋義詞彙，即**第一階(直接)釋義詞彙組合**；X^2 爲 X^1 釋義詞彙再經過辭典釋義文字擴充並符合過濾條件詞彙，即**第二階釋義詞彙組合**(釋義文字的釋義)；同理，Y^2 爲 Y^1 符合的釋義詞彙，再經辭典釋義擴充後且符合過濾條件詞彙集合。而 $X^{(1+2)}$ 爲 X^1 第一階(直接)與 X^2 第二階(釋義文字的釋義)釋義詞彙的集合總合；同理，$Y^{(1+2)}$ 爲 Y^1 第一階與 Y^2 第二階釋義詞彙的集合總合。而 X^1、X^2 與 Y^1、 Y^2 共同擁有的釋義文字的集合則表示爲 $X^{(1+2)} \bigcap Y^{(1+2)}$ ， 使前述的第二階共有釋義關聯值計算則修改成 $CoAP^2_{x,y}(x) = \frac{(X^1 + X^2) \bigcap (Y^1 + Y^2)}{(X^1 + X^2)}$ 以表示所有用來釋義 *x* 詞條的第一階與第二階文字集合，與所有用來釋義 *y* 的第一階與第二階文字集合，兩者之間交集共同出現的擴充釋義字詞的比例。以此類推，則第 n 階層釋義關聯的一般式即可寫成：

$$CoAP^n_{x,y}(x) = \frac{\sum\limits_{i=n} X^i \bigcap \sum\limits_{i=n} Y^i}{\sum\limits_{i=n} X^i} \tag{3}$$

從而用來計算兩者之間的第 n 階釋義關聯數值則修改成爲下列一般式：

$$SRD^n_{x,y} = 2\frac{CoAP^n_{x,y}(x) CoAP^n_{x,y}(y)}{CoAP^n_{x,y}(x) + CoAP^n_{x,y}(y)},\ 0 \le SRD^n_{x,y} \le 1 \tag{4}$$

透過前述多階層反覆釋義並計算共同出現的釋義文字比例，在先前 SRD 的比較方法研究中已討論過，除了可以發掘詞彙深層的釋義詞彙，亦可使釋義關聯值不受釋義文字

些微修改而大幅影響釋義關聯值。在計算共同擁有的釋義文字在反覆的階層釋義過程中，也可利用反覆釋義而出現的詞彙累計，突顯出權重高的佔有詞彙並加入計算。在多階層反覆釋義的過程，能找出兩詞彙之間深層的共有釋義詞彙，並且因爲能被擴充釋義的詞彙都已經透過辭典釋義過，所以反覆釋義有益於穩定計算關聯的結果(即所有釋義成份所佔百分比會傾向一定的組成比率)。最後在前先的研究中亦建議多階層釋義關聯值，可以取用第四階層的語義關聯程度進行討論(即反覆進行釋義處理至 X^1、X^2、X^3、X^4，並將所有結果加總計次)。

3. 《同義詞詞林（擴展版）》

哈爾濱工業大學信息檢索研究室 (HIT IR Lab) 所提供的公開版本《擴展版》，是整理自《詞林》(梅家駒等，1983)，除了刪除舊詞與罕用詞外，並依新聞語料加入常用新詞。在梅版的《詞林》中，收錄來自詞素、詞組、成語、方言詞與古語等詞共五萬三千多詞彙數，並且依照同義詞分類涵義有系統地區分爲人、物、時間、空間、抽象事物、特徵、動作、心理活動、現象與狀態、關聯、語助、敬語等十二組大類(以 A 到 L 標記類別的第一位英文字)以及若干中類與小類。同類型詞語依照「相對、比較」的排序原則(梅家駒等，1983)依同義/近義程度在同類型中，單行詞語由左自右排列，詞語所屬類別與列舉位置則隱含有作者們的巧思。《擴展版》對原始的分類也擴展到五層，其中加入「相等、同義」(=)、「不等、同類」(#)及「自我封閉、獨立」(@)等相關涵義。

表 1.《同義詞詞林（擴展版）》例

Bp20B03=招子**幌子**市招
Dd15A09=**幌子**招牌牌子旗號金字招牌
Aa01B03#良民順民
Bg02B07#超聲波低聲波聲波
Aa01C05@眾學生
Bg03A01@火

在表 1 中可看得出，《擴展版》都保留了分類類別、字彙及同義詞彙，並沒有針對該類別給予明確的類別涵義定義，亦沒有對所類別中的詞彙給予明確定義。以 Bp20B03=來說明，雖依《詞林》編撰原則—同義詞在前、近義詞在後—說明“*招子*”與 Bp20B03=應屬同義，在《國語辭典》中分爲六個釋義，其分別爲*「招牌、廣告、海報」*、*「門票。」*、*「亦稱爲花招、招兒」*、*「死刑犯就刑時，插於背後的紙標，用來揭示犯人的罪狀、姓名。」*、*「隨風招展的長布簾子。」*、*「眼睛。多用於江湖人物間。」*。而排序在第二個詞*幌子*，在《國語辭典》中分爲二個釋義，分別爲*「掛在店鋪門外，用來招徠顧客的招牌。」*與*「表現在外用以蒙蔽他人的言行。」*；*市招*的釋義僅只有*「商店門外標示其名號及所賣貨物的招牌或標誌。」*。雖然分類類別是屬 B 類(物)，但從分類架構或比較同類型詞組，亦無法從《擴展版》的分類之中得知在 Bp20B03=是屬“*招子*”的六個釋義中那一個定義。此外，一詞多義(Homographs)在《擴展版》之中則會分列在不同的類

別之中，如前述的“*招子*”則分別被歸在 Bp20B03=(B 類，物)與 Dk15A03= (D 類，抽象事物)之中。此分類結果不僅進行同義詞歸類時會造成模糊，亦有可能會在分析文本時造成所表達的義涵辨別錯誤。由於《詞林》僅將詞彙高度概化後，再將詞彙置放於相對的類別之中(鮑克怡，1983)，雖然經過增、刪、修改之後已較符合時下用語，但這樣人爲的相對分類原則很難讓後人把研究的語料歸納到《詞林》分類之中。

4. 同義概念的相關研究

在討論詞彙的同義概念的研究上，大概可區分爲兩種面向進行：詞義訓詁與機器學習。詞義訓詁中的同義相訓工作即是在建立同義詞關係，如《爾雅・釋言》：「宵，夜也。」與《說文》：「麗爾，猶靡麗也。」。然而使用人工方法進行詞義訓詁時，如《爾雅》在訓釋散佈於各處的資料中，難以尋找在同一基礎概念上構築的同義詞彙，使訓詁工作需要專業人士經過長時間累積 (王建莉，2012)。此外，在已由前人歸類完成的同義詞組內涵中，亦是由於沒有明確說明詞組內容，從而使後人會有理解上的差異。如范紅麗(2011) 對"*拜*"、"*揖*"、"*稽首*"、"*頓首*"、"*稽顙*"、"*拜稽首*"等詞群，以《左傳》爲資料分析在同義詞群中的異同。而在現代詞彙的研究之中，由於無古文可訓，從而使用語料庫方法進行詞彙同義比較。全文奕與郭聖林(2012)對現代漢語中的"*計程車*"、"*出租汽車*"、"*計程車*"、"*德士*"同義詞群的來源義涵與競爭進行討論，並輔以北大 CCL 語料庫驗證，以討論外來譯語存在的分佈情況與其限制。

如前述提及的機器學習方法，基於統計資料分析則可區分爲監督式與非監督式的學習方法兩類。監督式的學習方法以詞語規則、文法規則或概念規則中擷取計算詞彙之間的相似度。曾慧馨、劉昭麟、高照明、陳克健(2002)使用了詞彙中用字的組合規則，和結構與概念之間相似度對動詞進行未知詞的同義分類。而非監督式的學習方法則是將詞彙成對，透過在文本之中的出現字串的『相對頻率』(relative frequency)、『互見資訊』(mutual information)和『上下文依附』(context dependency) 等各種統比率(林頌堅，2004)計算統計上的相似程度。舉例來說，使用詞彙出現頻率的餘弦(cosine)關係進行圖書資訊檢索中的詞彙擴展(陳光華與莊雅蓁，2001)與互斥資訊熵原則 PMI-IR (Pair-wise Mutual Information-Information Retrieval)學習並預測在 TOEFL 考試中的同義詞考題(Turney, 2001)。

5. 研究方法

本研究旨在使用辭典釋義關聯計算同義字群中的共有概念擷取，並提供除了同義詞組外更明確的辭典釋義文字，以利後續的詞義消歧工作。在此我們則使用《教育部重編國語辭典修訂本》釋義作爲計算的基礎，計算在《同義詞詞林（擴展版）》之中的所有分類項後，提取出該分類項中共同有的釋義文字、交叉比對釋義關聯比例與平均最大釋義關聯詞(average dictionary definition relationship)，以呈現在同義詞群之中最能表達該群的概念。

5.1 語料準備

本研究所使用的語料主要有 HIT IR Lab 的《擴展版》與教育部的《國語辭典》。其中《擴展版》的資料是以簡體字碼編寫，因此我們使用了維基百科的繁簡分歧詞表進行繁簡轉換。維基百科的繁簡分歧詞表包括了大陸、台灣、香港與新加坡各地的漢語編碼與詞彙互換原則，例如 hardware 一詞在大陸稱作*硬件*、台灣則稱做*硬體*。另一方面使用《國語辭典》作爲釋義計算的基礎，因此需要將辭典內的詞條去除掉詞目、正形、注音、引證和案語等資訊，只保留釋義的部份再送至中研院 CKIP 斷詞系統進行詞性標記(P.O.S. tagging)。經過處理資訊總計處理條目數爲 156296，共計 278793 種不同的釋義。

5.2 詞彙間釋義關聯計算

爲了解詞彙之間共同有的釋義詞彙，我們使用釋義關聯對詞彙之間的釋義做交互比對，取得到最大釋義關聯值的文字作爲兩詞彙之間的同義概念。在進行釋義關聯計算時，則使用較多語義的動詞詞類 ('VA', 'VAC', 'VB', 'VC', 'Vi', 'Vt', 'VCL', 'VD', 'VE', 'VF', 'VG', 'VH', 'VHC', 'VI', 'VJ', 'VK', 'VL', 'V_2') 與名詞詞類 ('Na', 'Nb', 'Nc', 'Ncc', 'Nd', 'N') 進行釋義關聯的計算。在釋義關聯計算原則上，我們修正前多階層釋義關聯計算原則。由於先前提出的多階層釋義關聯計算原則 (趙逢毅與鍾曉芳，2011)在進行釋義階層擴展時，會局限在特定的部首的概念中，使階層較高的釋義權重與低階層權重相同。不同於先前的計算原則，在本研究中的同義詞釋義關聯計算是以**無特定概念的方向擴展**，因此深階層釋義的概念權重應比較低階層釋義權重低(即釋義權重與階層深度成反比)。爲了使第一階層(直接)釋義文字的權重能高於第二階層(釋義文字的釋義)，我們則將每次的釋義計算過程中，累計低階層中的計次數值，使第一階層中出現的釋義詞彙權重，會較第二階層中出現的釋義詞彙重(不一定成倍數關係。因爲在第一階層中出現詞彙不一定只會出現一次，不一定在每一階層都會出現。)。若 x, y 爲兩待測詞彙，X', Y' 爲包涵 x, y 兩詞彙與其釋義的各別集合，則第 n 階層之釋義詞彙則爲 $X'^{n+1} = X'^{n} + X^{n+1}$ 且 $Y'^{n+1} = Y'^{n} + Y^{n+1}$：

$$CoAP'^{n}_{x,y}(x) = \frac{\sum_{i=n} X'^{i} \bigcap \sum_{i=n} Y'^{i}}{\sum_{i=n} X'^{i}} \tag{5}$$

$$SRD'^{n}_{x,y} = 2\frac{CoAP'^{n}_{x,y}(x)CoAP'^{n}_{x,y}(y)}{CoAP'^{n}_{x,y}(x) + CoAP'^{n}_{x,y}(y)}, \quad 0 \le SRD'^{n}_{x,y} \le 1 \tag{6}$$

由於計算關聯的原則不同，因此我們先觀察在計算不同階層結果以下，是否如先前研究結果一樣，可以取第四層爲基準進行兩詞彙之間的關聯比較。在此則以*「漣漪」*與*「鱗波」*兩詞，計算修改後版本的詞彙間釋義關聯，結果呈現由第一階層到第八階層的結果(詳見下頁表 2)。

從表 2 的結果可以知道，當階層越深則釋義關聯越高，且因爲使用反覆的辭典釋義擴充參與計算的詞彙，因此較深層的義涵在第三、四層之間，共有的釋義文字佔有的比

例產生較明顯的差異。在低階層釋義關聯(淺層釋義)中，因爲詞彙並沒有經過太多解釋(階層數少)，因此共有文字相較於高階層釋義關聯(深層釋義)較爲明確具體，如：*水面*、*波紋*、*魚鱗狀*等等。當詞彙經多次釋義之後，共同出現的詞彙則會依知識概念 (concept) 較高的層級堆疊累積而清楚說明，而開始將明確具體的共有詞彙所佔的比例降低，從而出現*表面*、*部份*等文字，直到計算第八階層的計算結果出現共有釋義詞彙如*個體*、*事物*等文字。雖然在本次的實驗之中，我們因爲加入了**無特定概念的趨向**(先前的研究中，局限在屬「目」字部的字與詞彙)的多階層擴充原則，而增加了低階層的共有釋義在計算過程之中的比例。在上述的結果顯示，這樣的計算方法可以擷取出較低階層中(明確具體)與較深階層中(一般性概念)的共有釋義詞彙。在此需要特別說明的是，共有釋義文字*水面*、*波紋*等在第一、二、三、四階層中，這些詞彙擁有高權重的原因是多數釋義文字都共同使用到，而使*水面*、*波紋*詞彙的概念主宰支配(dominate)著該詞條的主要涵義。但在高階層的釋義關聯計算結果中，*水面*、*波紋*並非沒有出現，由於這此詞彙所佔比例太低且落於前 20 名單之後，取而代之的主宰支配共有文字則爲*表面*、*部分*、*事物*、*個體*等一般性概念詞彙。

表 2.漣漪與鱗波兩詞由第一階層到第八階層釋義關聯結果及共用的釋義文字比例 (每階層只取前釋義詞彙所佔權重的 Top 20 列表)

關聯層級深度	第一 階層	第二 階層	第三 階層	第四 階層	第五 階層	第六 階層	第七 階層	第八 階層
釋義關聯值(SRD)	0.4	0.714	0.9	0.964	0.989	0.997	0.999	≒1
(釋義文字所佔權重)	(釋義文字出現比例)							
水面	**20.00%**	**23.81%**	**20.00%**	**12.97%**	6.97%	3.39%	1.59%	-
波紋	**20.00%**	**14.29%**	**8.00%**	**3.60%**	-	-	-	-
表面	-	4.76%	**9.00%**	**9.73%**	**7.71%**	**5.21%**	**3.29%**	**2.05%**
部分	-	-	2.00%	**4.68%**	**6.14%**	**6.17%**	**5.46%**	**4.55%**
上	10.00%	4.76%	3.00%	2.34%	1.94%	1.65%	1.43%	1.26%
紋理	-	4.76%	6.00%	4.86%	3.14%	1.85%	-	-
稱爲	-	4.76%	6.00%	4.68%	2.94%	1.79%	-	-
水	-	4.76%	6.00%	4.68%	2.86%	1.58%	-	-
水皮兒	-	4.76%	6.00%	4.68%	2.80%	1.46%	-	-
漣漪	10.00%	4.76%	2.00%	-	-	-	-	-
似	10.00%	4.76%	2.00%	-	-	-	-	-
魚鱗狀	10.00%	4.76%	2.00%	-	-	-	-	-
鱗波	10.00%	4.76%	2.00%	-	-	-	-	-
細微	10.00%	4.76%	2.00%	-	-	-	-	-
物體	-	-	2.00%	3.60%	3.71%	3.02%	2.28%	1.70%
形成	-	4.76%	5.00%	3.24%	1.63%	-	-	-

事物	-	-	1.00%	**2.16%**	**2.74%**	**2.86%**	**2.81%**	**2.70%**
外在	-	-	1.00%	2.16%	2.57%	2.30%	1.79%	1.28%
微浪	-	4.76%	4.00%	2.16%	-	-	-	-
外界	-	-	-	2.16%	2.57%	2.31%	1.81%	1.34%
現象	-	-	-	2.16%	2.57%	2.30%	1.79%	1.30%
個體	-	-	-	-	1.97%	**2.50%**	**2.51%**	**2.25%**
實質	-	-	-	1.98%	2.17%	1.80%	1.30%	-
接觸	-	-	-	1.80%	1.83%	1.42%	-	-
中	-	-	-	-	-	1.52%	1.67%	1.67%
構成	-	-	-	1.62%	1.60%	1.28%	-	-
認識	-	-	-	-	-	1.33%	1.52%	1.52%
呈現	-	-	1.00%	1.62%	1.57%	-	-	-
整體	-	-	-	-	-	-	1.46%	1.48%
人	-	-	-	-	-	-	1.30%	1.58%
某些	-	-	-	-	-	-	1.26%	1.19%
部下	-	-	-	-	-	-	1.26%	1.18%
秩序	-	-	-	-	-	-	1.26%	1.18%
局部	-	-	-	-	-	-	1.26%	1.18%
部屬	-	-	-	-	-	-	1.26%	1.18%
空間	-	-	-	-	-	1.29%	-	1.07%
紋路	-	-	-	-	1.60%	-	-	-
指	-	-	-	-	-	-	-	0.95%

本文研究旨在尋找在同義詞彙之中，以其共同擁有的概念爲內容。綜合上述分析，使用多階層釋義關聯若以深階層的計算結果爲主(如第八階層)，則實驗結果會使所有詞彙都落於一般性概念詞彙上，從而無法突顯同義詞彙較具體的概念涵義；反之，若使用淺階層(如第一階層)釋義關聯，則共同釋義詞彙又會因字型不符而無法歸類在一起。故從而則參考先前的研究結果，選定以第四階層釋義關聯進行計算，並取出現比率較高的前 20 釋義文字作爲同義詞彙的概念釋義。

5.3 多義詞處理與最大平均釋義關聯詞

在辭典釋義中，單一詞彙往往會有許多釋義協助辨析該詞條的主要涵義、用法及來源典故。但在關聯計算上，雖然可以透過處理多階層釋義解決多詞同義的問題，但這個方法卻不能解決一詞多義的問題。爲了能確定多義詞彙在同義詞組所適用的涵義，在本研究中，我們則將不同的釋義分開來計算，以一對一的詞條釋義比對原則來尋找能得到最大釋義關聯值的釋義，作爲在該同義詞組的最佳同義說明。以前述在《擴展版》中同義詞組“Bp20B03=*招子 幌子 市招*”爲例(相關在《國語辭典》中的釋義，請參照前述)。經

過釋義處理後，各別可供計算的釋義文字如下表 3。

表 3.經處理過後的同義詞組 Bp20B03=《國語辭典》釋義，及關聯計算結果

同義詞彙	經處理後的《國語辭典》釋義	釋義關聯值 SRD^4		
		幌子-1	幌子-2	市招
招子	招子-1. 花招 招兒	0.380	0.475	0.475
	招子-2. 死刑犯 刑 時 插於 背後 紙標 用來 揭示 犯人 罪狀 姓名	0.470	0.566	**0.636**
	招子-3. 招牌 廣告 海報	0.549	**0.675**	0.620
	招子-4. 隨風 長布 帘子	0.421	0.541	0.505
	招子-5. 門票	0.398	0.481	0.499
	招子-6. 用於 江湖 人物 間	0.506	0.581	0.597
	招子-7. 眼睛	0.427	0.524	0.499
幌子	幌子-1. 表現 在外 蒙蔽 他人 言行 幌子			0.662
	幌子-2. 掛 店鋪 門 外 用來 招徠 顧客 招牌			**0.813**
市招	商店 門 外 標示 其 名號 賣 貨物 招牌 標誌 市招	N/A	N/A	

在表 3 中同時比較七個詞義的“*招子*”、二個詞義的“*幌子*”與單一詞義的“*市招*”，並透過多階層釋義關聯計算列出對應表並加總求最大值。從表 3 中可以知道最適合詞的詞彙語義關係建立在“*招子*”與“*幌子*”之間，因爲“*市招*”與“*幌子-1*”、“*幌子-2*”之間在第四階層的釋義語義關聯下仍無法產生關係(沒有 *SRD*^{4} 數値 N/A)，而釋義關係最大値是建立在“*招子-3*”與“*幌子-2*”之間。在這兩兩成對的關係之中，我們計算各別平均的關係並取其最大値平均値作爲此類別的代表詞彙，則“*招子*”爲(0.675+0.636)/2=0.655、**“*幌子*”爲(0.675+0.813)/2=0.744**、“*市招*”爲(0.813+0.636)/2= 0.724，故“*幌子*”平均値在釋義概念上爲其它詞彙的釋義關聯最高，很適合作爲此分類的代表詞彙。接著，我們將出現在這三個詞彙裡第四階層共同擁有的釋義文字權重中，取最高的前 20 個釋義詞彙列表，作爲該同義詞彙之中主宰支配(dominate)整個同義詞組的主要釋義詞彙。而經計算後選取出各詞彙之間第四階層共有文字釋義，以統計各別文字的分佈(同 5.2 處理原則)，例如前 20 個較高釋義文字，順序如下(由高到低排列)：

*共有釋義文字分佈：**掛=3.12%人=1.69%招牌=1.54%牌子=1.47%***
招子=1.46%有=1.40%上=0.95%標識=0.86%揭示=0.82%
用來=0.81%單位=0.79%表示=0.78%懸吊=0.74%廣告=0.71%
賣=0.70%刑=0.68%物品=0.65%獻=0.61%計算=0.59%團體=0.59%
#總釋義文字：152296

雖然從這些釋義文字中可分析出，在此同義詞之中較多共用釋義的概念爲何。我們能羅列上述 20 個主要主宰支配這個同義詞組的共有釋義詞詞彙，但卻無法將上述的詞彙組合成爲文字，除了尚有約 15 萬個字詞沒有列出外，在使用多階層釋義關聯計算時，我們僅只保留了動詞與名詞詞性，忽略了文法結構。此外我們亦無法將所取得的共有釋義文字組合成爲單一精簡句型，但我們亦可以透過比較相似的同義字群組，幫助我們了解不同的同義詞組之間的差別。分析在《詞林》中擁有相同詞彙“*幌子*”的同義詞組“*Dd15A09=幌子招牌牌子旗號金字招牌*”同義詞組的結果(見表 4)，我們依前述方法，逐一比較同義詞組之中各詞彙的每一項釋義內容，除了使用與其它同組詞彙的第四階釋義語義關聯最大值平均值，與共有釋義詞彙最大涵蓋數來決定最適合代表的詞彙之外，表 4 中亦列出共有釋義詞彙之中的 Top 20 作爲比較。從表 4 中我們可知，“*幌子*”在此群組之中與“*招牌*”的共有釋義詞彙涵蓋度較高，在第四階釋義詞彙語義關聯值 0.975，但是與其它詞彙間的關聯就相對低。再比較表 3 之結果，我們可以說“*幌子*”釋義關聯與“*招牌*”的關係較與“*市招*”的關係較接近，接著才是“*金字招牌*”與“*招子*”。

表 4.同義詞類 Dd15A09=之各詞間釋義關聯表

	幌子	*招牌*	*牌子*	*旗號*	*金字招牌*	平均值
幌子		0.975	0.548	0.583	0.738	0.711
招牌	0.975		0.789	0.527	0.830	0.780
牌子	0.548	0.789		0.381	0.637	0.588
旗號	0.583	0.527	0.381		0.638	0.532
金字招牌	0.738	0.830	0.637	0.638		0.710
平均值	0.711	0.780	0.588	0.532	0.710	
max Average:招牌 0.780 *共有釋義文字分佈：**人=2.87%掛=2.45%**有=1.91%**牌子=1.66%表示=1.08%*** *調子=1.07%上=0.84%**招牌=0.82%**某=0.73%指=0.70%單位=0.64%他人=0.63%* *種=0.62%事物=0.61%一=0.60%懸吊=0.57%名義=0.55% 獻=0.55%具有=0.53%* *標識=0.52%* *#總釋義文字：426366*						

然而表 3 與表 4 之間亦可從與其它同組詞彙的第四階釋義語義關聯最大的平均值，與共有釋義詞彙最大涵蓋數來決定最適合代表的詞彙。在表 3 中“*幌子*”釋義爲“(1) 表現在外用以蒙蔽他人的言行。 (2) 掛在店鋪門外，用來招徠顧客的招牌。”在表 4 中“*招牌*”釋義爲“(1) 商店機構作爲標識的牌子。 (2) 演藝人員或團體揭示其所獻技藝有關事項的牌子。(3) 拿手的，可作爲標識的。(4) 比喻騙人的幌子。”從釋義之中，我們可以理解共有釋義的處理原則是將“*招牌-4*”中所出現的“*幌子*”釋義文字納入計算而得的最大值來關聯，但“*招牌*”對其它 Dd15A09=同義詞組中的詞彙關聯卻不從“*招牌-4*”而來，而是“*招牌-1*”或“*招牌-3*”。另一方面，“*幌子*”能與 Dd15A09=同義詞組中的

詞彙產生釋義關聯的，僅僅只有“*幌子-2*”釋義中，較為明確的“*招牌*”釋義而來，因此在此一同義群組之中得到的平均數值就會低許多。透過上述的說明我們可以知道，雖然最適合同義詞組的代表詞彙必需是滿足釋義語義關聯最大值平均值與共有釋義詞彙最大涵蓋數。但要同時達到這兩項條件，則該詞彙必須包涵的各階層釋義文字集合，且大多都要能出現在同一辭組之中其它詞彙的釋義文字裡，也就是要「被用來說明辭典裡的某個辭條」時，才會可以成為該辭組的最合適同義詞彙。

在計算上，並非以深層釋義關聯(如第八階層)所能取得的一般性概念詞彙進行計算，而是詞層概念層級在有限度(本研究以第四階層進行計算)的概化擴充之後，仍要滿足前述的條件。此外，詞彙間是 part-and-whole 的隱喻關係的“*金字招牌*”與“*招牌*”兩詞，在《詞林》的依同義/近義程度在同類型中由左至右排列原則之下，我們也可以得出相對於“*招牌*”一詞，“*金字招牌*”可視為近義詞而非同義關係。這項結果也可以從表 4 中看出“*招牌*”一詞可以完全可主宰此同義詞組共有釋義，而“*金字招牌*”卻無法主宰看出 (因為兩詞彙在高度釋義相關的條件下，應同時能主宰該同義詞組。若否，在一詞多義的條件之下，則可推論兩者相關聯的釋義並非此同義詞組的主宰釋義)。 最後，從共同擁有的釋義權重 Top 20 中，兩同義詞組所表現的釋義詞彙亦是不同的。可以看出，Bp20B03=之中從較具體的“*招牌*”，而在 Dd15A09=則從抽象的“*牌子*、*表示*”等概念涵意，此與《詞林》之分類規則上(B 類為物、D 類是抽象事物)是相同的。

雖然同義詞組經過上述計算可以得到詞組之中最適合用來表達的同義詞彙，與共同擁有釋義文字的權重比例(在此僅列出 Top 20)，但此方法亦有限制。由於此方法需透過辭典反覆對詞彙進行釋義、斷詞再擴充釋義文字，所以當無法取得釋義內容時，則會造成無法計算釋義關聯的窘境。以下參考維基百科為例。描述「訓詁學」定義一組同義詞“*訓詁訓故故訓古訓解故解詁*”，其計算結果如下表 5。從表中可以看出“*訓故*”與“*故訓*”三詞在所使用的教育部《國語詞典》中找不到釋義，因此無法計算釋義關聯，此為缺點一。而從其它三個字的四階層釋義關聯計算之後的結果，最大釋義關聯文字可以使用“*解詁*"代表，因為經過計算之後最大平均值是 0.741。這項數值與“*解故*”的 0.740 之間雖然只差千分之一，但仍無法直接將“*解詁*”與"*解故*”視為相同詞彙或概念義涵相同(因為本研究所用方法僅止於計閱釋義而非詞彙概念)。最後釋義文字比重前二十個字詞亦能表現出此同義字組的共有釋義文字，如：“*指*”、“*解釋*”、“*古代*”、“*文字*”、“*說明*”、“*分析*”等，但卻不能組織並架構成一句精簡的同義詞釋義(前文已討論過)。儘管此方法還有許多尚待改善的缺陷，但能提到詞彙之間客觀的比較基準與可供參考的釋義文字權重，相對人工訓詁方法，仍能在眾多的資料之中提供快速同義關聯參考依據。

表 5. 自建同義詞組"訓詁"及各詞間釋義關聯表

	訓詁	訓故	故訓	古訓	解故	解詁	平均值
訓詁		N/A	N/A	0.814	0.575	0.569	0.652
訓故	N/A		N/A	N/A	N/A	N/A	N/A
故訓	N/A	N/A		N/A	N/A	N/A	N/A
古訓	0.814	N/A	N/A		0.677	0.684	0.725
解故	0.575	N/A	N/A	0.677		0.971	0.740
解詁	0.569	N/A	N/A	0.684	0.971		0.741
平均值	0.652	N/A	N/A	0.725	0.740	0.741	

max Average:解詁 0.741
共有釋義文字分佈：指=4.96%解釋=2.73%手=1.98%某=1.97%古代=1.79%
個=1.72%文字=1.51%說明=1.47%人=1.40%分析=1.26%部分=1.15%
一=1.07%原因=1.04%事=0.94%希望=0.92%指示=0.92%中=0.91%
理由=0.84%消除=0.83%直立=0.82%
#總釋義文字：93720

5.4《擴展版》詞類義涵分析

在《擴展版》中將同義詞類區分爲「相等、同義」(=)共計 9995 類 55844 字/詞數、「不等、同類」(#)爲 3445 類 29893 字/詞數與「自我封閉、獨立」(@)有 4377 類 4377 字/詞數。從資料上可以看出，在「自我封閉、獨立」分類下的同義詞均爲單一字/詞彙分類，而「不等、同類」的詞彙，如「*Bg02B08# 麥浪松濤煙波*」，不僅難從辭典釋義之中找到關聯（“*麥浪*”釋義爲“*麥田中的麥苗遭風吹拂時起伏如浪的樣子*”；“*松濤*”釋義爲“*風吹松樹所發出像波濤般的聲音*”；“*煙波*”釋義爲“*雲煙瀰漫的水面。*”)，甚至在詞意概念關聯上仍需要人爲的語感協助才能分別。

因此，在本研究中僅以「相等、同義」(=)分類別進行分析。總計處理詞類 8311 類，其中同類詞組中找不到任何《國語辭典》釋義詞組計有 502 組、僅能找到單一詞彙之同義詞組爲 1182 組，共計處理的字/詞數爲 44872，釋義總計爲 126605 組。處理《擴展版》的方式如前所述，單一合適的「相等、同義」詞組取得詞彙的釋義之後，並將釋義內容送至中研院斷詞系統處理後留取合適的動詞與名詞，並反覆進行四次以計算第四階層釋義語義關聯。最後將釋義內容一對一交互比對，取得最合適的代表詞組詞彙與 Top 20 共有釋義詞彙權重，再依表 1 中《擴展版》的資料羅列如下列表 6 排版。表 6 僅列出各大項(A 組到 L 組之中的取樣結果)。

表 6.分析《擴展版》同義詞類範例結果：摘錄 A 組到 L 組各組一筆

<SYNONYM_TAG><SYNONYMS> | <WORDS in ZH-TW DICTIONARY> | <MOST REPRESENTING WORD> | <MOST APPEAR WORD IN DICTIONARY EXPLANATION>

Ab01C01=男女 士女 兒女 紅男綠女 男男女女 少男少女|男女 士女 兒女 紅男綠女 男男女女|男女=0.87 |有 女 表示 大 男女 中 人 作 者 色 相對 前 多 男 指 某 用於 一 表 事物

Bf06B01=閃電 電 銀線 電閃|閃電 電|電=0.90 |種 有 中 人 表示 其 時 事物 一 電 內 兩 單位 計算 姓 多 帶 使 指 某

Bq03C03=絨褲 衛生褲| x | x |

Cb06B01=附近 就近 鄰近 近處 一帶 內外 左右 左近 前後 近水樓台 就地 近旁 跟前 不遠處|附近 就近 鄰近 近處 一帶 內外 左右 左近 前後 就地 近旁 跟前|附近=0.82 |中 表示 其 有 言 大 稱 人 左右 處 某 一 指 姓 句 個 地方 屬 時 單位

Dk16B01=電報 電 報|電報 電 報|電報=0.94 |種 有 人 地 表示 其 時 事物 電 一 單位 內 得 中 計算 指 兩 使 上 種子

Df01B01=感想 感 感觸 感受|感想 感 感觸 感受|感觸=0.90 |中 心 內 有 人 感 種 事物 一 表示 某 影響 三 稱爲 之一 思想 姓 內心 上 指

Ef06C01=空閑 悠閑 幽閑 安閑 清閑 輕閑 閑暇 空暇 閑空 悠然 閑 空 逸 暇 空餘 得空 有空 沒事|清閑 閑暇 空暇 悠然 閑 空 暇 得空 有空 沒事|空=0.61 |有 表示 養 閑 一 某 時間 時候 閒 事物 前 樣子 指 人 存在 正面 餘數 地方 相對 用於

Fa04B01=捧 掬 端 端面 端平|捧 掬 端|捧=0.94 |手 兩 人 單位 計算 一 某 等於 種 公斤 量詞 斤 事物 詞 雙 公制 臺 表 做 部首

Ga17B01=感動 感 觸 動容 感觸 百感叢生 催人淚下 令人感動 動感情 動人心魄|感動 感 觸 動容 感觸 動人心魄|感動=0.89 |人 指 有 使 事物 物 感應 一 某 種 稱爲 表示 影響 物體 姓 個 現象 中 電 感動

Hb02C03=開戰 開仗 開火 動武 用武 動干戈 宣戰|開戰 開仗 開火 動武 用武 動干戈 宣戰|開火=0.78 |指 一 一點 中 個 開始 開戰 雙方 開火 手 稱爲 某 兩 事物 人 爭鬥 開仗 部分 有 指示

Ig04B01=循環 輪迴 循環往複 周而復始 大循環 巡迴|循環 輪迴 周而復始 大循環|循環=0.76 |有 指 人 事物 一 中 一切 承 組織 血液 動物 表示 各 佛教 種 物 稱爲 個 內 眾生

Jd08A03=失去 失掉 失卻 失 去 奪 錯過 錯開|失去 失掉 失卻 失 奪 錯過|失掉=0.47 |奪 失 錯過 遺落 做 到 失去 決定 挫 機會 脫漏 失掉 丟掉 眩目 耀眼 漏掉 爭取 人 強取 衝過

Ka21A01=借故 託故 假託 推託 借口 假說 託詞 託辭|借故 託故 推託 借口 託詞 託辭|託詞=0.83 |藉口 理由 某 人 推託 假借 有 借口 借用 字 話 自己 語言 事 別人 我 爲 事物 作爲 論說

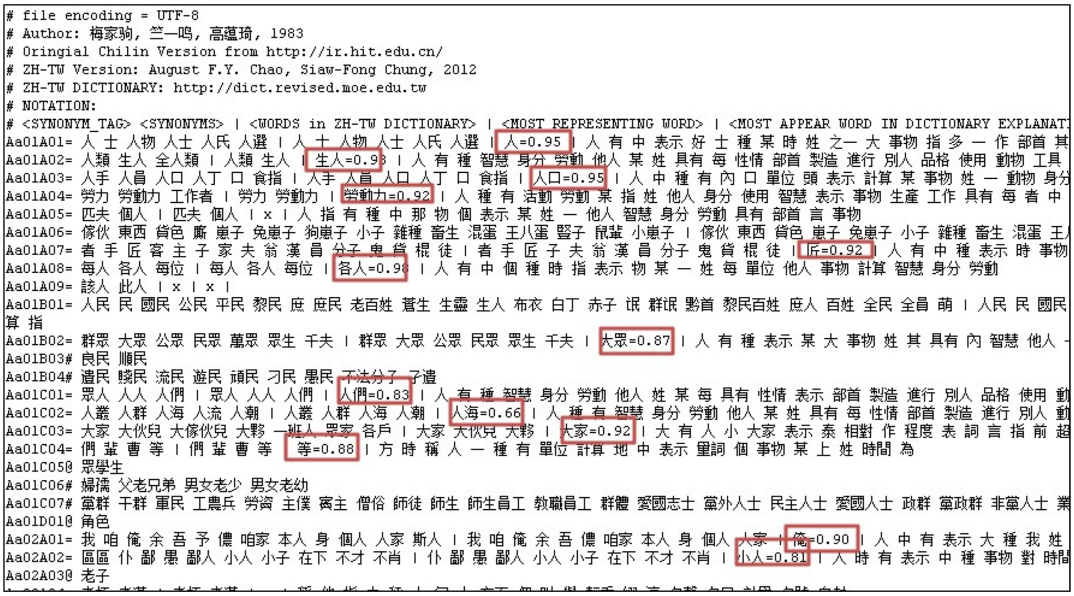
La04C01=抱歉 對不起 對不住|抱歉 對不起 對不住 |對不住=0.98 |中 心 人 有 表示 之一 姓 內 種 某 一 端 進行 性情 宿 部首 兩 名 星 事物

在表 6 之中，我們依表 1《同義詞詞林-擴展版》的原始資料保留在每行的前兩個欄位，標記爲<SYNONYM_TAG><SYNONYMS>，隨後我們將能在《國語辭典》中找到的詞彙列於<WORDS in ZH-TW DICTIONARY>欄位、最合適的代表該詞組的詞彙列於<MOST REPRESENTING WORD>，最後擷取用於計算的共有釋義詞彙權重最高的 Top

20 一併依序羅列於最後欄位之中<MOST APPEAR WORD IN DICTIONARY EXPLANATION>。從表 6 觀察在列表的資料，其中<MOST REPRESENTING WORD>一欄結果爲詞彙在表達該同義詞組字彙 ci 涵蓋率高，且其平均值也是最高的結果中，Ab01C01= 是以“*男女*”第四階釋義語義關聯值爲 0.87、Bf06B01=爲“*電*”0.90、Cb06B01=爲“*附近*”0.82、Dk16B01=爲“*電報*”0.94、Fa04B01=爲“*捧*”0.94、Ga17B01=爲“*感動*”0.89 、Hb02C03=爲“*開火*”0.78、Ig04B01=爲“*循環*”0.76、Ka21A01=爲“*託詞*’0.83、 La04C01=爲“*對不住*” 0.98 等，這些詞彙在表達該詞組上從字面上即可以了解該同義詞組的主要包涵的概念。至於無法得到較高釋義關聯數值的同義詞組，如 Ef06C01=“*空*’0.61 與 Jd08A03=“*失掉*’0.47 兩辭組，主因爲該同義詞組中包括了概念層級很上位的詞彙(Ef06C01=裡的“*空*”使其該詞彙在釋義的涵蓋度很高)，且在詞組之中加入詞彙的釋義與其它詞彙釋義的概念領域之間交集程度較低的近義詞彙(如 Jd08A03= “*奪*’有八組釋義爲“*強取*’、“*削除*、*使失去*。”、“*爭取*。”、“*做決定*。”、“*錯過*。”、“*衝過*。”、“*耀眼*、*眩目*。”、“*脫漏*、*漏掉*。”，此八組釋義與 Jd08A03=同義詞彙能建立較高釋義關聯的共有釋義詞彙是“*錯過*。”；從而使 Jd08A03=同義詞組中“*失去*”、“*失掉*”、“*失卻*’、“*失*”與“*奪*’、“*錯過*’可區分爲兩個子辭組，後者爲前者的近義詞。)。另外要說明的是，在詞典之中，同義詞找不到任何詞條釋義就無法進行關聯運算，因此在資料則會如上表 6 中的 Bq03C03= 同義詞類以"| *x* | *x* |"表示；同理，若僅只找一個詞彙則會無法計算釋義關聯，則亦無法決定最大平均釋義關聯詞，故以“| *x* |“表示。而在最後一欄的<MOST APPEAR WORD IN DICTIONARY EXPLANATION>詞組是羅列共同擁有的釋義文字權重 Top20 的詞彙，並依權重高低順序排列，雖然不見得就能構成爲可讀的句子，如表 6 之中的 Bf06B01=“*電*”與 Dk16B01= “*電報*’兩詞組在共有釋義詞彙 Top20 的重覆程度太高，因而難以區分。但可提供輔助詮釋該同義詞組更多資訊，在共有釋義詞彙 Top20 的中就可以很明顯的區別出。Ga17B01= “*感動*’及其共有釋義詞彙 Top20 中獨有的八個詞彙“*使*”、“*物*”、“***感應***’、“***物體***’、“*個*’、“*現象*’、“*電*’、“*感動*’，相對比較 Df01B01=“*感觸*’及其獨有的八個共有釋義詞彙：“***心***”、“*內*’、“*感*’、“*三*”、“*之一*”、“***思想***’、“***內心***”、“*上*”，亦可以說明《詞林》的大分類中 D 大類指是抽象事物，而 G 大類是心理的結果。

綜合前述的資料結果，我們將《同義詞詞林-擴展版》的原始資料，附加本次研究的內容結果放在 Google Code 的 tw-synonyms-chilin 的專案之中，網址爲 http://code.google.com/p/tw-synonyms-chilin/(或使用 http://goo.gl/H6YRK 直接下載處理後的文本)供其它研究人員在網站的軟體庫存專案中下載。爲了處理上述資料，我們使用了 2 台 Unix 電腦(Unbuntu: IntelCore2 Duo 2.80GHz; FreeBSD: AMD Sempron 1.8GHz)，輔以 NLTK(Loper & Bird, 2002)工具開發計算工具，進行同義詞組的多階層釋義關聯計算，與辭組之間多釋詞彙之間的交互比較，總處理時間大約 32 小時，總計處理同義詞類 8311 類，結果文件約爲 2.2MB，內容列於下表 7。

表 7.取自 tw-synonyms-chilin 之摘要結果。(方框爲最合適代表該詞組的詞彙)



```
# file encoding = UTF-8
# Author: 梅家驹, 竺一鸣, 高蕴琦, 1983
# Oringial Chilin Version from http://ir.hit.edu.cn/
# ZH-TW Version: August F.Y. Chao, Siaw-Fong Chung, 2012
# ZH-TW DICTIONARY: http://dict.revised.moe.edu.tw
# NOTATION:
# <SYNONYM_TAG> <SYNONYMS> | <WORDS in ZH-TW DICTIONARY> | <MOST REPRESENTING WORD> | <MOST APPEAR WORD IN DICTIONARY EXPLANATI
Aa01A01= 人 士 人物 人士 人氏 人選 | 人 士 人物 人士 人氏 人選 | 人=0.95 | 人 有 中 表示 好 士 種 某 時 姓 之一 大 事物 指 多 一 作 部首 其
Aa01A02= 人類 生人 全人類 | 人類 生人 | 生人=0.93 | 人 有 種 智慧 身分 勞動 他人 某 姓 具有 每 性情 部首 製造 進行 別人 品格 使用 動物 工具
Aa01A03= 人手 人員 人口 人丁 口 食指 | 人手 人員 人口 人丁 口 食指 | 人口=0.95 | 人 中 種 有 內 口 單位 頭 表示 計算 某 事物 姓 一 動物 身分
Aa01A04= 勞力 勞動力 工作者 | 勞力 勞動力 | 勞動力=0.92 | 人 種 有 活動 勞動 某 指 姓 他人 身分 使用 智慧 表示 事物 生產 工作 具有 每 者 中
Aa01A05= 匹夫 個人 | 匹夫 個人 | x | 人 指 有 種 中 那 物 個 表示 某 姓 一 他人 智慧 身分 勞動 具有 部首 言 事物
Aa01A06= 傢伙 東西 貨色 廝 崽子 兔崽子 狗崽子 小子 雜種 畜生 混蛋 王八蛋 豎子 鼠輩 小崽子 | 傢伙 東西 貨色 崽子 兔崽子 小子 雜種 畜生 混蛋 王
Aa01A07= 者 手 匠 客 主 子 家 夫 翁 漢 員 分子 鬼 貨 棍 徒 | 者 手 匠 子 夫 翁 漢 員 分子 鬼 貨 棍 徒 | 匠=0.92 | 人 有 中 種 表示 時 事物
Aa01A08= 每人 各人 每位 | 每人 各人 每位 | 各人=0.98 | 人 有 中 個 種 時 指 表示 物 某 一 姓 每 單位 他人 事物 計算 智慧 身分 勞動
Aa01A09= 該人 此人 | x | x |
Aa01B01= 人民 民 國民 公民 平民 黎民 庶 庶民 老百姓 蒼生 生靈 生人 布衣 白丁 赤子 氓 群氓 黔首 黎民百姓 庶人 百姓 全民 全員 萌 | 人民 民 國民
算 指
Aa01B02= 群眾 大眾 公眾 民眾 萬眾 眾生 千夫 | 群眾 大眾 公眾 民眾 眾生 千夫 | 大眾=0.87 | 人 有 種 表示 某 大 事物 姓 其 具有 內 智慧 他人 一
Aa01B03# 良民 順民
Aa01B04# 遺民 賤民 流民 遊民 頑民 刁民 愚民 不法分子 孑遺
Aa01C01= 眾人 人人 人們 | 眾人 人人 人們 | 人們=0.83 | 人 有 種 智慧 身分 勞動 他人 姓 某 每 具有 性情 表示 部首 製造 進行 別人 品格 使用 動
Aa01C02= 人叢 人群 人海 人流 人潮 | 人叢 人群 人海 人潮 | 人海=0.66 | 人 種 有 智慧 身分 勞動 他人 某 姓 具有 每 性情 部首 製造 進行 別人 動
Aa01C03= 大家 大伙兒 大傢伙兒 大夥 一班人 眾家 各戶 | 大家 大伙兒 大夥 | 大家=0.92 | 大 有 人 小 大家 表示 泰 相對 作 程度 表 詞 言 指 前 超
Aa01C04= 們 輩 曹 等 | 們 輩 曹 等   等=0.88 | 方 時 稱 人 一 種 有 單位 計算 地 中 表示 量詞 個 事物 某 上 姓 時間 為
Aa01C05@ 眾學生
Aa01C06# 婦孺 父老兄弟 男女老少 男女老幼
Aa01C07# 黨群 干群 軍民 工農兵 勞資 主僕 賓主 僧俗 師徒 師生 師生員工 教職員工 群體 愛國志士 黨外人士 民主人士 愛國人士 政群 黨政群 非黨人士 業
Aa01D01@ 角色
Aa02A01= 我 咱 俺 余 吾 予 儂 咱家 本人 身 個人 人家 斯人 | 我 咱 俺 余 吾 儂 咱家 本人 身 個人 人家 | 俺=0.90 | 人 中 有 表示 大 種 我 姓
Aa02A02= 區區 仆 鄙 愚 鄙人 小人 小子 在下 不才 不肖 | 仆 鄙 愚 鄙人 小人 小子 在下 不才 不肖 | 小人=0.81 | 人 時 有 表示 中 種 事物 對 時間
Aa02A03@ 老子
```

在表中亦將使用第四階釋義語義關聯值，以選出的最合適代表該詞組詞彙並以方框標示，從而可以比較在原始《詞林》編排方式下，同義詞群群首與最適合代表該詞組詞彙之間表達該辭群的義涵。如 Aa01A04=中，群首詞彙爲“*人手*”(有“*他人的手。辦事的人。*”兩釋義)與最適合代表詞彙“*人口*”(有“人。家族或家中的人數。人的嘴巴。指言語議論。一定時間內一地區具有戶籍身分的全部居民。”五釋義)；Aa01A07=中，群首詞彙爲“*者*”(有“*人或事物的代稱。指示形容詞。用於句中，表示停頓。用於句末，表示語氣結束。表比擬。*”六釋義)與最適合代表詞彙“*匠*”(有“*泛稱各種技術工人。尊稱在某方面有特殊造詣的人。技藝靈巧、構思巧妙。*”三釋義)。

5.5 與 Sketch Engine 比較

Sketch Engine (http://the.sketchengine.co.uk) (Kilgarriff *et al*., 2004)是語料庫處理系統，主要的功能是使用 KWIC (key word in context) 出現頻率及分佈，結合中文語法關聯 (grammatical relations, gramrels) 分析，計算文字共同出現行爲的統計結果，以提供索引 (concordance)、詞彙列表 (word list)、詞彙速描 (word sketch)、同近義詞 (thesaurus) 等功能 (Huang *et al*., 2005)。我們以“*招牌*”爲例，使用 zhTenTen 語料庫取得其同義詞，並比較本研究所使用的釋義語義關聯原則所表示的關聯程度，與使用 Sketch Engine 中語法模式及文字共同出現行爲所尋找的同義詞彙進行比較。選用的 zhTenTen 語料庫是由程式自動抓取網路上簡體字中文文本後，使用 Stanford Chinese Word Segmenter 及 Chinese Penn Treebank standard 模式所建立的 Stanford Log-linear Part-of-speech Tagger 原則處理的語料庫，目前約有 20 億個字，合計 17 億不重覆字在語料庫中。在 Sketch Engine 之中，我們以“*招牌*”詞彙查詢 Thesaurus 功能中的 Find Similar Words，“*招牌*”在 zhTenTen

之中出現詞頻爲 10185，且得到相似詞彙共 60 個。接著我們使用相同的維基百科的繁簡分歧詞表進行繁簡轉換，將取得的簡體詞彙轉換成繁體詞彙後，使用《國語辭典》進行釋義，並一一與“*招牌*”計算第四階層釋義語義關聯值，區別 Sketch Engine 所得到的相似詞彙與本研究中所指的釋義語義相關之間的差異(結果見下表 8)。

表 8. Sketch Engine 與釋義關聯計算結果比較 (部分刪除)

Sketch Lemma	Sketch Scores	Sketch Freq.	SRD-Scores	SRD-共有辭典釋義字出現比率 Top 10
牌匾	0.21	5421	0.79	有 字 表示 題 前 單位 提 記錄 人 題目
廣告牌	0.15	4839	0	
標語	0.15	19510	0.43	宣傳 宣布 廣告 宣揚 說明 講解 傳達 大眾 公布 文字
橫幅	0.12	15051	0.35	繪畫 吊掛 懸掛 書法 字畫 筆 作品 文字 藝術 筆畫
喜歡	0.12	247564	0.3	高興 事情 決定 根據 歡喜 快樂 興致 興趣 愉悅 愉快
牌子	0.12	18442	0.79	調子 音 牌子 大調 說話 程度 高低 表示 音調 時
可口	0.12	6800	0.57	美 有 人 使 好 野 表示 變 事物 好看
條幅	0.11	5848	0.66	組 指 單位 組織 物 機關 人事 事物 人 中
喜愛	0.11	47276	0.08	愛好 喜歡 高興 喜好 喜愛 快樂 自愛 事情 事物 根據
櫥窗	0.10	7917	0.62	事物 物 者 人 媒介 指 中 一切 現象 各
名片	0.10	18497	0.3	電影 影片 人 膠片 名聲 供 活動 底版 響亮 人物
標牌	0.10	5443	0	
海報	0.10	15672	0.62	指 大家 人 一 稱爲 有 個 中 上 種
廣告	0.10	212258	0.67	事物 有 人 一 上 觀念 精神 表示 指 意識
門面	0.10	5216	0.65	人 體面 門面 個人 指 身分 有 面子 上 一
餐館	0.09	13639	0.73	人 供 者 有 受 說 別人 表示 他人 智慧
~~過長刪除~~~				
字號	0.07	13197	0.93	人 牌子 招牌 商店 名稱 標識 字號 獻 有 號碼
出名	0.07	10320	0.38	名 具名 出名 出面 人 單位 簽名 計算 稱號 署名
~~過長刪除~~~				
做	0.07	1650329	0.68	前 做 人 某 我 進行 事物 製造 自稱 詞
特色	0.07	566599	0.66	地方 事物 物 當地 人 某 客觀 一切 指 存在
熟悉	0.07	141431	0.22	知道 明白 詳細 道理 仔細 細節 瞭解 明曉 詳情 熟悉

對聯	0.07	6769	0.62	兩 單位 計算 一 人 個 量詞 物 等於 公斤
獨特	0.07	135938	0.07	獨有 占有 特殊 僅有 指 特別 意思 不同於 所有 占據
歡迎	0.07	196262	0.62	他 指 人 方面 個 綴 這 別的 第三人 中
亮點	0.07	64220	0	
促銷	0.07	39704	0.67	有 人 存 使 事物 表示 某 各 意識 令
精緻	0.07	23820	0.28	細密 東西 仔細 文明 精緻 周詳 小史 紅樓夢 精深 東
形象	0.07	274352	0.72	人 指 事物 有 個 中 實體 物 一 表示
吃	0.07	433702	0.65	說話 說 事物 樣子 短 唐 講 事情 今 一

在表 8 中，前三欄分別爲 Sketch Engine 所提供語法模式及文字共同出現行爲，找出,與“*招牌*”相似的還原字詞(Sketch Lemma)、分數(Sketch Score)與頻率(Sketch-Freq.)。後兩欄爲使用(第四階層)多階層釋義關聯計算的結果(SRD Score)與 SRD-共有辭典釋義字出現比率。表 8 是依 Sketch Score 由大而小排列，所以可以看到在 Sketch Score 相對較高的詞語中，詞彙行爲雖與「*招牌*」相似，但字義上與“*招牌*”無關的詞彙，如：“*喜歡*、*可口*”等。而 SRD Score 值較高的詞彙(粗體線外框)，則可看出與“*招牌*”義涵上有較高的同義，且共有的辭典釋義字亦羅列於後。如“*招牌*”與“*字號*”產生釋義關係詞彙較高權重值的 Top 10 是“*人牌子招牌商店名稱標識字號獻有號碼*”等，更清楚地知道兩詞彙是透過“*牌子*”、“*招牌*”等進行釋義關聯。

雖然前述兩種方法都能取得兩詞彙相似關聯參考指標，但使用辭典釋義字進行同義關聯探究的方法與Sketch Engine方法不同在於：(1) 兩者的相似關係所建立的原則不同：Sketch Engine 中的同義關聯的建立是透過詞彙在許多語料庫之中的出現頻率、文法結構樣式與造句行爲來決定，而本研究所使用的方式是以辭典爲基礎的共用釋義字詞觀點出發，同義詞彙則是建構在使用相同釋義字所佔的比率多寡而決定。但比較下，使用多階層釋義語義關聯計算原則下的詞彙相關性，相較於使用 Sketch Engine 所得到的結果，更能令人直覺理解期涵義。因爲計算過程中釋義不斷進行詞彙擴充，並計算兩者間的共有釋義詞彙，因此尋找出的詞彙也較 Sketch Engine 易於了解。(2) 相似詞表產生方式：Sketch Engine 透過文法模式與統計值產生相似詞表，可以依模式尋找符合的詞彙而產生列表。然而使用多階層語義關聯計算原則，因受階層變數決定所計算的兩詞彙之間概念深淺而定，且需經過詞彙間交互比較計算後才可得到結果，即需要將辭典中的所有詞條進行交互比較。若以目前實驗中所使用的《國語辭典》爲例，是在 15 萬條詞條、27 萬個不同釋義之中交互比較，且在階層計算原則下會產生指數增加的計算負荷(computing loading)，所以這確實一樣大工程。在本實驗中要計算 Dd15A09=第一到第四階層大約都在 1 秒之內、第五階層約在 1.24 分鐘而第六階層則約需要 5 分鐘。雖然釋義語義關聯計算方法不適合用在產生同義字表上，但在計算已知的同義詞或兩生詞之間的關聯，因爲能提供兩詞彙之間較能令人理解的釋義關聯，相較 Sketch Engine 方法是更爲合適的。

6. 結論與討論

在本文中，我們已經討論多階層釋義關聯在計算同義詞彙上的應用，也比較與現有的Sketch Engine 中 Thesaurus 計算原則上的差別。在本研究中，爲了避免釋義詞彙在多階層的詮釋後而太過發散(無法收斂)，從而使用修正後的多階層釋義關聯計算方法。將較淺階層的釋義詞彙權重增加，以減少在深層釋義之中泛一般性的概念詞彙影響，以突顯兩詞彙之間的共同擁有釋義詞彙的特色，並建立關聯。關聯值的計算，是將兩詞彙間共同擁有的釋義字詞出現佔有比率，來表示共同釋義文字概念的交集，並使用反覆釋義的多階層原則，以減少釋義文字同義不同型的問題，同時利用階層釋義文字比率作爲釋義詞彙參與計算中的權重。此多階層統計中的共有釋義文字權重，可視爲解釋詞彙之間共同擁有的釋義內涵，作爲兩詞彙間關聯描述使用。

在完成多階層釋義關聯原則後定義，我們以處理詞彙間一詞多義的情況，使用多階層釋義關聯的最大值，以及共有釋義文字的涵蓋程度，來決定哪一組釋義內容適合作爲同義詞組間一詞多義代表。並且以此方式將《擴展版》中「相等、同義」詞類進行實驗，其結果雖然受限於《國語辭典》無法完全釋義《擴展版》中的各項詞彙，但我們將現有的資料進行標記，區別出《擴展版》中《國語辭典》所擁有的辭條，並將最適合詮釋同義詞組與共有釋義詞彙權重最高的 Top 20 筆資料，整理出 8311 筆合併於《擴展版》中，並放於網路上供研究者參考。

最後，爲更了解多階層釋使用在同義詞組的計算上有什麼區別，我們亦比較透過計算龐大語料庫 (如：Sketch Engine) 所取得的同義詞彙，與本研究方法的同義關聯之間的差別。雖然多階層釋義語義關聯方法無法如 Sketch Engine 進行大語料庫中詞語的計算後產生相似字表以供查詢，但可以作爲 Sketch Engine 計算取得結果之後，同義詞後的釋義比較。

多階層釋義關聯基於辭典釋義計算，雖然辭典內容的編寫用字會影響計算結果，但因中文釋義能提供詞彙語義中所包括的概念內容，從而使已知的兩詞彙間進行同義概念的探討時，釋義語義關聯相較於文法、詞頻與共現次數，較能取得更好的同義結果。雖然概念表達是由字/詞彙組合而成，而釋義內容所使用的文字雖不必然會與概念組成文字相同，但透過多階層的釋義比較下，亦能對多義詞進行同義歸納。辭典與《詞林》的編撰都是艱巨的文學語料工作，在過去的人工的相互訓詁工作之中，我們希望能應用電腦輔助語料工具方法，協助編撰者能對釋義內容進行整理、校對。亦希望利用人工釋義的內容，能與知識概念，如 HowNet 或 Chinese WordNet，進行交互比對，使釋義關聯計算能直接使用釋義概念進行比較，從而讓詞彙之間的同義關係可以更加清楚。

References

Huang, C.R., Kilgarriff, A., Wu, Y., Chiu, C.M., Smith, S., Rychly, P., Bai, M.H., & Chen, K.J. (2005). Chinese Sketch Engine and the Extraction of Grammatical Collocations, In

Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju island, Korea, 48-55.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D.(2004). The Sketch Engine, *Information Technology Research Institute Technical Report*, ITRI-04-08.

Loper, E. & Bird, S. (2002). NLTK: The Natural Language Toolkit, In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, *1*, 63-70.

Thesaurus Entry，https://trac.sketchengine.co.uk/wiki/SkE/Help/PageSpecificHelp/Thesaurus, last visited 2012/6/30.

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL, In *Proceedings of the Twelfth European Conference on Machine Learning*, 491-502.

王建莉(2012)。論爾雅的同義詞詞典性質。*辭書研究*，(02)，60-65。

中研院斷詞系統，http://ckipsvr.iis.sinica.edu.tw/, last visited 2012/6/27.

仝文奕，郭聖林(2012)。“淺談的士”及其同義詞群的競爭與選擇。*前沿*，02，153-154。

《同義詞詞林》擴展版，http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162, last visited 2012/6/17.

周亞民，黃居仁(2005)。漢字意符知識結構的建立。*第六屆漢語詞彙語義學研討會論文集*。

范紅麗(2011)。《左傳》中跪拜義同義詞群考察，*西南科技大學學報(哲學社會科學版)*，*28*(5)，93-97。

林頌堅(2004)。基於術語抽取與術語叢集技術的主題。*Computational Linguistics and Chinese Language Processing*, *9*(1)，97-112。

重編國語辭典修訂本，http://dict.revised.moe.edu.tw/, last visited 2012/6/17.

陳光華、莊雅蓁(2001)。應用於資訊檢索的中文同義詞之建構。*中國圖書館學會會報*，*67*，93-108。

梅家駒、竺一鳴、高蘊琦與殷鴻翔(1983)。編纂漢語類義詞典的嘗試-《同義詞詞林》簡介。*辭書研究*，*1983*(01)，133-138。

黃侃述、黃悼編(1983)。《文字聲韻訓詁筆記》，*上海古籍出版社*，1983 年 4 月版，190。

曾慧馨、劉昭麟、高照明與陳克健(2002)。以構詞與相似法爲本的中文動詞自動分類研究。*International Journal of Computational Linguistics and Chinese Language Processing*，*7*(1)，1-28。

趙逢毅與鍾曉芳(2011)。基於辭典詞彙釋義之多階層語義關聯程度計量-以「目」字部爲例。*中文計算語言學期刊*，*16*(3-4)，21-40。

維基百科，繁簡分歧詞表，http://zh.wikipedia.org/zh-hant/Wikipedia:繁简分歧词表，last visited 2012/6/27.

劉挺、車萬翔。中文語義處理，http://ir.hit.edu.cn/，last visited 2012/12/06.

鮑克怡(1983)。漢語類義詞典探索《同義詞詞林》編後。*辭書研究*，(02)，64-70。

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.

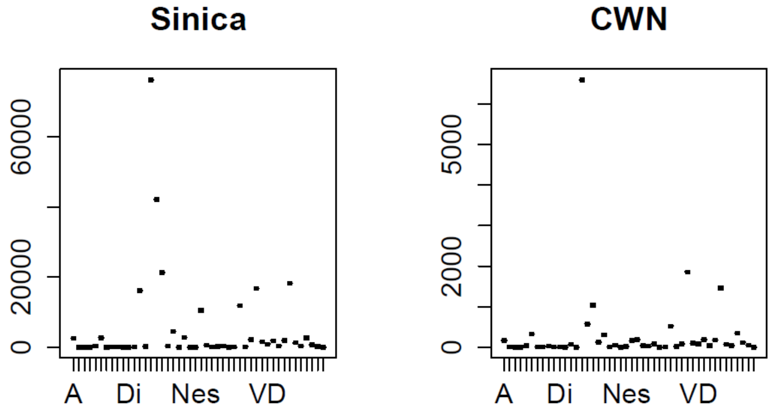


Figure 1. The POS distribution of the Sinica Corpus and the Chinese Wordnet

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* (2nd 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

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：__________________

Name：______________________________________ Date of Birth：_________________

Country of Residence：______________________ Province/State：______________________

Passport No.：_______________________________ Sex: __________________________

Education(highest degree obtained)：____

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：_________________________________ Fax No：______________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/_____/_____ （Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.- （NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號			（由本會填寫）
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
			申請人:		（簽章） 中　華　民　國　　年　　月　　日

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: ___________________________________(Please print)   Date: _______________

Please debit my credit card as follows: US$ _____________

❑ VISA CARD   ❑ MASTER CARD   ❑ JCB CARD   Issue Bank:__________________

Card No.: ________ -_________-________ -_________   Exp. Date:____________(M/Y)

3-digit code: _________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: ___________________________________

Phone No.: _________________________E-mail: ___________________________________

Address: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: ______

US$ _____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: ______

US$ _____ ❑ Publications:________________________________

US$ _____ ❑ Text Corpora: ______________________________

US$ _____ ❑ Speech Corpora:_____________________________

US$ _____ ❑ Others: __________________________________

US$ _________ ❑ Membership Fees   ❑ Life Membership   ❑ New Membership ❑Renew

US$ _________ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP
％ IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會
信用卡付款單

姓名: ______________________________(請以正楷書寫)　　日期:：_____________________

卡別：□ VISA CARD　　□ MASTER CARD □ JCB CARD　　發卡銀行：_______________

信用卡號：__________-__________-__________-__________　　有效日期：__________(m/y)

卡片後三碼：_______________（卡片背面簽名欄上數字後三碼）

持卡人簽名：____(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：______________________________E-mail：____________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$___________ □ 中文計算語言學期刊(IJCLCLP) ________________

NT$___________ □ Journal of Information Science and Engineering (JISE)

NT$___________ □ 中研院詞庫小組技術報告______________________

NT$___________ □ 文字語料庫 ______________________________________

NT$___________ □ 語音資料庫 ______________________________________

NT$___________ □ 光華雜誌語料庫1976~2010

NT$___________ □ 中文資訊檢索標竿測試集/文件集

NT$___________ □ 會員年費：□續會　　　　□新會員　　　□終身會員

NT$___________ □ 其他: ______________________________________

NT$________________ =　合計

填妥後請傳真至 02-27881638 或郵寄至:
11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：______________________________ 收據抬頭：____________________________________

地　　址：____

電　　話：________________________________ E-mail: __________________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright： It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.18 no.2 June 2013

Contents

Special Issue Articles:

Chinese Lexical Resources: Theories and Applications

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing