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 Core, Linguistics Abstracts, and ACL Anthology.

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.16 No.3-4 September/December 2011 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

Advisory Board

Jason S. Chang
National Tsing Hua University, Hsinchu

Hsin-Hsi Chen
National Taiwan University, Taipei

Keh-Jiann Chen
Academia Sinica, Taipei

Sin-Horng Chen
National Chiao Tung University, Hsinchu

Ching-Chun Hsieh
Academia Sinica, Taipei

Chu-Ren Huang
The Hong Kong Polytechnic University, H. K.

Lin-Shan Lee
National Taiwan University, Taipei

Jian-Yun Nie
University of Montreal, Montreal

Richard Sproat
University of Illinois at Urbana-Champaign, Urbana

Keh-Yih Su
Behavior Design Corporation, Hsinchu

Chiu-Yu Tseng
Academia Sinica, Taipei

Hsiao-Chuan Wang
National Tsing Hua University, Hsinchu

Jhing-Fa Wang
National Cheng Kung University, Tainan

Kam-Fai Wong
Chinese University of Hong Kong, H.K.

Chung-Hsien Wu
National Cheng Kung University, Tainan

Editorial Board

Yuen-HsienTseng **(Editor-in-Chief)**
National Taiwan Normal University, Taipei

Kuang-Hua Chen ***(Editor-in-Chief)***
National Taiwan University, Taipei

Speech Processing

Hung-Yan Gu **(Section Editor)**
National Taiwan University of Science and Technology, Taipei

Berlin Chen
National Taiwan Normal University, Taipei

Jianhua Tao
Chinese Academy of Sciences, Beijing

Hsin-Min Wang
Academia Sinica, Taipei

Yih-RuWang
National Chiao Tung University, Hsinchu

Information Retrieval

Pu-Jen Cheng **(Section Editor)**
National Taiwan University, Taipei

Chia-Hui Chang
National Central University, Taoyuan

Hang Li
Microsoft Research Asia, Beijing

Chin-Yew Lin
Microsoft Research Asia, Beijing

Shou-De Lin
National Taiwan University, Taipei

Wen-Hsiang Lu
National Cheng Kung University, Tainan

Shih-Hung Wu
Chaoyang University of Technology, Taichung

Linguistics & Language Teaching

Shu-Kai Hsieh **(Section Editor)**
National Taiwan University, Taipei

Hsun-Huei Chang
National Chengchi University, Taipei

Meichun Liu
National Chiao Tung University, Hsinchu

James Myers
National Chung Cheng University, Chiayi

Jane S. Tsay
National Chung Cheng University, Chiayi

Shu-Chuan Tseng
Academia Sinica, Taipei

Natural Language Processing

Jing-Shin Chang **(Section Editor)**
National Chi Nan University, Nantou

Sue-Jin Ker
Soochow University, Taipei

Tyne Liang
National Chiao Tung University, Hsinchu

Chao-Lin Liu
National Chengchi University, Taipei

Jyi-Shane Liu
National Chengchi University, Taipei

Jian Su
Institute for Infocomm Research, Singapore

Executive Editor: *Abby Ho*

English Editor: *Joseph Harwood*

The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

The editing of this journal is subsidized by Center for Humanities Research National Science Council in 2011.

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

© The Association for Computational Linguistics and Chinese Language Processing

International Journal of Computational Linguistics and Chinese Language Processing is published four issues per volume by the Association for Computational Linguistics and Chinese Language Processing. Responsibility for the contents rests upon the authors and not upon ACLCLP, or its members. Copyright by the Association for Computational Linguistics and Chinese Language Processing. All rights reserved. No part of this journal may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical photocopying, recording or otherwise, without prior permission in writing form from the Editor-in Chief.

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. 16, no. 3-4, September/December 2011

Contents

Papers

English Article Errors in Taiwanese College Students' EFL Writing………………………………………………………………. 1
Neil Edward Barrett and Li-mei Chen

基於辭典詞彙釋義之多階層釋義關聯程度計量—以「目」字部爲例……………………………………………........................…. 21
趙逢毅、鍾曉芳

Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction........……..…….….........................…. 41
Bor-Shen Lin and Yi-Cong Chen

Intent Shift Detection Using Search Query Logs………………… 61
Chieh-Jen Wang and Hsin-Hsi Chen

Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents...……..…….…....…. 77
Darren Hsin-Hung Lin and Shelley Ching-Yu Hsieh

Reviewers List & 2011 Index..……..…………………....…....…. 107

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 1-20
© The Association for Computational Linguistics and Chinese Language Processing

English Article Errors in Taiwanese College Students' EFL Writing

Neil Edward Barrett* and Li-mei Chen*

Abstract

The English articles, *the, indefinite a/an*, and *zero* can often be troublesome for English language learners to master, especially in longer texts. Thomas (1989) demonstrated that English as a second language (L2) learners from first languages (L1) that do not have the equivalent of an article system encounter more problems using articles. Ionin and Wexler (2004) found that such learners fluctuate between the semantic parameters of definiteness and specificity. This study examines English L2 article use with Taiwanese English learners to determine the potential factors influencing English article substitution and error patterns in their academic writing. This corpus-based analysis used natural data collected for the Academic Writing Textual Analysis (AWTA) corpus. A detailed online tagging system was constructed to examine article use, covering the semantic (specific and hearer knowledge) as well as the other features of the English article. The results indicated that learners overused both the definite and indefinite articles but underused the zero article. The definite article was substituted for the indefinite article in specific environments. Although no significant difference existed between specific and non-specific semantic environments in zero article errors, a significant difference emerged between plural and mass/non-count nouns. These results suggest that, in regard to writing, learners need to focus on the semantic/pragmatic relationships of specificity and hearer (or reader) knowledge.

Keywords: Definite Article, Indefinite Article, Zero Article, Hearer Knowledge.

1. Introduction

The use of cohesive devices in writing is a well-researched topic in second language acquisition research, taking on a greater significance in recent years as increasingly more

* Department of Foreign Languages and Literature, National Cheng Kung University, 1 University Rd., Tainan, Taiwan. Telephone: (06)2757575 ext. 52231
E-mail: leemay@mail.ncku.edu.tw
The author for correspondence is Li-mei Chen.

students are being asked to present their work in English, thus pointing to the need for greater accuracy and cohesion. Errors within the article system (*i.e.*, *a*, *an*, *the*, and the zero article) have been noted in studies examining L2 learners' writing, and such errors are present in advanced learners' texts as well (Lee, 2007). To put this issue into perspective, a corpus study of 668 TOEFL essays from Chinese, Japanese, and Russian students found that 13% of sentences-or 1 in every 8 noun phrases-had article errors (Han, Chodorow, & Leacock, 2006).

In written discourse, the omission of an article or the use of the wrong article may cause some ambiguity for the reader, especially when the writer wants to identify a noun anaphorically/cataphorically or assume reader/writer familiarity. Halliday and Hasan (1976), in their time-honored investigation into cohesion, pointed out that, "Whenever the information is contained in the text, the presence of an article creates a link between the sentence in which it occurs and that containing the referential information; in other words, it is cohesive" (p. 74). Therefore, the use of articles creates an understanding between the writer and reader, enabling the reader to locate where a noun or noun phrase is located as well as identify if it is already understood as known by the reader.

In addition to the above, the use of the English article involves the integration of semantic, pragmatic, and grammatical functions, as no one-to-one form-function mapping exists for each article, creating a large number of rules for students to master. In terms of native Mandarin-speaking English learners, article errors have been found to be cohesive writing errors in research by Chen (2002), Chiang (2003), and Ting (2003).

Research into article errors has revealed that English article errors may be due to an inability to acquire the semantic feature of specificity (Ionin & Wexler, 2004; Snape, 2006), resulting in the overuse of the definite article in specific environments. On the other hand, it may be a pragmatic deficit (Diez-Bedmar & Papp, 2008) when learners overuse the definite article due to extra-linguistic features, such as world knowledge. Other studies (Goto-Butler, 2002; Snape, 2008; Yoon, 1993) have investigated noun countability in terms of its influence on article errors.

Although previous research has examined Mandarin English L2 article use in spoken discourse (Moore, 2004; Robertson 2000) or article use in a cloze test (Lee, 2007; Snape, 2009) , only Diez-Bedmar and Papp (2008) have investigated texts from native Mandarin speaking English learners. They concluded that native Mandarin-speaking English learners have both a grammatical and a pragmatic deficit. Nevertheless, in their study, the essays were completed with time restrictions placing constraints on the writer, which may have resulted in more article errors.

The aim of this study is to identify the features that influence students' article use or misuse. We first noticed the frequency of article errors in undergraduate writing while tagging cohesion errors for the Academic Writing Textual Analysis (AWTA) corpus, an online corpus of Taiwanese undergraduate writing. Although the article errors did not seriously impair communication, they interrupted the cohesion of the writing. Consequently, it was felt that the reasons for these errors deserved further attention.

To investigate the factors that influence article errors, this study asks the following questions:

1. What is the influence of specificity and definiteness on the English article substitution and error patterns in the academic writing of Taiwanese college students?
2. What other potential factors influence English article substitution and error patterns?

2. Literature Review

English has three articles, the definite, indefinite, and zero, which have a wide range of semantic and syntactic functions in discourse (Moore, 2004). A widely-used theory related to English article use is the semantic wheel (Bickerton, 1981). According to Bickerton, English noun phrases (NPs) can be classified according to two semantic features: specific reference [+/-SR] and hearer knowledge [+/-HK]. Table 1 illustrates the four NPs. Many studies have shown that the failure to recognize [HK] has led to article errors in article production tasks (Lee, 2007; Robertson, 2000) and cloze tests (Goto-Butler, 2002; Trenkic, 2008).

Table 1. Bickerton's noun phrase environments (Goto-Butler, 2002, p. 478)

Noun phrase environment	Example
[−SR, +HK], (*the, a, zero*): Generics.	*A* cat likes mice. *The* whale is a mammal. *(zero)* Language is a great invention of human kind.
[+SR, +HK], (*the*): Unique, previously mentioned, or physically present referents.	When I found a red box in front of my house, it was too late. *The* box blew up with a terrific explosion. This book did not sell well even though *the* author was a famous writer.
[+SR, −HK], (*a, zero*): First-mention NPs or NPs following existential "has/have" or "there is/are."	There is *a* new version of the I-phone. Did you see it? I keep sending *(zero)* messages to him.
[−SR, −HK], (*a, zero*): Equative NPs or NPs in negation, question, or irrealis mood.	He used to be *a* lawyer. *(zero)* Foreigners would come up with a better solution to this problem.

A more recent development in article system research was proposed by Ionin and Wexler, (2004). Based on their studies of Russian and Korean-two languages that do not have an article system-these authors proposed that articles are governed by semantic parameters. Their theory is known as the article choice parameter.

Evidence for this comes from languages such as Samoan, which has different articles to indicate if a NP is specific or non-specific. English does not have the [+/-specific] setting, but instead has the definiteness setting [+/- definiteness]. Samoan uses the article *le* with specific noun phrases and *se* with non-specific, but does not mark definiteness (Ionin & Wexler, 2004).

The Samoan data analyzed by Ionin and Wexler demonstrate that definiteness may be irrelevant in languages like Samoan. Thus, the authors proposed the article choice parameter, which states that, "A language that has two articles distinguishes them as follows: The Definiteness Setting: Articles are distinguished on the basis of definiteness; The Specificity Setting: Articles are distinguished on the basis of specificity" (Ionin & Wexler, 2004, p. 12).

For [–] article languages, the authors proposed the fluctuation hypotheses, which states that learners fluctuate between the two parameter settings until they have enough input and the settings stabilize. Moreover, L2 learners may adopt parameter settings not found in their L1 or their L2 because, if an L2 learner lacks articles in his/her L1, no language transfer should occur as there should be no parameter preference (Ionin & Wexler, 2004). Thus, if languages, such as Mandarin Chinese, are seen as having neither articles for definiteness nor specificity, learners should fluctuate between the two settings for definite and specific reference. Based on this, Ionin & Wexler (2004) made specific predictions for [–] article L2 learners (see Table 2).

Table 2. Definite and Indefinite Fluctuation Hypothesis Predictions (Snape, 2009, p. 32)

Semantic type	+ definite	-definite
+ specific	Correct use of *the*	Overuse of *the*
-specific	overuse of *a*	Correct use of *a*

Although studies indicate that the fluctuation hypothesis correctly predicts L2 output (Snape, 2009), it has been criticized for several reasons. First, the fluctuation hypothesis does not take the zero article into account. For many first mention mass and plural nouns, specificity-as in first mention singular nouns-can be a semantic feature of zero article NPs, so the fluctuation hypothesis should also be able to predict these error types. Furthermore, Snape (2008) pointed out that, in both his and in Ionin and Wexler's studies, individual patterns among participants do not fit into either the definiteness pattern used by L1 English or the proposed fluctuation patterns. Instead, individual learners showed a miscellaneous pattern, whereby article errors occur in all four semantic types [+/-definite, +/-specific].

2.1 The Definite Article in English

Hawkins (1978) initially based his location theory on previous article studies and subsequently revised his theory (Hawkins, 1991). Hawkins identified eight different types of definite articles. By using *the,* a writer or speaker asks the reader/listener to locate the referent using knowledge that is available in the text (anaphoric and associative anaphoric use), can be sensed in the vicinity (visible and immediate situation use), or is available from local or general knowledge (immediate and local situation use). The other types of use-what Hawkins (1978) called 'structural information,' which refers to prepositional phrases, relative clauses, or adjectives-help locate the referent.

In 1991, Hawkins revised his location theory based on theories of pragmatics developed by Grice (1989). According to Hawkins, the referents are located in pragmatic sets (p-sets) that are available to the speaker/hearer via discourse sets that contain information about a certain situation or event. These p-sets are associated knowledge shared by the discourse participants and can be accessed from present or prior discourse, the local environment, shared knowledge, or general knowledge. The main point of the p-sets is that they allow the hearer or reader to accept information as definite.

2.2 Definiteness in Mandarin Chinese

A major difference between English and Mandarin Chinese is that English is a language that uses articles to show that a noun phrase is definite/or indefinite whereas Mandarin generally lacks articles (Snape, 2009). In Mandarin Chinese, a bare noun (with no classifier, demonstrative, or numeral) can be definite, indefinite, or generic. Classifiers can be defined as: "One of a set of specialized grammatical form constituents of certain types of noun phrases, especially those containing numerals, the choice of classifier being determined by the semantic characteristics of the head noun" (Trask, 1995, p.44). For example, classifiers include *ge* 個 and *ke* 棵 and they are a salient feature of Mandarin Chinese.

According to Cheng and Sybesma (2005) this semantic reading is dependent on the predicate. The following examples illustrate this (Cheng & Sybesma, 2005):

1. *Hu2fei1 mai3shu1 qu4 le* 胡飛買書去了 Hufei buy book go = Hufei went to buy a book/books (indefinite).

2. *Hu2fei1 he1 wan2-le tang1* 胡飛喝完了湯 Hufei drink-finished soup = Huefei finished the soup (definite).

3. *Wo3 xi3huan1 gou3* 我喜歡狗 I like dog = I like dogs (generic).

In preverbal position, bare noun phrases receive a definite or generic interpretation. Noun phrases with a classifier but no numeral only receive a nonspecific interpretation.

4. *wo3 xiang3 mai3 ben3 shu1* 我想買本書 I want buy CL book = I would like to buy a book

(any book, nonspecific).

Noun phrases with both a number and a classifier can have either a specific or non-specific reading.

5. *wo3 xiang3 mai3 yi1-ben3 shu1* 我想買一本書 I want buy one-CL book = I would like to buy a book. (non specific)

6. *Ta1 he1-wan2-le yi1-wan3 tang1* 他喝完了一碗湯 He drink-finished one-CL soup = a finished one/a bowl of soup (specific). (Cheng & Sybesma, 2005).

Definiteness in Chinese can be marked by a demonstrative and a numeral (Li & Thompson, 1981), which also gives the noun phrase a deictic function (Wu & Bodomo, 2009). The following examples illustrate this.

22. *Nei4 ben3 shu1 wo3men dou1 du2gou4* 那本書我們都讀過 That CL book we all read = as for the/that book, we have all read it. (Wu & Bodomo, 2009).

Definiteness is also marked in Mandarin by word order, as Mandarin is a topic-prominent language. This means the topic appears sentence initial and shows either known information or generic uses, such as referring to an entire class of objects. The second part of the sentence is the comment, which contains new information (Moore, 2004). If a noun is preverbal, it is usually definite regardless of the use of the demonstratives *na4* 那 (“that”) or *zhe4* 這 (“this”). In addition, nouns that take the classifier *yi1* 一 (“one”) usually do not appear in the topic position, making the sentence indefinite. If a subject is post verbal and without the demonstratives, it is indefinite; therefore, if a subject appears before the verb without a demonstrative, it is perceived to be definite (Moore, 2004).

The issue of noun countability for many Chinese dialects has divided scholars, with some claiming that all Mandarin nouns are mass nouns (Chierchia, 1998; Wu & Bodomo, 2009), while others argue that Mandarin Chinese has both mass and count nouns. Chierchia (2008) argues that all nouns are treated as mass nouns; therefore noun countability would have to be learnt. Others (Cheng & Sybesma, 2005; Zhang, 2007) argue that Mandarin has both count and mass nouns with count and mass classifiers.

2.3 English Articles in Second Language Acquisition Studies

Numerous studies in second language acquisition (SLA) research have examined English articles, starting with Brown (1973). Research has indicated that both young L1 children and L2 learners tend to associate the definite article with specific contexts rather than hearer/discourse knowledge. This became known as *the* flooding, whereby a beginning learner overuses the definite article in all article contexts. Chaudron and Parker (1990) found evidence that English learners misused articles in specific, discourse-first locations. Using Huebner’s (1983) noun types, Thomas (1989) investigated whether L2 learners overused *the* in [+SR-HK] (first mention) contexts. Interestingly, unlike earlier L2 article acquisition studies (Huebner, 1983), Thomas's participants did not show any significant signs of '*the* flooding,' but the learners did overuse the definite article. The learners also over-generalized the zero article, although it was not clear whether the learners had failed to use this article or had made an explicit article choice, as the difference was impossible to detect without interviewing the participants. Master (1997) investigated how acquisition differed between English L2 learners from article-less L1s (Japanese) and L1s with articles (Spanish). He found that *the* flooding was more dominant in the Japanese subjects; *a/an* acquisition was also delayed for these subjects. These studies further demonstrated that L1 had an effect on article acquisition and that learners with L1s lacking articles had more difficulty acquiring the English article system. This has been confirmed in studies by Trademan (2002) and by Diez-Bedmar and Papp (2008).

In terms of native Mandarin-speaking English learners, Moore (2004) found that intermediate/advanced learners tended to overuse the indefinite article in both a cloze test and an oral narration task. Most of the indefinite errors occurred in *a for the* errors during the cloze test, but *zero/the* accuracy was almost the same during the narration task. Lee (2007) investigated advanced English L2 Mandarin speakers studying at the PhD level in the United States. Unlike other studies, Lee's research looked at error patterns in an online forum and a cloze test based on the findings of the online forum. The learners tended to omit rather than overuse the indefinite and definite articles in the online forum, but overused the definite article in the cloze test. For definite article error types, *the for Ø* errors were more common in front of unique common nouns and in specific contexts. More recently, (Diez-Bedmar & Papp, 2008) carried out a corpus study into article acquisition in Spanish and Chinese English L2 learners. The definite article was overused in specific contexts, but the zero article was also overused, demonstrating issues related to noun countability. They suggested that the overuse of the definite article was a pragmatic problem, as the writers did not consider the readers knowledge, while noun countability was seen as a grammatical problem.

2.4 Noun Countability and English Articles in SLA Studies

Noun countability has been an issue in article acquisition, especially for languages that do not use an article system (Goto-Butler, 2002; Hua & Lee, 2005; Lee, 2007; Master, 1997; Moore, 2004; Snape, 2008; Yoon, 1993). Using a cloze test, Yoon (1993) found that Japanese learners had problems with *indefinite for zero* errors, especially with mass nouns. Goto-Butler (2002) found that noun countability was also a source of errors with Japanese participants. Lower proficiency participants encountered problems with mass and count nouns, but higher level participants also had problems with countability-especially in nouns where the countability was context-dependent. Goto-Butler (2002) suggested that these errors with context-dependent nouns often cause definite article errors, as the listener depends on noun countability to determine if a noun is unique within a set. The noun *culture* is an example. Goto-Butler (2002) pointed out that culture is often indivisible, so it can be seen as uncountable. Nevertheless, the NP *old culture* belongs to a set of old cultures that need to be identified. When introducing the NP *old culture*, it must be introduced using the indefinite article. Goto-Butler's participants often introduced the phrase "Japan has **an** old culture" with the definite article (*i.e.*, "Japan has **the** old culture"), believing that Japanese culture was identifiable as definite (Goto-Butler, 2002). This problem with abstract nouns may lead to problems with definite article use.

Snape (2008) found that native Japanese-speaking English learners made more errors with the definite article within plural and mass contexts compared to singular contexts. For Mandarin L1s, Hua and Lee (2005) found that participants were able to distinguish between countable and uncountable nouns in English L2 and were more accurate with abstract nouns. Lee (2007) did not find any relationship between definite article errors and noun countability with her Mandarin-speaking high level participants, but did find errors with indefinite articles and noun countability. The learners often failed to use an indefinite article with countable singular nouns and failed to judge if a noun had a countable or uncountable reading.

3. Methodology

A total of 30 students participated in this study. The subjects were third-year university students who had attended writing class with the same instructor for four semesters. These participants were chosen for several reasons. Participants who had taken a writing class with the same instructor were needed in order to avoid the effect of differing writing instruction. In addition, all participants had received the same length of writing instruction. Although an earlier pilot study had shown no longitudinal effect, some of the essays may have been too short to provide an adequate amount of tokens; thus, it is possible that longitudinal changes could affect article accuracy. In order to control for this, the participants had to be students who had a similar amount of exposure to writing instruction.

The corpus consisted of 30 argumentation essays, with a total of 28,020 words. Only 30 essays were coded due to time limitations. The article types and error types had to be coded manually, as no automatic parsing had been developed to deal with the multiple functions of the article system. The pilot study revealed that shorter essays did not contain enough articles and article errors. Therefore, argumentation essays were coded, because their lengths ranged from 789 to 1,449 words, resulting in a mean of 980 words per essay. The original drafts of the essays were coded because they had not been corrected by the student, instructor, or peers.

To explore article use and error patterns in Taiwanese students' EFL academic writing, a coding scheme was developed to annotate the data with linguistic information. The coding scheme was based on a modified version of the one used by Moore (2004). Other corpus-based article coding schemes were examined, such as Han *et al*. (2006), who used the syntactic position of the NP in an automated system. Although Han *et al*.'s approach provided a general account of the errors over a large corpus; it did not include the semantic environment of the noun phrases, making it unsuitable for the current investigation. Neff *et al*. (2007) used the Spanish International Corpus of Learner English (SPICLE) corpus to investigate definite, indefinite, and zero articles, and this effectively described the general differences between the three articles; however, it was not comprehensive enough for the current study because it did not provide information on the semantic and pragmatic features of the English articles. Diez-Bedmar and Papp (2008) used Huebner's (1983) semantic environments to investigate article use in Spanish and Mandarin speakers' English writing; although their study bears some similarities to the present one, it did not investigate the use of the eight definite article types described by Hawkins (1978), which were needed for a related study into English article accuracy.

After investigating these other schemes, Moore's taxonomy (2004) was found to be the most comprehensive system, as it was based on article research conducted by Hawkins (1978) and Robertson (2000). This coding scheme has many advantages over the other schemes used in SLA article research because it combines the semantic environments, the definite article types identified by Hawkins (1978), and the language transfer features described by Robertson (2000). Although this scheme follows the procedure described by Moore (2004), it was sometimes necessary to make some adaptations or collapse some of Moore's categories.

In terms of coding, Figure 1 shows a brief diagram of the actual data as they would appear in the window of the corpus. The tagging system and AWTA corpus are described in detail in Kao and Chen (2009). The first pair of brackets indicates the meta-linguistic tag used in the corpus, and the annotation shows either the article type or the error type after the equal sign. The original text is in the arrowed brackets, followed by the meta-linguistic information to make the tags clear in the reviewing process. The tagging system works as follows. Inside the brackets is the name of the article (*e.g.*, the semantic or article type); information regarding whether it is used correctly is indicated by the letters Y or N, which represent correct and incorrect use. This is followed by a number indicating the general error type. For example, in<tag D PN N annotation="2">, the D is a definite article, PN stands for plural noun, which is the error type, N indicates an article error, and 2 is the code number for *definite for zero specific* errors. In this way, the article error can be identified first and meta-linguistic information can be added afterward. Figure 1 is an extract taken from the AWTA corpus. The tag <tag D IA N annotation="5"> indicates a definite for indefinite article substitution. The D is a definite article, IA stands for indefinite article, which is the error type, and 5 is the code number for *the for specific indefinite a/an* errors.

Many studies have showed that it would be better for the hearing disabled to have ＜tag D IA N annotation="5"＞the＜/tag＞ cochlear implant at an early age. Also, if implanted the cochlear implant at the age one to two, their language learning could come out of great improvement. However, the situation now seems that the elder people who are more than 55 years old, are not suitable to have the cochlear implant. They are usually told only to use ＜tag D PN N annotation="2 "＞the＜/tag＞ hearing aids for that most people think it would be too late for them to have the implantation.

Figure 1. Annotating meta-linguistic information.

To deal with the repetition of an NP, which is often necessary in writing due to its cohesive function (Trademan, 2002)-although some overuse or repetition can be interpreted as an immature writing style-a types/token distinction was used. Here, token counts refer to the frequency of a particular word or phrase whereas type refers to the occurrence of a distinct word or phrase in a text. In terms of errors, token counts would record the same error throughout the text, whereas type frequency would only record a mistake once. Therefore, if tokens were classified as errors, it would present an inflated picture. This paper only coded the types to avoid inflating the number of errors.

Once the coding procedures were decided, the data were coded for errors, as article errors are often discourse-dependent, making it necessary to read the essays first without the distraction of tagging every English article. All of the errors were highlighted and subsequently coded according to their error type. Next, the essay was coded for article use, starting with the definite article, followed by the indefinite, and finally the zero article. This was done to collect information for related research into L2 English article use. The annotation system consisted of two main parts: the semantic and pragmatic relations of each article and a description of the common error patterns.

Article error types can tell a researcher a lot about what kind of articles the participants were using in their writing (Lu, 2001). The most important contribution is that they can indicate if any patterns of underuse or overuse exist or if the errors are purely random. Altogether, 37 possible error types were identified. Article errors in the text that could not be tagged according to the error system were labeled “unclassified”; these included definite and indefinite articles that were erroneously used outside the NP, meaning they were general errors, not errors within the article system. Furthermore, it was presumed that these were writing mistakes, as there was no pattern to the errors.

Cohen’s Kappa analysis was used to measure inter-rater reliability. In the inter-rater procedure, only two coders were used due to time and financial restrictions. Both coders were linguistics graduate students and experienced English teachers. The coders were trained to use the corpus over three essays. If agreement was not reached, the two coders discussed the coding problems, and extra training was provided when necessary. In this study, 20% of the data was randomly selected from the argumentation essays and coded by the two raters. The Kappa statistic was calculated to be 0.332, which indicates a fair level of agreement between the two raters.

4. Results

This section explains the rationale and formulas for reporting accuracy and presents the accuracy of the three articles. Following this is a description of the distribution patterns of each article, including the semantic and structural functions. After the essays had been tagged, the data was checked for inter-rater reliability, and the raw frequency counts for each error type were computed.

In order to report the frequency of the article errors, the data had to be normalized to allow data from different texts to be accurately compared. As the lengths of the essays differed between participants, reporting the raw frequency counts would not present an accurate account of the errors. In a longer text, there are more opportunities for errors to occur, so 'normalization' is a formula that adjusts the raw frequency counts so texts of different lengths can be compared (Pica, 1983). In normalization, the raw frequency counts are divided by the number of words in the text then multiplied by the mean essay lengths for the 30 essays, which are 980 words per essay. The following example illustrates the normalization formula:

definite for zero errors 26 /1020 x 980 = 24.98 *definite for zero* errors per 980 words.

In this formula, there are 26 *definite for zero* errors in one essay. This is divided by the total number of words in the essay then multiplied by the mean essay length, giving a total of 24.98 errors per 980 words.

Table 3 reports the distribution of the article types and article errors throughout the corpus. It is presented as a matrix table and it is read from left to right. The article type *the* on the horizontal axis shows the definite article, and reading the column from left to right indicates where the definite article is substituted for another article. If the table is read from left to right, starting with the definite article, it indicates where the definite article is being substituted in place of another. For example, reading the matrix from left to right indicates that 9.16% *the for a* substitution errors occurred. The highest frequency is *definite for zero* errors at 28.45%. Countability errors occurred when the indefinite article was substituted for the zero article or vice-versa. The results illustrate that 6.77% *zero for a/an* errors occurred, and 2.45% *a for zero* errors occurred. The number of *zero for the* and *a for the* errors are low at 5.33% and 0.79%, respectively, indicating that the frequency of definite article underuse is low. Table 3 indicates that overuse of the indefinite and zero article is low, but more errors are made with the definite article, while countability errors are relatively lower. In other words, there are far more semantic or pragmatic errors than grammatical errors. Grammatical errors are due to noun countability errors where the writer must assign the indefinite article to singular nouns and the zero article to plural or mass/non-count nouns.

Table 3. Article error distribution

Article	The		Indefinite a/an		Zero	
	freq	%	freq	%	freq	%
The	922.71	**93.87**	**48.41**	**9.16**	**120.36**	**28.45**
Indefinite a/an	**7.81**	**0.79**	443.95	**84.07**	**10.4**	**2.45**
Zero	**52.43**	**5.33**	**35.65**	**6.77**	292.21	**69.13**
Total	982.95	100	528.01	100	422.97	100

The next section analyzes the influence of semantic NP environments and countability in order to determine their impact on article errors, as the effects of specificity and countability have been well-documented as factors that influence L2 learners' article errors.

Table 4 illustrates the frequency of the main article errors according to NP environment and countability. The highest frequency of errors can be found in *definite for zero specific plural* errors followed by *the for specific indefinite a/an* errors. These descriptive results suggest that specificity influences the frequency of the *for indefinite a/an* errors, as more errors occur in specific NPs. The frequency of *zero for a* errors is low at 10.58% of total errors, but suggests that some participants have trouble using the correct article with singular and plural nouns. The number of *a for zero* and *zero for the* errors was not reported as their frequencies were very low, indicating that this was not a problem for the participants. The frequency of *definite for zero* errors in both specific and non-specific environments suggests that specificity may not be the only influence on *definite for zero* errors.

Further statistical analysis was needed to investigate the influences on error types. It has been predicted that, for English L2 learners with no article system in their L1, more errors are found in specific indefinite noun phrase environments. To determine the effect of specificity on definite for indefinite errors, a paired sample *t*-test was conducted. As there are only two independent variables, a *t*-test could show if the difference between specific and non-specific *the for indefinite a/an* errors is significant. It revealed a significant difference between the two groups (t (29) = 6.94, $p < .001$). The mean of the specific indefinite errors was significantly higher (m = 1.36, sd = 1.03) than the mean of the non-specific errors (m = 0.25, sd = 0.46), indicating that specificity influences definite article errors in indefinite specific environments. In other words, the definite article is being substituted for the indefinite article in specific environments, as predicted by Ionin and Wexler's (2004) fluctuation hypotheses. The implications of this are discussed in Section 5.

Table 4. Error types across the corpus per 980 words

Error type	Freq. of errors	% of total errors*
Zero for A	**35.65**	10.58
Definite for zero		
Non-count specific	23.98	9.62
Plural specific	**45.12**	**18.10**
Plural non-specific	**35.12**	**14.09**
Definite for zero non-count non-specific	16.14	6.47
The for A		
Specific	**41.09**	16.48

Note. N = 30 (N = shows the size of the data pool which is 30 subjects.)

* Other error types are not included in this table.

Table 5 presents the descriptive statistics for definite article for zero article errors, where the four independent variables are *definite for zero specific plural* errors, *definite for zero non-specific plural* errors, *definite for zero non-count/mass specific* errors, and *definite for zero non-count/mass non-specific* errors. Some researchers (Goto-Butler 2002; Yoon, 1993) believe that, in addition to semantic environments, the difference between count and mass/non-count nouns may have an influence on article errors. Due to this, more errors are expected with mass/non-count nouns than with plural nouns. Also, due to the fluctuation hypothesis (Ionin & Wexler, 2004), which states that specificity influences article errors, more errors are expected in specific NPs. It was suggested that a repeated measure ANOVA would be able to show any significant differences between NPs environments and would also reveal any differences between plural and mass noun errors.

