International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

Vol.19 No.2 June 2014 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
Eduard Hovy
University of Southern California, U. S. A.
Chu-Ren Huang
The Hong Kong Polytechnic University, H. K.
Jian-Yun Nie
University of Montreal, Canada
Richard Sproat
University of Illinois at Urbana-Champaign, U. S. A.
Keh-Yih Su
Behavior Design Corporation, Hsinchu
Chiu-Yu Tseng
Academia Sinica, Taipei
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

Yuan-Fu Liao **(Section Editor)**
National Taipei University of Technology, Taipei
Berlin Chen
National Taiwan Normal University, Taipei
Hung-Yan Gu
National Taiwan University of Science and Technology, Taipei
Hsin-Min Wang
Academia Sinica, Taipei
Yih-Ru Wang
National Chiao Tung University, Hsinchu

Information Retrieval

Ming-Feng Tsai **(Section Editor)**
National Chengchi University, Taipei
Chia-Hui Chang
National Central University, Taoyuan
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
Hao-Jan Chen
National Taiwan Normal University, Taipei
Huei-ling Lai
National Chengchi University, Taipei
Meichun Liu
National Chiao Tung University, Hsinchu
James Myers
National Chung Cheng University, Chiayi
Shu-Chuan Tseng
Academia Sinica, Taipei

Natural Language Processing

Richard Tzong-Han Tsai **(Section Editor)**
Yuan Ze University, Chungli
Lun-Wei Ku
Academia Sinica, Taipei
Chuan-Jie Lin
National Taiwan Ocean University, Keelung
Chao-Lin Liu
National Chengchi University, Taipei
Jyi-Shane Liu
National Chengchi University, Taipei
Liang-Chih Yu
Yuan Ze University, Chungli

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

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. 19, no. 2, June 2014

Contents

Papers

Social Metaphor Detection via Topical Analysis..... 1
Ting-Hao (Kenneth) Huang

Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem……..... 17
Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen

Resolving the Representational Problems of Polarity and Interaction between Process and State Verbs…….. 33
Shu-Ling Huang, Yu-Ming Hsieh, Su-Chu Lin, and Keh-Jiann Chen

不同母語背景華語學習者的用詞特徵：以語料庫為本的研究... 53
張莉萍

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 2, June 2014, pp. 1-16
© The Association for Computational Linguistics and Chinese Language Processing

Social Metaphor Detection via Topical Analysis

Ting-Hao (Kenneth) Huang*

Abstract

With the massive amount of social media data becoming available, there is a rising interest in automatic metaphor detection and interpretation from open social text. One of the most well-known approaches to this subject is identifying the violation of selectional preference. The basic concept of selectional preference is that verbs tend to have semantic preferences of their arguments and that violations of these preferences are strong indicators of metaphorical language use. Nevertheless, previously, few works have focused on metaphor detection of social media data. In response to this problem, we propose a three-step framework that is based on the technology of selection preference modeling to detect metaphors in social media data. We conduct a pilot study of this framework on the data of a real-world online support group. Furthermore, to improve our approach, we also leverage topical analysis techniques in our framework. As a result, we address the challenges of the task of metaphor detection in social media data, provide qualitative analysis for our experiments, and illustrate our insight based on the results.

Keywords: Metaphor, Cluster, Selectional Preference, Social Media Data.

1. Introduction

With massive social media data, *e.g.*, comments, blog articles, or tweets, becoming available, there is a rising interest in automatic metaphor detection from open social text. One of the most well-known approaches to this subject is detecting the violation of selectional preference. The idea of selectional preference is that the predicates (*i.e.*, mostly verbs) tend to have semantic preferences of their arguments. For instance, the verb "flex" has a strong preference of "muscle" and "bone" as its object. If we find that, in some text, the object of "flex" is not of the semantic class of "muscle" and "bone," it is very likely to be a metaphorical use.

Previously, researchers have studied metaphor identification by modeling selectional preference (Loenneker-Rodman & Narayanan, 2010; Shutova *et al.*, 2010; Shutova, 2010;

* Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
E-mail: windx@cmu.edu

Resnik, 1997; Shutova & Teufel, 2010; Calzolari *et al*., 2010; Preiss *et al*., 2007), while few papers have focused on social media data. In our work, we call the metaphors occurring in social media "social metaphor" to emphasize their different properties and difficulty.

Furthermore, to improve the technology of metaphor detection, we also leverage topic analysis techniques in our approach. The intuition behind combining metaphor identification and topic analysis is that both verbs and arguments exhibit strong tendencies towards a few specific topics, and this topical information provides additional evidence to facilitate identification of selectional preference among text. For instance, in the topic of sports, the subjects of "flex" are mostly humans; but in the topic of finance or politics, the subjects of "flex" are mostly organizations or countries, *e.g.*, "*China to flex its financial muscles at US meeting*." In this paper, we study how the metaphor detection technique can be influenced by topical analysis techniques.

The problem of automatic social metaphor detection poses two main challenges. First, as social media data is usually noisy, how to effectively preprocess the input texts before an actual detection component is employed should be studied carefully. We should estimate empirically the performance of existing NLP tools, especially lemmatizers and POS taggers. Second, how to apply and evaluate the proposed approach on a real world data set is not straight-forward. As there is neither an existing data set nor benchmark to evaluate metaphor detection, we need to create a benchmark that can show the performance difference effectively.

Furthermore, incorporating topical analysis into metaphor detection has another layer of challenges: how to automatically discover the topical distribution for each term (including verbs and nouns) within open text, which is not a trivial problem. Moreover, we need to study how to leverage the topical distribution of each verb and noun to metaphor detection.

In this paper, we will define the problem before proposing our 3-step approach for meta-phor detection. Specifically, we first preprocess the input text by extracting tokens and further clustering nouns then detect selectional association outliers. Finally, we apply a selectional preference strength filter to extract metaphor-embedded text snippets.

We then conduct experiments on a real-world social media data set. The LDA model is applied to partition the input corpus based on topics, and we adopt the 3-step approach both on the whole corpus and on every single topic data partition. Finally, we compare the metaphor detection results between those with and without the influence of topics, and we observe which one performs better.

The rest of the paper is organized as follows: In Section 2, we summarize related work for metaphor detection based on selectional preference detection. In Section 3, we formally de-fine the problem of automatic social metaphor detection. Then, in Section 4, we conduct a

preliminary test to compare two technologies for metaphor detection and choose one to establish the 3-step framework we will describe in Section 5. In Section 6, we further discuss the details of topic analysis. Finally, we demonstrate the experiment in Section 7, discuss the results in Section 8, and conclude the work in Section 9.

2. Related Work

In this section, we briefly survey papers that investigate approaches to detect metaphors in text.

2.1 Automatic Metaphor Detection

There have been many computational approaches in the field of natural language processing toward modeling metaphors. Based on Shutova *et al.* (2010), the research of modeling meta-phors could be divided into two sub-fields: metaphor detection and metaphor interpretation. In this paper, we focus on metaphor detection. In this field, the first challenge is how to define a metaphor. As mentioned in Loenneker-Rodman and Narayanan (2010), "*there is rich continuing theoretical debate on the definition and use of metaphor*." In our work, we limited the scope of our research in that we only aim to detect a "non-conventionalized metaphor," which usually has low frequency and could reasonably be considered as an outlier of selectional preferences. For instance, conventional metaphors like "Life is a journey" or "Time is running out," which would not strongly violate the selectional preference, are considered to be out of scope of this work.

In the field of metaphor detection, the Met* System (Fass, 1991) can be considered the first attempt to explore this field, and the following approaches include (Goatly, 1997), (Peters & Peters, 2000), CorMet System (Mason, 2004), and TroFi System (Birke & Sarkar, 2006). Most of them adopt the concept of selectional preference that we mentioned above, along with some hand-coded knowledge base, *e.g.*, VerbNet. VerbNet contains information about the constraint of arguments of verbs. By matching the text with the verb and its argument, we are able to detect the violation of arguments. Nevertheless, in this paper, we apply a different approach that learns the violations directly from statistics based on natural texts. One advantage of this approach is that we do not need any hand-coded knowledge base, so it could be ported to other languages more easily.

2.2 Topical Analysis

Many topical analysis techniques have been developed, *e.g.*, latent semantic analysis, proba-bilistic LSA, NMF, and LDA. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) models documents using a latent topic layer. In LDA, for each document d, a multinomial distribution θ_d over topics first is sampled from a Dirichlet distribution with

parameter α. Second, for each word w_{di}, a topic z_{di} is chosen from this topic distribution. Finally, the word w_{di} is generated from a topic-specific multinomial distribution φ_{zdi}. Accordingly, the generating probability of word w from document d is:

$$P(w \mid d, \theta, \phi) = \sum_{z \in T} P(w \mid z, \phi_z) P(z \mid d, \theta_d)$$

Basically, we will use this approach as our topical analysis component to discover the un-derlying topic distribution for nouns, verbs, and adjectives.

3. Problem Definition

In this section, we formally define the problem of the social metaphor detection via topic diversity identification.

Social Metaphor detection: We aim to recognize non-conventionalized metaphors in so-cial media text by a fully automatic approach, where the input would be real text from social media. Based on the word distribution among the input data, we aim to detect metaphors without using any external knowledge resources.

There are many sub-categories of metaphors. In this work, we only focus on "non-conventionalized metaphors," which reasonably could be considered as an outlier of language behavior. One advantage of non-conventionalized metaphors is that the approach can be language-independent and there is no need of external knowledge resources. This type of framework reasonably could be ported to other languages.

We will present how to tackle the problem by our proposed 3-step framework and discuss how to take the advantage of topical analysis for metaphor detection. We will also show how to quantitatively calculate these values in the next section.

4. Preliminary Test

As mentioned above, one of the most important approaches of metaphor detection is to identify the violation of selectional preference. Nevertheless, none of the other approaches are proposed as a baseline model to compare with the proposed model. In this section, to investigate the reliability of selectional preference modeling, we adopted another possible approach for metaphor detection, *i.e.*, the semantic outlier word detection, and run a preliminary test to compare their effectiveness.

4.1 Semantic Outlier Word Detection

Intuitively, for a certain topic, people tend to use the words that are "semantically related" to the topic. Therefore, we can assume that the set of words that are used frequently to describe a certain topic are more strongly related to each other than to the words used to describe other

topics. For instance, the words used to describe "finance," *e.g.*, bank, money, and business, are semantically more similar (or related) to each other than to the words used to describe "entertainment," *e.g.*, movie, music, and star. Based on this idea, we can detect the "semantic outlier" in a chunk of text, which can indicate the words that are borrowed from other topics to establish metaphors.

In this paper, we basically followed the method proposed by (Inkpen & Désilets, 2005) to detect the semantic outlier words. For a chunk of an input sentence, we first use the DISCO[1] package to calculate the pair-wise semantic similarities between any two words within the text, before calculating the average of the three greatest similarities of each word as its "semantic coherence (SC)." Finally, the semantic outliers tend to have obviously lower semantic coherence than other words, so we just set an empirical threshold to capture those outliers.

4.2 Selectional Association Outlier Detection

Selectional preference (also referred to as selectional association or selectional restriction) describes the semantic preference of predicates to noun classes in a given grammatical rela-tion. For instance, the predicate "eat" prefers the noun class of "food" as its *direct object* more than the noun class "building" and also prefers the noun class of "human" and "animal" as its *subject* more than the noun class "vehicle". Modeling selectional preference could help us to find the anomaly grammatical argument, which is an important clue to metaphorical language.

In this paper, for a given predicate p and a semantic noun class c, we adopt the measure of selectional association (SA), which was proposed by Resnik (1997), to present the selectional preference value between them. The selectional association equation can be calculated similar to point-wise mutual information, as follows:

$$A_R(p,c) = \frac{1}{S_R(p)} \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)}$$

A_R is the selectional association value between a given predicate p and a semantic noun class c. SR is the selectional preference strength of p, which can be formally defined similar to the K-L divergence between prior and posterior, as follows:

$$\begin{aligned} S_R(p,c) &= D(\Pr(c \mid p) \,||\, \Pr(c)) \\ &= \sum_c \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)} \end{aligned}$$

Finally, similar to Section 4.1, the selectional preference outliers tend to have obviously lower SA value than others, so we set an empirical threshold to capture those outliers. Note

[1] http://www.linguatools.de/

that, for this preliminary test, we only focus on the direct-object (*dobj*) and subject (*subj*) grammatical relations.

4.3 Experiment and Discussion

Since labeling metaphor embedded sentences is laborious, we conduct experiments on a relatively small benchmark corpus, which contains 122 sentences extracted from the Web, where 61 (50%) of them contain metaphors and 61 of them do not contain metaphors.

We apply both approaches on this data set. For the selectional association outlier detection, the best resulting F-1 score is 0.58, with precision of 0.60 and recall of 0.56. On the other hand, for the semantic outlier word detection, regardless of which value of threshold we set, the performance remains very low. This method returns a huge number of false positive semantic outliers, which is mainly caused by two reasons.

First, the semantic coherence can be affected easily by very general words, which usually have very high similarities and occur very often. If one sentence has more than one very general context word, *e.g.*, "take," "put," or "get," the semantic coherences of all other words could be systematically increased, and thereby fail to present the outlier words. We believe this is the main reason this method cannot detect the semantic outliers we expected.

Second, the measure of semantic similarity between word pairs is not very reliable for in-frequent words. The similarities calculations that are based on the text of a large corpus usually have this problem – being reliable on high frequency words, but not on low frequency words, which are exactly what we aim to capture.

To conclude, the selectional association outlier detection method outperformed the semantic outlier word detection in the preliminary test. Therefore, in this paper, we only focus on selectional association to develop our technology.

5. 3-Step Framework of Metaphor Detection

In this section, we introduce our approach to the problem of social metaphor detection.

In particular, our approach consists of three steps: (1) word extraction and building noun clustering, (2) selectional association outlier detection, and (3) selectional preference strength filtering. The first step deals with noisy input social media data, and it produces relatively clean output with richer NLP information labeled on the text. In the second step, we use a statistical method to calculate the selectional association scores of particular types of token pairs, based on the tokens and noun clusters extracted from the first step. Finally, as a post-process step, the output generated from the first step will be further analyzed and false positives will be filtered out via an empirical threshold.

5.1 Step 1: Word Extraction and Noun Clustering

Different from well-phased corpora, *e.g.*, Wall Street Journal or Wikipedia pages, which are used by other metaphor detection works, social metaphors tend to be embedded in noisy social media texts, *e.g.*, blog and forum texts. The goal of word extraction is to filter out the noise from grammatically structured phrases and tokens.

We first use a POS tagger to label the tokens with part-of-speech tags. Nevertheless, since the POS taggers are unlikely to produce high quality results on noisy data, we only select nouns with word frequency greater than 5 and greater than 70% of the overall occurrences as a noun. For adjectives and verbs, more strictly, we require a word frequency greater than 50 and over 80% of all occurrences should be adjectives or verbs. All of these parameterized thresholds are set experimentally.

Then, based on the nouns we extracted, we build a set of semantic noun clusters, which is the foundation for modeling the selectional preference. In this work, we apply the spectral clustering algorithm as follows.

1. For each noun W_N, we use the DICSO toolkit, which uses Wikipedia as the knowledge source, to generate its top 100 semantically similar nouns. For the first similar word W_{S1}, the similarity weight $Sim(W_N, W_{S1})$ is set to 1/2; for the second word, $Sim(W_N, W_{S2})$ is 1/3; for the third word, $Sim(W_N, W_{S3})$ is 1/4, and so on.
2. For all nouns, the first step will generate an asymmetric graph of word similarity. Based on the graph, we run the spectral clustering algorithm on it and get the noun cluster.

Note that, although the DISCO toolkit calculates word similarity based on Wikipedia, which is a reliable corpus, we only focus on the nouns actually occurring in the input data set, *i.e.*, the social media data. Namely, if a certain noun appears in the extracted "top 100 semantically similar nouns" but never occurs in the input data, we just ignore it. Moreover, we ignore the similarity score produced by the toolkit and calculate the similarity based on the similarity ranking. This is because, for the top 100 similar words, we tend to trust the ranking more than the scores, which is a common engineering trick for a clustering problem.

5.2 Step 2: Selectional Association Outlier Detection

Based on the formula mentioned in Section 4.2 and the semantic noun clusters built in Step 1, we measure the selectional associations for the most frequent verbs we extracted, particularly on the three kinds of grammatical relations, namely, adjective modifier (*amod*), direct object (*dobj*), and subject (*subj*).

In this work, we intentionally include the adjective modifier (*amod*) relation. When speak-ing of the selectional preference, most previous works have focused only on verbal predicates. Nevertheless, in the grammatical relation of adjective modifier, the modifier can also be considered as a predicate and the words being modified are mostly also nouns. Therefore, we also apply our approach on the *amod* relation and see if the method effectively captures adjective metaphors as well.

We considered the relations with negative SA values as "SA outliers," and we labeled the sentences containing "SA outliers" as metaphors.

5.3 Step 3: Selectional Preference Strength Filter

As mentioned in Section 4.3, the selectional preference strength of a predicate is defined as the K-L divergence between the prior and the posterior of noun clusters. For the predicates with strong preference, *e.g.*, "filmmake," it significantly affects the posterior probability distribution of noun clusters. In the case of the direct object of "film-make," the probability of the "movie/film" noun class is increased considerably. On the other hand, some "light verbs," *e.g.*, "get," "put," or "take," have quite weak preferences toward their direct object or subject.

The idea of selectional preference strength filtering was first proposed by Shutova *et al.* (2010) and suggests that the predicates with less strong selectional preference would rarely "violate" their own weak preference. Therefore, if we filter out the predicates with weak se-lectional preference, the false positives of metaphor detection will be reduced, and the preci-sion will increase significantly. In our framework, we apply this filtering method as the final step. Note that, due to the lack of a training and development data set, we just set the same threshold, which is 1.32, as suggested in Shutova *et al*. (2010).

6. Topic Model Analysis

We use LDA to model the topical distribution of words and documents of corpora, and we want to observe the changes of selectional preferences among various topics. The steps are as follows.

1. We train an LDA topic model with *k* various topics based on the whole input data set, *i.e.*, social media corpus.
2. For each document *d* in the input data set, we assign *d* to its favorite topic. Namely, we partition the corpus into *k* document collections, based on topics.
3. Run the 3-step process mentioned in Section 5 on the whole data set and on the *k* dif-ferent document collections.
4. Compare the SA outlier detection results among the data with and without topic modeling.

The underlying hypothesis in this comparison is that the selectional preference would increase for certain predicates in certain topics; thus, the outlier of SA values would be further emphasized. In that case, the metaphor detection technique could be improved.

7. Experiment

7.1 Data and Setting

Our method requires a fully-parsed data set, so we decided to choose a relatively small size of social media data. We collected the whole text of posts from a large online breast cancer support community, Breastcancer.org, which also is used in Wen *et al*. (2013). We have collected all of the public posts, users, and their profiles on the discussion board platform from October 2001 to January 2011. During this period, there were a total of 90,242 unique users who posted 1,562,459 messages. We then parsed it by the Stanford Parser toolkit[2]. In our word extraction step, we extracted 55,511 distinct nouns, 3,242 distinct adjectives, and 1,827 distinct verbs.

In the noun clustering step, we experimentally set the number of clusters (k) as 2,000. Note that we also manually removed the following three clusters to avoid some systematic parsing errors of the Stanford parser:

- hours, minutes, times, days, weeks, months, seconds, …
- yourselves, oneself, somebody, everybody, someone, anything, everything, anyone, …
- boy, girl, child, woman, children, guy, kid, person, …

In the topic model analysis phase, we adopted the JGibbLDA[3] toolkit to build the model and set the number of topics (k) as 20.

7.2 Results and Case Study

For the whole data set, the top 10 sample detected selectional association outliers[4] (of the three grammatical relationships) are listed in Table 1. We also demonstrate the result of one

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] A Java Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference: http://jgibblda.sourceforge.net/

[4] For each pair of predicate and noun cluster, we try to select the most "metaphor-like" usage if multiple outliers are detected. To protect the privacy of forum users, we also skip all the examples which contain name entities.

out of 20 topic document collections in Table 2 for comparison. Note that example usages are lightly disguised based on the techniques suggested by Bruckman (2006).

We found out that the strength of selectional preference of each predicate was actually in-creased in split topics. Nevertheless, the increase had no clear benefits to metaphor detection in our results. It successfully detected "outliers," but those outliers were not necessarily metaphors.

Take the results of direct object for example. Without topic analysis, the top outliers we detected were *(accomplish, Bianca)*, *(defy, breast)*, *(occupy, breast)*, and *(sprinkle, germ)*. Most of them are just rarely used verb-object combinations, but not metaphors. With topic analysis, we picked one topic out of 20 as an example, and the top outliers we detected were *(celebrate, cancer)*, *(join, skin)*, *(draw, brow)*, and *(play, head)*. We can observe that the verbs and nouns are actually more concentrated. In this case, the topic seems like celebration/play/event/play. Nevertheless, those pairs are rare, but not metaphors.

8. Discussion

Though the final result is not very promising, we gain some valuable experience in this work.

First, a parsing error is lethal for our approach. It would hurt our performance in at least two aspects: putting incorrect nouns in the noun cluster, which is the foundation of the whole method, and creating a significant amount of noise in the data, thereby impacting the statistical modeling phase. Therefore, the pre-processing is critical. After we added the strict word extraction strategy into our system, the quality of output was improved.

Second, from our experiments, we found that the strength of selectional preference is actually increased when clustering the documents by topic modeling. In each topic's document collection, we collected documents by word co-occurrences. Therefore, predicates are more concentrated on their preferred grammatical arguments. Nevertheless, the enhancement of selectional preference strength turned out not strong enough to improve metaphor detection. For some certain topics, the top SA outliers were even worse than those of the whole set, because selectional association is a linguistic phenomenon with high data sparsity. Partitioning would further reduce the amount of data and affect the reliability of the model.

Finally, we noticed that our fundamental hypothesis might not be accurate. We found that the SA outliers are not necessarily metaphors. Some of the outliers just rarely-used language, or some "weird" usage, *e.g.*, *(hug, multiply)* in "*the hugs we are storing will multiply*" of Ta-ble 1, or the *(play, head)* in "*It keeps playing through my head now*" of Table 2. In the future, we might need to reconsider the hypothesis we adopted.

Table 1. Examples of Selectional Association Violation Identified without Topical Analysis

Relation (arg0, arg1)	SA(10^{-3})	Example Usage	Analysis
amod			
amod(breast, yearly)	-2.7306	"yearly breast MRI"	Parsing Error
amod(skin, circular)	-2.7079	"circular skin patches"	Non-metaphor
amod(skin, greasy)	-2.6896	"greasy skin"	Non-metaphor
amod(head, administrative)	-2.6864	"the administrative head of this institute"	Weak metaphor
amod(hug, weary)	-2.6461	"…get weary. Hugs to you all…"	Sentence Segmentation Error
amod(breast, uncertain)	-2.6138	"The breast dimpling and uncertain mammography…"	Parsing Error
amod(kiss, french)	-2.5970	"…about French kiss…"	Non-metaphor
amod(breast, slim)	-2.5752	"My breasts are not slim but not fat..."	Non-metaphor
amod(tomorrow, crisp)	-2.5636	"…it's expected to be a crisp 72 tomorrow."	Parsing Error
amod(wing, seasoned)	-2.5510	"seasoned chicken wings"	Non-metaphor
dobj			
dobj(defy, breast)	-2.5893	"gravity defying breasts"	Parsing Error
dobj(occupy, breast)	-2.5749	"…(cancer) occupy the whole breast…"	Non-metaphor
dobj(sprinkle, germ)	-2.5350	"sprinkle wheat germ"	Non-metaphor
dobj(ooze, skin)	-2.5260	"oozing skin"	Non-metaphor
dobj(circulate, breast)	-2.5157	"…let air circulates around patient's breast."	Parsing Error
dobj(win, tomorrow)	-2.5095	"If John win tomorrow night, …"	Metonymy
dobj(hire, dvd)	-2.4972	"hire the dvd"	Non-metaphor
dobj(defy, cancer)	-2.4773	"…to defy the cancer and smile…"	Non-metaphor
dobj(float, cancer)	-2.4380	"…cancer cells float around in my blood…"	Non-metaphor
dobj(shut, head)	-2.4141	"…shut my head off…"	Metaphor
nsubj			
nsubj(cleanse, breast)	-2.5783	"breast cleanse"	Parsing Error
nsubj(metabolize, tumor)	-2.5513	"Tumors metabolize …"	Non-metaphor
nsubj(deny, adjuster)	-2.4950	"The claims adjuster denied this claim …"	Non-metaphor
nsubj(occupy, head)	-2.4827	"…keep my head occupied …"	Weak metaphor

nsubj(multiply, hug)	-2.4617	"…the hugs will multiply."	Metaphor
nsubj(constipate, hug)	-2.4286	"… hugs … that percocet is constipating."	Parsing Error
nsubj(overtake, belly)	-2.3276	"… my belly has overtaken the boobs …"	Metaphor
nsubj(multiply, treatment)	-2.2361	"…treatment for.. , multiply that by…"	Weak metaphor
nsubj(pay, patient)	-2.2164	"…patients pay for…"	Non-metaphor
nsubj(manufacture, expander)	-2.2056	"…ask the expander manufactures come up with better tissue expander."	Parsing Error

Table 2. Examples of Selectional Association Violation Identified Based on Topical Analysis (for one Particular Topic)

Relation (arg0, arg1)	**SA(10^{-3})**	**Example Usage**	**Analysis**
amod			
amod(head, gray)	-2.5469	"gray head"	Metonymy
amod(belly, former)	-2.5462	"your former belly"	Non-metaphor
amod(carcinoma, vaginal)	-2.5452	"… vaginal squamous cell carcinomas …"	Non-metaphor
amod(cancer, unilateral)	-2.5144	"unilateral breast cancer"	Non-metaphor
amod(breast, unilateral)	-2.4714	"unilateral breast"	Non-metaphor
amod(lesion, bilateral)	-2.3713	"bilateral lesions"	Non-metaphor
amod(treatment, immediate)	-2.3687	"immediate treatment"	Non-metaphor
amod(flyer, weekly)	-2.3064	"weekly flyer"	Non-metaphor
amod(symptom, bilateral)	-2.2976	"bilateral symptoms"	Non-metaphor
amod(tumor, enlarged)	-2.2626	"enlarged malignant tumor"	Non-metaphor
dobj			
dobj(celebrate, cancer)	-2.7801	"…celebrate my 10th cancer free year."	Parsing Error
dobj(weigh, head)	-2.7256	"So many questions … is weighing my head."	Metaphor
dobj(join, skin)	-2.7097	"…join the skin together…"	Non-metaphor
dobj(draw, nose)	-2.4197	"…drew a nose on it."	Non-metaphor
dobj(play, cheek)	-2.3255	"…play up my eyes…"	Non-metaphor
dobj(join, slew)	-2.1792	"Mary joined a slew of women …"	Non-metaphor
dobj(play, tomorrow)	-2.1190	"Playing golf tomorrow…"	Parsing Error

dobj(apply, forehead)	-2.0029	"…apply directly to the forehead."	Non-metaphor
dobj(pay, cancer)	-1.9471	"…price to pay for surviving cancer…"	Non-metaphor
dobj(regain, head)	-1.9457	"…regained a full head of hair…"	Parsing Error
nsubj			
nsubj(specialize, patient)	-2.3001	"…specializes in working with breast cancer patients, …"	Parsing Error
nsubj(pay, treatment)	-2.2237	"…get the treatment and self pay, …"	Parsing Error
nsubj(cover, cheek)	-2.0421	"…my cheeks covered with…"	Non-metaphor
nsubj(pay, head)	-1.8908	"…you're drinking safe and only your head is paying the price."	(Weak) metaphor
nsubj(pay, homeschooling)	-1.7228	"…the homeschooling paid off."	Non-metaphor
nsubj(build, expander)	-1.3925	"... an expander to build ..."	Parsing Error
nsubj(cover, melatonin)	-1.3865	"…melatonin covers the need for…"	Non-metaphor
nsubj(cover, wife)	-1.2500	"…so his wife should be covered…"	Non-metaphor
nsubj(cover, nurse)	-1.1849	"…the nurses talking about the insurance would cover it."	Parsing Error
nsubj(cover, dose)	-1.1708	"…do the single big dose to cover 2 weeks…"	Non-metaphor

9. Conclusion and Future Work

In this paper, we tried to leverage one of the most well-known approaches in detecting the violation of selectional preference with topical analysis techniques. The idea of selectional preference is that verbs tend to have semantic preferences of their arguments, while topical information provides additional evidence to facilitate identification of selectional preferences among text. Although our experimental results show that topics do not have strong impact on the metaphor detection techniques, we analyzed the results and presented some insights from our study.

As our next step, to reconsider our hypothesis, we need to quantitatively compare our re-sults to the gold-standard benchmark. Another interesting experiment might be to cluster the predicates, similar to nouns, as in our experiments, because the predicates still suffer from the sparsity issue.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily repre-senting the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

We would also like to thank Zi Yang for his help of the topical analysis experiments, Teruko Mitamura and Eric Nyberg for their instructions, and Yi-Chia Wang and Dong Ngu-yen for the work of data collection.

References

Birke, J., & Sarkar, A. (2006). A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*, *6*, 329-336.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993-1022.

Bruckman, A. (2006). Teaching students to study online communities ethically. *Journal of Information Ethics*, *15*(2), 82-98.

Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. eds. (2010). In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17-23 May 2010, Valletta, Malta. European Language Resources Association.

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, *17*(1), 49-90.

Goatly, A. (1997). *The language of metaphors*. volume 3. Routledge London.

Inkpen, D., & Désilets, A. (2005). Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*. October 6-8, 2005. Vancouver, British Columbia, Canada. NRC 48278.

Loenneker-Rodman, B., & Narayanan, S. (2010). Computational approaches to figurative language. *Cambridge Encyclopedia of Psycholinguistics*.

Mason, Z. (2004). Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, *30*(1), 23-44.

Ng, A., Jordan, M., Weiss, Y., *et al*. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *2*(8), 49-856.

Peters, W., & Peters, I. (2000). Lexicalised systematic polysemy in wordnet. In *Proc. Secondt Intnl Conf on Language Resources and Evaluation*.

Preiss, J., Briscoe, T., & Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 45, 912.

Resnik, P. (1997). Selectional preference and sense disambiguation. In P*roceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, 52-57. Washington, DC.

Shutova, E., & Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *Proceedings of LREC*.

Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 1002-1010. Association for Computational Linguistics.

Shutova, E. (2010). Models of metaphor in nlp. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 688-697. Association for Computational Linguistics.

Wen, M., Zheng, Z., Jang, H., Xiang, G., & Rose, C. (2013). Extracting Events with Informal Temporal References in Personal Histories in Online Communities. In *ACL'13*.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 2, June 2014, pp. 17-32
© The Association for Computational Linguistics and Chinese Language Processing

Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem

Yi-Ching Zeng[*], Tsun Ku[+], Shih-Hung Wu[*],

Liang-Pu Chen[#], and Gwo-Dong Chen[+]

Abstract

The paper addresses an opinion mining problem: how to find the helpful reviews from online consumer reviews via the quality of the content. Since there are too many reviews, efficiently identifying the helpful ones earlier can benefit both consumers and companies. Consumers can read only the helpful opinions from helpful reviews before they purchase a product, while companies can acquire the true reasons a product is liked or hated. A system is built to assess the difficulty of the problem. The experimental results show that helpful reviews can be distinguished from unhelpful ones with high precision.

Keywords: Helpful Opinion Mining, Online Consumer Review, Online Customer Reivew, Text Quality.

1. Introduction

Online consumer (or customer) review is a very important information source for many potential consumers to decide whether to buy a product or not. Li *et al.* (2011) shows that, compared to an expert product review, "the consumer product review in the online shopping environment will be perceived by consumers to be more credible." This fact makes opinion mining of consumer reviews more interesting since it shows that opinions from other consumers are more appreciated than those from experts. Nevertheless, some reviews are not

[*] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan, R.O.C
E-mail: st9506522@gmail.com; shwu@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.
[+] Department of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: cujing@gmail.com; chen@csie.ncu.edu.tw
[#] Institute for Information Industry, Taiwan
E-mail: eit@iii.org.tw

very helpful, as we can see from the voting results on each consumer review from readers on Amazon.com.

This paper will address an opinion mining problem: how to find the helpful reviews from online consumers' reviews before mining the information from them. This task can benefit both consumers and companies. Consumers can read the opinions from useful reviews before they purchase a product, while companies can acquire the true reasons a product is liked or hated. Both save time from reading meaningless opinions that do not show good reasons. Figure 1 shows a clip image of an Amazon.com customer review. Each review has been labeled with stars by the author and people who found the review helpful and has been labeled with the number of total votes. A three-class classification problem is defined to model this application. A system is designed to find the helpful positive reviews for finding good reasons to buy a product; to find the helpful negative reviews for finding reasons not to buy a product; and to filter out the unhelpful reviews, no matter whether they are positive or negative.

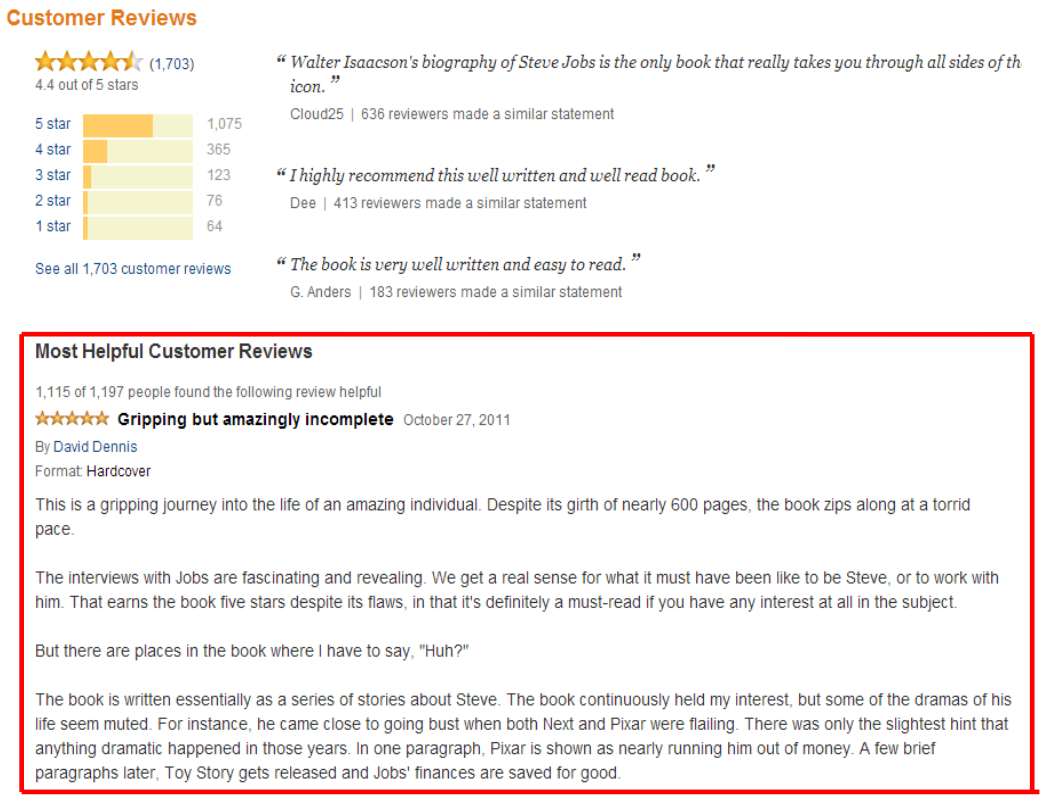


Figure 1. A clip image of an Amazon.com customer review.

The paper is organized as follows. Section 2 describes the related works. Section 3 describes the features that can be used to classify the reviews as helpful or unhelpful. Section

4 describes the data collection of this study. Section 5 reports and discusses the experiment. The final section gives conclusions and future work.

2. Related Works

Early works on opinion mining focused on the polarity of opinion, positive or negative; this kind of opinion mining is called sentiment analysis. Another type of opinion mining focused on finding the detailed information of a product from reviews; this approach is a kind of information extraction (Hu & Liu, 2004). Recent research has focused on assessing the review quality before mining the opinion. Kim *et al*. (2006) explored the use of some semantic features for review helpfulness ranking. They found that some important features of a review, including length, unigrams, and stars, might provide the basis for assessing the helpfulness of reviews. Siersdorfer *et al*. (2010) presented a system that could automatically structure and filter comments for YouTube videos by analyzing dependencies between comments, views, comment ratings, and topic categories. Their method used the SentiWordNet thesaurus, a lexical WordNet-based resource containing sentiment annotations. Moghaddam *et al*. (2011) proposed the Matrix Factorization Model and Tensor Factorization Model to predict of the quality of online reviews, and they evaluated the models on a real-life database from Epinions.com. Lu (2010) exploited contextual information about authors' identities and social networks to improve review quality prediction. Lu's method provided a generic framework to incorporate social context information by adding regularization constraints to the text-based predictor. Xiong and Litman (2011) investigated the utility of incorporating specialized features tailored to peer-review helpfulness. They found that structural features, review unigrams, and meta-data combination were useful in modeling the helpfulness of both peer reviews and product reviews.

3. Classification Features

3.1 Observation

Observation is necessary to find features for the helpful/unhelpful classification. Connors *et al*. (2011) gave a list of common ideas related to helpfulness and unhelpfulness, shown in Table 1, which was collected from 40 students, with each student reading 20 online reviews about a single product and giving comments on the reviews. The study provided 15 reasons people think a consumer review is helpful and 10 reasons why it is unhelpful. These ideas can be considered as features for a classifier. Nevertheless, some of them are difficult to implement and require clear definition. For example, mining comparative sentences from text requires considerable knowledge of the language. (Jindal & Liu, 2006).

Table 1. The 15 reasons that people think a customer review helpful and the 10 reasons they think it to be unhelpful (Connors et al., 2011).

Helpfulness	**Times Mentioned**
Pros and Cons	36
Product Usage Information	30
Detail	24
Good Writing Style	13
Background Knowledge of Product	12
Personal Information about Reviewer	12
Comparisons	10
Layman's Terms	9
Conciseness	8
Lengthy	7
Use of Ratings	7
Authenticity	5
Honesty	5
Miscellaneous	4
Unbiased	4
Accuracy	3
Relevancy	3
Thoroughness	3
Unhelpfulness	**Times Mentioned**
Overly Emotional/Biased	24
Lack of Information	17
Irrelevant Comments	9
Not Enough Detail	6
Poor Writing Style	6
Using Technical Language	6
Low Credibility	5
Problems with Quantitative Rating	5
Too Much Detail	5

3.2 Features

Table 2 lists the features that we implement in this study. Compared with the features used in Kim *et al.* (2006), we add more features, based on the observation of Connors *et al.* (2011), especially the degree of detail. The first three features are common n-grams used between a review and the corresponding product description. We believe that they are effective since a good review should contain more relevant information and use exact terminology. The fourth feature is the length of the review. A very short review cannot give much information, and a long review might give more useful information. The fifth feature is whether or not the review makes a comparison among things. A good review should compare similar products. Our program detects whether the string "compare to/with" or the pattern "ADJ+er than" exists in the review or not, with the help of a list of comparative adjectives. The sixth feature is the degree of detail, which is a combination of length and n-gram. The degree of detail has not been defined well in previous works. Our definition is only a tentative one. We define the degree of detail of a review as:

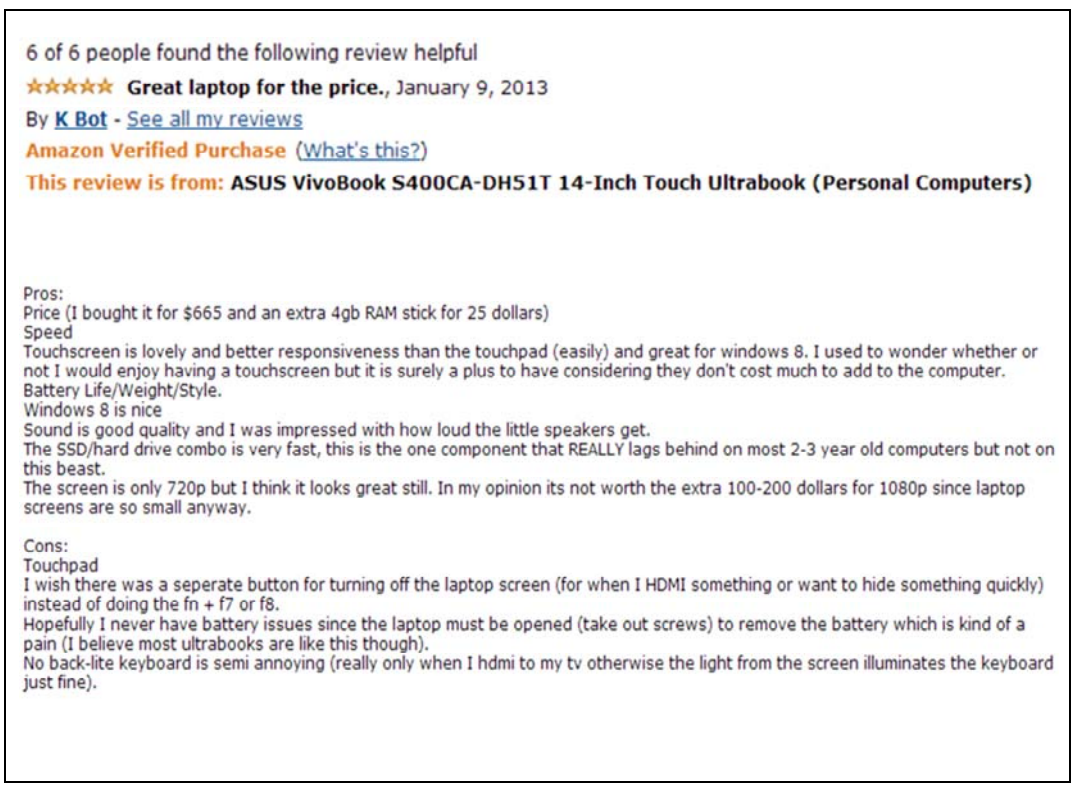
$$log_{10}(\text{Unigram+Bigram+Trigram+Length}) \quad (1)$$

where unigram, bigram, and trigram are the common n-grams between a review and the corresponding product description. Length is the length of the review. The seventh feature is the number of stars given by the review author. The eighth feature is whether the review contains "Pros" and "Cons" or not. Our system detects whether the string "Pros" and "Cons" exist in the review or not.

Table 2. Eight Features used in our system.

Feature	**Description**
Unigram (Product Description)	The number of unigrams used between the review and the corresponding product description
Bigram (Product Description)	The number of bigrams used between the review and the corresponding product description
Trigram (Product Description)	The number of trigrams used between the review and the corresponding product description
Length	The length of a review
Comparisons	The review uses the string "compare to" or "ADJ + er than"
Degree of detail	Defined by formula (1)
Use of Ratings	The "Star" ratings of the review
Pros and Cons	The review contains exact the strings "Pros" and "Cons"

We use an example to show the eight feature values. Consider the review in Figure 2, where the "pros_cons" value is 1, since we can see the author explicitly lists the pros and cons. The "Detail" value is 1.17760, as defined in Formula (1). The "Length" value is 568, which is the number of words in the review. The "Compare" value is 4, because the author really makes a comparison of this product with other products. The "Star" value is 5, since the author gave five stars to the product. The "Unigram" value is 15. The "Bigram" value is 0, since we found no common bigrams between the review and the corresponding product description (not shown here). Hence, the "Trigram" value is also 0.


6 of 6 people found the following review helpful

★★★★★ **Great laptop for the price.**, January 9, 2013

By **K Bot** - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: ASUS VivoBook S400CA-DH51T 14-Inch Touch Ultrabook (Personal Computers)

Pros:
Price (I bought it for $665 and an extra 4gb RAM stick for 25 dollars)
Speed
Touchscreen is lovely and better responsiveness than the touchpad (easily) and great for windows 8. I used to wonder whether or not I would enjoy having a touchscreen but it is surely a plus to have considering they don't cost much to add to the computer.
Battery Life/Weight/Style.
Windows 8 is nice
Sound is good quality and I was impressed with how loud the little speakers get.
The SSD/hard drive combo is very fast, this is the one component that REALLY lags behind on most 2-3 year old computers but not on this beast.
The screen is only 720p but I think it looks great still. In my opinion its not worth the extra 100-200 dollars for 1080p since laptop screens are so small anyway.

Cons:
Touchpad
I wish there was a seperate button for turning off the laptop screen (for when I HDMI something or want to hide something quickly) instead of doing the fn + f7 or f8.
Hopefully I never have battery issues since the laptop must be opened (take out screws) to remove the battery which is kind of a pain (I believe most ultrabooks are like this though).
No back-lite keyboard is semi annoying (really only when I hdmi to my tv otherwise the light from the screen illuminates the keyboard just fine).

Figure 2. Example of review

4. Data Collection

In order to test the idea, we collected online customer reviews manually from Amazon.com in March and April 2013. The reviews were from eight different product domains: Book, Digital Camera, Computer, Food & Drink, Movie, Shoes, Toys, and Cell phone. Without any special selection criterion in each domain, we collected the first available 1000+ reviews with an equal number of reviews of one to five stars. The average length was 80.63 words. The summary of our data collection is listed in Table 3.

Table 3. The summary of our data collection of 8 classifications and 8,690 reviews.

Product	Reviews	Total Reviews Words	Average Length	s.d.
Book	1,065	93,497	87.79	1.8
Digital Camera	1,028	93,404	90.85	2.7
Computer	1,067	83,708	78.45	2.1
Foods & Drink	1,025	71,027	69.29	1.7
Movies	1,097	94,037	88.13	2.5
Shoes	1,000	75,237	75.23	1.6
Toys	1,100	85,196	77.45	1.7
Cell Phone	1,308	101,957	77.88	2.0
Total / Average	8,690	884,964	80.63	2.02

The helpfulness score is given by the readers. As shown in Figure 1, the reviewer labeled the number of stars and other users voted the review as helpful or unhelpful. We take the confidence in being helpful as an index to sort the reviews. Figure 3 shows the distribution of polarity (from 1 to 5 stars) and the helpful/unhelpful confidence, where the y-axis is the confidence score. Note that the confidence score in previous works has been defined as:

$$\text{Confidence}=100\% \times \left(\frac{\#\text{of Think helpful vote}}{\#\text{of Total vote}} \right) \tag{2}$$

Nevertheless, since there are some high confidence reviews with very little support, the reviews might not be very helpful. We discount the confidence of them by redefining the confidence score as the log-support confidence (LSC):

$$\text{LSC}=log_{10}\left[\begin{array}{l} \#\text{of Think Help ful vote} * \\ (\#\text{of Think Help ful vote}/\ \#\text{ of Total vote}) \end{array} \right] \tag{3}$$

Figure 3 shows the data distribution. The positive reviews (with 4 or 5 stars) get higher helpfulness confidence in most product categories. This fact shows that readers think other consumers are credible. The confidence of helpfulness is lower for the negative reviews. The average LSC confidence scores for each product category are listed in Table 4.

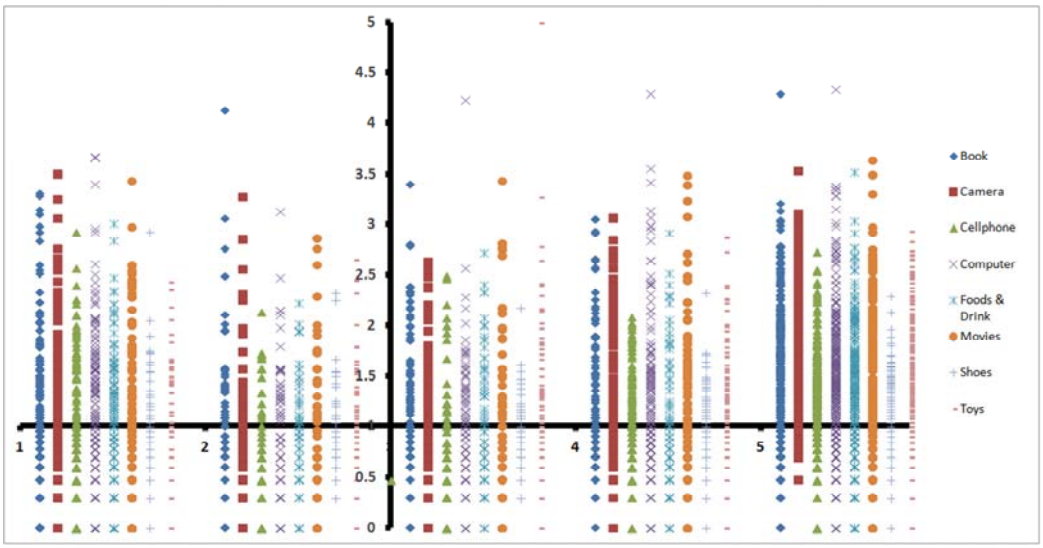


Figure 3. Stars vs. helpfulness distribution of our data collection. The x-axis is the number of stars of customer reviews; the y-axis is the confidence score LSC.

Table 4. The average LSC Confidence scores of the eight product categories.

Product	**Average LSC Confidence score**
Book	1.134
Digital Camera	1.373
Computer	1.140
Foods & Drink	0.932
Movies	1.116
Shoes	0.808
Toys	0.807
Cell Phone	1.005
Total average	1.039

4.1 The Three-class Classification Problem

Instead of finding the correlation between the ranking of helpfulness and the prediction, we define the problem as a three-class classification problem. The three classes are: the helpful

positive reviews, for finding good reasons to buy a product; the helpful negative reviews, for finding reasons not to buy a product; and the unhelpful reviews.

Since there is no distinct boundary between the helpful and the unhelpful and since one purpose of the system is to filter out the most unhelpful reviews, the sizes of the three classes can be adjusted by setting different thresholds. A higher threshold filters out more data. We can control the filtering level by setting different thresholds.

In our experiments, Class 1 includes positive reviews with 4 or 5 stars and the helpfulness confidence higher than the threshold. Class 2 includes negative reviews with 1 to 3 stars and the helpfulness confidence higher than the threshold. Class 3 is the remaining reviews, which are regarded as unhelpful, where the helpfulness confidence is lower than the threshold.

5. Experiments

The goal of the experiment is to test the filter accuracy of the three-class classification problem with different thresholds. We use the libSVM[1] toolkit to build the classifier, based on the features described in Section 2.2.

5.1 Experimental Design

We divide the data into a training set and test set, consisting of 7,690 reviews and 1,000 reviews, respectively. The class distribution of the test data are balanced to one third for each class. The different thresholds tested in our experiment are 1.039, 1.5, and 2.0. The first threshold is the average confidence score in Table 5, which filters out 56.1% of the reviews as unhelpful; the second threshold 1.5, filtering out 79.6%; and the third threshold 2.0, filtering out 91.0%. The numbers of useful (both positive and negative) reviews of each product domain to the three thresholds are listed in Tables 5, 7, and 9. The sizes of classes corresponding to the three thresholds are shown in Tables 6, 8, and 10.

Table 5. Number of reviews over the threshold "1.039"

Product	Reviews
Book	522
Digital Camera	698
Computer	532
Foods & Drink	404
Movies	521

[1] http://www.csie.ntu.edu.tw/~cjlin/lib

Shoes	246
Toys	318
Cell Phone	571
Total Reviews	