Table 5. Descriptive statistics for definite article for zero article errors

Substitution type	M	SD
Definite for zero specific plural errors	1.50	1.67
Definite for zero specific non-count/mass errors	0.79	0.99
Definite for zero non-specific plural errors	1.19	1.51
Definite for zero non-specific non-count/mass errors	0.53	0.73

Note: N = 30

Table 6 shows the repeated measure ANOVA results for the definite article for zero article errors. A significant effect was found (*F* (3, 87) = 5.66, *p* < .005). Follow-up protected *t*-tests revealed a significant difference between *definite for zero plural* (*m* = 2.70, *sd* = 2.67) and *definite for zero non-count/mass* substitution errors (*m* = 1.33, *sd* = 1.54), showing an effect with noun countability on *definite for zero* errors. In other words, more *definite for zero* substitution errors are found with plural nouns indicating that, for these participants, mass/non-count nouns do not have a significant influence on definite article errors. The follow-up protected *t*-tests between *specific definite for zero* (*m* = 2.37, *sd* = 2.31) and *non-specific definite for zero* errors (*m* = 1.73, *sd* = 2.09) revealed no significant difference between specific and non-specific zero, indicating that specificity is not a significant influence in *definite for zero* article errors. The implications of this are discussed in Section 5.

Table 6. ANOVA results for definite article for zero article errors

	df	F	η^2	*p*
Between subjects				
Definite for zero subs	3	5.66	.003	.001**
Within-group error	87	(0.96)		

Note: N = 30; ***p*<. 001

5. Discussion

The results indicated that the participants in this study had problems using the English article in terms of distinguishing between a definite and indefinite noun phrase. Correct article use in terms of noun countability was not a major problem for these writers. This section discusses the influence of specificity on article error patterns. First, the indefinite article is discussed, followed by the zero article.

Errors with specificity may stem from some participants' identification of a specific noun clause as definite, as predicted by the fluctuation hypothesis (Ionin & Wexler, 2004). The results of this study support the view that the definite article is overused in specific noun phrases with indefinite a/an, as the results of the *t*-test show a significant difference between *the for indefinite a/an* errors, with more errors occurring in specific NPs. Nevertheless, the fluctuation hypothesis also predicts overuse of the indefinite article with definite non-specific nouns (*i.e.*, *a for the generic* errors). No such errors were evident in the results of this study, although only 54 generic indefinite noun types were counted in the data. This is a result of the low frequency of generic indefinite noun types in the writing samples.

Zero articles not taking a generic, proper noun, or idiomatic reading can be specific or non-specific, in accordance with Lu's (2001) specifications. The repeated measure ANOVA and follow up protected *t*-tests revealed no effect of specificity on *definite for zero* errors, as no significant difference was found between specific and non-specific errors. Thus, unlike the indefinite article, specificity was not the only influence on the overuse of the definite article with zero articles. A misrepresentation of the pragmatic functions of the definite article is a possible reason for these errors, and this will be discussed below.

The results demonstrate that the learners in this corpus lacked accuracy with regard to the zero article, regardless of semantic type. As a result, the participants often compensated for this by using the definite article. The indefinite article cannot be used for plural nouns or mass or non-count nouns due to countability rules. Thus, a writer has two article options: the zero or the definite. Although the fluctuation hypothesis may explain the errors in specific environments, it cannot explain definite article overuse in non-specific environments; thus, the effects of other influences need to be considered-particularly mass/non-count nouns or the hearer knowledge [HK] feature of definite articles.

In English, the context-namely, the speaker's and hearer's knowledge of the context-determines whether an NP can be located by both participants. If the writer believes that the hearer is aware of the noun, the definite article is used. In other words, as Diez-Bedmar and Papp (2008) pointed out, a writer often takes the readers' knowledge into account when using the definite article.

According to Hawkins (1991), using the definite article enables the hearer to access the NP in a p-set (a set of knowledge known by the hearer/reader as being definite). The speaker/writer should use the definite article when he/she is confident that the other party knows that the NP is definite. A communication breakdown will occur if the speaker/writer uses the definite article erroneously or mistakenly believes that the hearer has such knowledge. The writers in this corpus have not been falsely assuming that the reader had definite knowledge-this would signal a lack of pragmatic awareness-but the writers may not have acquired how the definite article signals this knowledge. Thus, errors with the definite article could be classified as errors regarding the acquisition of the pragmatic functions of the English definite article.

The results reveal that participants made significantly more errors with plural nouns than with mass/non-count nouns. Errors involving the definite article with mass/non-count nouns have been found in other studies with Japanese L1s (Goto-Butler, 2002; Snape, 2008), although the results in this study reveal plural errors have a greater effect on error patterns. A *t*-test indicated a significant difference between mass/non-count nouns and plural nouns, indicating that definite article errors with mass/non-count nouns are less frequent than errors with plural nouns. This differs from what Goto-Butler (2002) found with their Japanese participants, who made more errors with mass/non-count nouns. In other words, for the participants in this study, the influence of mass/non-count nouns is not a significant factor in

English definite article errors.

Although noun countability has been seen as a problem for English L2 learners, especially learners whose L1 does not have an article system, for the participants in this study, the number of errors in *zero for a* and *a for zero* contexts was relatively low (9.51% and 2.87% of the total errors, respectively). A total of 18 *zero for a* and *a for zero* errors occurred with count nouns, indicating that the writers may be influenced by their L1.

Example 1. Every citizen is suitable by the law. No one is exception if he or she committed *crime.*

Example 2. For some losers may bankrupt and then rob *bank* in order to win back.

These examples indicate that zero articles were substituted for the indefinite article. One reason for this is that the writer applied his/her L1 rule instead of using an article with singular nouns because, in Mandarin Chinese, nouns do not always need a classifier, demonstrative, or numeral.

5.1 Pedagogical Suggestions

This section will offer suggestions to the language teacher based on the results of this study. It has already been pointed out the English articles are extremely difficult words to teach for two reasons. First, the definite article stacks multiple functions onto one word, making it cognitively more demanding for a learner to process. Second, as article errors do not cause communication breakdowns in daily conversation, they may be subject to fossilization in a learner's interlanguage (Brender, 2002). Although many researchers have looked at ways to teach all of the articles under one system (Bitchener, 2008; Master, 1990; 1994), the results of the current study demonstrate that the most frequent errors occur with the definite article in two main areas: *the for zero*, and *the for specific indefinite*. As most of the errors involved the definite article, the semantic environment of [+/-HK] and [+/- SR] are effective parameters for helping learners determine whether an NP needs the hearer's knowledge element or whether it is just a specific noun. In this way, both specificity and hearer knowledge can be brought into focus, as this study found it was the influence of both factors that resulted in more than 80% of the article errors.

Research on teaching article use (Bitchener, 2008; Brender, 2002; Master, 2002) has shown that explanations in the form of mini-lessons-along with group work and meaning-focused activities-are more suitable for this type of language feature. As much of the information about hearer knowledge is found in discourse or is non-linguistic, activities that incorporate the communication aspect of definiteness would also be beneficial for article errors.

Finally, this study helped with our understanding of the influence of specific knowledge, hearer knowledge and noun countability on English article errors in writing. Given that the participants were all undergraduate English majors, it would be beneficial to design a cross-linguistic study involving higher level and lower level learners to observe the changes as learners' writing improves with ability and exposure to academic reading and writing. This would allow the researcher to design article teaching systems for all levels of learners based on the frequency of error types for each level.

References

Bickerton, D. (1981). *Roots of language*. Ann Arbor, MI: Karoma Press.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, 17, 102-118.

Brender, A. (2002). *The effectiveness of teaching articles to (-ART) students in EFL classes using consciousness raising methods.* Doctoral dissertation, Temple University.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition*, 12(1), 43-64.

Chen, Y. (2002). The problems of university EFL writing in Taiwan. *Korea TESOL Journal*, 5(1), 59-81.

Cheng, L.-S., & Sybesma, R. (2005). Classifiers in four varieties of Chinese. In G. Cinque & R. Kayne (Eds.), *The Oxford handbook of comparative syntax* (pp. 259-292). Oxford: Oxford University Press.

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, 31, 471-484.

Chierchia, G. (1998). Reference to kinds across languages. *Natural Language Semantics 6,* 339-405.

Diez-Bedmar, M. B., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, M. B. Diez-Bedmar, & S. Papp (Eds.), *Linking up contrastive and learner corpus research* (pp.147-175). New York: Cambridge University Press.

Goto-Butler, Y. (2002). Second language learners' theories on the use of English article: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition*, 24(3), 451-480.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

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

Han, N. R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article use by non-native speakers. *Natural Language Engineering*, 12(2), 115-129.

Hawkins, J. A. (1978). *Definiteness and indefiniteness*. London: Croom Helm.

Hawkins, J. A. (1991). On (in) definite articles: Implicatures and (un)grammaticality predictions. *Journal of Linguistics*, 27(2), 405-442.

Hua, D., & Lee, H. (2005). Chinese ESL learners' understanding of the English count-mass distinction. In D. Laurant, R. A. Sprouse, & A. Liljestrand (Eds.), *Proceedings of the 7th Generative Approaches to Second Language Acquisition Conference (GASLA 2004)* (pp. 138-149).

Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor: Karoma.

Ionin, T., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, 12, 3-70.

Kao, T. & Chen, L. M. (2008). Coherence in Chinese students' English writing: An initiative to a learner's corpus. In Y. Leung, & H. Chang (Eds.), *Selected Papers from the Seventeenth International Symposium on English Teaching* (pp. 167-175). Taipei: Crane Publishing.

Lee, E. H. (2007). *English article usage in online graduate forums by non-native EFL teachers*. Doctoral dissertation, Indiana University.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese. A functional reference grammar*. Los Angeles: University of California Press.

Lu, F. C. (2001). The acquisition of English articles by Chinese learners. *Working Papers in Second Language Studies*, 20, 1-36.

Master, P. (1990). Teaching the English articles as a binary system. *TESOL Quarterly*, 24, 461-478.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System,* 25(2), 215-232.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System,* 25(2), 215-232.

Master, P. (2002). Information structure and English article pedagogy. *System*, *30*(3), 331-348.

Moore, J. M. (2004). *Articles and proper names in L2 English*. Doctoral dissertation, Northwestern University.

National Cheng Kung University. (n.d.). *Academic Writing Textual Analysis (AWTA) corpus*. http://awta.csie.ncku.edu.tw/

Neff, J., Ballesteros, F., Dafouz, E., Martinez, F., Rica, J. R., Diez, M., & Prieto, R. (2007). A contrastive functional analysis of errors in Spanish EFL university writers' argumentative texts. A corpus based study. In E. Fitzpatriz (Ed.), *Corpus linguistics beyond the word. Corpus research from phrase to discourse* (pp.203-227). Amsterdam: Rodopi.

Pica, T. (1983). The article in American English: What the textbooks don't tell us. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition* (pp. 222-233). Rowley, MA: Newbury House.

Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2), 135-172.

Snape, N. (2006). L2 acquisition of definiteness and specificity in English by advanced Japanese and Spanish learners. In A. Belletti, A. Bennati, C. Chesi, E. Di Domenico, & I. Ferrari (Eds.), *Language acquisition and development. Proceedings of the Generative Approaches to Language Acquisition Conference* (pp. 591-596). Cambridge, UK: Cambridge Scholars Press/CSP.

Snape, N. (2008). Resetting the nominal mapping parameter in L2 English: Definite article use and the count-mass distinction. *Bilingualism: Language and Cognition*, 11, 63-79.

Snape, N. (2009). Exploring Mandarin Chinese speakers' article use. In N. Snape, Y. K. I. Leung, & M. Sharwood Smith (Eds.), *Representational deficits in SLA: Studies in honor of Roger Hawkins*. (pp. 27-51). Amsterdam: John Benjamins.

Thomas, M. (1989). The acquisition of English articles by first- and second-language learners. *Applied Psycholinguistics,* 10, 335-355.

Ting, F. (2003). An investigation of cohesive errors in the writing of PRC tertiary EFL students. *SIETS Language and Communication Review*, 2(2), 1-8.

Trademan, J. E. (2002). *The acquisition of the English article system by native speakers of Spanish and Japanese: A cross-linguistic comparison.* Doctoral dissertation, University of New Mexico.

Trask, R. L. (1995). *A dictionary of grammatical terms in linguistics.* New York: Routledge.

Trenkic, D. (2008). The representation of English articles in second language grammars: Determiners or adjectives? *Bilingualism: Language and Cognition,* 11(1), 1-18.

Wu, Y & Adams, B. (2009). Classifiers#determiners. *Linguistic Inquiry,* 40, 487-503.

Yoon, K, K. (1993). Challenging prototype descriptions: Perception of noun countability and indefinite vs. zero article use. *International Review of Applied Linguistics,* 31(4), 269-289.

Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics,* 16, 43-59.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 21-40
© The Association for Computational Linguistics and Chinese Language Processing

基於辭典詞彙釋義之多階層釋義關聯程度計量—以「目」字部爲例[1]

A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations

趙逢毅[*]、鍾曉芳[+]

F. Y. August Chao, Siaw-Fong Chung

摘要

本研究使用辭典中詞語與釋義的關係，透過計算機進行詞語的多階層釋義關聯(Multi-Level Semantic Relation)量化計算，以比較字與詞組之間的釋義關聯程度。本研究從臺灣「教育部重編國語辭典修訂本」之中，取出屬「目」字部之所有字/詞共計 4549 個字與詞彙進行直接試驗，以避免人爲語意辨別模糊與使用不同語料庫定義不同的缺點。經過不同的試驗說明多階層釋義關聯的特色及使用方式，其中包括將屬「目」字部所有的字與「目」進行多階層釋義關聯計算，表現出漢字意符的特色。最後並與常用的 MI Value 及 t-score 比較多階層釋義關聯之異同。

關鍵字：釋義關聯，多階層釋義關聯網路，辭典，語料庫，「目」字部

[1] 本論文出處於中國蘇州舉辦之「第十一屆漢語詞彙語義學研討會(CLSW2010)」增修版本。

[*] 國立政治大學資訊管理研究所，台灣台北市文山區指南路二段 64 號
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

Abstract

In this study, we utilize a quantitative method measuring the Multi-Level Semantic Relations based on 4549 Mandarin lexemes containing the radical *mu4* (目). The research is carried out by first extracting all dictionary definitions for all lexemes containing this radical. Then, we consider the different layers of definitions (e.g., the definitions of the keywords in a definition) and measure whether two different *mu4* (目) lexemes are related in meanings. It was found that both width (the number of lexemes covered) and depth (the number of levels to be calculated) contribute to the measurement of semantic relatedness. Some seemingly unrelated *mu4* (目) lexemes are found related when the depth of definitions increases. The study also compares two sets of results - one based on MI value and the other based on t-score. Our findings show that our measurement based on multi-level semantic relations produces better results than MI value does, as a collocation measurement like MI value is less suitable for analyzing semantically related dictionary entries.

Keywords: Definition relation, Multi-Level Semantic Relation, Dictionary, Corpus, Mandarin radical *mu4* (目)

1. 前言

「辭典」乃依據詞彙體系及一定的編輯體例蒐集詞、詞組、短語等資料，加以解釋以備查索、參考的工具書。在編輯辭典時，除了要考量到查詢者所擁有的先備知識(Prior Knowledge)以撰寫其能理解的釋義之外，還要使用簡單的釋義文字說明，才能使查詢者瞭解該字所俱備的義涵。中文辭典的排版則是透過具有意符表徵的部首，將所有漢字與詞語聯繫並形成知識架構(周亞民、黃居仁，2005)。黃居仁(2005)亦認爲詞是「語言中表達意義的最小獨立單位」，在辭典中則是透過不同的詞彙組織成爲釋義文字以說明查詢的辭彙。因此從辭典的釋義文字、詞組與查詢字(詞)之間的釋義關聯，可以發現釋義說明所用的單一詞語會包括被解釋字(詞)的部份或完全的涵義，以及對該字(詞)認知和意義延伸的比喻涵義，並且使用屬於知識層級上較爲通俗的語句或概念撰寫釋義詞語，從而讓具有一般知識水準的大眾都能很快理解釋義的說明。臺灣的「教育部重編國語辭典修訂本」即是在提供各界人士及中小學生檢索、查閱的目的之下而編修的一本實體紙本工具書，因此編輯釋義內容均使用大眾能理解的文字及簡單的詞語說明辭典之中的字詞。

在分析字(詞)之間的釋義關聯方法中，除了透過人爲語感對詞彙進行詞意訊息上的討論外(黃居仁等 2005, 2008)，便是使用已經建立好的詞彙工具進行語意的探索，如：知網(Hownet)(Dong & Dong, 2003)、英文的 Wordnet (Miller, 1995)及中研院的中文詞彙網路(Chinese WordNet, CWN)(Huang, Chang, & Lee, 2004)。雖然現有語料工具定義詞彙語意十分詳細，卻無法如辭典釋義般詳述詞彙語意的內容。本論文透過使用 Chung, Chen, 與 Chao (2009) 所建立的隱喻關聯程度計算，直接計算中文辭典之中的釋義文字，同時比較

字(詞)之間的釋義關聯(dictionary definition relationship)。除了建立釋義關聯網路與同義字彙集群之外，更進一步擴展此關聯計算以揭露中文詞彙之間，在相同部首之中深層的釋義關聯。

2. 相關研究

中文字本身就存在很強的義形(plyph)與概念(concepts)的聯結關係。在編撰辭典時，參與編輯人員亦會仔細研究該字要表達的意義與所歸屬的部首之間概念上是否恰當(周亞民、黃居仁，2005)。鄭文泉(2004)更從符號學的角度指出「漢字藉由方位隱喻足以使人們領略它的肖像的符號涵意」。其它學者則從不同的觀點討論此一課題，如黎傳緒(2004)對“相”的解析，討論「本義」與「引義」之間的發展關係。另一方面，祝清(2009)從詞類上著手，探討漢語獨立詞類—動名詞，提出動名詞實為語法的隱喻，是詞類的去範疇化與再範疇化的結果。

中文辭典則是將不同的辭句，透過文字部首架構起來匯集成為辭典。辭典釋義的編撰會因操作視角策略與適用的領域不盡相同而造成釋義不同，但都是透過文字紀錄下適合於當代的定義與闡釋的語料庫(羅益民，2007)。從語料庫的觀點而言，透過部首分類的字或詞組應可在其所包含的義涵之中尋找到概念上的聯繫。黃居仁等(2005, 2008)則以人工的方式，從知識概念的層級上建立中文部首與文字之間的關係，透過物質(formal)、組成(constitutive)、功用(telic)、事件(participating)、參與者(participator)、描述狀態(descriptive)與產生(agentive)等七類衍生面向，探討漢字知識表達的層面與目、耳、口、鼻、舌等五官類漢字意符(漢字構字要件)的語意關聯。為能確保詞彙語意的品質與一致性，黃居仁、蔡柏生等(2003)則認為參與編撰人員需對詞義判準有一致性的準則，避免不同人對某一詞彙在語境中，依據不同的直覺而有不同的區分方式。

另一方面，近十多年來語料工具發展對詞彙語意探討, 與詞彙語意之間關係的規範有很大的幫助。常見的語料工具有：Wordnet (Miller, 1995)，一個由普林斯頓大學所發展出的英文語料工具，其中定義了詞彙的涵義、上/下位關係詞、部份/全體關係詞與同義詞；Hownet (Dong & Dong, 2003)，以常識(common sense)定義中文字彙的義原，並討論義原之間的關聯所建立的詞彙知識庫；不同於 Hownet 的義原關聯建立原則，中研院的中文詞彙網路(Chinese WordNet, CWN)(Huang, Chang, & Lee, 2004)使用中文詞義(sense)區分資料，並且將中文詞義與英文詞義(Wordnet)建立關聯。透過使用前述不同的中文/英文語料工具，高照明(2007)同時整合不同辭典中文詞彙語意訊息，建立詞彙關係擷取系統。然而前述中文語料工具所涵蓋的字與詞彙相較於辭典字彙仍有不足外，語料工具所建立的知識領域亦有所局限，而且使用英文語料工具還有語言平行對譯等問題。

不同於前述，本研究直接探討字與詞彙間釋義的描述關聯。延伸自 Chung *et al.* (2009) 對知識本體與語料庫之中的字詞，以字詞共同出現的知識概念出現比例，決定隱喻關聯的強度，藉此建立辭典字彙之間的釋義關聯網絡。

3. 研究方法

本研究收集提供給廣泛使用者查詢的「教育部重編國語辭典修訂本」資料(網路版)，透過統計詞彙之間共同出現的釋義字彙，以了解詞彙之間多階層釋義關聯程度，並說明此關聯程度所表達的意思之間的概念關聯。首先呈現資料收集原則與比較釋義關聯計算，並說明辭典資料透過此關聯度而表現出不同程度數値的釋義關聯內容。接著詳述延伸自隱喻關聯計算原則的多階層釋義關聯，並驗證黃居仁(2005, 2008)部首與中文字之間的意符關聯。最後則將比較多階層釋義關聯與互見信息值 (Mutual Information Value)、t-score間的不同。

3.1 資料收集與實作關聯計算

在中文辭典中，部首的分類原則在民國初年已經過許多人的改制，目前較通用的檢字規則是教育部頒訂的「部首檢字法」(改自林語堂「上下形檢字法」)，將中文字所表達的意符分類到不同部首中。因此本文在資料收集的過程之中，爲求所收集到的辭典釋義能將含括特定意符的知識概念都包含在詞彙的釋義之中，因此在收集的字詞選擇上，參考薛榕婷(2003)對部首與文化之間的研究，選定依「目」字部爲主，並收集所有包括屬「目」字部首的字詞共計 4549 字(字數 115 與詞彙數 4434)，其中亦包括異體字(如瞅、眀、眥等同爲「目」字的異體字與變形部首，如：罜與睘)。在取得釋義之後，將資料透過中研院所開發的中文斷詞系統(http://ckipsvr.iis.sinica.edu.tw/)以獲得釋義文字之中各詞語的詞類。在取得字詞類的釋義後，本研究保留知識概念涵義較大的動詞詞類('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')進行隱喻關聯的計算。此外亦將標記後句中含有分號(COLONCATEGORY)的句子刪除，因在「教育部重編國語辭典修訂本」的釋義中，使用到分號的句子爲來源或例句的說明。接著以「瞄」、「瞄準」二字爲例說明處理過程：

*「瞄」爲「**注視**」；*

*「瞄準」爲「用眼睛**注視**目標，使發射、投射的動作準確。」*

此二組字詞之間因包含「注視」而產生了釋義關聯。所取得的釋義在經過斷詞系統處理及詞類過濾後，保留下的釋義分別爲：

*「瞄」爲「**注視(VC)**」；*

*「瞄準」爲「眼睛(Na)**注視(VC)**目標(Na)使(VL)發射(VC)投射(VC)動作(Na)」*

從上述的資料可知，「瞄」與「瞄準」共同使用的釋義詞爲「注視」。本研究的釋義關聯計算，參考知識概念隱喻計算(Chung *et. al.*, 2009)，以共同出現的二度隱喻詞百分比表示 (“Percentage of co-appearance of 2nd degree relations”, CoAP) 。其中令 *x* 爲來源詞語的字詞，其過濾後符合的字詞集合爲 *X*，而 *y* 爲目標詞語，其過濾後符合的字詞集合爲 Y；$X \cap Y$ 則爲兩組字詞共同所使用的詞組，*x*、*y* 個別的二度隱喻詞百分比爲：

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

在 *X* 所有釋義中與 *Y* 共同出現的字詞的比例，同理

$$CoAP(y)=\frac{X\bigcap Y}{Y}\text{，} \tag{2}$$

即 *Y* 所有釋義中與 *X* 共同出現的字詞的比例。

依前例，「瞄」字的釋義爲："*注視(VC)* "、「瞄準」的釋義僅保留："*眼睛(Na) 注視(VC) 目標(Na) 使(VL) 發射(VC) 投射(VC) 動作(Na)* "，則

$$CoAP(\text{瞄})=\frac{(\text{注視})}{(\text{注視})}=1$$ ，同理

$$CoAP(\text{瞄準})=\frac{(\text{注視})}{(\text{眼睛、注視、目標、使 、發射、投射、動作 })}=\frac{1}{7}$$

「瞄」字對「瞄準」的釋義關聯大於「瞄準」對「瞄」的釋義關聯。意即「瞄」所含的概念(即共有的概念詞彙，"*注視(VC)*")能被「瞄準」所包括，而「瞄準」的概念卻無法透過「瞄」的概念完整詮釋。最後再透過網路分析工具 Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) 進行釋義關聯網路的繪製，如圖 1。圖中表示「瞄」與「瞄準」(深色方型，辭典詞彙)在釋義之中都有使用到"*注視*" (淺色菱型，辭典釋義中的釋義關聯媒介文字)。由此可得知，「瞄」與「瞄準」在辭典釋義之中都會使用"*注視*"來描述兩者涵義。

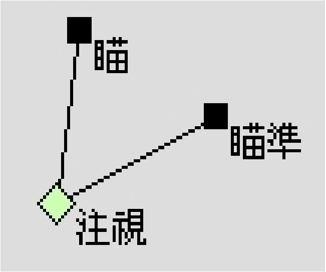


圖 1.「瞄準」與「瞄」之語意媒介網路(菱型爲釋義關聯媒介，方型爲辭典詞語)

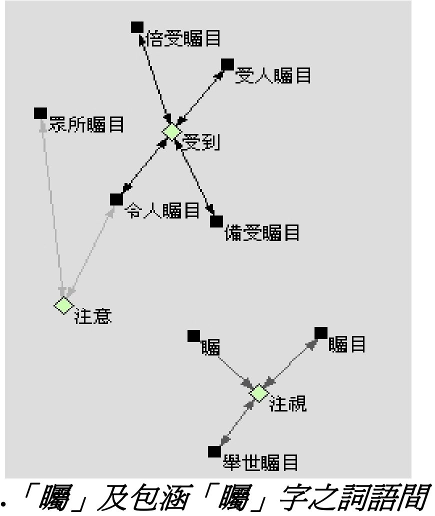


圖 2.「矚」及包涵「矚」字之詞語間交互語意媒介網路

同理，計算「矚」與其相關字詞(包括「令人矚目」、「倍受矚目」、「備受矚目」、「受人矚目」、「引人矚目」、「眾所矚目」、「矚望」、「矚目」、「舉世矚目」、「遠矚」、「駭矚」、「高瞻遠矚」、「麗矚」)等 14 個詞語，將其中產生語意關聯的 8 個詞語繪成語意媒介網路爲圖 2。從圖 2 可知「矚」及使用到「矚」字的 8 個詞語之中，依釋義的動詞及名詞的使用情況可以區分爲兩個群，即圖 2 中的左下方群集與"*注視*"有釋義關聯

的「矚」群，及屬於"*注意*"-"*受到*"釋義的群。爲了能更清楚地了解集群中的義涵，我們試著從釋義中了解中介節點的涵義如下："*受到*"爲單一動詞；"*注意*"的釋義爲：關注留意；"*注視*"的釋義爲：集中視線，凝目而望。因此"*注視*"群相較於"*注意*"-"*受到*"群保有實際上的視線、凝望等釋義關聯，而"*注意*"-"*受到*"釋義的辭典詞彙群則有接受及注意的釋義關聯。由此可知，應用知識概念隱喻計算於辭典釋義之中，再藉由詞彙語意媒介網路的集群具體呈現，結合詞彙釋義的深入了解，我們可以更清楚地區別辭典如何詮釋字彙的涵義，也更明白近義詞彙之間的不同形式。

進一步分析字詞時，由於釋義關聯媒介文字的比例會隨辭典詞彙的比例增加，使得釋義關聯無法將高度關聯的詞彙表現出來。因此參考 Chung *et. al.* (2009) 的隱喻關聯計算，將兩詞語相互之間共同詞彙所擁有之詞彙比例進行平均數計算，以表示來源及目標釋義關聯相互之間的釋義關聯程度 (Definition Relation Degree, DRD)：

$$DRD_{xy} = 2\frac{CoAP(x)CoAP(y)}{CoAP(x)+CoAP(y)}, 0 \le DRD_{xy} \le 1 \qquad (3)$$

承前例，「瞄」與「瞄準」兩詞之間的釋義關聯程度即爲：

$$DRD_{瞄,瞄準} = 2\frac{CoAP(瞄)CoAP(瞄準)}{CoAP(瞄)+CoAP(瞄準)} = 2\frac{1(\frac{1}{7})}{1+\frac{1}{7}} = 0.25 = DRD_{瞄準,瞄}$$

從「瞄」與「瞄準」的釋義之中可知，雖然兩者共同使用「注視」爲釋義詞，但因爲「瞄準」的釋義詞使用較多名詞與動詞，即「瞄準」的概念較「瞄」的概念複雜，從而降低了兩詞彙間的釋義關聯程度。

表 1.「矚」及包涵「矚」字之詞間交互釋義關聯程度表

詞(一)	詞(二)	詞(一)對詞(二) 之 CoAP	詞(二)對詞(一) 之 CoAP	詞(一)與詞(二) 的 DRD
矚	矚目	1	1	1
備受矚目	受人矚目	1	0.5	0.67
倍受矚目	備受矚目	0.5	1	0.67
倍受矚目	受人矚目	0.5	0.5	0.5
令人矚目	備受矚目	0.2	1	0.3
令人矚目	受人矚目	0.2	0.5	0.29
倍受矚目	令人矚目	0.5	0.2	0.29
眾所矚目	令 矚目	0.3	0.2	0.25
舉世矚目	矚目	0.125	1	0.22
矚	舉世矚目	1	0.125	0.22

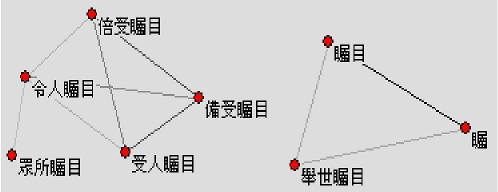


圖 3.「矚」及包括「矚」字之釋義關聯網路
(顏色深淺表 DRD 數值，顏色深表數值高)

依上述原則，計算「矚」及包括「矚」之辭典詞彙(見表 1)，並透過釋義關聯程度(DRD)值計算，簡化成單純釋義關聯網路，並去除釋義關聯網路中間的媒介釋義字詞(見圖 3)。在表 3 中，詞(一)與詞(二)為來源與目的辭典條目，透過「詞(一)對詞(二)之 CoAP」與「詞(二)對詞(一)之 CoAP」分別計算兩者間共用釋義媒介詞的占有比率後，再取兩 CoAP 去向性的平均值為釋義關聯值(DRD)。

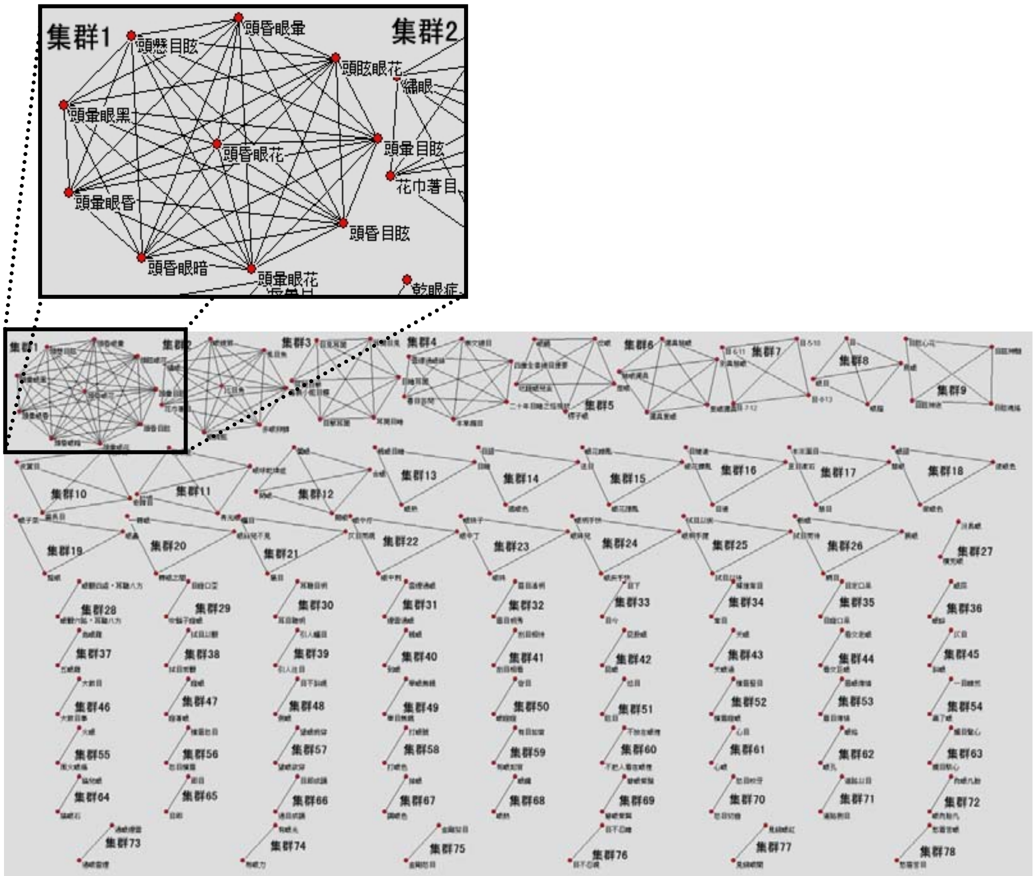


圖 4.「目」、「眼」及包括二字的高釋義關聯集群

釋義關聯網路可以藉由群集的結果，瞭解辭典詞彙之間大體上的釋義關聯情況及相關的程度。以「矚」及包括「矚」之辭典詞彙爲例，圖 3 左側「倍受矚目」、「備受矚目」及「受人矚目」，與右側的「矚」及「矚目」的釋義關聯程度都較其它辭典詞彙高(表 1 中，DRD 值高於 0.5)。從釋義文字中得知(見圖 2)，「眾所矚目」及「令人矚目」的釋義關聯之中還包括了"*注意*"的文字概念，而在另一個集群之中的「舉世矚目」，因釋義文字多於「矚」與「矚目」，從而使其釋義關聯程度(在圖 3 呈現) DRD 值降低。因此從資料中我們可以說明，「矚」-「矚目」與「倍受矚目」-「備受矚目」-「受人矚目」這兩個群組的釋義(涵義)上有高度的關聯，而「舉世矚目」的釋義較接近「矚」-「矚目」群而非「倍受矚目」-「備受矚目」-「受人矚目」群。

我們更進一步從前述釋義關聯分析「眼」字與「目」字及包括二字的辭典詞彙之釋義關聯，並分析兩兩成對的釋義關聯程度值 DRD 介在 0.9~1 的高釋義關聯部份，最後藉由社會網路分析軟體 Pajek 中的網路圖佈局集群將此關聯呈現視覺化圖形。結果得到上圖 4 共 78 個集群 (詳見表 2)。

表 2.「目」、「眼」及包括二字的詞語集群關聯媒介文字表，DRD>0.9

集群別	釋義關聯媒介文字	集群別	釋義關聯媒介文字	集群別	釋義關聯媒介文字
集群 1	頭腦、昏沉、視覺、模糊	集群 11	病名	集群 21	注視
集群 2	動物、名	集群 12	閉上、眼睛	集群 22	比喻、痛恨、人
集群 3	親自、看到、聽到	集群 13	看見	集群 23	眼球
集群 4	書名	集群 14	目、示意	集群 24	眼光、銳利、動作、敏捷
集群 5	北平、方言	集群 15	形容、眼睛、昏花、心緒、迷亂	集群 25	擦亮、眼睛、等待
集群 6	具有、特殊、眼光、見解	集群 16	佛教、人名	集群 26	眨眼
集群 7	異體字	集群 17	眼睛	集群 27 之後爲兩個辭典詞彙，且爲高度釋義關聯，故不再贅述。	
集群 8	眼睛	集群 18	眼睛、示意		
集群 9	形容、見、情景、令、人、驚異	集群 19	植物、名		
集群10	哺乳綱	集群20	眼睛、轉		

其中集群 1 爲最大，包括了十個辭典詞語(「頭眩眼花」、「頭暈目眩」、「頭昏眼暗」…等等)，且都是相似的詞義；集群 2 則是表示動物的辭典詞彙；集群 3 則是以「聽到」、「看到」爲釋義關聯媒介的集群；集群 4 是書名；集群 5 是北平方言；集群 6 是以「特殊」、「眼光」及「見解」所組成的；集群 7 爲異體字集；集群 8 即爲眼睛…等等。由此分析可知，DRD 值高即爲同義詞，因爲用以釋義的字詞幾乎相同。在此要特別

說明的是，透過釋義關聯程度 DRD 值所呈現出的同義詞集群，並非由辭典在編撰時特別說明爲同義詞組；而是經由釋義之中所擁有共同詞彙的百分比計算而得。因此，獲得集群後必需再進一步了解各個集群之中的釋義媒介文字爲何，才能說明該集群的特徵。

在了解高關聯度 DRD 之後，爲討論中度釋義關聯值，我們以「目所未睹」與目字所有詞組進行分析，並取出 DRD 值界在 0.3~0.6 的中度關聯 DRD 詞彙，共九個詞彙列於表 3。

表 3.「目所未睹」與同屬「目」字部之詞組比較（DRD 值介於 0.3~0.6 間）

詞彙	釋義	釋義過濾後的詞組	與「目所未睹」相同釋義詞組	DRD 值
目所未睹	從來沒見過，形容極爲罕見。	見,形容,罕見	n/a	n/a
目下有臥蠶	形容眼瞼浮腫，下瞼有臥蠶樣。多見於腎炎病人。	形容,眼瞼,浮腫,有,臥蠶樣,見,於,腎炎, 病人	形容 見	0.33
目連	人名。見「目犍連」條。	人名,見,目犍	見	0.33
目眩神迷	形容所見情景令人驚異。亦作「目眩神搖」。	形容,見,情景,令,人,驚異,作,目眩神搖	形容 見	0.36
目斷飛鴻	極目遠望，直至飛雁不見。常形容離別的悲悽之情。	遠望,直至,飛雁,見,形容,離別,悲悽,情	形容 見	0.36
目眩魂搖	形容所見情景令人驚異。亦作「目眩神搖」。	形容,見,情景,令,人,驚異,作,目眩神搖	形容 見	0.36
目瞪口歪	形容非常生氣的樣子。	形容,生氣,樣子	形容	0.33
目眩神馳	形容所見情景令人驚異。亦作「目眩神搖」。	形容,見,情景,令,人,驚異,作,目眩神搖	形容 見	0.36
目下十行	形容看書看得快。見「一目十行」條。	形容,看書,看,快,見,一目十行	形容 見	0.33
目視雲霄	形容眼界高闊。	形容,眼界,高闊	形容	0.33

從表 3 可知，九個詞彙分別透過釋義媒介文字"*形容*"、"*見*"與辭典條目「目所未睹」產生釋義關聯；但由於與「目所未睹」所共用的詞組太少，而且九個詞彙的釋義又多，從而降低了 DRD 值。使用較多的釋義文字說明詞彙，即表示該詞彙的概念較爲複雜。從表 3 的資料中可看到，與「目所未睹」產生釋義關聯的詞彙爲「形容」與「見」，雖然此二釋義詞並非詮釋該辭條的主要用意，但可知此中度釋義關聯的主要集群概念都爲比喻用詞彙。

同理在我們分析 DRD 值更小的如「眼」與「眼鏡蛇」兩個詞彙(表 4)。因「眼鏡蛇」是屬於表示特定動物的下位詞，從而在辭典編撰中即會使用包含較多上位詞概念的釋義文字以闡述此一詞彙的概念，即該辭彙所涵蓋的概念多且複雜，從而降低了兩個詞彙之間的 DRD 值。

表 4.「眼」與「眼鏡蛇」之詞組比較

詞彙	釋義	釋義過濾後的詞組	釋義中與「眼」相同詞組	DRD 值
眼	目，動物的視覺器官。	目 Na 動物 Na 視覺 Na 器官 Na	n/a	n/a
眼鏡蛇	動物名。爬蟲類有鱗目。長四、五尺，頸部有一對有白邊黑心的環狀斑紋，形如眼睛。背褐色，腹青白色，毒牙有溝，可注毒液，怒則頸膨大，昂首直立，晝伏夜出。俗稱爲「飯匙倩」。	動物 N 名 N 爬蟲類 N 有 Vt 鱗目 N 長 Vt 頸部 N 有 Vt 有 Vt 白邊 N 黑心 Vi 環狀 N 斑紋 N 形 N 如 P 眼睛 N 背 Vt 褐色 N 腹 N 青白色 N 毒牙 N 有 Vt 溝 N 注 Vt 毒液 N 怒 Vi 頸 N 膨大 Vi 昂首 Vi 直立 Vi 晝伏夜出 Vi 俗稱爲 Vt 飯匙倩 N	動物	0.05

綜合上述分析，直接使用釋義關聯程度(DRD)所得的值，因受詞彙所使用的釋義文字長短或涵蓋的知識概念多寡，而有不同的影響。此外若所使用的釋義文字沒有完全的對應，則無法納入相同詞組的計算。最後釋義關聯計算單元是以字與辭彙所含的「知識概念」，因所有過濾後的辭彙字詞是相等權重，因此在計算上則不需爲個別知識概念加權。反觀釋義文字在編撰時，會因該詞彙概念內容特殊而特別偏重在幾個詞上。因爲 DRD 的釋義關聯計算上，是將取自釋義的詞組視爲包括部分或全部的知識概念單元進行計算，從而無法完整表達個別釋義詞組所含蓋知識概念的權重。

3.2 多階層釋義關聯

前述已討論過，應用基於辭典釋義的釋義關聯程度值 DRD 計算時，會因釋義中單詞的權重與解釋文句長短而影響 DRD 值。爲了改善釋義關聯計算使其更加精確，本研究提出多階層釋義關聯(Multi-level Definition Relation Degree, MDRD)。黃居仁指出(2005)漢語的詞是知識表達的最基本單位，因此在釋義之中所使用的詞組也可以再透過辭典的釋義，擴充共同使用的字詞而併入計算的權重中。本研究所建立的辭典語料是以部首屬「目」的字詞，因此進行釋義文字擴充時會以現有的辭典語料進行擴充。

令 X^1 爲條目 x 所有符合過濾條件的辭典釋義詞彙；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；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 $X^{(1+2)} \cap Y^{(1+2)}$ 則分別爲 X^1、X^2 與 Y^1、Y^2 共同擁有的釋義文字的集合。 而 $CoAP_{x,y}^{2}(x) = \dfrac{\left(X^1 + X^2\right) \cap \left(Y^1 + Y^2\right)}{\left(X^1 + X^2\right)}$，表在 x 所有擴充釋義中與 y 共同使用的擴充釋義字詞的比例，則第 n 階層釋義關聯的一般式即可寫成：

$$CoAP_{x,y}^{n}(x)=\frac{\sum_{i=n} X^{i}\bigcap\sum_{i=n} Y^{i}}{\sum_{i=n} Y^{i}} \tag{4}$$

$$DRD_{x,y}^{n}=2\frac{CoAP_{x,y}^{n}(x)CoAP_{x,y}^{n}(y)}{CoAP_{x,y}^{n}(x)+CoAP_{x,y}^{n}(y)},\quad 0\le DRD_{x,y}^{n}\le 1 \tag{5}$$

當釋義關聯計算加入擴充的釋義文字之後，釋義的字詞片段會依所搜集來自辭典語料而增加該字詞的權重與增加計算的詞語廣度，進而產生且增加兩者釋義語意上的關聯程度。以「目的」與「瞄準」爲例(參考次頁表 5)，使用 DRD 計算「目的」與「瞄準」之第一階共同釋義字百分比與釋義關聯值：

$$CoAP(\text{瞄準})=\frac{1}{3}\text{ ，}CoAP(\text{目的})=\frac{1}{9}\text{ ，}DRD_{\text{目的,瞄準}}=2\frac{(\frac{1}{3})\times(\frac{1}{9})}{(\frac{1}{3})+(\frac{1}{9})}=0.167$$

在計算第二階層釋義關聯時，二字的釋義都經過部首屬「目」的字與辭彙進行擴充後並納入計算，則兩者之間的第二階層釋義關聯 $SRD^{2}_{\text{目的,瞄準}}$ 爲：

$$CoAP^{2}(\text{瞄準})=\frac{27}{40}=0.675\text{ ，}CoAP(\text{目的})=\frac{27}{28}=0.964$$

$$DRD^{2}_{\text{目的,瞄準}}=2\frac{(0.675)\times(0.964)}{(0.675)+(0.964)}=0.794$$

從上述的計算可知，透過將第一階層釋義文字內的「目標」一詞的釋義擴充後，「目的」與「瞄準」兩詞之間的釋義關聯提高到 0.794。其中，共同擁有的釋義文字，在第一階段僅只有"*目標*"一個媒介詞。但在經過二階段擴充之後，"*達到*"與"*注視*"兩字會因釋義有共同交集而使權重增加。媒介文字"*達到*"在「目的」的第一階段的釋義文字與第二階段透過「目標」擴充後，釋義文字產生加權效果；同理"*注視*"在「瞄準」的釋義之中亦產生了加權效果。此外，兩者之間的釋義關聯，也因爲「目標」一詞的擴充而加強了彼此的釋義關聯強度。在此特別說明，此擴充之釋義計算僅涵蓋現有辭典語料，即「目」字部所有字辭。

雖然共有的詞組數都爲 27，但「目的」的共有詞組表中，較「瞄準」多計算了分別落於第一階釋義(「目的」的直接釋義)與第二階釋義詞語 (擴充自「目的」釋義詞組中的"*目標*"釋義詞組)的"*達到*"一詞。而相反的，「瞄準」則較「目的」多計算了一組"*注視*"的釋義詞語(分屬於「瞄準」的直接釋義與「目標」的擴充釋義)。由此可進一步推論，「目的」相較於「瞄準」有地理位置轉移的概念義涵；「瞄準」相較於「目的」，則著重在視覺專注的狀態或動作。而從 $CoAP^{2}$ 的數值上看來，「目的」的概念義涵較能表達「瞄準」概念，但「瞄準」的概念涵義則無法表達「目的」的義涵，因「瞄準」一詞的義涵較複雜。

表 5.「目的」與「瞄準」之多階層釋義表 (共同出現的釋義已特別標示)

詞彙	釋義	釋義過濾後的詞組	共有的詞組數
第一階段釋義			
目的	想要達到的目標。如:「人生以服務爲目的。」	想要,*達到*,***[目標]***	1
瞄準	用眼睛注視目標，使發射、投射的動作準確。	用,眼睛,*注視*,***[目標]***,使,發射,投射,動作,準確	
第二階段釋義			
目標	可爲目力的標準或目力能注視的地方。工作或計畫中擬訂要達到的標準。軍事上運用軍隊所望達成的最終目的，或攻擊行動所望殲滅的敵軍或攻占的地區或地點。	爲,目力,標準,目力,注視,地方,工作,計畫,擬訂,*達到*,標準,軍事,上,運用,軍隊,望,達成,目的,行動,望,殲滅,敵軍,攻占,地區,地點	
眼睛	動物身上觀察外物的視覺器官。	動物,身,觀察,外物,視覺,器官	
合併計算之釋義詞組及權重 (字詞後的數字爲計次)			
目的	**行動 1 攻占 1 計畫 1 運用 1 望 2 地區 1 殲滅 1 敵軍 1 爲 1 工作 1 上 1 地方 1 [注視 1] 軍事 1 [達到 2] 軍隊 1 標準 2 目力 2 目標 1 目的 1 地點 1 擬訂 1 達成 1** 想要 1		27
瞄準	**行動 1 攻占 1 計畫 1 運用 1 望 2 地區 1 殲滅 1 敵軍 1 爲 1 工作 1 上 1 地方 1 [注視 2] 軍事 1 [達到 1] 軍隊 1 標準 2 目力 2 目標 1 目的 1 地點 1 擬訂 1 達成 1** 投射 1 發射 1 身 1 觀察 1 準確 1 視覺 1 眼睛 1 外物 1 器官 1 用 1 使 1 動作 1 動物 1		27

由於此辭典的釋義只限於部首屬「目」的字與辭彙，因此在表 5 第一級釋義裡僅有"*目標*"與 "*眼睛*"兩個詞是會被擴充的。其它不屬於部首「目」的字詞不會再經釋義擴充，因此這些不再擴充釋義字詞的權重相對於能擴充的字詞來說權重較低，也就是透過字詞的擴充決定不同釋義詞的權重。另一方面，雖然釋義關聯的計算仍以字形爲主的計算比較原則，但中文經過斷字詞之後的字詞組保有釋義文字要傳遞的知識單元，因此再經過許多階層釋義擴充之後，便能透過使用相同的釋義文字而說明兩個辭彙之間的關聯。依據前述的計算方式，我們計算「目的」與「目」、「目的」與「眼」之間的釋義關聯(表 6)。在「目的」第一階層的辭典釋義之中並沒有包括任何直接與"*眼睛*"一詞有關的字詞，但過濾後的詞「目標」經多階層釋義擴充之後則揭露出"*眼睛*"的概念，並與「目」與「眼」詞產生語意上的關聯。

表 6.「目的」與「目」、「目的」與「眼」第 1~4 階層釋義關聯值

$x,\quad y$	$DRD^{1}_{x,y}$	$DRD^{2}_{x,y}$	$DRD^{3}_{x,y}$	$DRD^{4}_{x,y}$	第 1~4 階層 共有的媒介釋義字
目的，目	0	0	0.049	0.298	身 觀察 視覺 眼睛 外物 器官 動物
目的，眼	0	0	0.044	0.297	身 觀察 視覺 眼睛 外物 器官 動物

接著討論多階層釋義關聯中的階層特性。我們使用相同的語料庫進行擴充多階層的釋義，並計算從第一階到第七階層、第一百階的「眼腦」與包括「眼」字的雙字詞釋義關聯值(圖 5)。圖中所設定的第一百階層釋義關聯值，是爲了確保所有釋義的字詞都已經透過語料庫擴充，也就是多階層釋義關聯值在本語料庫之中已經呈穩定狀態(釋義關聯值不會因爲再擴充而改變)。

「眼腦」的釋義爲"*眼睛*"，但在使用單階層的釋義關聯計算時，會無法與「眼睛」的釋義文字進行概念的計算，從而使 DRD 值爲 0。同樣的情況發生在後續許多雙字詞之中，如「眼圈」、「眼眉」、「眼裡」等 33 個雙字詞。但隨著多階層語意的擴展後，將個別雙字詞與「眼腦」之間共字使用的釋義字詞納入計算後，其 DRD 數值也隨之增加。另一方向，隨著納入計算的擴充釋義字詞的增加，會直接影響釋義字詞占 DRD 權重，DRD 值會突然提高（如「眼神」），使得兩者之間的釋義關聯形成偏差。但這樣的現象在所有擴充釋義詞步驟都完成之後(在本例中爲 Lv7)，所有的 DRD 值就會趨於穩定。也就是在教育部所訂的國語辭典中，在「目」字部所有字詞能解釋的範圍下，所有「眼腦」與包括「眼」字雙字詞的釋義關聯都能從圖 5 Lv7 中得到結果。

從圖 5 中亦可看出詞義「眼腦」與「眼睛」的概念相似。雙字詞的多階層 DRD 值較高者，表示則其概念比較明確，而且「眼睛」的語意概念占的比例較高，如「眼語」、「眼科」「眼罩」等等。多階層 DRD 值在中程度的雙字詞相較於「眼睛」來說，一部分著重在"*器官*"的概念上，如「眼眶」(0.5)、「眼孔」(0.5)等；另一部分雙字詞的概念則是概念延伸自"*目力*"、 "*眼力*"，如「眼格」(0.5)、「眼辨」(0.3)等。

多階層 DRD 值小的雙字詞則是含蓋「眼睛」與其它不同的概念於一詞裡，如「眼庫」雖然與「眼腦」共同使用了 "*外物*"、 "*器官*"、 "*動物*"、 "*身*"、 "*觀察*"、"*視覺*"、"*眼睛*" 七個釋義詞，但「眼庫」釋義原文"*國際獅子會中華民國總會爲推動眼角膜移植，接受病人身後捐獻，以協助有眼病的人恢復視力，於民國五十七年十一月十日正式成立的眼角膜儲備設施。*"經過濾後保留的字詞很多，且能經多階層擴充的釋義文字少，從而無法增加釋義文字的權重，因而降低了 DRD 值。

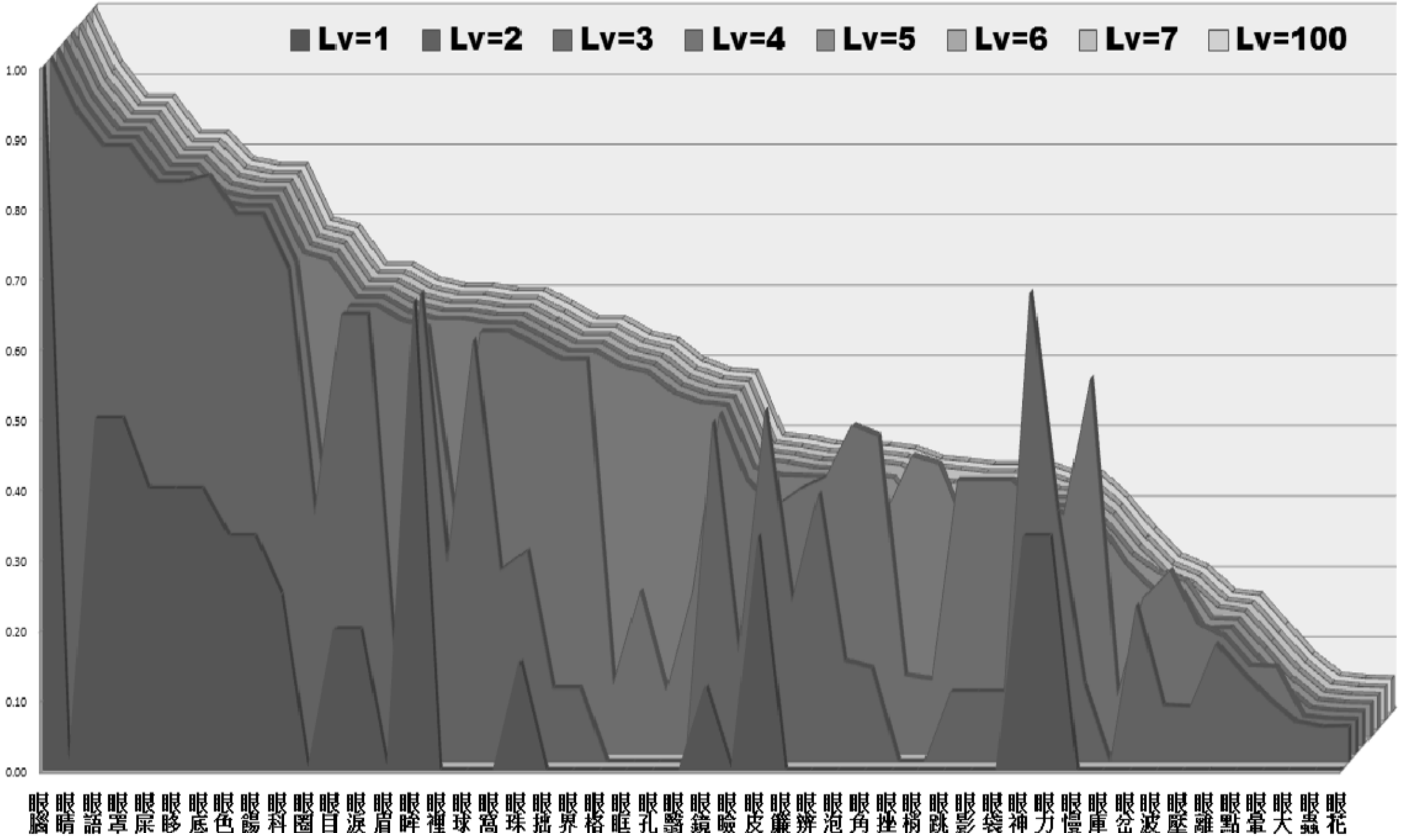


*圖 5. 包括「眼」字的雙字詞對於「眼腦」之多階層釋義關聯值比較***(Lv1-7,** *與* **Lv100)**

接着我們以上述多階層釋義關聯計算原則，進一步分析部首屬「目」的所有 115 個字，以了解與「目」字之間的多階層釋義關聯，表 7 詳列由高至低的多階層釋義關聯值 (其中 33 個字是無法與「目」字產生釋義關聯)。高度的釋義關聯，是使用相同的釋義字詞擁有符合過濾釋義字詞中的比例（已於 3.1 說明）。從表 7 中可以看出「目」與「眼」二字的釋義關聯很高，而「目」與「睛」的釋義關聯卻只有「目」與「眼」的一半，其因在於「睛」的釋義爲"*眼珠。眼睛。*"，其釋義的概念延伸了眼睛器官的概念，而「眼」的釋義"*目，動物的視覺器官。*"則是直接使用了「目」字當作釋義，因此「目」與「眼」釋義關聯十分高(0.9)。 其它與「目」釋義關聯較低的字，如「看」、「相」等都是透過"*觀察*"與「目」產生釋義上的關聯。

表 7. 部首屬「目」的字與「目」之多階層(100 階層)關聯計算表

$DRD_{x,目}^{100}$	屬「目」的字	字數
0.9	眼	1
0.8	睜 眺	2
0.7	瞭 晛 睆 眵 眢 矐 瞋 盱	8
0.6	盹 瞶 矘 瞍 眚	5
0.5	睛 睡 眶 眸 睚 睅 瞚 瞬 瞳 眥	10
0.4	眯 瞇 瞌 眨 矓 矊 睫 睏 眹 瞼	10
0.3	眉 瞪 盯 盲 昧 睨 睞 睽 眴 眩 眭 眠 眙 眝 眊 眇 矔 睒 瞀 瞠 瞽 盳 瞢 盻 矇	25
0.2	瞎 睥 盼 瞜 瞟	5
0.1	看 睪 睢 矉 瞅 瞧 瞑	7
< 0.1	相 睎 睇 眷 睠 瞥 瞰 瞻 眄	9
0	睄 瞄 真 瞿 睋 睹 睿 睾 睩 睬 督 睦 睟 睊 眔 眲 眕 眈 省 矚 矕 矗 矍 睮 瞂 瞏 瞞 瞫 瞯 瞵 盾 眽 直	33

在分析這 33 個無法與「目」字產生釋義關聯的字(即 DRD 值爲 0)，後可將其區分成三類：(1)釋義文中使用到部首屬「見」的字詞，如「見」或「視」等字詞。這類字包括了「睄」、「瞄」、「睩」、「睋」、「睟」、「睮」、「睊」、「眲」、「眈」、「矚」、「瞏」、「瞞」、「瞫」、「瞯」、「瞵」、「眽」、「矕」、「督」、「睹」及「矍」等 20 字；(2)「教育部重編國語辭典修訂本」沒有特別將《說文解字》的釋義納入。「直」與「矗」，這兩個字都有「直」的義涵。《說文解字》釋「直」爲“*正見也*”；「瞞」，"*平目也*"；「真」，"*僊人變形而登天也*"。段玉裁認爲「真」字於字形中包涵了「目」的概念，"*獨言目者，道書云養生之道，耳目爲先。*"「睿」爲"*深明也。通也。*"段玉裁認爲此字是取「目」字的，"*从口。从目。故曰深明*"。「睦」爲"*目順也*"。「督」爲"*察也*"。「眔」爲"*多也*"，段玉裁認爲此字爲"*从乑目*"。「省」爲"*視也*"。「眕」

爲"*目有所恨而止也*"。「睪」爲"*目視也。从橫目，从幸*"。「瞿」爲"*鷹隼之視也*"。「盾」與「瞂」爲"*所以扞身蔽目*"。上述字若在釋義之中納入《說文解字》的釋義，則 DRD 值則不爲 0。(3)無法判定爲與「目」有關聯的字。「睬」字的釋義 "*理會*"，爲聲符字（即「睬」的聲符是「采」，形符是「目」），其義源於「偢睬」(同「瞅睬」、「偢采」)爲"*理睬*"。上述除了「睬」字較爲特殊之外，其它部首屬「目」的字都能透過釋義與「目」產生關聯。多階層釋義關聯無法爲 33 個字建立釋義關聯，是因爲本研究中沒有收集部首屬「見」的字詞釋義與《說文解字》之中的釋義。雖然使用於計算的辭典釋義有所局限，但透過多階層釋義關聯計算可發現多數屬「目」部的字與「目」字之間存在有字義上的關聯，進一步的分析也可證明屬「目」部的字與屬「見」部的字詞間存在釋義上的釋義關聯。

3.3 多階層釋義關聯程度與其它語料統計值的比較

我們於 3.2 闡釋了多階層釋義關聯程度值的原則、計算方式及應用，接下來將以語料統計方法中，使用大樣本進行計算的 MI Value 與 t-score 爲例進行比較。以 Mutual Information 計算得到的 MI Value 是在估計兩個字之間的關係(associations)(Church & Hanks, 1990)，MI 的計算爲：

$$I(x,y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{6}$$

其中，*x, y* 爲兩個待測的字詞，其使用的機率爲 *P(x)*、*P(y)*，*I(x, y)*表兩者相互共有的訊息。*I(x, y)*在計算上是以聯合機率(joint probability)的方式計算兩個待測字詞之間的獨立觀測機率。透過上述的 MI Value 計算，得到的數值決定兩字詞之間的關聯：(1)若大於 3，則具顯著關聯(Hunston, 2002)；(2)若接近 0 則無顯著關聯；(3)若小於 0，則爲互補關聯。

另一方面，t-score 則是爲了計算兩個詞之間是否具統計上的顯著(Gao & Somers, 1998)，其計算方式爲：

$$t \approx \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}} \tag{7}$$

上述兩個待測的字詞 *x, y*，其個別出現的機率爲 *f(x)*、*f(y)*，*f(x,y)*則爲兩個詞共同出現的次數；而 N 則是所有出現的字詞數。Hunston(2002)指出，當 t-score 大於 2 時，則視此二字具顯著關係。在此我們則以「眼腦」與包括「眼」字雙字詞的關聯計算，來比較多階層釋義關聯、MI Value 與 t-score 值三者的差異(圖 6)。在圖 6 中，y 軸爲計算所得的值。我們可以看到 MI Value 與 t-score 兩者趨勢是與 DRD 值相反，且都呈現正相關。從字義上來看，「眼腦」使用了「眼睛」的釋義；但在 MI Value 與 t-score 值上卻比「眼腦」與「眼蟲」的關聯較高，也就表示 MI Value 與 t-score 在比較兩字詞的關聯度時是不適用的。

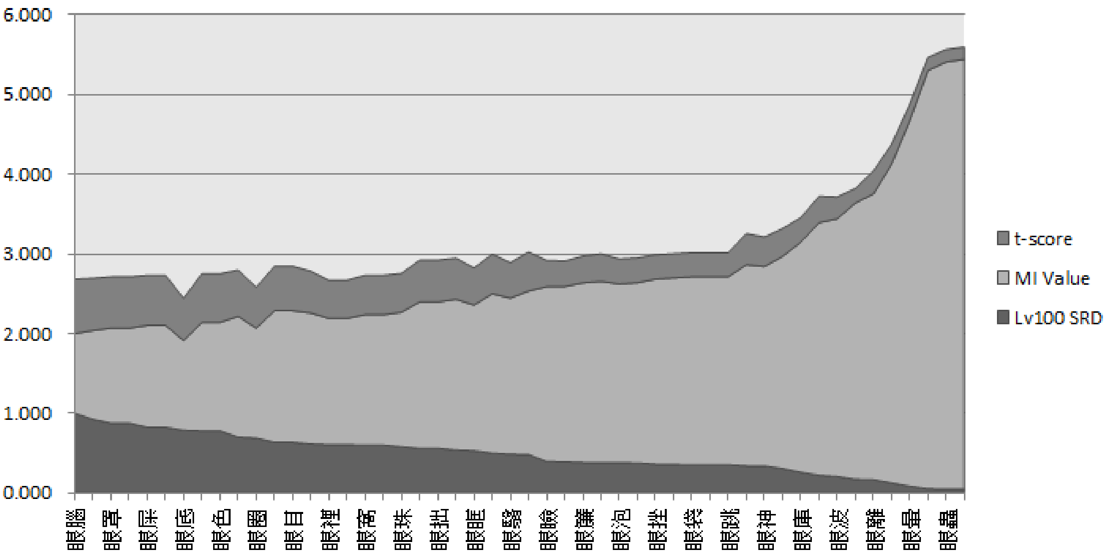


圖 6.「眼腦」與包含「眼」雙字詞之多階層釋義關聯、MI Value 與 t-score 比較

從上述的 MI 值與 t-score 的計算原義來看，二者雖然都是區分兩個字詞在語料庫之中同時出現的頻率的機率比較，但並非是從詞義概念的角度計算兩個字詞之間的關聯。兩者的計算原理與多階層釋義關聯的共用詞所占有的比率雖然相同，但詮釋上卻不同：MI 值是從資訊熵(information entropy)區分兩個字詞所含的釋義字詞是否有共同資訊熵；而 t-score 則是從統計上的顯著性，來區分兩個詞之間的差異。本文提出的多階層釋義關聯的計算是以辭典字詞的釋義做爲語料庫的計算基礎，不同於過去 MI 值與 t-score 計算上使用文本(texts)爲語料庫基礎的計算。多階層釋義關聯計算的原則不僅可以透過語料庫的內容進行多階層的釋義與擴充外，擴充的字詞的概念亦可從辭典的收集來控制釋義的概念延伸。如前述 3.2 透過同屬「目」的字詞擴展後，計算所有部首屬「目」的字與「目」字之間的關聯。

4. 結論

漢語的知識交流單位爲字詞，而且字詞組合的釋義文字亦有知識概念的義涵。本論文試延伸概念隱喻關聯計算至中文的字詞釋義關聯計算，透過辭典對條目的解釋說明文字，提出一套計算詞語釋義關聯的原則。不同於人工的語感詞意分類與語料庫工具計算方式，本文將辭典中的釋義字詞視爲計算基準，並進行多階層次擴充以加強釋義字詞的權重，最後在不同的關聯程度之下說明其集群應用與語言上的義涵。本文實作的語料庫來源-「教育部重編國語辭典修訂本」，是提供大眾查尋字詞的工具。雖然此辭典自民國八十二年編輯完成後，即很少增改其中內容，但從中取得字詞釋義語料進行計算的結果亦能表現出多階層釋義關聯計算的特色。在本篇文章中，我們討論了單一階層的釋義關聯計算，將與「眼」相關及屬「目」的所有字詞進行計算後，結合社會網路分析工具分析高度釋義關聯集群，不同的集群會依其釋義中共同使用的字詞進行群聚，即同義詞的群聚。

之後我們分析中度的釋義關聯計算中，思考如何處理釋義詞裡與不同概念的權重如何訂定，從而發展出多階層釋義關聯的原則，並說明多階層的釋義對釋義關聯值的變化。接著則分析所有部首屬「目」的字與「目」字之間的多階層釋義關聯，以詮釋漢字義符的特色。並從進一步的分析中得到「目」部字集與「見」部字集的關聯，以及「教育部重編國語辭典修訂本」內的文字釋義裡，沒有特別將字的來源加入，了解多階層釋義關聯在分析漢字與部首之間的使用方式。在本文的最後我們又比較了多階層釋義關聯與 MI Value、t-score 之間的差異，說明了多階層 DRD 值是從概念層級計算兩個字詞之間的交集程度。

本文所介紹的多階層釋義關聯雖然能表現出兩個字詞共同使用的釋義字詞交集程度，但因釋義詞來源爲辭典且經過濾而得，因此仍無法完全表達該字或詞的完整概念涵意。此外本研究進行的多階層釋義擴充僅保留動詞與名詞的釋義，且受限於所取得的辭典語料中部首屬「目」的字與詞彙。雖然此一限制於擴充時，能維持「目」部字的字義概念發展方向相同，但仍無法完整擴充所有的字詞。並且在辭典編寫釋義的過程中，若作者沒有將符合條目的所有涵義都編入釋義中，使用多階層釋義關聯計算亦無法完全將詞義的概念均列入運算。縱有上述限制，但透過多階層釋義關聯計算，能將前人所制訂的辭典轉換成爲語言計算的工具。也可再進一步與其它計算語言學的方法及網路分析原則結合，以探討漢字字義與詞義之間的概念及釋義集群中的語意關聯。

參考文獻

Chung, S.F., Chen, C.H., & Chao, F.Y.A. (2009). Building a Database of Related Concepts of Mandarin Metaphors Based on WordNet and SUMO. In *IEEE International Conference on Semantic Computing*, Berkeley, CA, 378-383.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), 22-29.

Dong, Z., & Dong, Q. (2003). HowNet-a hybrid language and knowledge resource. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 650-655.

Huang, C.R., Chang, R.Y., & Lee, S.B. (2004). Sinica BOW (bilingual ontological wordnet): Integration of bilingual WordNet and SUMO. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Miller, G.A. (1995). WordNet: a lexical database for English. Communiations of the ACM, 28(11), 39-41.

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*。

段玉裁(1808)。*說文解字注*。

祝清(2009)。語法隱喻視角的漢語詞典名動詞實證研究。*四川:內江師範學院學報*。

高照明(2007)。中文詞彙語意資料的整合及擷取：詞彙語意學的觀點。臺北:第十九屆自然語言與語音處理研討會論文集。

許慎(100)。*說文解字*。

黃居仁(2005)。漢字知識表達的幾個層面：字、詞與詞義關係概論。*漢字與全球化國際學術研討會*。臺北。

黃居仁、陳聖怡、楊雅君(2008)。意符知識系統研究:[五官類] 意符的概念衍生與知識表徵。*第九屆漢語詞彙語義學研討會*。新加坡。

黃居仁、蔡柏生、朱梅欣、何婉如、黃麗婉(2003)。詞義與義面:中文詞彙意義的區辨與操作原則。In *Proceedings of the Fourth Chinese Lexical Semantics Workshop*，23-25。

鄭文泉(2004)。立象盡意 從符號學的角度論楷體漢字的形上學。*國立中央大學哲學所博士論文*。

黎傳緒(2004)。“相”字的解析。*北京教育學院學報*，18(4)，17-29。

薛榕婷(2003)。說文解字人與自然類部首之文化詮釋。*淡江大學中國文學研究所碩士論文*。

羅益民(2007)。詞典內外的釋義系統。*中國辭書學會雙語詞典專業委員會第七屆年會*。重慶。

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 41-60
© The Association for Computational Linguistics and Chinese Language Processing

Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction

Bor-Shen Lin[*] and Yi-Cong Chen[*]

Abstract

With the evolution of human lives and the spread of information, new things emerge quickly and new terms are created every day. Therefore, it is important for natural language processing systems to extract new words in progression with time. Due to the broad areas of applications, however, there might exist the mismatch of statistical characteristics between the training domain and the testing domain, which inevitably degrades the performance of word extraction. This paper proposes a scheme of word extraction in which histogram equalization for feature normalization is used. Through this scheme, the mismatch of the feature distributions due to different corpus sizes or changes of domain can be compensated for appropriately such that unknown word extraction becomes more reliable and applicable to novice domains.

The scheme was initially evaluated on the corpora announced in SIGHAN2. 68.43% and 71.40% F-measures for word identification, which correspond to 66.72%/32.94% and 75.99%/58.39% recall rates for IV/OOV, respectively, were achieved for the CKIP and the CUHK test sets, respectively, using four combined features with equalization. When applied to unknown word extraction for a novice domain, this scheme can identify such pronouns as “海角七號” (Cape No. 7, the name of a film), “蠟筆小新” (Crayon Shinchan, the name of a cartoon figure), “金融海嘯” (Financial Tsunami) and so on, which cannot be extracted reliably with rule-based approaches, although the approach appears not so good at identifying such terms as the names of humans, places, or organizations, for which the semantic structure is prominent. This scheme is complementary with the outcomes of two word segmentation systems, and is promising if other rule-based approaches could be further integrated.

[*] Department of Information Management, National Taiwan University of Science and Technology,
Tel: (886)-2-2703-1225 Fax: (886)-2-2737-6777
E-mail: bslin@cs.ntust.edu.tw; m9709104@mail.ntust.edu.tw

Keywords: Unknown Word Extraction, Word Identification, Machine Learning, Multilayer Perceptrons, Histogram Equalization.

1. Introduction

With the evolution of human lives and the accelerated spread of information, new words are created quickly as new things emerge every day. It is then necessary for natural language processing systems to identify and learn new words to progress with time. Chinese word segmentation systems, for example, typically utilize large dictionaries collected over a long period of time. No matter the size of the vocabulary for the dictionaries, it is hardly possible for them to include all of the words or phrases that have been invented so far in the extensive knowledge domains, not to mention to predict in advance new terms to appear in the future. Therefore, it is more practical for Chinese word segmentation systems to use dynamic dictionaries that can be updated quickly and frequently with the new words found in the corpora of the desired domains. Hence, unknown word extraction is actually essential for quite a few natural language processing systems. It is also useful for exploring hot or new terms for desired knowledge domains or internet communities.

The approaches to unknown word extraction can be roughly divided into two categories, rule-based approaches and statistical approaches. For rule-based approaches, semantic rules for specific types of words, such as the names of humans, places, and organizations, normally are specially designed (Sun *et al.*, 1994). For statistical approaches, statistical features in corpora typically have been computed and used for the decision in the threshold test. Occurrence frequency, for example, is a widely used feature (Lu *et al.*, 2004). In such approaches, the threshold is often obtained heuristically and might depend highly on the corpus. In addition, statistical approaches and rule-based approaches can be combined. Some approaches have used statistical features obtained from the corpus and have designed rules for various types of unknown words based on these features, through which even the unknown words with low occurrence frequency can be extracted (Chen *et al.*, 2002). For most of the approaches, the decision rules are obtained from the training corpus heuristically, and perhaps cannot be applied to the testing domain. Therefore, use of machine learning approaches with more general features is suggested in order to obtain the decision boundary by learning automatically. Liang, for example, proposed a tri-syllable filter for screening the word candidates and the artificial neural network with statistical features for the final decision (Liang *et al.*, 2000). Nevertheless, the trained artificial neural network is not shown to be able to be applied to novice domains. Besides, Goh *et al*. made use of the character features (the POS and position) in support vector machine to extract new words (Goh *et al*., 2003).

To reduce the dependency of the word extraction scheme on the training corpus so that use in diverse or novice domains becomes possible, this paper utilizes the machine learning approaches to combine the statistical features. Histogram equalization for statistical features was further introduced to compensate for the mismatch between the training and testing corpora that might come from the difference in corpus size or the change of the domain. It is then unnecessary to retrain the model parameters, and the extraction approach becomes more general for new domains. This scheme was first evaluated on SIGHAN2 corpora for traditional Chinese provided by Chinese Knowledge Information Processing Group (CKIP) and City University of Hong Kong (CUHK). When combing four heterogeneous statistical features, DLG, AV, Link, and PreC, and applying histogram equalization for DLG, the F-measures of 68.43% and 71.40% for within-domain CKIP corpus and cross-domain CUHK corpus, respectively, can be achieved. This scheme was finally used to explore unknown words in a novice domain of a news event. When compared with the words extracted by two word segmentation systems provided by CKIP and Institute of Computing Technology Chinese Academy of Science (ICTCAS), it was found that this approach is complementary with the other two. Such terms as “海角七號”(Cape No. 7, the name of a film), “蠟筆小新”(Crayon Shinchan, the name of a figure in a cartoon), “金融海嘯”(Financial Tsunami), and so on, with prominent statistical characteristics but less structure in semantics, can be extracted successfully by the proposed approach only. These terms are hard to identify using rule-based approaches because it is difficult to draw semantic rules from such terms. Without using semantic rules, however, this extraction approach seems less robust for extracting the names of humans, places, or organizations with prominent structure. This, however, could be overcome by integrating the proposed scheme with the rule-based approaches.

2. Statistical Features

Every sentence in a Chinese corpus contains a sequence of characters. If every combination of adjacent characters in a sentence must be considered as a word candidate, there would be huge number of word candidates where a large portion would be redundant. Therefore, every combination of adjacent characters, denoted as “character group” in this paper, needs to be screened first so the total number of word candidates can be reduced to a manageable size and the statistics could be computed. The occurrence count for each character group, *i.e.* the character n-gram, is computed and used as one of the screening criteria. Those character groups with length less than eight and with occurrence count more than or equal to five are accepted as word candidates. For each word candidate, the statistical features are computed as below.

2.1 Logarithm of Character N-Gram (*LogC*)

$$LogC(T_i) = log\left(C(T_i)\right) \tag{1}$$

T_i : the word candidate with index *i*.

$C(T_i)$: the occurrence count for the word candidate T_i.

Since words tend to appear repeatedly in the corpora, those word candidates with high occurrence count are more probable to be words. Nevertheless, there are often quite a few false alarms when occurrence count is the only decision feature.

2.2 Description Length Gain (*DLG*)

$$DLG(T_i) = L(X) - L(X[@ \rightarrow T_i]) \tag{2}$$

$$L(X) = -\left|X\right| \sum_{x \in V} p(x) log_2 p(x)$$

X : all sentences in the corpus.

$X[@ \rightarrow T_i]$: all sentences in the corpus with T_i replaced as "@"

$L(\cdot)$: the entropy of the corpus.

$|X|$: the total number of characters in the corpus.

V: the set consisting of all characters in the corpus.

Description length gain was proposed by Kit *et al.* to measure the amount of information for every word candidate according to the degree of data compression (Kit *et al.*, 1999). In Equation 2, $L(X)$ is the entropy of the corpus containing the word candidate T_i, while $L(X[@\rightarrow T_i])$ is the entropy of the corpus with T_i replaced by the token "@". Therefore, $DLG(T_i)$ indicates the entropy reduction due to the elimination of the word candidate T_i in the corpus, or equivalently the information gain of the corpus contributed by including the word candidate T_i. The more information a word candidate contributes, the higher the probability that it is a word.

2.3 Accessor Variety (*AV*)

$$AV(T_i) = min\{ L_{AV}(T_i), R_{AV}(T_i)\} \tag{3}$$

$L_{AV}(T_i)$: the number of different left-context characters for the candidate T_i

$R_{AV}(T_i)$: the number of different right-context characters for the candidate T_i

Access variety was proposed by Feng *et al.* to estimate the degree to which a character group occurs independently in the corpus (Feng *et al.*, 2004). The access variety for a character group is evaluated by counting the number of different characters in its left or right context. If the access variety is high, it implies the character group is often used independently in diverse contexts and tends to be a word. On the contrary, low access variety implies that the character group is often used together with specific characters, and thus tends to be a part of a word instead of being a word itself. Hence, the larger the access variety is, the more probable the character group is a word.

2.4 Logarithm of Total Links (*Link*)

The feature *LogC* defined in Eq. 1 considers the occurrence count of a word candidate but does not take its internal structure into account. Since the occurrence counts of partial character sequences for a word candidate (denoted as *links* here) might also provide some evidence in support of this candidate being a word, a novel feature for estimating such links is proposed as follows.

$$Link(T_i) = log\left(\sum_{k \le 1} C(S(T_i;k,l))\right) \tag{4}$$

S(T_i;k,l): a partial character sequence of the word candidate T_i from position *k* through position *l*.

The word candidate “行政院長” (meaning *executive director*), for example, has the partial character sequences “行政,” “行政院,” “行政院長,” “政院,” “政院長,” and “院長,” in which the first three and the last one are also known words. The occurrence counts of these internal links can be accumulated, and the logarithm of the summation can be taken to obtain this feature.

2.5 Independence of Prefix Character(*PreC)*

In the Chinese language, some characters are frequently used and co-occur with other words as prefixes. The preposition “在” (meaning *at*), for example, might co-occur with the words “台北” (Taipei), “拍攝” (take a photo) or “學校” (school), and so on. Since such prefix characters are of high frequency, their combinations with other words (*e.g*. “在台北”, “在拍攝” or “在學校”) might also be of high frequency. This induces quite a few false alarms when only occurrence count is used for word extraction. To alleviate such problems, a novel feature is proposed here to measure the independence of the prefix character for a word candidate, which is defined as the average of the occurrence counts for all the character groups with the same prefix character.

$$\overline{C}(F) = \sum_{x \in S(F)} C(x_{1L}) \tag{5}$$

$$PreC(T_i) = \begin{cases} \frac{1}{|S(F)|}\overline{C}(F) & if\,|T_i| > 2 \\ C(T_i) & elsewhere \end{cases}$$

F: the prefix character of the word candidate T_i.

S(F): the set consisting of the character groups with the prefix character *F* and with length larger than two.

|S(F)|: the number of the character groups in the set *S(F)*.

x_{1L}: the partial sequence of a character group *x* after eliminating its prefix character *F*.

For the prefix character "在," the independence is computed according to the occurrence counts of those character groups whose first character is "在," such as "在台北," "在學校," and "在拍攝". If the average of these occurrence counts is high, it means this prefix character has high variety of context and should be separated from the other characters in a word candidate. In such a case, every word candidate with this prefix character is less probable to be a word. In other words, the higher the independence of the prefix character, the less probable that the candidate is a word.

2.6 Normalization

As the statistical features defined above are computed from the corpus, the dynamic range of the features for the training and the testing corpora might be different when the corpus is obtained from different domains and has a different size. Therefore, the statistical features need to be normalized before being used as the inputs of the classifier. In this paper, the following formula is utilized to normalize the features onto the range of 0 to 1.

$$F(v) = \frac{v - Min(y)}{Max(y) - Min(y)} \tag{6}$$

v: the input value of the feature.

y: the type of the feature.

Min(y): the minimum value of the feature *y*.

Max(y) : the maximum value of the feature *y*.

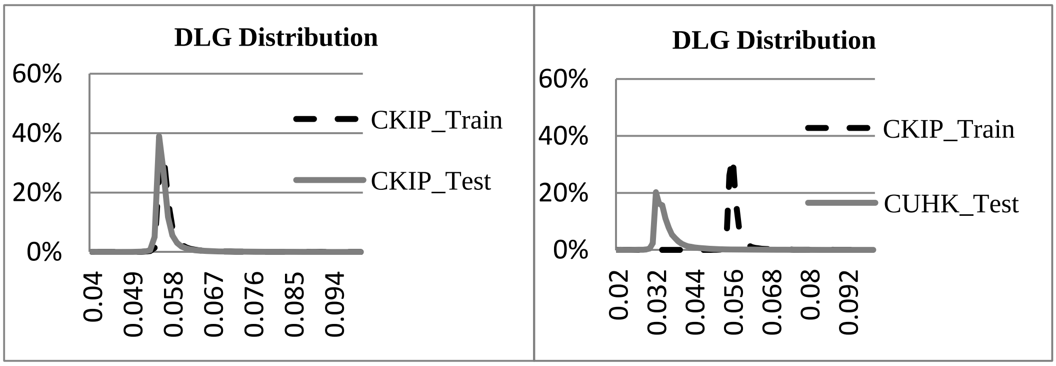
F(v) : the output value of the feature after normalization.

3. Word Extraction Method

3.1 Distribution of Statistical Features

Since the statistical features in this paper are obtained from the corpora, both the dynamic range and the distribution for the features might change. Although a normalization formula is introduced in Section 2.6 to deal with the problem, it is probably not sufficient for compensating for the mismatch of the feature distributions between the training and testing corpora, which often leads to performance degradation when the statistical approach is applied to new domains. In this section, we analyze how the histograms for the statistical features might differ between various domains.

The SIGHAN2 corpora, provided by CKIP and CUHK, respectively, are used for analysis here. First, the CKIP corpus was randomly and equally divided into two sets, named as the CKIP_Train set for training and the CKIP_Test set for testing, respectively. CKIP_Test can be regarded as the within-domain test set. The corpus provided by CUHK is used as the cross-domain test set, and named as CUHK_Test set. The histograms of *DLG* feature for the CKIP_Train and CKIP_Test sets are depicted in Figure 1(a), while those for the CKIP_Train and CUHK_Test sets are shown in Figure 1(b). In Figure 1(a), it could be observed that, for corpora in the same domain with compatible sizes, the dynamic ranges of the DLG feature are very close, while the distributions still differ a little. It can also be noticed that, in Figure 1(b), the histograms for CUHK_Test set and the CKIP_Train set differ more prominently. Not only the dynamic ranges but also the shapes of the distributions differ for the two sets. If the classifier is trained with the CKIP_Train set and tested with the CUHK_Test set, the *DLG* feature appears useless without being further calibrated. More sophisticated normalization schemes will be discussed in the following section.



(a) Within-domain ***(b) Cross-domain***

Figure 1. DLG distribution for different corpora.

3.2 Advanced Normalization Schemes

When the mismatch between the training set and the testing set is significant, the classifier generally fails to classify the testing data reliably. Since we hope to use the classifier for word extraction to explore novice domains, such a problem is inevitable. To handle this problem, a typical normalization scheme, mean standard deviation weight (denoted as MSW here) was often used, as defined below.

$$X_d = M_d + \sigma_d \left(\frac{X_s - M_s}{\sigma_s} \right) \qquad (7)$$

d : the destination domain.

s : the source domain.

M_d: the mean of the distribution for the destination domain.

M_s: the mean of the distribution for the source domain.

σ_d: the standard deviation of the distribution for destination domain.

σ_s: the standard deviation of the distribution for source domain.

X_s: the feature value obtained from the source domain.

X_d: the feature value for the destination domain.

Note that the source domain denotes the testing domain, while the destination domain denotes the training domain. This is because the classifier was trained with the training corpus, so the features for the testing corpus should be transformed back to the training domain to match the distribution of the training data as much as possible. MSW is a linear normalization scheme according to the distance between the feature value and the mean measured with the standard deviation. When the shapes of the distributions differ largely between the source and the destination domains, such a mismatch cannot be compensated for simply by linear shift or scaling, and MSW might not be effective enough.

Another normalization scheme, histogram equalization, denoted as HEQ here, was first introduced in image processing community and used for enhancing the contrast of an image (Hummel *et al.*, 1977; Efford 2000). As HEQ is a common technique for adjusting the statistics of the features via transformation, it can be used to compensate for the mismatch between different domains. This technique was successfully applied to such areas as speech or music processing for compensating for the mismatch of statistical features between the training and the testing domains (Ángel de la Torre *et al.*, 2005; Gallardo-Antolín *et al*. 2010). The transfer function of histogram equalization is described as follows.

$$X_d \;=\; P(X_s) \cdot (X_{MAX} - X_{MIN}) + X_{MIN} \tag{8}$$

X_s: the input feature from source domain.

X_d: the output feature of destination domain.

$P(X)$: the cumulative distribution function in the source domain.

$P_{EQ}(X)$: equalized cumulative distribution function in the destination domain.

X_{MAX}: the maximum value for the feature.

X_{MIN}: the minimum value for the feature.

Figure 2 illustrates how histogram equalization is performed. *P(X)* is the cumulative distribution function (CDF) of feature X in the source domain, as denoted by the solid curve, while *$P_{EQ}(X)$ is the equalized* cumulative distribution function in the destination domain, as denoted by the dashed line. The transfer function between the input feature X_s and the output feature X_d has to make the equality, $P(X_s) = P_{EQ}(X_d)$, hold, which leads to Equation 8. Since the heuristic cumulative distribution function of the output feature, $P_{EQ}(X_d)$, is desired to be linear, the corresponding probability density function, i.e. the histogram, needs to be uniform (equalized). Both HEQ and MSW have monotonic transfer functions, but the transfer function for HEQ could be nonlinear, and its output features in the destination domain will fall into the same dynamic range from X_{MIN} to X_{MAX} as the input features in the source domain.

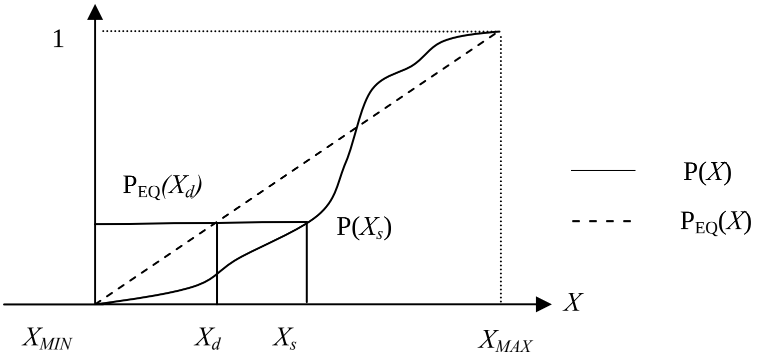


Figure 2. Histogram equalization.

When applying HEQ to word extraction, the cumulative distribution functions of the features for the training and testing domains need to be computed first, and are here denoted as $P_{TRAIN}(X)$ and $P_{TEST}(X)$, respectively. In the training phase, the features obtained from the training domain (with CDF $P_{TRAIN}(X)$) need to be transformed to the equalized domain according to Equation 8 so as to obtain the features for training. That is, the classifier is trained with the equalized features. In the classification phase, the features obtained from the testing domain (with CDF $P_{TEST}(X)$) also need to be transformed to the equalized domain, and the classifier then performs classification for equalized features.

It should be noted that, for either MSW or HEQ, such statistics as mean, standard deviation, and CDF need to be computed first so the transformation of the features can be performed accordingly. This imposes an extra limitation to batch mode for the statistical approaches since the testing corpus for computing statistics needs to be collected beforehand.

3.3 Classifier Based on Multilayer Perceptrons

The structures and rules for word formation in the Chinese language are so sophisticated that it is quite difficult to perform word identification based on a single feature. Occurrence count, for example, is not a reliable enough feature because quite a few fragments (*e.g*. “是非常,” meaning *is very*) occur frequently, but should not be regarded as words. If multiple decision features with complementary characteristics could be combined appropriately, better performance could be obtained in general. In this paper, a classifier based on multilayer perceptrons (MLP) is used for word verification. MLP is a machine learning approach based on nonlinear regression. In order to minimize the square errors, the gradient descent algorithm is applied, and the connection weights in the network are updated iteratively according to the errors propagated backwards till the estimation error converges. Figure 3 is the proposed word verification scheme for combining multiple features with an advanced normalization scheme. In this paper, the MLP classifier contains 3 layers of neurons and is trained 50000 times iteratively. The hidden layer contains five neurons, while the number of neurons for the input layer depends on how many features are used for training and testing. For every word candidate, the features *LogC*, *AV*, *Link*, *PreC*, and *DLG*, were computed with Equations 1 through 5 and normalized with Equation 6. The feature *DLG* was further processed with the advanced normalization scheme, HEQ or MSW, because of the significant difference between the histograms for the training and the testing data. In the selection module, features were combined to form the input vector $\boldsymbol{x}$ of the MLP classifier, whose output y is between 0 and 1. Through a threshold test on the output y, it can finally be decided whether the word candidate is accepted as a word. This word verification scheme, together with the screening process for word candidates as proposed in Section 2, can be used as a preprocessing stage for such NLP tasks as chunking or word segmentation to explore new terms quickly and efficiently for novice domains, such as news events or emerging communities.

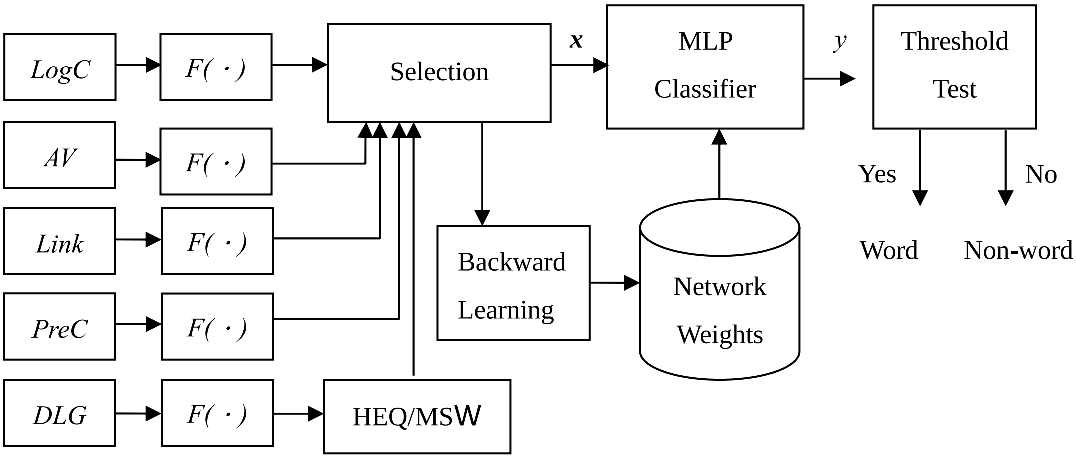


Figure 3. Word verification scheme based on MLP classifier.

4. Experiments and Analysis

In this section, the corpora CKIP_Train, CKIP_Test, and CUHK_Test, as described in Section 3.1, were used for experiments. They contain 361,691, 363,382, and 54,511 sentences and contain 222,446, 224,929, and 149,160 word candidates, respectively. In addition, the numbers of words for them are 33,429, 33,661, and 22,913, respectively, according to the sentences with word segmentation in every corpus. The segmentation results were used to label the ground truths for word verification, *i.e.*, whether a candidate can be a word or not in general without considering its usage context. That is, no information regarding to the local context of a candidate is tracked or used in word verification. The details of the experimental corpora are depicted in Table 1.

Table 1. Details of experimental corpora.

Corpus Name	Purpose	No. of Sentences	No. of Word Candidates	No. of Words
CKIP_Train	Training (Inside Test)	361,691	222,446	33,429
CKIP_Test	Within-domain Test	363,382	224,929	33,661
CUHK_Test	Cross-domain Test	54,511	149,160	22,913

First, the basic experiments of word verification were conducted using at least four types of features based on the architecture in Figure 3, but the advanced normalization scheme, MSW or HEQ, was not applied in this test. Here, CKIP_Train was used for training while CKIP_Train, CKIP_Test, and CUHK_Test were used for testing, respectively, and the results were shown in Table 2. In Table 2, ALL denotes all of the five features defined in Section 2 were used, while the others denote one of the five features was excluded. “No_LogC,” for example, means that the feature *LogC* was excluded, so the other four features were used as the input features. It can be seen in Table 2 that nearly optimal F-measures of 60.09%, 60.03%, and 63.03% can be achieved for the three corpora, respectively, through combining the four features *DLG*, *AV*, *Link*, and *PreC* where *logC* is excluded (No_LogC). This implies the feature *LogC* appears to be redundant and relatively replaceable. This is partly because the occurrence count is included in the more informative feature, *Link*, defined by Eq. 4. Therefore, in later experiments, the feature *LogC* was not used. We can also find in Table 2 that the recall rates for OOV words were worse than the In-Vocabulary (IV) words for CKIP_Test and CUHK_Test sets. This is because the classifier is trained by IV words and non-words in CKIP_Train. We hope the word detector trained with more IV words can achieve more sufficient training and grasp the major stochastic characteristics of IV words such that it could detect novel terms with similar stochastic characteristics in novice domains. Of course, it is theoretically possible to build an OOV detector directly with machine learning approaches. Nevertheless, the number of OOV words is much fewer than IV words such that the insufficient training often makes the classifier suffer from the over-fitting problem. For the same reason, the performance indices for word identification are mainly adopted instead of OOV detection in the development process.

Table 2. Verification performance by F-measure and R_{IV}/R_{OOV} for various testing corpora.

		CKIP_Train (inside test)	CKIP_Test (within-domain)	CUHK_Test (cross-domain)
F	No_LogC	60.09%	60.03%	63.03%
	No_DLG	57.11%	57.21%	62.59%
	No_AV	51.57%	51.74%	54.89%
	No_Link	48.15%	48.06%	42.62%
	No_PreC	53.39%	53.19%	56.76%
	ALL	59.74%	59.69%	63.91%
R_{IV}/R_{OOV}	No_LogC		65.75%/37.51%	74.84%/56.33%
	No_DLG		64.78%/34.99%	73.94%/55.11%
	No_AV		59.40%/41.35%	64.67%/52.22%
	No_Link		60.17%/31.97%	45.44%/41.46%
	No_PreC		72.20%/36.39%	67.60%/43.41%
	ALL		65.15%/35.42%	70.61%/56.18%

Then, the experiments were conducted with the advanced normalization scheme (MSW or HEQ) further applied individually. Figure 4 shows the verification performances with MSW (denoted as MSW), with HEQ (denoted as HEQ), and without MSW/HEQ (denoted as No_Equ), respectively. As can be seen in this figure, after applying HEQ for the feature DLG, the optimum F-measure rises significantly from 60.03% to 68.43%, but there is hardly any improvement achievable by applying MSW. The reason is probably that the dynamic ranges of *DLG* for the training and testing domains, as depicted in Figure 1(a), are almost the same, so MSW, which simply shifts and scales the features, becomes unhelpful. The shapes of the histograms in Figure 1(a), however, differ a little, so HEQ can help improve the performance through nonlinear transformation.

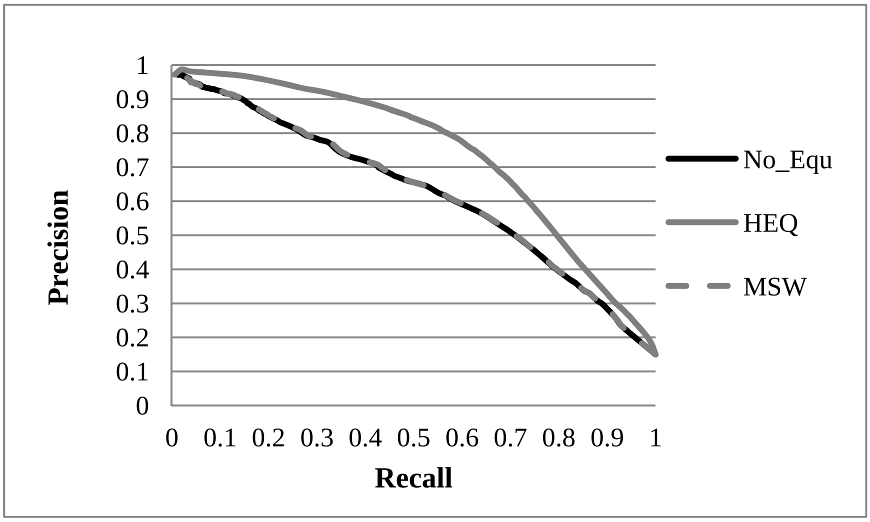


Figure 4. Performance improvements by MSW/HEQ for within-domain test.

Note here that only the performance of HEQs for *DLG* is depicted because the other features are not helpful when integrated with HEQ in our auxiliary experiments. For *Link* and *AV*, the histograms for the training and testing domains almost coincide such that it can hardly get benefits from equalization, while the extra nonlinear transformation of the feature might degrade the performance. This is similar to the robustness issue, where the robustness approaches for compensating for the mismatch between two environments usually incur the side effect of degrading the performance if the mismatch does not exist. In addition, for *PreC*, the histograms are very sparse and jerky with many zeros in bins, since many word candidates might share the independence of a prefix character. Therefore, it is not easy to model the cumulative distribution functions smoothly so HEQ can be well applied, and the results are not shown here.

Further, the same experiments were conducted for cross-domain testing data. That is, the CKIP_Train corpus was used for training while the CUHK_Test corpus was used for testing. The experimental results are shown in Figure 5. As can be observed in this figure, both MSW and HEQ can help improve the verification performance, but apparently the improvement for HEQ is more prominent. When comparing Figure 5 with Figure 4, it can be observed that the trend of MSW for the cross-domain test differs from that for the within-domain test. This is because, in Figure 1(b), the values of the cross-domain features become significantly smaller, which leads to rejections and degrades the verification performance. Such a problem can be alleviated slightly by MSW, which shifts the values, though the improvement is limited. HEQ, on the other hand, can adjust for the features of not only the dynamic range but also the shape of the distribution, so the improvement of the performance is more significant, with the F-measure increased from 63.03% to 71.40%.

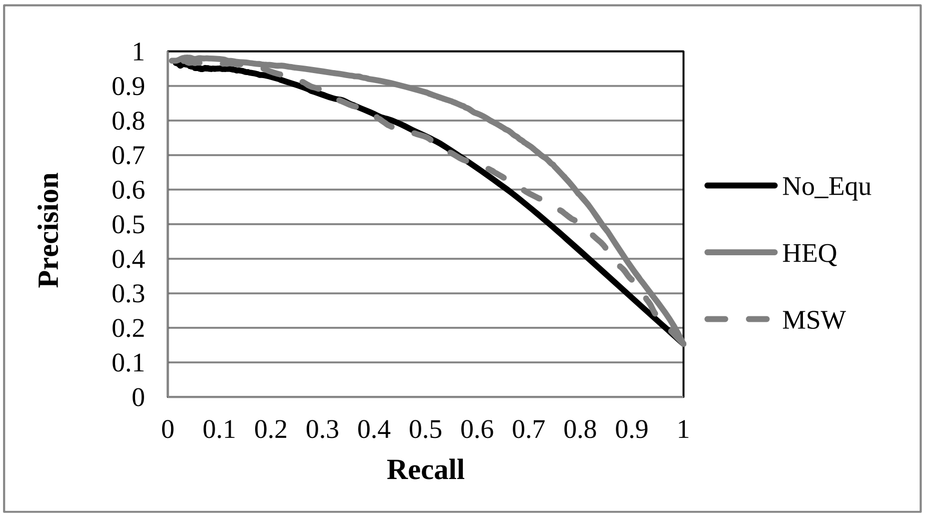


Figure 5. Performance improvements by MSW/HEQ for cross-domain test.

The above results show that HEQ can compensate for the mismatch of the *DLG* feature between the training and testing domains effectively, and the improvements achieved for cross-domain test and within-domain test are compatible. Besides, HEQ can improve the performance more significantly than MSW. This is because, the nonlinear equalization of HEQ works effectively even if the shapes of the distributions between different domains are quite different, but the equalization of MSW by shifting or scaling can work well only when the distributions have close shapes.

The above word verification scheme is further applied to the statistical classifier based on Gaussian mixture models (GMM) for comparison. That is, the MLP-based classifier in Fig. 3 is replaced with the GMM-based classifier. The F-measures for MLP and GMM classifiers with and without HEQ are shown in Table 3. As can be seen in this table, HEQ can also help to improve the verification performance for GMM-based classifier, but here MLP-based classifier achieves significantly better performance. GMM-based classifier is more sensitive to domain change, which can be compensated for appropriately through histogram equalization.

Table 3. F-measures and R_{IV}/R_{OOV} for MLP-based and GMM-based classifiers.

	CKIP_Train		CKIP_Test		CUHK_Test	
	No_HEQ	HEQ	No_HEQ	HEQ	No_HEQ	HEQ
MLP (F)	60.09%	68.70%	60.03%	68.43%	63.03%	71.40%
MLP (R_{IV}/R_{OOV})			65.75% 37.51%	66.72% 32.94%	74.84% 56.33%	75.99% 58.35%
GMM (F)	64.37%	64.65%	59.05%	64.44%	33.47%	66.34%

5. Unknown Word Extraction for Novice Domain

In this section, the word extraction scheme was applied to a novice domain. First, the corpus of the novice domain was collected from a news website in 2009 using the keyword of a news event, “八八水災” (the flood on August 8th). This corpus was used for test and denoted as UKW_Test. From 32,207 sentences in the corpus, 81,447 word candidates were extracted in accordance with the screening criteria in Section 2. Statistical features for these candidates were then computed and used as the input for the classifier trained with the corpus CKIP_Train. The system architecture is similar to that in Figure 3, but the threshold test for each candidate is not applied here. Instead, the classifier outputs between 0 and 1 for all the candidates sorted so as to obtain the 10,000 out of 81,447 candidates with the highest scores. The 10,000 candidates then were regarded as those words accepted by the classifier, while the other 71,477 candidates were regarded as rejected.

5.1 Labeling the Ground Truths

The main problem in applying word extraction in a novice domain is that there is no consensus about the definition of words for the Chinese language; therefore, it is difficult to decide the ground truth for every candidate. The CKIP corpus and CUHK corpus, for example, have different criteria for word defined by the organizations. Table 4 displays some sentences in which the definitions of some words are different. As can be observed in this table, “奶粉錢” (fee for milk) is regarded as a word in the CUHK corpus, but segmented into “奶粉”(milk) and “錢” (fee) in the CKIP corpus. Since such discrepancies exist, our strategy for labeling the ground truths here is to use two word segmentation systems to segment the test corpus, and accept a word candidate as a word if a consensus between the two systems can be reached. Those terms with discrepancy between the two systems then are inspected manually and labeled. Here, the two systems used in this paper are the web services of word segmentation provided by CKIP and ICTCAS, respectively.

Table 4. Examples of discrepancies between CKIP and HKCU.

Words		Original Sentences	
CKIP	CUHK	CKIP	CUHK
奶粉 錢	奶粉錢	**奶粉錢**也有點需要	爲了賺**奶粉錢**和教育基金
別 無 選擇	別無選擇	那自然**別無選擇**	除此**別無選擇**
混 日子	混日子	懶懶散散的**混日子**	以做肉串**混日子**
身 陷	身陷	則可能**身陷**其中無法自拔	**身陷**逃兵醜聞的韓星宋承憲
紐約 市長	紐約市長	**紐約市長**魯迪	朱利安尼當上**紐約市長**後

The UKW_Test corpus is first segmented with the two systems, respectively, and the vocabulary set of the segmented words for each system is generated. Each of the 81,447 word candidates obtained previously was checked to see if it was included in the vocabulary set. For CKIP and ICTCAS systems there were 11,290 and 10,642 candidates included in the two vocabulary sets, respectively, as depicted in Figure 6(a). This means 11,290 candidates were accepted as words by the CKIP system while 10,642 candidates were accepted by the ICTCAS system. The intersection of the two sets, containing 9,802 word candidates, were confirmed as words and used to label the basic ground truths since each of them was agreed upon by both systems. After the ground truths were labeled, the 10,000 words previously extracted with our approach can be compared with the 9,802 words in the intersection, *i.e*. the set of referenced answers. When the two sets were compared, it could be found 6,577 words were successfully identified in our approach, which corresponds to 67.09% recall rate provided the 9,802 words are used as target. The other 3,423 extracted words were accepted by our approach but not contained in the referenced answers. After manual inspection, 1,179 out of the 3,423 extracted words were labeled as acceptable answers while the others (2,224) were regarded as non-words, as shown in Figure 6(b). As a consequence, totally 7,756 (6,577 + 1,179) out of the 10,000 extracted words are either equal to the referenced answers or regarded acceptable, which corresponds to 77.56% precision rate. Such results are compatible with the F-measure for the CUHK_Test corpus obtained in Section 4.

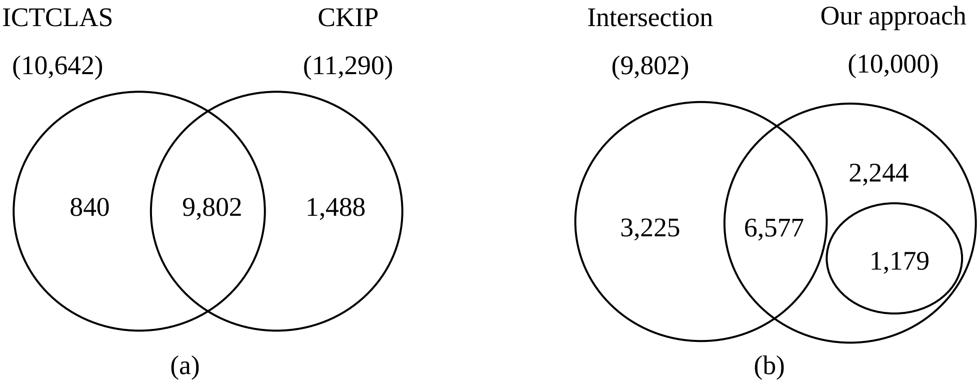


Figure 6. Ground truth labeling and test result.

5.2 Analysis of Unknown Word Extraction

Since the words extracted by our approach are different from those extracted from the two word-segmentation systems, the unknown words extracted by all of these approaches are compared in this section. Note that the unknown words are defined as those words unseen in the training corpus CKIP_Train. Table 5 shows the number of unknown words for every approach after removing those already appearing in the training corpus from the set of words accepted by that approach. Out of the 7,756 labeled words depicted in the previous section, 1,486 unknown words were obtained by our approach, while 2,404 and 1,477 unknown words were obtained by the CKIP system and the ICTCAS system, respectively. Note that the "words" (or extracted terms) for the CKIP/ICTCAS systems are based their respective results of word segmentation instead of consentient ground truths, since, for new domains with novice terms, it is not easy to reach consensus among all systems. This does not matter, however, since our concern here is how many extra terms outside the training corpus each system can extract and how they differ. In addition, for our approach, the number of extracted words can be controlled by adjusting the verification threshold, and set as 10,000 here because this is compatible with the number of words obtained by the CKIP/ICTCAS systems and manageable for laborious manual inspection. If more unknown words are desired, this can also be accomplished by simply lowering the threshold so as to accept more words in word verification.

Table 5. Numbers of unknown words for all approaches.

Approach of unknown word extraction	No. of extracted words (accepted)	No. of unknown words
This paper	10,000	1,486
CKIP System	11,290	2,402
ICTCAS System	10,642	1,477

Figure 7 further displays how the unknown words obtained by our approach differ from those obtained by the two systems. As can be seen in this figure, there were 522 common unknown words, while 897, 316, and 530 mutually exclusive unknown words can be obtained by the CKIP system, the ICTCAS system, and our approach, respectively. This implies that there are quite a few words for which consensus among the approaches cannot be reached. Some examples in the mutually exclusive results for these approaches are shown in Table 6 for illustration. Table 6(a), for example, lists some words that were extracted by our approach but not by the other two. As can be seen in Table 6(a), some hot or novice words, such as "海角七號" (Cape No. 7, the name of a film), "蠟筆小新" (Crayon Shinchan, the name of a figure in a cartoon), "金融海嘯"(Financial Tsunami), "批踢踢"(PTT, the name of a web site), can be successfully extracted only by our approach. These words are popular pronouns whose patterns are very dynamic and do not have apparent semantic structure. Thus, it is difficult to extract these words using semantic rules only. Nevertheless, since they have prominent stochastic characteristics, they can be extracted more reliably by the machine learning approach with HEQ using multiple statistical features including the novel features proposed in this paper. The capability of identifying such words is quite crucial for exploring novice domains.

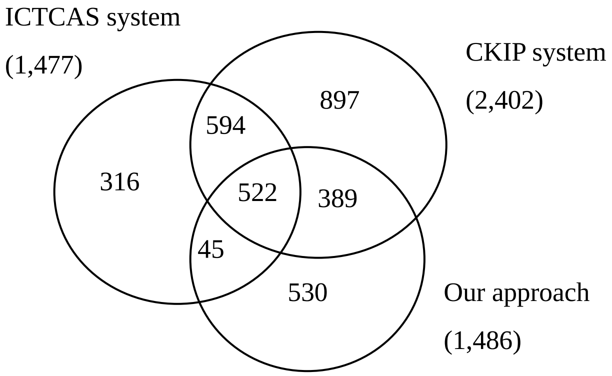


Figure 7. Comparison for differences of unknown words.

Although the proposed approach is distinguished in extracting some hot or novice words, it is more vulnerable than the other two in extracting those terms whose patterns are more static with specific structure, such as "蘇縣長" (Su, the head of the county), "經發局" (the bureau of economic development) or "光林村" (Kuang-lin village), as can be seen in Table 6. Many of these belong to the types of humans, places, organizations, numbers, and so on, which can be more readily extracted with semantic rules. Therefore, it is promising to integrate this approach with other rule-based ones such that they can complement each other.

Table 6. Mutually exclusive unknown words for three approaches.

(a) Our approach

海角7號	功夫灌籃
小巨蛋	批踢踢
佳暮英雄	綠豆椪
蠟筆小新	焦糖哥哥
紙教堂	龍眼乾
語音信箱	金融海嘯
那瑪夏鄉	劍湖山

(b)ICTCAS system

陳添勝	新發大橋
林政助	二手衣
夢工場	簡志忠
南迴公路	消費券
光林村	梅山鄉
總執行長	泰武村
義賣品	馬總統

(c) CKIP system

救難隊	凱達格蘭
平安米	秀姑巒溪
馬政府	監察院長
蘇縣長	正大光明
秋節禮品	毀於一旦
頂呱呱	副駕駛
張瑞賢	經發局

6. Conclusion

This paper proposes a more reliable word extraction scheme by combining multiple statistical features based on machine learning approaches. Since the formation and the structure for Chinese words are sophisticated, it is generally not robust enough to extract words simply according to single feature. This paper combined four features of the word candidates with diverse statistical characteristics to achieve the optimal performance, including the *DLG* that conveys the information for entropy gain with respect to the corpus, the *AV* for the usage context, the *Link* for the evidences of the internal structure, and the *PreC* for the independence of the prefix character. This scheme was initially verified on the CKIP corpus announced in SIGHAN2, and the performance of F-measure at 60.03% was achieved for within-domain test.

This scheme was further applied to the study of statistical mismatch problem between the training the testing domains. The difference of corpus size and the change of domain might lead to difference of dynamic range or distribution for the features, which inevitably degrades the verification performance. Histogram equalization proposed in this paper can compensate for the mismatch of DLG features effectively; thus, it is unnecessary to rebuild the training data for every desired testing domain or to worry about the incompatibility of the feature distributions due to different sizes of corpora. When this scheme of word extraction was evaluated on the within-domain test corpus provided by CKIP and cross-domain test corpus by CUHK, the F-measures can be improved from 60.03% and 63.03% to 68.43% and 71.40%, respectively, by equalization.

Finally, this scheme was used to explore a novice domain for a news event of the flood in Taiwan on Aug. 8th 2009. We proposed a strategy of labeling the ground truths for novice domains according to the consensus between two word segmentation systems. Experimental results show that this scheme can successfully identify some pronouns with prominent statistical tendency but without apparent semantic structure, which cannot be reliably identified with rule-based approaches. This scheme, however, is less robust for extracting those terms whose patterns are static with prominent semantic structure, since it is based on the statistical features instead of the semantic rules. Due to the functional complementation, it is promising to integrate this scheme with other rule-based approaches.

References

Sun, M.S., Huang, C. N., Gao, H.Y., & Fang, J. (1994). Identifying Chinese Name in Unrestricted Texts. *Journal of Chinese Language and Computing*, 4(2), 113-122.

Lu, X. Q., Zhang, L., & Hu, J. F. (2005). Statistical Substring Reduction in Linear Time. *Lecture Notes in Computer Science*, 3248, 320-327.

Zhao, H., & Kit, C. Y. (2008). An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *Proceedings of*

The 3nd International Joint Conference on Natural Language Processing(IJCNLP), 9-16.

Chen, K. J., & Ma, W. Y. (2002). Unknown Word Extraction for Chinese Documents. *Proceedings of The 19nd International Conference on Computational Linguistics (COLING)*, 169-175.

梁婷, 葉大榮 (2000). 應用構詞法則與類神經網路於中文新詞萃取. *Proceedings of Research on Computational Linguistics Conference XIII (ROCLING)*, 21-40.

Goh, C. L., Asahara, M., & Matsumoto, Y. (2003). Chinese Unknown Word Identification Using Character-based Tagging and Chunking. *Proceedings of The 41nd Annual Meeting on Association for Computational Linguistics*, 2, 197-200.

Kit, C. Y., & Wilks, T.(1999). Unsupervised Learning of Word Boundary with Description Length Gain. *Proceedings of Workshop On Computational Natural Language Learning CoNLL*.

Feng, H., Chen, K., Deng, X., & Zheng, W. (2004) Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1), 75-93.

Ji, L., Sum, M., Lu, Q., Li, W., & Chen, Y. (2009). Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, 62-74.

Hummel, R. (1977). Image Enhancement by Histogram Transformation. *Computer Graphics and Image Processing*, 6, 184-195.

Efford, N. (2000). *Digital Image Processing: A Practical Introduction Using Java*. Pearson Education Limited.

De La Torre, Á., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., & Rubio, A. J. (2005). Histogram Equalization of Speech Representation for Robust Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3), 355-366.

Gallardo-Antolín, A., & Montero, J. M. (2010). Histogram Equalization-Based Features for Speech, Music, and Song Discrimination. *IEEE Signal Processing Letters*, 17(7), 659-662.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 61-76
© The Association for Computational Linguistics and Chinese Language Processing

Intent Shift Detection Using Search Query Logs

Chieh-Jen Wang[*], and Hsin-Hsi Chen[*]

Abstract

Detecting intent shift is fundamental for learning users' behaviors and applying their experiences. In this paper, we propose a search-query-log based system to predict users' intent shifts. We begin with selecting sessions in search query logs for training, extracting features from the selected sessions, and clustering sessions of similar intent. The resulting intent clusters are used to predict intent shift in testing data. The experimental results show that the proposed model achieves an accuracy of 0.5099, which is significantly better than the baselines. Moreover, the miss rate and spurious rate of the model are 0.0954 and 0.0867, respectively.

Keywords: Intent Shift Detection, Intent Analysis, Search Query Logs Analysis.

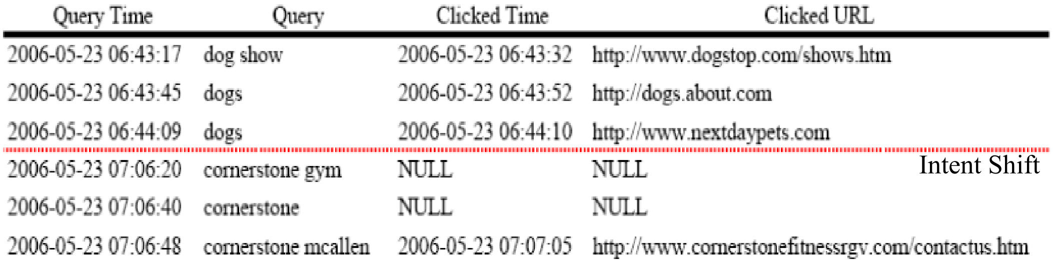
1. Introduction

Understanding behavior in users' search sessions is important because of the multiple potential applications, such as query recommendation, web page re-ranking, and advertisement arrangement. Several approaches have been proposed to define a session in a sequence of actions between a user and a search engine. For example, the time-based approach employs a time threshold to partition the queries in a fixed time period into a session. Determining a suitable threshold is the major problem of this approach. Information will be lost when a large threshold is adopted. In contrast, noise will be introduced when a small threshold is adopted. The query-based approach postulates that an information need is satisfied with a fixed number of queries. This suffers from a problem similar to the time-based approach. A large or a small threshold will introduce too little or too much information.

Clarifying a boundary in a sequence of queries to form an intent-coherent session is a fundamental task in mining users' behaviors on the World Wide Web. One of the possible approaches to accomplish this task is determining intent shifts in a sequence of queries. An intent shift occurs when the intent of the current query is different from the original search intent during information access. Given a sequence of queries, q_1, q_2, ..., and q_n, there is an intent shift between q_i and q_{i+1} if the intent of q_{i+1} is different from that intent of q_1, q_2, ..., q_i.

[*] Department of Computer Science and Information Engineering, National Taiwan University
E-mail: cjwang@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Take Figure 1 as an example, which is selected from a real session in a query log dataset. An intent shift occurs at q_4 (the 4th query in the session) because the search intent from q_1 to q_3 is about dogs and the search intent of q_4 is about a gymnasium, which is significantly different from the original search intent.



Query Time	Query	Clicked Time	Clicked URL
2006-05-23 06:43:17	dog show	2006-05-23 06:43:32	http://www.dogstop.com/shows.htm
2006-05-23 06:43:45	dogs	2006-05-23 06:43:52	http://dogs.about.com
2006-05-23 06:44:09	dogs	2006-05-23 06:44:10	http://www.nextdaypets.com
2006-05-23 07:06:20	cornerstone gym	NULL	NULL
2006-05-23 07:06:40	cornerstone	NULL	NULL
2006-05-23 07:06:48	cornerstone mcallen	2006-05-23 07:07:05	http://www.cornerstonefitnessrgv.com/contactus.htm

Figure 1. An intent shift example.

On the web, users complete their information needs through searching and browsing. They submit queries and click URLs in a session to represent their intents. The interactions between users and search engines are kept in search query logs. As similar search behaviors demonstrate similar intents, such a log dataset is a good resource to investigate common users' behaviors. In this paper, we will mine users' behaviors from search query logs and use them for detecting whether there exist intent shifts or not.

The challenging issue is that there may be more than one search intent in a session. Multiple-intent sessions may have negative effects on the clustering performance and impact later intent shift detection. Detecting intent shifts in a session will result in sessions of better quality and have positive effects on the clustering. This is a chicken and egg problem. In this paper, we propose some strategies to sample sessions in search query logs, disambiguate the ambiguous queries and clicked URLs, extract features from different intent representations, generate intent clusters by different cluster algorithms, and explore different intent shift detection models.

The remainder of this paper is organized as follows. In Section 2, we compare our research with others. Section 3 gives an overview of the proposed system and the major resource used in this work. Section 4 describes the session sampling strategies and an algorithm for query and URL disambiguation. Section 5 describes how to assemble the sessions of similar intent in a cluster. In addition, we apply intent clusters generated by different clustering models to detect intent shifts. Section 6 analyzes the performance of different clustering models and discusses the findings along with their implications. Section 7 concludes the remarks.

2. Related Work

Given a sequence of queries, an intent boundary detection algorithm divides it into several sub-sequences. Each sub-sequence of queries, containing a single information need, forms a session. Jansen, Spink, Blakely, & Koshman (2007) specify that a session is a series of users' interactions toward a single information need. A session is a basic unit for intent clustering because clustering models put together user behaviors of the same intent into a cluster.

Generally speaking, a session is usually identified by hard or soft segmentation. In hard segmentation, a session is segmented by users' actions, such as open/close a browser or login/logout of a search engine, or by some heuristic methods, such as time cutoffs (Silverstein, Henzinger, Marais, & Moricz, 1998; Montgomery & Faloutsos, 2001) or mean session lengths (Silverstein *et al*., 1998; Jansen, Spink, Blakely, & Koshman, 2007). On the contrary, soft segmentation identifies an intent boundary according to topic shift in query streams (He & Harper, 2002), some category of user intent (Ozmutlu & Cavdur, 2005) or dynamic comprehension time (Wang, Lin, & Chen, 2010). Several algorithms have been proposed for detecting the intent shifts or identifying intent boundaries (Cao *et al*., 2008).

The task of query classification is to classify the queries into some predefined categories. Queries, however, are usually short and ambiguous. To realize the meanings of queries, researchers have introduced the concept of user intent behind queries (Broder, 2002). Queries were classified by the searcher's intent, such as navigational query, whose immediate intent is to reach a particular web site, informational query, whose intent is to acquire some information, and transactional query, whose intent is to perform some web-mediated activity. Queries are characterized along four general facets, *i.e*., ambiguity, authority, temporal sensitivity, and spatial sensitivity (Nguyen & Kan, 2007). Manshadi and Li (2009) classify queries into finer categories. Shen *et al*. (2005) employ the Open Directory Project (ODP) taxonomy to represent clicked URLs and investigate the topic transition. Hence, queries have to be disambiguated if we want to know their exact meanings.

3. System Overview

An intent shift detection system is outlined in Figure 2. The MSN Search Query Log excerpt (RFP 2006 dataset) (Craswell, Jones, Dupret, & Viegas, 2009) is the main resource of this work. This data set consists of 14.9 million queries and 12.2 million clicks during a one-month period in May 2006. The MSN Search Query Log excerpt is separated into two files, one named *query* and the other named *click*. *Query* file is described by a set of attributes, including Time, Query, QueryID, and ResultCount, and the *click* file contains attributes like QueryID, Query, Time, URL, and URL Position. Note that these two files are linked through QueryID. In total, there are 7.4 million sessions, which contain the activities of users from the time of the first query submission to the time of a timeout between their web browser and the search engine.

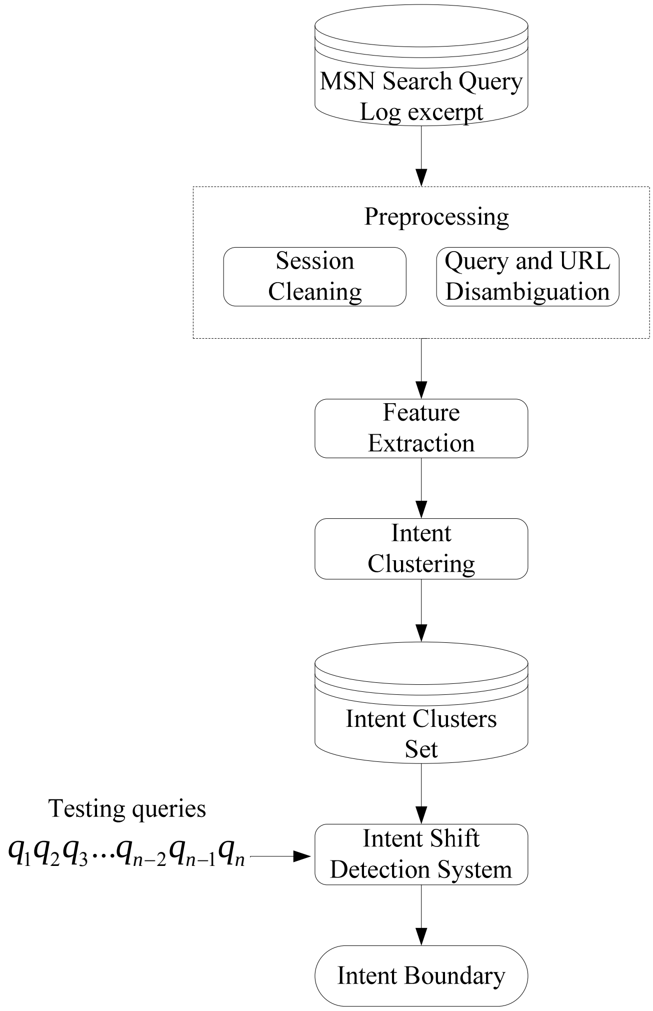

MSN Search Query Log excerpt
Preprocessing
Session Cleaning
Query and URL Disambiguation
Feature Extraction
Intent Clustering
Intent Clusters Set
Testing queries
$q_1q_2q_3 \dots q_{n-2}q_{n-1}q_n$
Intent Shift Detection System
Intent Boundary

Figure 2. A system overview of intent shift detection.

Not all sessions are suitable for constructing the intent shift detection system because of the different backgrounds of users. Besides, a session may contain noise or insufficient information. For these reasons, the search query logs are purified. We take several strategies to get a reliable data set for this study. Queries and URLs are disambiguated based on the ODP (The Open Directory Project, 2002), which is the largest, the most widely distributed human-compiled, and the most comprehensive taxonomy of the websites.

The ODP contains more than 4.5 million websites organized into more than 600 thousand paths. A path (Perugini, 2008) is defined as an ordered hierarchical structure of hyperlink labels from the root category of a directory to a leaf in the ODP. For example, the Academia Sinica website (http://www.sinica.edu.tw/) is assigned a path (Top/Regional/Asia/Taiwan/Education). The root category is "Top" and "Regional" is a sub-category of "Top". The ODP not only contains website annotations edited by volunteers collaboratively, but also provides a natural language description of categories and websites.

After preprocessing, the purified sessions are partitioned into intent clusters by two hierarchical cluster algorithms with queries, clicked URLs, and their corresponding ODP categories. We use the sets of the intent clusters generated by various clustering models to construct the intent shift detection system. To evaluate the performance of the intent shift detection methods on various intent representations, we manually prepare a ground truth. A total of 500 sessions are sampled and annotated for testing. Given a sequence of queries in a testing session, the intent shift detection system will identify whether intent shifts occur.

4. Preprocessing

For creating a reliable dataset, we first clean the MSN Search Query Log excerpt by session cleaning and category disambiguation before intent clustering. Session cleaning filters out potential noise in the MSN Search Query Log excerpt in Section 4.1. The most preferable ODP categories for a clicked URL are selected by a category disambiguation in Section 4.2.

4.1 Session Cleaning

In search query logs, longer sessions containing more queries and clicked URLs tend to contain noise because users have higher probability of changing intents. On the other hand, smaller sessions may not be complete enough to describe the whole information need. In this work, we aim to capture the user intent embedded in sessions as much as we can. Balancing the noiselessness and the completeness is a basic issue for session cleaning.

At the session cleaning stage, we employ the following four filtering strategies.

Filter Strategy #1: Sessions longer than one hour are removed, because a long session may have multiple intents in it. According to the statistics of the MSN Search Query Log excerpt, the longest duration time of a session is more than 99 hours and several intents are observed. Intuitively, it is unlikely for a user to interact constantly with a search engine and maintain only a single intent. As mentioned in related work, several time cutoffs are proposed to segment a session. In specific, Anick (2003) adopts 60 minutes as a time cutoff to segment a session. Therefore, we postulate that the original intent of a user will shift if s/he queries a search engine for more than 60 minutes.

Filter Strategy #2: Users may accomplish their goals with few queries in a small session. For example, a user may submit a navigation query (*e.g*., Academia Sinica), click the official website, and stop the search. A small session does not provide enough information to know his or her exact intent. In this example, the intent may be finding a job or searching for a research institute in Academia Sinica. We consider total number of queries in a session as a filtering strategy. Sessions with less than *n* queries are regarded as small sessions and are removed. As mentioned in related work, a variety of length cutoffs have been proposed by different research projects. Additionally, Silverstein, Henzinger, Marais, & Moricz, (1998) reported the most probable range is between 2 to 3 queries. In our experiments, *n* is set to 3.

Filter Strategy #3: Users may click other search engines during searching and browsing. Nevertheless, how users interact with the search engines is not recorded, so that we have no information about the subsequent actions. As a result, we remove sessions which contain the clicked URLs to other search engines.

Filter Strategy #4: Since our system utilizes ODP categories of URLs for queries and URLs disambiguation, we keep only the sessions where all clicked URLs appear in the ODP.

Table 1 lists the number of remaining sessions after each combination of strategies. After all of the filtering processes, a total of 14,242 sessions remain. The number of the remaining sessions is a low percentage (*e.g*. 0.19%) of the original sessions, but that is not a problem since a huge collection of query logs is available in the real world. More sessions will be generated if more logs are available. Pure sessions are important to generate *correct* results. The basic idea of filtering strategies is avoiding the *garbage in*, *garbage out* problem.

Table 1. Results of session filtering.

Filtering Strategies	Number of Sessions	Remaining Rate (%)
Original	7,468,628	100.00
1	2,815,843	37.70
1 + 2	1,063,906	14.24
1 + 2 + 3	419,383	5.62
1 + 2 + 3 + 4	14,242	0.19

4.2 Category Disambiguation

We postulate that the sequence of clicked URLs in a session is relevant and accomplishes an information need. Therefore, the clicked URLs are co-related and coherent. The clicked URLs and the surrounding URLs are employed to disambiguate the categories of the clicked URL and the associated queries.

For a URL, we consult the ODP to collect all possible paths. As a clicked URL may belong to more than one path, we must find its correct meaning. For example, Brookfield Zoo's website can be mapped to two paths. One specifies that Brookfield Zoo is a travel and tourism location (Top/Regional/North_America/United_States/Illinois/Localities/B/Brookfield/Travel_and_Tourism) and the other one regards it as a biology science location (Top/Science/Biology/Zoology/Zoos_and_Aquariums/North_America/United_States/Illinois). Without disambiguation, it is unclear whether a user clicking on Brookfield Zoo's URL plans a trip to the zoo or is interested in biological science.

The goal is to find the most semantically-coherent path of each URL in a session. Take Figure 3 as an example. A total of 3 URLs, i.e., URL1, URL2, and URL3, have been clicked in a session. URL1, URL2, and URL3 contain 1, 2, and 2 probable paths after consulting the ODP, respectively. There are four possible trails that make up the set of probable paths from each clicked URL. For example, "path11-path21-path31" is one possible trail. The score of a trail is the sum of similarity between a path and the other paths in the session. The similarity of two paths is the number of common categories between these two paths. The number of common categories among paths reflects the degree of intent coherency. Therefore, the trail with the highest score will be selected, and the trail contains the disambiguated ODP categories for the clicked URLs.

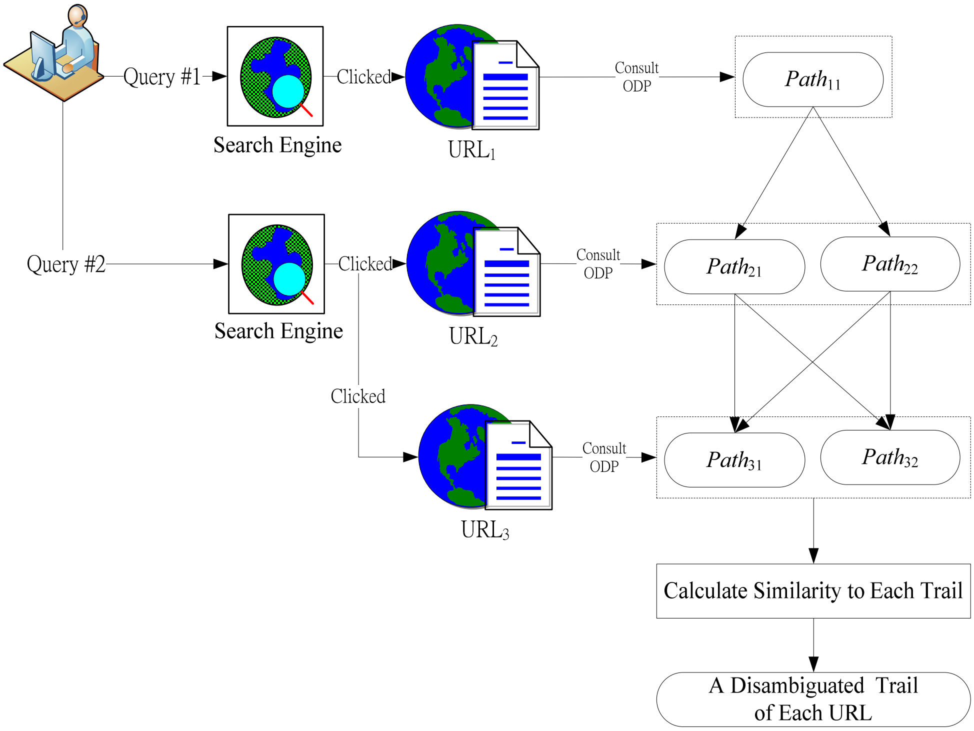


Figure 3. Category disambiguation procedure.

5. Intent Clustering

Sessions are similar if the search intents in them are the same. Sessions of the same intent will be put into a cluster to represent the common behaviors related to the intent. Three sets of features are extracted from each session, and the feature weight is determined by binary and *tf-idf* schemes. Two hierarchical cluster algorithms are used to cluster sessions of similar intent. A clustering model is defined by features and a hierarchical cluster algorithm. A cluster set created by the model is used to detect intent shifts. The performance of the intent shift detection system will be evaluated by three metrics, including *miss rate*, *accuracy*, and *spurious rate*.

5.1 Feature Extraction

Three sets of features, *i.e*., Query, clicked URLs, and path of clicked URLs with category disambiguation, are considered. The features are used to cluster sessions. Table 2 shows the details of each feature and feature combination. Query terms are the first type of features. In a query feature, query terms are transferred to lower case and stop words are removed, but not stemmed. In addition, a bag-of-words strategy is employed so that two queries consisting of identical terms in any orders are regarded as the same. A complete URL is the second type of features. A disambiguated path is the third types of features. That is, only the best ODP categories are selected as features in this type. Different combinations of the above 3 types of features are shown from the 4th -6th rows.

Table 2. Description of each feature set.

Feature Name	Description
Query	query terms as features
URL	URLs as features
Path	disambiguated ODP categories of URLs as features
Query+URL	query terms and URLs as features
Query+Path	query terms and disambiguated ODP categories as features
Query+URL+Path	query terms, URLs and disambiguated ODP categories as features

The weight of a feature is determined by two possible schemes: binary or tf-idf (Salton & Buckley, 1988). In the binary setting, the weight of a feature is set to 1 if the feature appears in the session, 0 otherwise. In the tf-idf setting, the weight of a feature is defined by Equation (1):

$$w_{i,s} = (0.5 + \frac{0.5\, freq_{i,s}}{\max_s\ freq}) \times log\frac{N}{n_i} \quad (1)$$

where $\max_s \mathit{freq}$ is the maximum feature frequency in session *s*, $\mathit{freq}_{i,s}$ is the frequency of feature *i* in session *s*, *N* is the total number of sessions, and n_i is the number of sessions where feature *i* appears.

5.2 Clustering Models

After preprocessing, there were 14,242 sessions for intent clustering. We randomly sampled 500 sessions used for the later evaluation. The remaining 13,742 sessions were clustered by the average link and the complete link hierarchical clustering algorithms with different sets of features. The similarity between two sessions was determined by Euclidean distance. In the experimental setup, there are a total of 24 clustering models since the combinations of 2 clustering hierarchical algorithms, 2 feature weight schemes, and 6 features form the clustering models.

The two hierarchical clustering algorithms need a similarity threshold to separate an agglomerative hierarchical cluster tree into clusters. The similarity between two sessions in the agglomerative hierarchical cluster tree is represented by Euclidean distance. Different intents may be put into a cluster when a loose similarity threshold is selected. On the other hand, a cluster may be divided into more than one smaller cluster when a tight similarity threshold is adopted. The tighter similarity threshold is selected in the experiment due to the fact that the resemblance of sessions is improved in an intent cluster. For all clustering models, setting the threshold to 1 produces more than one cluster and setting the threshold to 2 produces a cluster. We use Algorithm 1 to select the largest threshold between 1 and 2. A hierarchical cluster tree is separated into *n* clusters by a threshold *s* (Line 4). The search procedure is stopped if increasing a value of similarity threshold is accurate to 5 decimal points. Each clustering model constructs an intent cluster set by a similarity threshold. The

Algorithm 1. Searching a similarity threshold

```
Input: An agglomerative hierarchical cluster tree $t$
Output: A similarity threshold $s$
 1: $s \leftarrow 1$ and $d \leftarrow 0$
 2: while ($d<5$)
 3:     $d \leftarrow d + 1$
 4:     $n \leftarrow$ Cluster($t$, $s$)
 5:     $i \leftarrow 1$
 6:         While ($i \leq 9$)
 7:             $u = s + i \times 10^{-d}$
 8:             if ($n \neq$ Cluster($t$, $u$)) then $s \leftarrow u$
 9:             end if
10:             $i \leftarrow i + 1$
11:         end While
12: end while
13: return $s$ as the similarity threshold
```

number of clusters by the 24 clustering models ranges from 3,960 to 4,756, and the details are shown in Table 3.

Table 3. The number of intent clusters created by different clustering models.

Feature	Binary		*tf-idf*	
	Average Link	Complete Link	Average Link	Complete Link
Query	4612	4518	4756	4688
URL	4555	4455	4655	4566
Path	4481	4369	4578	4489
Query+URL	4352	4288	4429	4399
Query+Path	4222	4186	4295	4258
Query+URL+Path	4163	3960	4200	4145

5.3 Intent Shift Detection

Note that the goal of this work is to predict the intent shift. Given a query sequence, $q_1q_2...q_n$, an intent shift at q_j is defined as an intent switch between q_j and q_{j+1}. Algorithm 2 is an intent cluster based intent shift detection algorithm. Given a testing query sequence, q_1, q_2, …, q_n, an approximate strategy considers a span of *d* queries to form a potential segment. The *d* value in Algorithm 2 is set to 5, which is the average length of the 13,742 sessions. In this way, a rough segment, q_1, q_2, …, q_5, is derived.

Then, the following procedure finds the most similar intent cluster *c** to identify the intent boundary. The segment, q_1, q_2, …, q_5, is represented by queries, clicked URLs, and disambiguated ODP categories. Category descriptions, URL snippets in ODP, and anchor texts of the clicked URLs are also used if they exist. We measure the similarity of the segment with all intent clusters, and assign the intent cluster with the highest similarity to *c**. LuceneSim (Line 1) is adopted as the similarity function, which is derived from the vector space model[1]. A document whose vector is closer to the query vector gets a higher similarity score. The similarity function is shown as follows:

$$S(Q,D)=\frac{|Q\cap D|}{|Q|}\times\frac{1}{\sum_{t\in Q}(1+log\frac{N}{df(t)+1})^2}\times\sum_{t\in Q\cap D}\frac{\sqrt{c(t,D)}\times(1+log\frac{N}{df(t)+1})^2}{\sqrt{|D|}}$$

where $|Q\cap D|$ is the number of terms that occur in both query *Q* and document *D*, $|Q|$ is the length of query *Q*, $df(t)$ is the number of documents that contain term *t*, $|D|$ is the length of document *D*, *N* is the number of documents in the collection, and $c(t,D)$ is the number of

[1] http://lucene.apache.org/java/docs/.

occurrences of term t in document D.

We move the boundary from the first query q_1 and compute the similarity between the query vector represented by q_1 and the c^*. Then, we move the boundary to the right to include one more query q_2 and compute the similarity between the query vector represented by the enlarged segment (q_1 and q_2) and c^*. If the similarity becomes lower, q_2 does not belong to the same intent. The right movement procedure is repeated until the similarity becomes lower or the final query is considered. The main idea of the algorithm is: the similarity scores are calculated by a selected intent cluster and a query vector that is represented by queries, clicked URLs, and disambiguated ODP categories. If the intent of the enlarged query is different from the original query vector, the newly-formed query vector contains information different from the selected intent cluster, so the similarity becomes lower.

Algorithm 2. Detecting intent shift

```
Input: A query sequence $q_1q_2...q_n$, an intent cluster set H
Output: an intent shift at $q_{i-1}$
c* ← argmax$_{c \in H}$ LuceneSim($c$, $q_1q_2...q_d$)
$g$ ← 0 and $i$ ← 1
while ($i \leq n$)
    if (LuceneSim($c^*$, $q_1q_2...q_i$) ≥ $g$)
        $g$ ← LuceneSim($c^*$, $q_1q_2...q_i$)
    else
        break while
    end if
    $i$ ← $i$ + 1
end while
```

5.4 Evaluation

For evaluating the performance of intent shift detection, a test dataset is constructed as follows. A group of annotators assess the 500 testing sessions according to the intent purity and type. Of the 500 sessions, 355 sessions are annotated single intent, which means all of the queries and clicked URLs belong to the same intent. The intents of the other 145 sessions are tagged ambiguously. This means the intent type cannot be recognized in these sessions. We append each pair of single intent sessions, forming $355^2 = 126{,}025$ new testing query sequences. If two sessions in a pair contain different intents, the correct intent shift location is at the last query of the first session. If two sessions in a pair contain the same intent, the correct intent shift location is at the last query of the second session. Consider an example. Given two sessions, the first session contains m queries ($q_{1,1}$, $q_{1,2}$, …, $q_{1,m}$) and the second session contains n queries ($q_{2,1}$, $q_{2,2}$, …, $q_{2,n}$). Therefore, the test query sequence includes m+n queries

($q_{1,1}$, $q_{1,2}$, …, $q_{1,m}$, $q_{2,1}$, $q_{2,2}$, …, $q_{2,n}$). If the two sessions contain the same intent, then a correct intent shift location exists at $q_{2,n}$, otherwise at $q_{1,m}$.

As previously mentioned, a total of 24 intent cluster sets were created by different clustering models. Each of the 24 cluster sets was input into Algorithm 2 to predict intent shifts in the 126,025 testing query sequences.

We evaluate a testing query sequence according to the following evaluation metrics. Equations (2) and (3) define *miss distance* and *spurious distance* between the predicted intent shift location q_{sp} of the system and the intent shift location q_{gt} of ground truth, where *sp* and *gt* are integers and denote the location of a testing query sequence. If *sp* < *gt*, *i.e.*, the system-predicted intent shift is shorter than ground truth, we calculate a miss distance by Equation (2). In contrast, *sp* > *gt* means the system-predicted intent shift is larger than ground truth. In this case, we calculate a spurious distance by Equation (3). When *sp* = *gt*, *i.e.*, the system prediction and ground truth is exactly matched, we compute the accuracy.

$$miss\ distance = \frac{gt - sp}{gt} \tag{2}$$

$$spurious\ distance = \frac{sp - gy}{gt} \tag{3}$$

Table 4 shows the performance of three baselines using query cutoffs or topic shift. 3QueryCutoff, proposed by Silverstein, Henzinger, Marais, & Moricz (1998), considers a segment consisting of 3 queries. Avg#Queries regards 5 queries as a segment, which shows the average number of queries in sessions of the MSN Search Query Log excerpt. TopicShift is an algorithm proposed by He, Göker, & Harper (2002) to detect intent shifts. They proposed eight search patterns to formulate the topic transformations and calculate a probability of topic transformation to detect intent shifts.

The *Miss Rate*, *Accuracy,* and *Spurious Rate* are computed as Equations (4), (5), and (6), respectively. Decreasing *Miss Rate* and *Spurious Rate* and increasing *Accuracy* are the goal of our prediction system. TopicShift achieves the best performance on *Accuracy* among the three baselines. The *Miss Rate*, *Accuracy,* and *Spurious Rate* of TopicShift are 0.2676, 0.3087, and 0.0956, respectively.

$$miss\ rate = \frac{\sum_{i=1}^{n} miss\ distance}{n} \tag{4}$$

$$spurious\ rate = \frac{\sum_{i=1}^{n} spurious\ rate}{n} \tag{5}$$

$$accuracy = \frac{\#of\ \ testing\ data\ sp = gt}{n} \tag{6}$$

Table 4. Performance of three baselines.

	3QueryCutoff	Avg#Queries	TopicShift
Miss Rate	0.2722	0.2329	0.2676
Accuracy	0.2999	0.1920	0.3087
Spurious Rate	0.0148	0.0190	0.0956

Table 5 and Table 6 show Miss Rate, Accuracy, and Spurious Rate of introducing intent clusters created by the average link clustering algorithms with binary and tf-idf feature weight schemes, respectively. Intent clusters generated by different clustering models are compared. Feature weight with binary is better than tf-idf. Using URL features is better than using Query features. This may be due to the fact that clicked URLs express clearer user intents and that users' queries may be ambiguous. Using Path feature performs better than using URL features. This meets our expectation because an ODP category is a conceptual representation of a URL. Using a Query together with URL/Path is better than using the URL/Path only. Using a Query together with URL+Path is better than using a Query together with URL/Path.

We also introduce intent clusters created by the complete link clustering algorithms with binary and tf-idf feature weight schemes, respectively. Table 7 and Table 8 show the results. The tendency is similar to Table 5 and Table 6. The intent clusters created by the clustering model integrating the features of Query, URL and Path in the complete link clustering algorithm perform the best on intent shift detection. The clustering model achieving accuracy 0.5099 is significantly better than the baselines in Table 4 (p-value<0.001). The results show that considering intent clusters are quite useful for intent shift detection.

Table 5. Introducing the intent clusters created by the average link clustering algorithm with binary weight scheme

Binary	Query	URL	Path	Query + URL	Query + Path	Query + URL + Path
Miss Rate	0.1711	0.1468	0.1493	0.1346	0.1362	0.1118
Accuracy	0.4122	0.4455	0.4602	0.4651	0.4752	0.4866
Spurious Rate	0.0934	0.0686	0.0720	0.0730	0.0751	0.0791

Table 6. Introducing the intent clusters created by the average link clustering algorithm with tf-idf weight scheme

tf-idf	Query	URL	Path	Query + URL	Query + Path	Query + URL + Path
Miss Rate	0.1539	0.1406	0.1317	0.1473	0.1279	0.1206
Accuracy	0.3834	0.4316	0.4386	0.4419	0.4579	0.4655
Spurious Rate	0.1258	0.0846	0.0879	0.0790	0.0990	0.0891

Table 7. Introducing the intent clusters created by the complete link clustering algorithm with binary weight scheme

Binary	Query	URL	Path	Query + URL	Query + Path	Query + URL + Path
Miss Rate	0.1889	0.1548	0.1469	0.1361	0.1306	0.0954
Accuracy	0.4142	0.4548	0.4629	0.4688	0.4795	0.5099
Spurious Rate	0.0984	0.0703	0.0742	0.0815	0.0763	0.0867

Table 8. Introducing the intent clusters created by the complete link clustering algorithm with tf-idf weight scheme

tf-idf	Query	URL	Path	Query + URL	Query + Path	Query + URL + Path
Miss Rate	0.1529	0.1536	0.1469	0.1431	0.1440	0.1359
Accuracy	0.4042	0.4395	0.4490	0.4559	0.4601	0.4690
Spurious Rate	0.1243	0.0662	0.0750	0.0728	0.0730	0.0699

6. Discussion

We measured the performance of the intent cluster sets produced by the 24 intent clustering models. The performance of a cluster set was determined by judging whether each cluster in the set has a coherent intent. Our evaluation strategy was to devise a cluster based intent shift detection system that utilizes an intent cluster set to perform intent shift detection. The model's accuracy depends on the intent coherency of the clusters. Hence, the cluster set that achieves the highest performance of intent shift detection has the most intent-coherent clusters.

Tables 5-7 show a common phenomenon: Miss Rate is higher than Spurious Rate. This shows that our system is conservative for enlarging the intent boundary. In Algorithm 2, the intent shift query is identified if the current similarity is lower than the previous one. Therefore, we may introduce a relaxation strategy that is determined by the similarity score multiplying a weight. This means the drop must be sharp enough to have at least a fixed percentage of the previous similarity. Along the way, the strategy will decrease Miss Rate but Spurious Rate may be increased. It trades off Miss Rate and Spurious Rate.

7. Conclusion and Future Work

This paper predicts intent shifts in the MSN Search Query Log excerpt. The intent clusters generated by using Query, URL, and Path are proven to be useful for this work. We evaluated the intent clusters created by different intent clustering models from the aspects of *Miss Rate*, *Accuracy*, and *Spurious Rate*. The experimental results show that the complete link cluster algorithm is better than the average link cluster algorithm in almost all evaluation metrics.

We will explore the uses of the intention clusters learned from search query logs in one language (e.g., MSN Search Query Log excerpt) to identify the intent shifts of query logs in another language (e.g., SogouQ Query Log), and we will compare the ways of expressing intent in different languages, different areas, and different cultures.

References

Anick, P. (2003). Using terminological feedback for web search refinement: a log-based study. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval,* 88-95. Toronto, Canada: ACM.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2), 3-10.

Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., & Li, H. (2008). Context-aware query suggestion by mining click-through and session data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 875-883. Las Vegas, Nevada, USA.

Craswell, N., Jones, R., Dupret, G., & Viegas, E. (2009). Proceedings of the 2009 workshop on Web Search Click Data, 95. Barcelona, Spain.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic web session identification. Inf. *Process. Manage.*, 38(5), 727-742.

He, D., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. INFORMATION *PROCESSING AND MANAGEMENT*, 38, 727-742.

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines: Research Articles. *Journal of the American Society for Information Science and Technology*, 58(6), 862-871.

Manshadi, M., & Li, X. (2009). Semantic tagging of web search queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2,* 861-869. Suntec, Singapore.

Montgomery, A. L., & Faloutsos, C. (2001). Identifying Web Browsing Trends and Patterns. *Computer*, 34(7), 94-95.

Nguyen, V., & Kan, M. (2007). Functional Faceted Web Query Analysis. In *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007).*

Ozmutlu, H., & Cavdur, F. (2005). Application of automatic topic identification on Excite Web search engine data logs. *Information Processing & Management*, 41(5), 1243-1262.

Perugini, S. (2008). Symbolic links in the Open Directory Project. *Information Processing and Management: an International Journal*, 44(2), 910-930.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5), 513-523.

Shen, X., Dumais, S., & Horvitz, E. (2005). Analysis of topic dynamics in web search. In *Special interest tracks and posters of the 14th international conference on World Wide Web,* 1102-1103. Chiba, Japan.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1998). Analysis of a Very Large AltaVista Query Log. *Technical Note: Digital Equipment Corporation*.

The Open Directory Project. (2002). About the Open Directory Project. Retrieved October 26, 2010, from http://www.dmoz.org/about.html

Wang, C., Lin, K. H., & Chen, H. (2010). Intent boundary detection in search query logs. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval,* 749-750. Geneva, Switzerland.

Computational Linguistics and Chinese Language Processing
Vol. 16, No. 3-4, September/December 2011, pp. 77-106
© The Association for Computational Linguistics and Chinese Language Processing

Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents

Darren Hsin-Hung Lin[*,+] and Shelley Ching-Yu Hsieh[+]

Abstract

This paper presents a corpus-driven linguistic approach to embodiment in modern patent language as a contribution to the growing needs in intellectual property rights. While there is work that appears to fill a niche in English for Specific Purposes (ESP), the present study suggests that a statistical retrieval approach is necessary for compiling a patent technical word list to expand learner vocabulary size. Since a significant percentage of technical vocabulary appears within the range of independent claim among claim lexis, this study examines the essential features to show how it was characterized with respect to the linguistic specificity of patent style. It is further demonstrated how the proposed approach to the term independent claim contained in the patent specification is reliable for patent application on an international level. For example, clausal types that specify how clauses are used in U.S. patent documents under co-occurrence relations are potential for patent writing, while verb-noun collocations allow learners to grip hidden semantic prosodic associations. In short, the research content and statistical investigations of our approach highlight the pedagogical value of Patent English for ESP teachers, applied linguists, and the development of interdisciplinary research.

Keywords: Intellectual Property Rights, Patent Document Processing, Corpus, Systemic Functional Linguistics, Co-Ocurrence.

[*] Graduate School of Decision Science and Technology, Tokyo Institute of Technology, Japan
E-mail: darryanlin@gmail.com
The author for correspondence is Darren Hsin-Hung Lin.

[+]Department of Foreign Languages and Literature, National Cheng Kung University, Taiwan
E-mail: shelley@mail.ncku.edu.tw

1. Introduction

In the knowledge economy age, intellectual property rights (IPR) are important assets. Especially to the knowledge industry, IPR is the key measure of a company competing with others.

As globalization has resulted in rapid greater economic growth, the challenges of interdisciplinary communication concerned with intellectual property and other significant sector encounters have increased. The recognition of this importance has brought intellectual property into the limelight. Resulting from such recognition, as well as the recent emphasis on using English as the lingua franca to apply patents on an international level, the application of technical vocabulary for the writing of professional patents has become an essential issue in applied linguistic research.

2. Literature Review

Law is a system of rules, carried out by lawyers, attorneys, jury, paralegals, and related legal institutes. It is not just a tool that shapes politics, economy, and society, but also it is a socially prominent medium applied to maintain social order. A large number of recently specialized areas, such as international trade, economics, finance, accounting, and electronic commerce, recently have been recruiting interdisciplinary specialists with expertise in both law and English proficiency to engage in legal workplace practice.

While the widespread use of law has naturally had impact on almost all fields of discipline, the increasing use of English has radically changed the way in which we perceive this language’s international function (Modiano, 2001). English for Legal Purposes (ELP), a growing trend in the field of English for Specific Purposes (ESP), therefore, has become a research topic (Dudley-Evans & St. John, 1998:7) and is used in either professional or legislative settings.

As the Internet shortens the distance between countries, patent information is readily available via online access. To protect novel inventions, intellectual property law is a developing domain across legal professions. The area of intellectual property law includes patent law, copyright law, trademark law, and trade secret law, together with some aspects of other branches of the law, such as licensing and unfair competition (American Bar Association, 2010). Intellectual property lawyers are required to have command of interdisciplinary knowledge as new developments in law generate the need for lawyers with specific backgrounds-patent law, technology law, business law, and economy economic law. It is worth mentioning that the demand for intellectual property lawyers has remained unusually high even though the global markets were affected by economic recession in the end of 2007 (World Intellectual Property Organization, 2009). As long as novel inventions continue to be created, there is a need for intellectual property law to be enforced to protect human rights and their invisible property for specific purposes.

Patent, known as interdisciplinary innovation, has drawn the attention of most lawyers. Tsai (2008) reported that patents are granted for innovations as they reflect economic growth of a country by illustrating creative activities and displaying the knowledge power of that particular country or region. The diversity of languages used in patent applications has boosted translation demand for patent right protection. Besides, many paralegals, such as patent attorneys, lawyer assistants, or translators, participate in legal circles for a living nowadays. It is important to equip them not only with background knowledge, but competency of professional writing for the job market. Accordingly, in the present study, the researchers look at the role of patent writing for research purposes.

2.1 Corpus-based Studies on Law

Corpus linguistics is often concerned with the study of natural language, which explores real and authentic language use by means of a corpus (McEnery & Wilson, 2001). At the present day, a corpus represents a wide variety of language use, both spoken and written language, by a collection of texts stored in a computer (Mudraya, 2006).

Biber, Conrad, & Reppen (1998) claimed a corpus-based analysis is characterized by four primary features. First, a corpus-based study is empirical, for it uncovers the natural patterns of real language use. A corpus-based study, however, relies heavily on computer-assisted tools. Computer-assisted tools, such as concordancers, enable researchers and practitioners to tag linguistic features, to code grammatical variants, and to carry out data capture and mark-up. Third, research data are analyzed either quantitatively or qualitatively in a corpus-based study. For example, the total frequency of the term independent claim is shown in a quantitative way. The concordancer can show the frequency of coded articles and average words per article. Analysis probing into observing linguistic phenomena of the term, such as polysemy or near-synonym, in turn, is qualitative. Finally, a corpus-based analysis is meaningful once research questions have been proposed. A corpus may be designed to characterize the use of an independent claim adopting a functional approach. Since the investigation is prompted to answer the research questions concerned with such design, the corpus-based analysis becomes meaningful. As corpus-based study is widely accepted and has become the norm in interdisciplinary social sciences (Ball, 1996; Chen, 2001; Lee & Swales, 2006), it further represents how language has been evaluated in prescriptive and descriptive ways in academic research (Dudley-Evans & St. John, 1998; Hyland & Tse, 2005; Nelson, 2006; Hyland, 2008).

In sum, a corpus-based study is insightful in that it is not only representative in social science research, but also it contributes to characterizing the legal language people associate with (Hsieh, 1998). Over the years, there has been corpus-based research on law in social science research (Feak, Reinhart, & Sinsheimer, 2000; Candlin, Bhatia, & Jensen, 2002; Badger, 2003; Chiu, 2008). Nevertheless, few works concerned with patents can be found. As corpus-based studies have been conducted widely in social science research, the application of corpus tools has been noticed in recent development. The present study is warranted by such trends for investigation into contemporary patents.

2.2 ESP Studies on Law

ESP is now well established as an important and distinct part of English teaching (Cheng, Sin, & Li, 2008:16). As English has acquired the status of lingua franca in almost any field of research, the teaching of ESP generally has been seen as a separate activity within English language teaching and ESP research has been seen as an identifiable component of applied linguistic research (Dudley-Evans & St. John, 1998).

The origins of ESP can be traced back to the 1960s, when there was a growing need for technological and business industries (Swales, 2000:59-61). ESP, the prime realization of applied discourse analysis, later evolved for every specialized area needing appropriate teaching materials. Recently, ESP has been utilized as an umbrella term with a multitude of acronyms denoting the various sub-fields (Dudley-Evans & St. John, 1998).

Under the ESP framework, there are two major sub-fields, English for Academic Purposes (EAP) and English for Occupational Purposes (EOP), which are distinguished by their research nature and pedagogical tradition (Robinson, 1991; Dudley-Evans & St. John, 1998). EAP is concerned with students’ needs to learn academic language, which constitutes the majority of ESP, whereas EOP comprises professional purposes in administration, medicine, law and business, and vocational purposes for non-professionals in work or pre-work situations (Dudley-Evans & St. John, 1998:7).

In the ESP domain, ELP is an important but comparatively uncultivated corner (Dudley-Evans & St. John, 1998:51). González and Vyushkina (2009) characterize English for Academic Legal Purposes (EALP) as being used in university degree programs, while English for Occupational Legal Purposes (EOLP) is used in training for practical skills in the workplace. Over the years, there has been continuing interest in the research of EALP (Bhatia, 1993; Bowles, 1995; Harris, 1997; Feak, Reinhart & Sinsheimer, 2000; Candlin, Bhatia, & Jensen, 2002; Badger, 2003; Du, 2009). Nevertheless, studies have been concerned mostly with material development, genre analysis, and curriculum design. Corpus-based studies on EOLP, in contrast, are relatively undeveloped. Badger (2003) once conducted a corpus-based study on law in the genre of newspaper law reports. He found that newspaper law reports serve the same function as law cases do, which facilitates law school students in identifying the reasoning of the legal decision of the case. His corpus-based study is innovative, but it is

EALP and is solely for reading. To be specific, corpus-based applications on EOLP are comparatively unseen and the voice that professional writing gathers in the workplace entails the directions for future research.

Accordingly, it is confirmed that while EALP is widely developed for law school students and academic purposes, there is an underlying need to build up EAOP, in particular, Patent English, for workplace needs.

2.3 Vocabulary Studies on Law

Writing for specific purposes requires familiarity with not only the content but also the language. Unfamiliarity with vocabulary in writing is perceived to be a challenging task for language learners. As the importance of teaching vocabulary in ESP has gained recognition (Swales, 1983), Coxhead & Nation (2001) have categorized vocabulary in ESP into four groups: high-frequency words, academic vocabulary, technical vocabulary, and low-frequency vocabulary.

Nation (2001) defines those words in the use of writing. High-frequency words refer to the most frequently used 2000 words of English used in all types of writing. Low-frequency words are the rarely used terms and cover only 5% of all words. Academic words, namely semi-technical or sub-technical vocabulary, are for academic purposes. This kind of vocabulary is common to a wide range of academic fields but is not what is known as high-frequency vocabulary and is not technical in that it is not typically associated with just one field (Chung & Nation, 2003:104). In contrast, technical words are the ones used in a specialized field that are considerably different from subject to subject. As Chung & Nation (2003:104) point out, technical vocabulary is largely of interest to and used by people working in a specialized field. In the genre of law, Mellinkoff (1963) suggests legal vocabulary highlighting those common words with uncommon meanings. For example, merger and acquisition bear the same literal meaning as ‘combination’ in general English. Nevertheless, in economic and financial law, *merger* depicts the acquisition of one company by another. This combination into a single legal entity will increase the benefits to each other and is semantically positive. As to *acquisition*, the combination often bears unequal treatment and is considered negative.

Since there is very little research on technical vocabulary in legal disciplines, Harris (1997) analyzed procedural vocabulary extracted from the area of English contract law. His research shows that technical words enhance legal reading and also strengthen text analysis skills. Denton (2009:5) covered frequently used legal vocabulary in his teaching. Specific meaning of vocabulary, such as merger and acquisition in economic law, is viewed as concept for him to teach. His research concludes that the learning of terminology for Legal English is the priority for participants to foster when they are learning vocabulary conceptually. In other words, learning legal vocabulary with concepts of the target context is essential in vocabulary development. Haberstroh (2009) developed the legal academic word list. His research enriches the well-established area of EALP at the present day; however, the rapidly growing trend of EOLP remains comparatively undeveloped.

In brief, a general conclusion can be drawn in that there is a need to prepare inter-disciplinary patent writing, but exploring technical vocabulary with corpus-driven approaches into such development has the higher priority.

3. Methodology for Corpus Creation

The present study adopted a corpus-based research approach to study patent technical words from the USPTO (United States Patent and Trademark Office) glossary[1] in the field of intellectual property, with an emphasis on their frequency and word associations in contemporary patents.

In assessing the proper coverage needed for a lexical study, the distribution of each IPR domain is taken into consideration beforehand. Figure 1 shows the results of the coverage of technical words of IPR from the USPTO glossary.

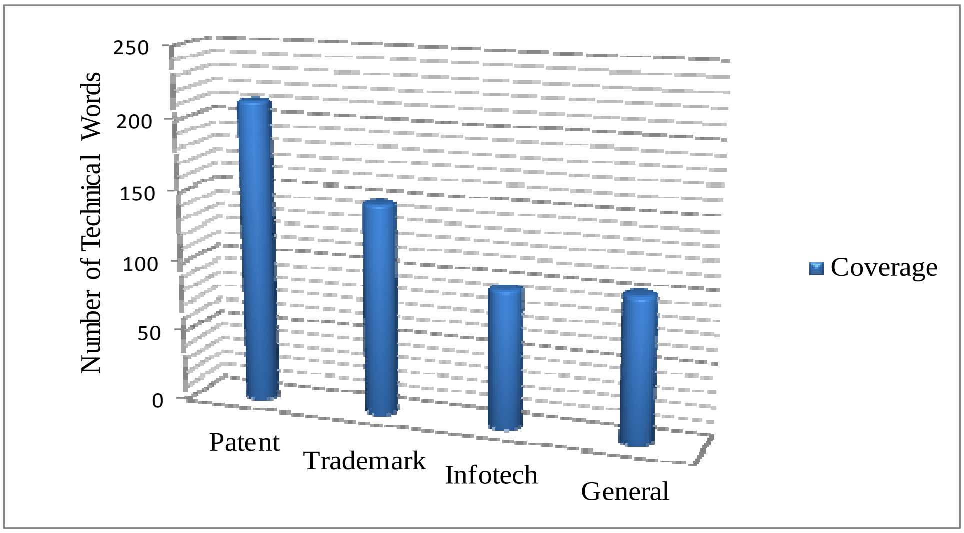


Figure 1. The coverage of technical words in intellectual property

The coverage was confined to the domains. As the USPTO glossary surveyed, four primary domains were outlined-patent, trademark, infotech (information technology), and general domain. Among the total 558 words of the glossary, 212 words are word items

[1] USPTO glossary is available at http://www.uspto.gov/main/glossary

included in the patent on a domain level, making up 38% of the total. In other words, the coverage of patent technical words was 38%, which is much higher than the 18.3% of the 102 words used in a general domain, as shown in Figure 1. Compared with the coverage of trademark (26.3%) and infotech (17.4%) domains, patent technical words are more widely covered in intellectual property. This suggests that there is a growing need in the area of patents and is consistent with the literature review, which suggests that patent plays a significant role in the genre of intellectual property.

3.1 Purpose

One of the major objectives in this section was to find the most frequently used technical words in patents. This aim was achieved by calculating the frequencies of each patent word in Figure 1. The frequency of the patent technical words has been listed according to the frequency of their occurrence in the USPTO Patent Full-Text and Image Database (PatFT)[2], and the distribution is presented in Table 1.

Table 1. Distribution of patent technical words in PatFT

Times of Occurrence	Number of Words	Percentage	Accumulative Percentage
≧ 1,000,000	7	53.20	53.20
1,000,000 ~ 999,999	23	42.56	95.76
10,000 ~ 99,999	21	3.62	99.38
1,000 ~ 9,999	39	0.53	99.91
100 ~ 999	40	0.07	99.98
1 ~ 99	53	0.02	100.00
0	29	0.00	100.00
TOTAL	212	100	100.00

Among the 212 patent technical words, 90 words (99.91%) occurred more than 1000 times in PatFT and are considered frequently used patent technical words. There were only 53 words (0.02%) that appeared less than 100 times and 29 words that did not appear at all, both of which are viewed as not frequently used technical words in patents. The other 40 words occurred less than 1000 times but more than 100 times in PatFT.

As can be seen, there were 7 words that occurred more than one million times. Among them, the most frequently used technical word was the verb ‘comprising,’ which appeared 3,785,213 times. Other technical word items, such as scope, patent, group and element,

[2] USPTO Patent Full-Text and Image Database (PatFT) is available at http://patft.uspto.gov/netahtml/PTO/search-adv.htm

consisting of, and drawing, occurred over one million times. The high-frequency of these words reflects the important role of technical vocabulary in patent texts.

With regard to word associations, Nattinger (1988) suggests that grouping of the words according to their meanings enhances vocabulary learning. He once mentioned that word grouping can be presented in the form of topic (situational sets). With a library, such words as book, shelf, borrow, loan, and so on can be taught together for teaching and learning. In order to get a clearer picture of the patent technical words for better use, the researchers here made a detailed analysis of the word associations based on topic.

The 212 patent technical words are considered to be statistically unusually frequent in their occurrence, but it was then noted that they seemed to fall into a limited number of recurring topic sets; therefore, six sections were proposed based on words in the same semantic network or field that share similar meanings or semantic features in PatFT: 'patent activity (99),' 'patent aid (25),' 'patent community (23),' 'patent claim (17),' 'patent description (30),' and 'people of the patent community (18)'. This was made not only on an intuitive basis, but also on the criteria of the produced data. The following illustrates the criteria the researchers set up for each section.

(1) What do patent-specific activities usually consist of? (Patent activity)

(2) What tools can be applied in a patent-specific context? (Patent aid)

(3) Where are patent-specific places in United States? (Patent community)

(4) What entities do patent applicants need for specific requests? (Patent claim)

(5) What specific entities can usually be found in patents? (Patent description)

(6) Who are in the patent-specific contexts? (People of the patent community)

Table 2 presents the top ten frequently used technical words in each of the six sections, which are arrayed according to their frequency of occurrence in descending order based on PatFT.

Table 2. Top 10 technical word items of six topic-based sections

Rank	Activity	Aid	Community	Claim	Description	People
1	patent	concept	Group	comprising	specification	representative
2	disclosure	doctrine of equivalents	Pubs	scope	sequence listing	person
3	application	file wrapper	TC	element	filing date	assignee
4	patent application	ADS	Technology Center	consisting of	serial number	applicant
5	continuation	mask work	ISA	drawing	application number	inventor
6	interference	EFS	IB	dependent claim	PLT	practitioner
7	demand	PAIR	RO	composed of	Control No.	attorney
8	restriction	OG	IPEA	independent claim	publication number	disclaimer
9	designation	PSIPS	GAU	benefit claim	issue date	CSR
10	divisional application	PALM	Group Art Unit	priority claim	patent number	lawyer

The keyword analysis made on a large number of words in the present study was not intended solely to keep interdisciplinary learners informed of the frequency of some word items, but also to awaken the learners to the influence of intellectual property and patent on lexical units, which might vary in accordance with different topics.

In addition to the top ten word items, the researchers calculated the total frequency and total words of each section. Table 3 shows the total frequency and total words of each topic section.

In the patent technical word list, patent claim accounts for 54%, followed by patent activity (making up 28%), patent community and people of the patent come next at 6%, and finally patent description (represented by 4.5%). Patent aid only constitutes 1.5% of all.

As patent law 35 U.S.C.§112 Paragraph 1 reads, “patent claim” is viewed as the specifications, containing a written description of the invention and the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same, and shall set forth the best mode contemplated by the inventor of carrying out his invention. That is to say, the patent claim of a published patent informs the public of the scope of rights that distinguish the invention. As it is technically dealt with, specific terms are used, allowing users to become familiarized with the invention an applicant owns.

Table 3. Frequency of the patent technical world list

Topic	Technical Words	Total Frequency	Percentage	Rank
Patent Activity	99	6,622,873	28	2
Patent Claim	17	12,695,484	54	1
Patent Community	23	1,455,693	6	3
People of the Patent	18	1,468,215	6	3
Patent Description	30	1,060,782	4.5	5
Patent Aid	25	342,988	1.5	6
TOTAL	212	23,646,035	100	-

As shown in Table 3, "patent claim," which has high priority, is valuable for corpus-based research. Besides, to build up a small-scale corpus for the present study, the researchers analyzed 'patent claim' based on parts of speech for further investigation. Table 4 shows the results.

Table 4. Distribution of patent claim in the patent technical word list

Group	Patent Technical Word	Total Frequency	Percentage
Noun	scope	2,459,656	55.24
	element	1,245,265	
	drawing	1,015,261	
	dependent claim	625,886	
	independent claim	587,926	
	benefit claim	437,599	
	priority claim	381,352	
	withdrawn claim	227,433	
	canceled claim	32,306	
	multiple dependent claim	494	
	rejoinder	80	
	claims	6	
Verb	comprising	3,785,213	44.76
	consisting of	1,165,427	
	composed of	617,353	
	consisting essentially of	114,211	
	having	16	
TOTAL		12,695,484	100

As can be seen, there are only two syntactic categories that can be found and noun words outperform verb words, making up a 55.24%. Accordingly, the researchers lay their attention on noun words of patent claim, and compile a patent technical word corpus.

As a first step motivated toward the establishment of a patent corpus for investigation, the present study was based on a corpus of U.S. patents, European patents, Patent abstracts of Japan, PCT[3] patents, and U.K. patents over a decade (2000 to 2009) gathered from LexisNexis[4], a corpus of a multitude of information for professionals in legal fields, in forms of case, newspaper, journal, and magazine reporting.

The Patent Technical Word Corpus (PTWC), made up of patent claim texts retrieved from LexisNexis, was created consisting of 16,101,256 word tokens.

Although LexisNexis does not have built-in patent claim subcorpora, the self-compiled PTWC adds significant strength to the development of claim language. Although an available specialized corpus contains an infinite amount data, constructing a small scale one would be needed for a profound linguistic study (Hsieh, 1998:48). Therefore, the PTWC was established for the present study.

3.2 Lexical Analysis Software

Owing to the size of the text collection, the quantitative analysis was computer-assisted, using WordSmith Tools 5.0 (Scott, 2008) to search for the word item as a string of letters to ascertain the absolute and relative frequency. The concordancer-tagged function of WordSmith 5.0 allowed us to calculate collocations and clusters around the search or node word.

With the help of such tools, we can find more discriminative linguistics patterns and structures of patents. Table 5 shows the result of citations from each level.

As can be seen, element, drawing, and scope were noun words that occurred over 10,000 times. In turn, claims occurred less than 10,000 times but more than 1,000 times on in-between levels. There were noun words that occurred less than 1,000 times-independent claim, dependent claim, multiple dependent claim, benefit claim, rejoinder, priority claim, withdrawn claim, and canceled claim. Compared with the other two levels, the current level covers noun words that were comparatively less used but more precisely employed. Dependent claim, for example, contains a reference to a claim previously set forth. Multiple dependent claim, in contrast, contains a reference to more than one claim previously set forth.

[3] PCT (Patent Cooperation Treaty).

[4] LexisNexis is available at https://www.lexisnexis.com

Table 5. Citations of patent claim

Patent Claim	Citation	Level
element	54,151	≧10,000 times
drawing	40,634	
scope	28,236	
claims	5,800	≦10,000 times
independent claim	249	≦1,000 times
dependent claim	59	
multiple dependent claim	58	
benefit claim	57	
rejoinder	22	
priority claim	3	
withdrawn claim	0	
canceled claim	0	
TOTAL	129,268	

As WordSmith tools provide a comprehensive view of noun words in patent claim texts, it is noted that more effort should be made to explore the possibilities on those that occurred less than 1,000 times. Therefore, the present study lays its focus on such words.

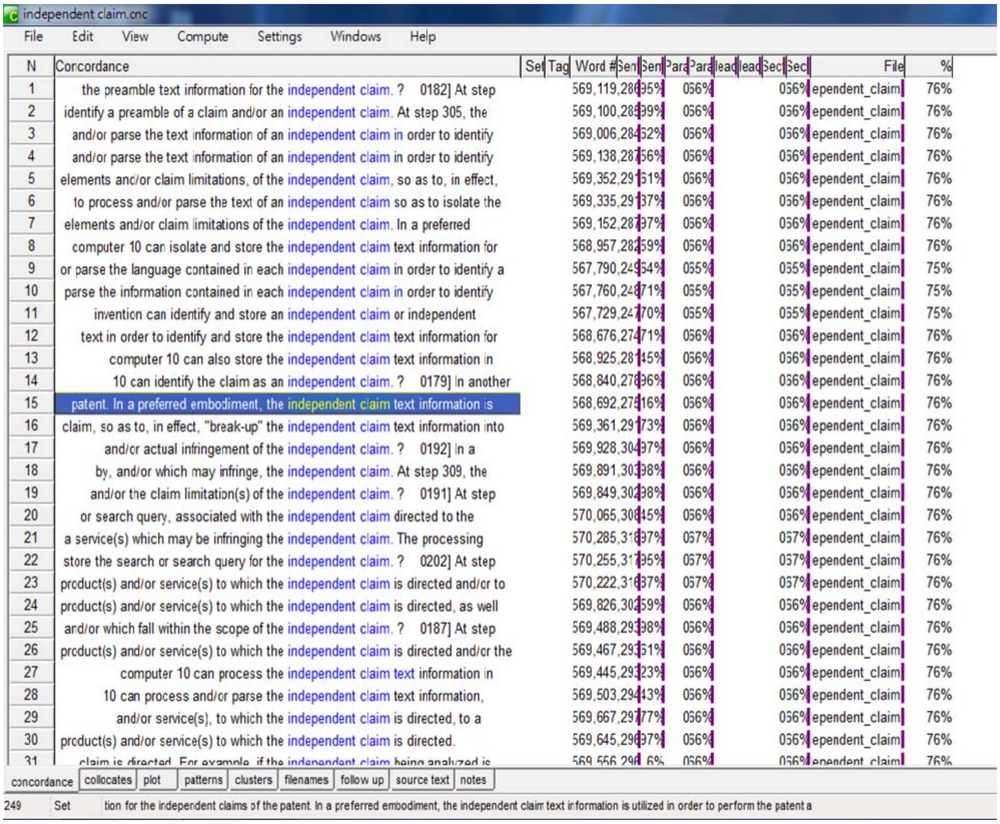
3.3 Data Selection Criteria

The researchers observed citation of each level from Table 5 and found those that appeared less than 1,000 times were more specific word items. Among them, ‘independent claim’ outperforms others with 249 occurrences.

Technically, an “independent claim” is a proper noun of patent that formally describes the invention in adding the essential features. In the patent application for a pencil, the independent claim might begin with “a device comprising a cylindrical piece of wood with a piece of graphite inserted into the center of the wood.” In such a case, the pencil was distinguished with regard to the shape (cylindrical) and the materials it was made of (wood and graphite). For the same pencil with another shape, it will not be taken into consideration for such an invention.

Patent claim is the precise legal definition of the invention, identifying the specific elements of the invention for which the inventor is claiming rights and seeking protection. A patent claim shares technical terminology with the rest of a patent but differs greatly in its contents and syntax (Sheremetyeva, 2003). Of patent claims, *independent claim* best describes the invention in adding essential features. Since the independent claim is specific in that it stands on its own and does not rely upon or refer to any other claims in a patent, the researchers chose "independent claim" as the target word for data analysis.

The corpus of the present study contains 98 English patents with 'independent claim' tagged in the patent specification[5], and is made up of 4,887,084 word tokens. The researchers use the concordance function to find the technical terminology 'independent claim,' with 249 citations generated. There is a list of the 249 examples of 'independent claim' with the words that preceded and followed. Figure 2 shows part of the concordance.



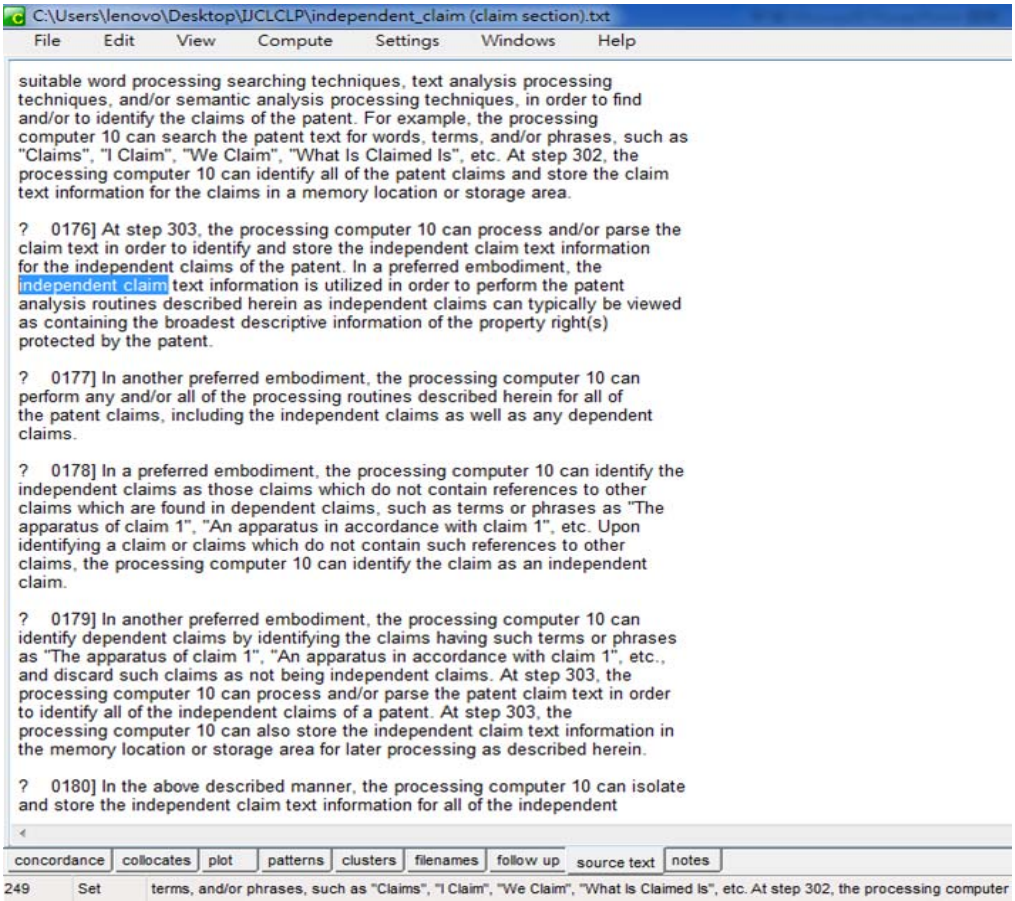
N	Concordance	Set	Tag	Word #	Sen	Sen	Para	Para	lead	lead	Sec	Sec	File	%
1	the preamble text information for the independent claim. ? 0182] At step			569,119,28		95%		056%				056%	ependent_claim	76%
2	identify a preamble of a claim and/or an independent claim. At step 305, the			569,100,28		99%		056%				056%	ependent_claim	76%
3	and/or parse the text information of an independent claim in order to identify			569,006,28		52%		056%				056%	ependent_claim	76%
4	and/or parse the text information of an independent claim in order to identify			569,138,28		56%		056%				056%	ependent_claim	76%
5	elements and/or claim limitations, of the independent claim, so as to, in effect,			569,352,29		61%		056%				056%	ependent_claim	76%
6	to process and/or parse the text of an independent claim so as to isolate the			569,335,29		37%		056%				056%	ependent_claim	76%
7	elements and/or claim limitations of the independent claim. In a preferred			569,152,28		97%		056%				056%	ependent_claim	76%
8	computer 10 can isolate and store the independent claim text information for			568,957,28		59%		056%				056%	ependent_claim	76%
9	or parse the language contained in each independent claim in order to identify a			567,790,24		54%		055%				055%	ependent_claim	75%
10	parse the information contained in each independent claim in order to identify			567,760,24		71%		055%				055%	ependent_claim	75%
11	invention can identify and store an independent claim or independent			567,729,24		70%		055%				055%	ependent_claim	75%
12	text in order to identify and store the independent claim text information for			568,676,27		71%		056%				056%	ependent_claim	76%
13	computer 10 can also store the independent claim text information in			568,925,28		45%		056%				056%	ependent_claim	76%
14	10 can identify the claim as an independent claim. ? 0179] In another			568,840,27		96%		056%				056%	ependent_claim	76%
15	patent. In a preferred embodiment, the independent claim text information is			568,692,27		16%		056%				056%	ependent_claim	76%
16	claim, so as to, in effect, "break-up" the independent claim text information into			569,361,29		73%		056%				056%	ependent_claim	76%
17	and/or actual infringement of the independent claim. ? 0192] In a			569,928,30		97%		056%				056%	ependent_claim	76%
18	by, and/or which may infringe, the independent claim. At step 309, the			569,891,30		98%		056%				056%	ependent_claim	76%
19	and/or the claim limitation(s) of the independent claim. ? 0191] At step			569,849,30		98%		056%				056%	ependent_claim	76%
20	or search query, associated with the independent claim directed to the			570,065,30		45%		056%				056%	ependent_claim	76%
21	a service(s) which may be infringing the independent claim. The processing			570,285,31		97%		057%				057%	ependent_claim	76%
22	store the search or search query for the independent claim. ? 0202] At step			570,255,31		95%		057%				057%	ependent_claim	76%
23	product(s) and/or service(s) to which the independent claim is directed and/or to			570,222,31		37%		057%				057%	ependent_claim	76%
24	product(s) and/or service(s) to which the independent claim is directed, as well			569,826,30		59%		056%				056%	ependent_claim	76%
25	and/or which fall within the scope of the independent claim. ? 0187] At step			569,488,29		98%		056%				056%	ependent_claim	76%
26	product(s) and/or service(s) to which the independent claim is directed and/or the			569,467,29		61%		056%				056%	ependent_claim	76%
27	computer 10 can process the independent claim text information in			569,445,29		23%		056%				056%	ependent_claim	76%
28	10 can process and/or parse the independent claim text information,			569,503,29		43%		056%				056%	ependent_claim	76%
29	and/or service(s), to which the independent claim is directed, to a			569,667,29		77%		056%				056%	ependent_claim	76%
30	product(s) and/or service(s) to which the independent claim is directed.			569,645,29		97%		056%				056%	ependent_claim	76%
31	claim is directed. For example, if the independent claim being analyzed is			569,556,29		6%		056%				056%	ependent_claim	76%

249 Set tion for the independent claims of the patent. In a preferred embodiment, the independent claim text information is utilized in order to perform the patent a

Figure 2. Concordance of independent claim tagged in the patent specification

[5] The patent specification is drafted both to satisfy the written requirements for patentability, as well as to define the scope of the claims. A typical patent specification has the following sections: (1) title of the invention, (2) field of the invention, (3) background of the invention, (4) object of invention, (5) statement of invention, (6) summary of invention, (7) description of the drawings, (8) detailed description of the preferred embodiments, (9) claims, and (10) abstract.

Out of the 249 examples of ‘independent claim,’ 5 were irrelevant to the researchers’ analysis because ‘independent claim’ was used without a subject in the present progressive tense-“identifying at least one independent claim of the patent.” Of the remaining 244 examples, all concordance entries for each were stored. Then, the source texts for each concordance line were manually analyzed one by one for further investigation. The authors provide a screenshot of the source text tool interface with technical terminology *independent claim* contained in the patent specification documentation in Figure 3.



suitable word processing searching techniques, text analysis processing techniques, and/or semantic analysis processing techniques, in order to find and/or to identify the claims of the patent. For example, the processing computer 10 can search the patent text for words, terms, and/or phrases, such as "Claims", "I Claim", "We Claim", "What Is Claimed Is", etc. At step 302, the processing computer 10 can identify all of the patent claims and store the claim text information for the claims in a memory location or storage area.

? 0176] At step 303, the processing computer 10 can process and/or parse the claim text in order to identify and store the independent claim text information for the independent claims of the patent. In a preferred embodiment, the independent claim text information is utilized in order to perform the patent analysis routines described herein as independent claims can typically be viewed as containing the broadest descriptive information of the property right(s) protected by the patent.

? 0177] In another preferred embodiment, the processing computer 10 can perform any and/or all of the processing routines described herein for all of the patent claims, including the independent claims as well as any dependent claims.

? 0178] In a preferred embodiment, the processing computer 10 can identify the independent claims as those claims which do not contain references to other claims which are found in dependent claims, such as terms or phrases as "The apparatus of claim 1", "An apparatus in accordance with claim 1", etc. Upon identifying a claim or claims which do not contain such references to other claims, the processing computer 10 can identify the claim as an independent claim.

? 0179] In another preferred embodiment, the processing computer 10 can identify dependent claims by identifying the claims having such terms or phrases as "The apparatus of claim 1", "An apparatus in accordance with claim 1", etc., and discard such claims as not being independent claims. At step 303, the processing computer 10 can process and/or parse the patent claim text in order to identify all of the independent claims of a patent. At step 303, the processing computer 10 can also store the independent claim text information in the memory location or storage area for later processing as described herein.

? 0180] In the above described manner, the processing computer 10 can isolate and store the independent claim text information for all of the independent

249 Set terms, and/or phrases, such as "Claims", "I Claim", "We Claim", "What Is Claimed Is", etc. At step 302, the processing computer

Figure 3. Source text tool interface with independent claim taggings

4. Transitivity Development of Independent Claim

To explore the embodiment, the researchers first looked at the transitivity in the corpus. Analyzing the transitivity patterns of independent claim, in turn, contributes to the understanding of how verbs measure the clausal function through Halliday’s systemic-functional view of language.

4.1 Transitivity

In Halliday's (2004:168) study, he distinguished six central processes that elicit transitivity to describe a whole clause, rather than the verb and its object. These six central processes, in turn, were material, mental, relational, behavioural, verbal, and existential processes. Each process type, however, constructs a particular experience that distinguished distinguishes clausal functions. Li (2010:3447) suggests transitivity analysis is a semantic perspective on the ideas expressed by a clause, a proposition about the world in which an event, situation, relation or attribute is predicted of some participants. Toward a functional view of language, the total set of functions used in interpreting the clause as representation can include the process types summarized in Table 6.

Table 6. Process types, their meanings and participants (Halliday, 2004:260)

Process type	Category meaning	Participants, directly involved	Participants, obliquely involved
Material action event	Doing doing happening	Actor, Goal	Recipient, Client; Scope; Initiator; Attribute
Mental perception cognition desideration emotion	Sensing seeing thinking wanting feeling	Senser, Phenomenon	-
Relational attribution identification	Being attributing identifying	Carrier, Attribute Identified, Identifier; Token Value	Attributor, Beneficiary Assigner
Verbal	saying	Sayer, Target	Receiver; Verbiage
Behavioural	behaving	Behaver	Behaviour
Existential	existing	Existent	-

As for the present study, transitivity analysis is applied as the 244 citations of independent claim were examined. The researchers first singled out each citation as a constructed clause. In this regard, the researchers conducted analysis at the clausal level to better reflect the actual process an independent claim was associated with. In this manner, the researchers elicited the verbs that distinguished each process type. For verification, the researchers derived nominals that represent participants in each clause. The researchers give an instance in (1).

(1) The processing computer can store the independent claim text information
Actor Process Goal

As shown in (1), ‘store’ outlines a material process in which the *processing computer* (Actor) accumulates *independent claim text information* (Goal). In such a case, *processing computer* which occurs with ‘store’ might provide selectional features[6] (Chomsky, 1965:111) of the knowledge of independent claim. It is noted that the verb-noun collocation ‘store + independent claim’ followed by *processing computer* is a subtle distinctive feature of independent claim that is expected to be known for such a grammatical pattern, which makes up knowledge of the grammar of patents in use. The investigation of this collocationally-fixed relationship will, in turn give insights to learners on how independent claim is used on a lexical level and prepare them for the actual business world they may need to work in or offer them the information regarding modern patent language where they already work.

4.2 Transitivity Structures of Independent Claim

Since independent claim describes the invention in adding the essential features, in this section, independent claim is annotated by three primary clauses of the total four clausal types found in the data. They are material, relational, and verbal clauses. The concept of clause as representation (Halliday, 2004) is applied to remind language users where to locate independent claim to produce correct sentences.

There are a total of four clausal types found in the data (see Table 7).

Table 7. Clausal types of patent specification tagged with independent claim

Clausal Type	Total Frequency	Percentage
Material Clauses	127	52.0
Relational Clause	65	26.6
Verbal Clause	48	19.7
Existential Clause	4	1.7
TOTAL	244	100

As Table 7 shows, material clauses make up the largest proportion of the total, accounting for 52%, with relational clauses coming next at 26.6%, followed by verbal clauses, making up 19.7%, and existential clauses at 1.7%. Nevertheless, behavioral clauses were not found as legal discourse in the Republic of China to address such phenomena. Tsai (2006:108) explains that law is essential as it elaborates the obligations of human beings. Behaviors such

[6] Chomsky (1965:111) defined selectional features as vocabulary knowledge requiring a number of qualified rules in which lexical items in question cannot appear. For example, *admire* only occurs with subject nouns that are human, such as man, not abstract ones, like faith.

as dream, cough, and cry, however, are basic instincts that human beings embrace. There is less importance to further develop such behaviors in the discourse of law. Although patent language and legislative language differ in their rationale, the fact that declarative sentences were favored in the examined clauses of the present study is in accordance with Tsai's (2006:109) research on legislative language.

In sum, it can be concluded from Table 7 that material clauses are the most frequent experience that independent claim shares, while existential clauses are the least. These clausal types of independent claim provide direction for the novice. They should learn material clauses first. As the distribution of independent claim involves different transitive processes, we make a further step to delineate how lexical items were generated with reference to the co-occurrence relations.

5. Lexical and Clausal-Specific Features of Independent Claim

In an attempt to characterize language-specific entities that could serve as a pedagogical base to help language awareness for patent writing, we explicate lexical and clausal-specific features of independent claim to promote discourse-level proficiency in modern patent language learning contexts.

5.1 Verb-Noun Collocation

The researchers investigated the verb-noun collocations in three primary clauses and introduce verb-noun collocations that make up the knowledge that learners need to be aware of in their learning. Verb-noun collocation here is defined as verbs with specific meaning that collocate with independent claim. The frequency of the verb-noun collocations then is annotated.

Technically, frequently used verbs in patents can be seen as concepts that carry meanings to specify the clauses for communication. Among the 244 examined clauses, the researchers found 23 verb-noun collocations from the data. Meanings of each collocating verb from the verb-noun collocations were carefully analyzed. Table 8 illustrates the results.

As Table 8 indicates, the auxiliary 'be' made up nearly 8.2%, while the rest constitutes 91.8%. 'Identify' and 'direct' were frequently used with independent claim, accounting for approximately 46%. 'Contain,' in contrast, was the third most remarkable (17.2%). These three verbs represent over 63% of the verb-noun collocations. There were five verb-noun collocations (identify, direct, contain, be, and correspond) that appeared over 10 times, making up 76.2%.

Table 8. Collocating verbs of patent specification tagged with independent claim

Verb	Verb Meaning	Total Frequency	Percentage
identify	to extract, recognize, discover, or find	61	25.00
direct	to request or enjoin with authority	51	20.90
contain	to have within	42	17.20
be	to state of having existence	20	8.19
correspond	to be in conformity or agreement	11	4.50
infringe	to encroach upon in a way that violates law or the rights of another	7	2.90
analyze	to determine the nature and relationship of the parts of by analysis	6	2.50
isolate	to set apart from others	6	2.50
perform	to carry out an action or pattern of behavior	6	2.50
generate	to bring into existence	5	2.00
process	to a series of actions or operations conducing to an end	4	1.64
store	to place or leave in a location	4	1.64
regard	to an aspect to be taken into consideration	4	1.64
exist	to have the functions of vitality	4	1.64
break up	to do away with	2	0.80
formulate	to develop a formula for the preparation	2	0.80
permit	to consent to expressly or formally	2	0.80
fall	to come within the limits	2	0.80
illustrate	to make clear	1	0.41
provide	to take precautionary measures	1	0.41
utilize	to turn to practical use or account	1	0.41
associate	to bring together or into relationship	1	0.41
exhibit	to show or display outwardly, especially by visible signs or actions	1	0.41
TOTAL		244	100

In most cases, ‘identify’ (to extract, recognize, discover, or find) collocates with an independent claim, making up 25% of the verb-noun collocations. Examples (2) to (4) demonstrate such collocations.

(2) The database can also contain any one or more of software programs and/or algorithms for parsing patent language in order to identify a claim or claims of a patent, software programs, and/or algorithms for parsing patent language in order to identify an independent claim or independent claims of a patent.

(3) Parsing claim information of the patent in order to identify at least one independent claim.

(4) The processing computer can identify and store the preamble text information for the independent claim.

As can be seen, in (2) to (4), "independent claim" is viewed as the Goal. For instance, Example (2) points out that database will parse the patent language to be identical in independent claim. Example (3) elaborates the behavior to parse information regarding patent claim to recognize independent claim. In (4), the processing computer enables the preamble textual information to be extracted with independent claim as the Goal. In (2) to (4), 'identify' is with the precise meaning "to cause something to become identical," implying that patent is a specific genre where the fixed verb meaning is embodied.

While vocabulary knowledge may involve a number of qualified rules of the kind Chomsky (1965) calls "selectional feature," a collocating verb has a selectional feature of its own. In other words, a collocating verb is a collocation-based feature of verb-noun collocation that maps the detailed contour of knowledge on clausal types. For each clausal type, the verb-noun collocations involved explain the grammar of words, the interaction between two associated participants, and the experience a particular clausal type has embraced. In this regard, verb-noun collocations elicited from the present study can equip learners with a better sense of the firmness of this collocational relationship.

5.2 Clausal Nominalization

As the verb-noun collocation 'independent claim + direct' shows a strong tendency in characterizing the passive structure of verbal clauses, the researchers found the nominalized *to which the independent claim is directed* functions as an adverbial constituent of the clauses and is unusually positioned clause-final. Based on this, 'independent claim + direct' is a selectional feature of clausal nominalization in verbal clauses as transitivity analysis is applied. Clausal nominalization, in turn, is a functional feature that elucidates mutual information shared in verbal clauses of the modern patent language. The following elaborates our finding.

Theme is a single constituent that happens to come at the beginning of a given clause that will label the function of the clause, while everything else in the clause is known as rheme. Example (5) illustrates the theme-rheme structure of the clause.

(5) What the duke gave to my aunt was this teapot

Theme Rheme

As Halliday (2004) elaborates, this kind of clause is known as a "thematic equative" because it sets up the theme-rheme structure in the form of an equation, where theme=rheme. According to Halliday, a form, such as what the duke gave my aunt, is an instance of a structural feature known as nominalization. In this case, theme is the primary element, while nominalization serves a thematic purpose for communication. Nevertheless, once the normal relationship is reversed, the nominalization becomes marked. In this fashion, it is called 'marked thematic equative,' as presented in Example (6).

(6) This teapot was what the duke gave my aunt

Theme Rheme

Syntactically, the theme-rheme structure constructs the topic of a clause and further helps learners identify the linguistic elements within, such as Goal and Actor of material clauses, Say and Verbiage of verbal clauses, or Identified and Identifier of relational clauses. In this regard, the researchers found verbal clauses in the data displayed marked thematic equative followed Halliday's research. Such kinds of nominalization of clausal or clause-like structures into a nominal one conform to Heyvaert's (2003) nominalization as functional reclassification. Based on Lehrmann (1988), such nominalization is the process wherein a clause is reduced so that it loses the properties of being a clause but acquires nominal properties that allows it to become a nominal or adverbial constituent of a matrix clause. In Halliday's (2004) term, such nominalization is known as a structural feature in which theme-rheme structure in the form of an equation occurred. In the following, the researchers examine clausal nominalization of verbal clauses and specify the syntactic environment where nominalized units were found.

Of the 48 verbal clauses, the researchers found that 48 (100%) were nominalized. Table 9 shows the findings.

Table 9. Clausal nominalization of verbal clauses

Item	Total Frequency	Percentage
Product	18	37.50
Product/service	15	31.25
Service	15	31.25
TOTAL	48	100

In the verbal clausal nominalization the researchers investigated, "to which an independent claim is directed" appears to be the adverbial constituent of the main clause nominalization. In this manner, product/service and service make up a similar proportion at

31.25%, whilst product represents 37.5%. Examples (7) to (9) illustrate such findings.

(7) A product to which the independent claim is directed.

(8) The product(s) and/or service(s) to which the independent claim is directed.

(9) A service to which the independent claim is directed.

As can be seen, these examples demonstrate not only 'marked thematic equatives' but also wh-cleft[7]. Based on this observation, the researchers found that rheme in verbal clauses of modern patent language states an authority to its target of product and/or service. In (7), for example, to which the independent claim is directed as rheme and the independent claim located requests for a particular product, a particular product is addressed by "to which the independent claim is directed" where the independent claim is within.

In short, the emergence of nominalization underlines the psychological phenomenon that human beings' verbal behavior (independent claim) is embodied in modern patent language. Further, since the verb-noun collocation 'independent claim + direct' has no other similar collocation in verbal clauses, "to which an independent claim is directed" was of mutual information value with the same rheme but alternative themes.

5.3 Semantic Prosody

As mentioned earlier, a verb-noun collocation has selectional features that associate it with a particular set of semantic contexts. A verbal clause, for example, shows a tendency to occur when a product collocates with 'independent claim + direct.' Based on this, it shows how a verbal clause is regularly found collocated with 'independent claim + direct' that share a semantic similarity-product. In this regard, the semantic context that attracts such a verb-noun collocation is considered 'semantic prosody'. Since the function of semantic prosody is to transfer communicative purposes (Stubbs, 2009:125), the researchers lay their attention on semantic prosody of the verb-noun collocations to further elucidate semantic associations in patent environment of independent claim.

Based on the verb-noun collocations the researchers examined, semantic prosodic associations of the technical terminology independent claim contained in the patent specification are elaborated in Table 10 below.

[7] 'Wh-cleft' involves the division and repacking of the information in a clause in two parts (Locks, 1996:238).

Table 10. Semantic prosodic relation of independent claim taggings

Prosodic Type	Semantic Prosody	Total Frequency	Percentage
Innovation	product, present invention	63	25.8
Technology	processing computer, processing device	59	24.2
Service	service	39	16.0
Knowledge	information	34	13.9
Tool	apparatus, database, vehicle	29	11.9
Function	search query, claim	16	6.60
Violation	infringement	4	1.60
TOTAL		244	100

From the corpus-based analysis, verb-noun collocations of independent claim were found to collocate mostly with prosodic type ‘innovation’ (25.8%), followed by ‘technology’ (24.2%), ‘service’ (16%), and ‘knowledge’ (13.9%), making up nearly 80% of the total. All of these prosodic types imply a positive semantic prosody-patents are important assets of human beings. Based on this, the researchers argue that semantic prosody is the exponent of a special correlation between the semantic structure and syntactic form they were put into. The distribution of the prosodic items, in turn, shows the extent of the syntactic forms expressed by semantic links of the grammar of words. The present study rated those over 20% as high frequency; less than 20% but more 10% as mid frequency; less than 10% as low frequency. It is noted that 1.6% were concerned with infringement. This is of lower percentage but of importance in that the public should draw their attention to the rise of potential perils as ‘violation’ (infringement), which bring about torts and plagiarism, were overlooked.

The researchers turned their focus on the low frequency level for an instance. In their opinion, aside from prosodic type “violation” which is on the low frequency level discussed earlier, there is a rate of 6.6% verb-noun collocations that co-occur with prosodic type “function” that might elicit the underlying mechanisms of independent claim. The researchers give examples in (10) to (11).

(10) A search query containing information corresponding to the at least one independent claim.

(11) An example of a search or search query, associated with the independent claim directed to the exemplary vehicle locating apparatus, can include the following search words terms and/or connectors.

From the above examples, search query is viewed as the semantic prosody that co-occurs with ‘correspond + independent claim’ and ‘associate + independent claim,’ respectively. In (10) and (11), search query is the shared prosody embraced by different collocating verbs. Since semantic prosody is a powerful linguistic device in that it stands for language universality (Lewandowska-Tomaszczyk, 1996:159), the result obtained from the low frequency level further highlights how it is shared by a particular syntactic category of collocating verbs, which may motivate the investigation into different frequency levels for future research.

5.4 Discourse Thematic Referentiality

Chen (2009:1666) proposed a discourse-functional approach “discourse thematic referentiality” to the referential use of NP. He points out such context-dependent referentiality is viewed as thematicity of referents or referentiality in terms of thematic importance of objects in discourse. Based on this, he holds the view that grammatical categories, such as nouns and verbs, are potential functional features to perform the referring function. He lays his attention on the noun group as we lay the focus on how semantic prosody associated with verb-noun collocations. He further emphasizes that the noun group is of genuine importance in that it highly represents thematic referentiality in the context of language use.

In the previous section, semantic prosody is considered referential of thematic importance in the discourse of independent claim. As for the present study, semantic prosody, however, only collocates with certain verbs unusually. Some share the same verbs; some share a unique verb on their own; some have both tendencies. In this section, semantically, we state the intimacy between semantic prosodies and independent claim taggings. Pragmatically, we address semantic prosodies that are referential when they were structured with collocating verbs that highlight the referring functions.

As Table 11 shows, discourse thematic referentiality shows a strong tendency of language specificity. It can be said of true condition in which conditions that must be satisfied by the world if an utterance of a declarative sentence is true. For example, the utterance “There is a cat on the table” is only true if there actually is a table with a cat on it at that time of the utterance (Hurford, Heasley, & Smith, 2007:252). Based on this, discourse thematic referentiality of material clauses can be realized only when the processing computer, processing device, present invention, product/service, search query, information, apparatus, database, or claim is associated with specific verb-noun collocations of independent claim. Once inappropriate elements, such as boy toy or gossip girl appear, it violates the truth condition because it goes with the wrong semantic prosody so as to hinder semantic presupposition (Levinson, 1983:201).[8] Further, once an inappropriate verb works with semantic prosody, it no longer satisfies the truth condition and infringes on semantic presupposition. For example, processing device only works with ‘identify’ and once either ‘analyze’ or ‘fall’ is adopted, the principle is not cooperated with; discourse thematic referentiality then is cancelled.

Table 11. Discourse thematic referentiality of material clauses

Theme (Semantic Prosody)	Referentiality (Verb)	Discourse (Genre)
processing computer	[+identify], [+be], [+break up], [+contain], [+formulate], [+generate], [+infringe], [+isolate], [+perform], [+process], [+regard], [+store]	independent claim
processing device	[+identify]	independent claim
present invention	[+identify]	independent claim
product/service	[+fall]	independent claim
search query	[+infringe]	independent claim
information	[+identify], [+correspond], [+provide]	independent claim
apparatus	[+identify], [+be], [+utilize], [+store]	independent claim
database	[+identify]	independent claim
claim	[+analyze], [+permit]	independent claim

Of the relational clauses, ‘contain’ addresses the function mostly as a product/service, information, and service, in turn, becoming thematically referential, as described in Table 12.

Table 12. Discourse thematic referentiality of relational clauses

Theme (Semantic Prosody)	Referentiality (Verb)	Discourse (Genre)
product/service	[+be], [+exhibit], [+contain]	independent claim
search query	[+correspond]	independent claim
information	[+contain], [+regard], [+correspond]	independent claim
service	[+be], [+contain], [+regard]	independent claim

Of verbal clauses, discourse thematic referentiality is maintained when semantic prosodies work with ‘direct.’

[8] Semantic presupposition is presupposition based on either truth conditional theory or semantic relations, which were defined in terms of semantic feature or atomic concepts.

Table 13. Discourse thematic referentiality of verbal clauses

Theme (Semantic Prosody)	Referentiality (Verb)	Discourse (Genre)
product/service	[+direct]	independent claim
product	[+direct]	independent claim
service	[+direct]	independent claim

As shown in Table 13, product/service, product, and service were referential once they were functioned with ‘direct.’ Further, ‘direct’ is specifically used in that it appears in only verbal clauses. The degree of discourse thematic referentiality, therefore, is therefore comparatively stronger than that of other clauses. It appears that product and service are basic prosodies that, when interacting with a semantic trigger, ‘direct,’ brings about discourse thematic referentiality. Based on clausal nominalization mentioned earlier, in Example (7) (“A product to which the independent claim is directed”), product and ‘direct’ were essential linguistic components that represent the relatively compositionality fixed relationship of verbal clauses.

In sum, discourse thematic referentiality accounts for how collocating verb, semantic prosody, and independent claim are constructed linguistically. Before closing, it is important to accentuate discourse thematic referentiality, which addresses how lexical units build up modern patent language, providing empirical evidence for the overall characterization of independent claim.

6. Conclusion and Future Work

There has been little investigation into modern patent language in applied linguistics research. Therefore, the present study fills the gap by compiling a patent technical word corpus. The researchers create a patent technical word list regarding frequently used word items of six primary patent areas. Such a word list is significant in that it can help learners expand their vocabulary by displaying the words they should learn. Further, since learners are especially deficient in verb-noun collocations (Chen & Tang, 2004; Liu, 1999), collocational patterns identified in the present study can equip learners with a better sense of verb-noun collocational relationships. For practitioners and researchers, the results of the present study are essential to be incorporated into the English for Occupational Purposes curriculum development.

On the one hand, practitioners and researchers can encourage the application of independent claim as a primer or beginner guide of English patent language. Based on the functional account of independent claim, teachers can show examples by means of clauses as the hidden context. Students can learn how clauses are used in patents under different situations. For example, a product to which an independent claim is directed in Example (7) is a verbal clause constructed by virtue of collocating verb ‘direct’ and semantic prosody product in which clausal nominalization occurred. Based on this, teachers can integrate verb-noun collocation ‘independent claim + direct’ to guide students to notice the overlooked prosodic relations. Moreover, embedded clausal nominalization can be taught for a better understanding of the rhetorical function. Finally, teachers can encourage students to apply and learn other technical vocabulary for the writing of professional patents.

On the other hand, vocabulary teaching needs to take account of semantic prosody (Hunston, 2002:142) because ESL/EFL textbooks or bilingual dictionaries do not explicitly represent the feature of semantic prosody or may provide inappropriate semantic prosodic information that can mislead language learners (Zhang, 2009:10). In this regard, teachers can choose a particular area that students familiar with or feel interested in to encourage the application of semantic prosody to further develop technical vocabulary for the writing of patents. Consequently, functional accounts of independent claim add relatively importance in the teaching of technical vocabulary for the writing of professional patents.

Although the ‘independent claim’ corpus in this study contains over 4.8 million running words, it is relatively small compared to the PTWC corpus (16 million running words) for Patent English. It is suggested that future works can examine other technical words such as ‘dependent claim’ or ‘beneficial claim’ in order to generalize the results.

Due to restricted time, the present study examines contemporary patents over a decade, 2000 to 2009. It is suggested that future work can further probe into different temporal periods so as to provide a more comprehensive point of view for this field.

Further, since the present study aims at exploring language-specific characteristics of independent claim, teachers can measure students’ familiarity from functional perspectives. It is suggested that future work can collect students’ writing and compare their use of verb-noun collocations, semantic prosody, and other linguistic features. The results may provide in-depth insights into how teachers can help students learn technical vocabulary in the EOLP-based courses.

As the extensive use of generic terms and vague expressions poses a great challenge in patent retrieval (Sarasúa, 2000), it would be essential to research on linguistic specificity of patent lexis for a better understanding of relational lexical semantics in modern patent language. In considering Sheremetyeva’s (2003) approach to analyzing patent claim texts with natural language processing (NLP) methodology which improved analyses robustness, our work, in contrast, pinpoints the preliminaries and peculiar associations in patent documentation. Aside from playing a role in modern patent language, the proposed approach and genre-based characteristic analysis is considered influential in bridging ELP to NLP for future research.

References

American Bar Association. (2010). *ABA section of intellectual property law: Introduction*. Retrieved from http://www.abanet.org/intelprop/intro.html

Badger, R. (2003). Legal and general: Toward a genre analysis of newspaper law reports. *English for Specific Purposes,* 22, 249-263.

Ball, C. (1996). *Tutorial notes: Concordances and corpora*. Retrieved from http://www.georgetown.edu/cball/corpora/tutorial.html

Bhatia, V. K. (1993). *Analyzing genre: Language use in professional settings*. New York: Longman.

Biber, D, Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bowles, H. (1995). Why are newspaper law reports so hard to understand? *English for Specific Purposes,* 14, 201-222.

Candlin, C. N., Bhatia, V. K., & Jensen, C. H. (2002). Developing legal writing materials for English second language learners: Problems and perspectives. *English for Specific Purposes*, 21, 299-320.

Chen, C. Y., & Tang, Y. I. (2004). Collocation Errors of Taiwanese College Students: Oral and Written Production. In *Proceedings of the Tenth International Symposium on English Teaching* (pp. 278-286). Taipei: Crane.

Chen, H. J. (2001). Taiwanese EFL learner corpus and interlanguage analysis. In *Proceedings of the Tenth International Symposium on English Teaching* (pp. 288-299). Taipei: Crane.

Chen, P. (2009). Aspects of referentiality. *Journal of Pragmatics*, 41, 1657-1674.

Cheng, L., Sin, K. K., & Li, J. (2008). A discursive approach to legal texts: Court judgement as an example. *The Asian ESP Journal*, 4(1), 14-28.

Chiu, S. H. (2008). WAR metaphor in legal discourse: A reminder of their perils. *The Seventh International Conference on Research and Applying Metaphor*, Cáceres, Spain. May 29-31.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, 15(2), 103-116.

Coxhead, A., & Nation, P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for Academic Purposes* (pp. 252-267). Cambridge: Cambridge University Press.

Denton, J. (2009). Content vs. concept: Two different focuses in the teaching of Legal English. In *Proceedings of ESP Seminar: English for Legal Purposes* (pp. 4-11). National University of Kaohsiung, Taiwan.

Du, J. (2009). Content and language integration in tertiary education in China: A case study in Wuhan Law College. *The Asian ESP Journal*, 5(1), 61-77.

Dudley-Evans, T., & St. John, M. (1998). *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Feak, C. B., Reinhart, S. M., & Shinshimer, A. (2000). A preliminary analysis of law review notes. *English for Specific Purposes*, 19, 197-220.

González, M., & Vyushkina, E. G. (2009). International cooperation in designing effective methods to prepare non-native EFL teachers for training and assessing Legal English skills. *Georgetown Law Global Legal Skills Conference IV*, Washington D.C., United States. June 4-6.

Haberstroh, J. (2009). The LAW List (The Legal Academic Word List). *Georgetown Law Global Legal Skills Conference IV*, Washington D.C., United States. June 4-6.

Halliday, M. (2004). *An introduction to functional grammar* (3rd Ed). New York: Oxford University Press.

Harris, S. (1997). Procedural vocabulary in law case reports. *English for Specific Purposes*, 16(4), 289-308.

Hayvarert, L. (2003). *A cognitive-functional approach to nominalization in English*. Berlin: Mouton de Gruyter.

Hsieh, S. K. (1998). *Characteristics of legal Mandarin: A corpus-linguistic approach to the criminal law*. Unpublished master's thesis, Fu-Jen Catholic University, Taipei, Taiwan.

Hurford, J. R., Heasley B., & Smith, M. (2007). *Semantics: A course book*. New York: Cambridge University Press.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K., & Tse, P. (2005). Hooking the reader: A corpus-based study of evaluative that in abstracts. *English for Specific Purposes*, 24, 123-139.

Hyland, K. (2008). Small bits of textual material: A discourse analysis of Swales's writing. *English for Specific Purposes*, 27, 143-160.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25, 56-75.

Lehmann, C. (1988). Toward a typology of clause linkage. In J. Haiman and S. A. Thompson (Eds.), *Clausal combining in grammar and discourse* (pp. 181-225). Amsterdam: John Benjamins.

Levinson, C. (1983). *Pragmatics*. New York: Cambridge University Press.

Lewandowska-Tomaszczyk, B. (1996). Cross-linguistic and language-specific aspects of semantic prosody. *Language Science*, 18 (1-2), 153-178.

Li, J. (2010). Transitivity and lexical cohesion: Press representation of a political disaster and its actors. *Journal of Pragmatics*, 42, 3444-3458.

Liu, C. P. (1999). An analysis of collocational errors in EFL Writings. In *Proceedings of the Eighth International Symposium on English Teaching* (pp. 483-494). Taipei: Crane.

Lock, G. (1996). *Functional English grammar: An introduction for second language teachers*. New York: Cambridge University Press.

McEnery T., & Wilson A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Mellinkoff, D. (1963). *The language of the law*. Boston: Little, Brown and Co.

Modiano, M. (2001). Linguistic imperialism, cultural integrity, and EIL. *ELT Journal*, 55 (4), 339-347.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25, 235-256.

Nattinger, J. (1988). Some current trends in vocabulary teaching. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 62-82). New York: Longman.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nelson, M. (2006). Semantic associations in business English: A corpus-based analysis. *English for Specific Purposes*, 25, 217-234.

Robinson, P. C. (1991). *ESP today: A Practitioner's guide*. New York: Prentice Hall.

Sarasúa, L. (2000). Cross lingual issues in patent retrieval. In *Online Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval*. Athens, Greece.

Scott, M. (2008). WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

Sheremetyeva, S. (2003). Natural language analysis of patent claims. In *Proceedings of the ACL 2003 Workshop on Patent Corpus Processing* (pp. 66-73). Sapporo, Japan.

Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics*, 30(1), 115-137.

Swales, J. M. (1983). Vocabulary work in LSP: A case of neglect. *Bulletin CILA*, 37, 21-34.

Swales, J. M. (2000). Language for specific purposes. *Annual Review of Applied Linguistics*, 20, 59-76.

Tsai, S. I. (2006). *Legal language used in laws of the Republic of China*. Unpublished doctoral dissertation. National Tsing Hua University, Hsinchu, Taiwan.

Tsai, Y. (2008). Supply and demand analysis of patent translation. *Translation Journal*, 12(3), http://accurapid.com/Journal/45patents.htm

World Intellectual Property Organization (2009). *WIPO intellectual property handbook: Policy, law and use*. Geneva: WIPO Publication.

Zhang, W. (2009). Semantic prosody and ESL/EFL vocabulary pedagogy. *TESL Canada Journal*, 26(2), 1-12.

The individuals listed below are reviewers of this journal during the year of 2011. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

June-Jei Kuo

Chao-Jan Chen

Lun-Wei Ku

Richard Tzong-Han Tsai

Jiun-Shiung Wu

2011 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 16

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2011.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

B

Barrett, Neil Edward
and Li-mei Chen. English Article Errors in Taiwanese College Students' EFL Writing; 16(3-4): 1-20

C

Chao, F. Y. August
and Siaw-Fong Chung. A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; 16(3-4): 21-40

Chen, Hsin-Hsi
see Wang, Chieh-Jen, 16(3-4): 61-76

Chen, Li-mei
see Barrett, Neil Edward, 16(3-4): 1-20

Chen, Yi-Cong
see Lin, Bor-Shen, 16(3-4): 41-60

Cheng, Jyun-Hua
see Liang, Tyne, 16(1-2): 15-26

Chiu, Wei-Yun
see Liu, Chao-Lin, 16(1-2): 27-46

Chung, Siaw-Fong
see Chao, F. Y. August, 16(3-4): 21-40

H

Hsieh, Shelley Ching-Yu
see Lin, Darren Hsin-Hung, 16(3-4): 77-106

Hsu, Chih-Chuan
and Shih-Hung Wu. Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; 16(1-2): 47-68

J

Jin, Guantao
see Liu, Chao-Lin, 16(1-2): 27-46

L

Lee, Hsin-Chieh
see Tsai, Wei-Ho, 16(1-2): 1-14

Liang, Tyne
and Jyun-Hua Cheng. Resolving Abstract Definite Anaphora in Chinese Texts; 16(1-2): 15-26

Lin, Bor-Shen
and Yi-Cong Chen. Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; 16(3-4): 41-60

Lin, Darren Hsin-Hung
and Shelley Ching-Yu Hsieh. Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; 16(3-4): 77-106

Liu, Chao-Lin
Guantao Jin, Qingfeng Liu, Wei-Yun Chiu, and Yih-Soong Yu. Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; 16(1-2): 27-46

Liu, Qingfeng
see Liu, Chao-Lin, 16(1-2): 27-46

T

Tsai, Wei-Ho
and Hsin-Chieh Lee. Performance Evaluation of Speaker-Indetification Systems for Singing Voice Data; 16(1-2): 1-14

W

Wang, Chieh-Jen
and Hsin-Hsi Chen. Intent Shift Detection Using Search Query Logs; 16(3-4): 61-76

Wu, Shih-Hung
see Hsu, Chih-Chuan, 16(1-2): 47-68

Y

Yu, Yih-Soong
see Liu, Chao-Lin, 16(1-2): 27-46

SUBJECT INDEX

A

Anaphora Resolution
Resolving Abstract Definite Anaphora in Chinese Texts; Liang, T., 16(1-2): 15-26

C

Chinese Historical Documents
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

Chinese Text
Resolving Abstract Definite Anaphora in Chinese Texts; Liang, T., 16(1-2): 15-26

Chinese Text Analysis
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

Collocation
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

Co-Ocurrence
Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; Lin, D. H.-H., 16(3-4): 77-106

Corpus
Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; Lin, D. H.-H., 16(3-4): 77-106

A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; Chao, F. Y. A., 16(3-4): 21-40

D

Definite Anaphora
Resolving Abstract Definite Anaphora in Chinese Texts; Liang, T., 16(1-2): 15-26

Definite Article
English Article Errors in Taiwanese College Students' EFL Writing; Barrett, N. E., 16(3-4): 1-20

Definition Relation
A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; Chao, F. Y. A., 16(3-4): 21-40

Dictionary
A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; Chao, F. Y. A., 16(3-4): 21-40

Digital Humanities
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

F

Feature Extraction
Resolving Abstract Definite Anaphora in Chinese Texts; Liang, T., 16(1-2): 15-26

G

General Query
Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; Hsu, C.-C., 16(1-2): 47-68

H

Hearer Knowledge
English Article Errors in Taiwanese College Students' EFL Writing; Barrett, N. E., 16(3-4): 1-20

Histogram Equalization
Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; Lin, B.-S., 16(3-4): 41-60

I

Indefinite Article
English Article Errors in Taiwanese College Students' EFL Writing; Barrett, N. E., 16(3-4): 1-20

Information Retrieval
Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; Hsu, C.-C., 16(1-2): 47-68

Intellectual Property Rights
Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; Lin, D. H.-H., 16(3-4): 77-106

Intent Analysis
Intent Shift Detection Using Search Query Logs; Wang, C.-J., 16(3-4): 61-76

Intent Shift Detection
Intent Shift Detection Using Search Query Logs; Wang, C.-J., 16(3-4): 61-76

K

Keyword Trends
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

M

Machine Learning
Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; Lin, B.-S., 16(3-4): 41-60

Mandarin Radical *mu4*(目)
A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; Chao, F. Y. A., 16(3-4): 21-40

Model Adaptation
Performance Evaluation of Speaker-Identification Systems for Singing Voice Data; Tsai, W.-H., 16(1-2): 1-14

Multilayer Perceptrons
Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; Lin, B.-S., 16(3-4): 41-60

Multi-Level Semantic Relation
A Measurement of Multi-Level Semantic Relations among Mandarin Lexemes with Radical *mu4*: A Study based on Dictionary Explanations; Chao, F. Y. A., 16(3-4): 21-40

N

Natural Language Processing
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

P

Patent Document Processing
Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; Lin, D. H.-H., 16(3-4): 77-106

Q

Query Expansion
Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; Hsu, C.-C., 16(1-2): 47-68

S

Search Query Logs Analysis
Intent Shift Detection Using Search Query Logs; Wang, C.-J., 16(3-4): 61-76

Singing
Performance Evaluation of Speaker-Identification Systems for Singing Voice Data; Tsai, W.-H., 16(1-2): 1-14

Speaker Identification
Performance Evaluation of Speaker-Identification Systems for Singing Voice Data; Tsai, W.-H., 16(1-2): 1-14

Systemic Functional Linguistics
Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents; Lin, D. H.-H., 16(3-4): 77-106

T

Temporal Analysis
Some Chances and Challenges in Applying Language Technologies to Historical Studies in Chinese; Liu, C.-L., 16(1-2): 27-46

Test Collection
Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; Hsu, C.-C., 16(1-2): 47-68

U

Unknown Word Extraction
Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; Lin, B.-S., 16(3-4): 41-60

W

Wikipedia
Evaluating the Information Retrieval Performance of Query Expansion Method and On-line Search Engine on General Query; Hsu, C.-C., 16(1-2): 47-68

Word Identification
Histogram Equalization on Statistical Approaches for Chinese Unknown Word Extraction; Lin, B.-S., 16(3-4): 41-60

Z

Zero Article
English Article Errors in Taiwanese College Students' EFL Writing; Barrett, N. E., 16(3-4): 1-20

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:________________

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

CARD HOLDER SIGNATURE : _____________________________________

Tel.: ____________________________________ E-mail: ____________________________________

Add: ____

PAYMENT FOR

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

Quantity Wanted: ______

US$ _____ ❑ Publications:_____________________________________

US$ _____ ❑ Text Corpora: __________________________________

US$ _____ ❑ Speech Corpora:_________________________________

US$ _____ ❑ Others: _______________________________________

US$ __________ ❑Life Member Fee ❑ New Member ❑Renew

US$ __________ = Total

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

ACLCLP
% Institute of Information Science, Academia Sinica
R502, 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：________________

卡號:_________-_________-_________- ________　有效日期：________________

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$__________□ 中研院詞庫小組技術報告

NT$__________□ 中文（新聞）語料庫

NT$__________□ 平衡語料庫

NT$__________□ 中文詞庫八萬目

NT$__________□ 中文句結構樹資料庫

NT$__________□ 平衡語料庫詞集及詞頻統計

NT$__________□ 中英雙語詞網

NT$__________□ 中英雙語知識庫

NT$__________□ 語音資料庫______________________________

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:
115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

			AIR	AIR		
		Surface	(US&EURP)	(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$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與説明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統説明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 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.16 no.3-4 September/December 2011

Contents

Papers

English Article Errors in Taiwanese College Students' EFL Writing…… 1
Neil Edward Barrett and Li-mei Chen

基於辭典辭彙釋義之多階層釋義關聯程度計量─以「目」字部爲例…… 21
趙逢毅、鍾曉芳

Histogram Equlaization on Statistical Approaches for Chinese Unknown Word Extraction…… 41
Bor-Shen Lin and Yi-Cong Chen

Intent Shift Detection Using Search Query Logs…… 61
Chieh-Jen Wang and Hsin-Hsi Chen

Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents…… 77
Darren Hsin-Hung Lin and Shelley Ching-Yu Hsieh

Reviewers List & 2011 Index…… 107

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing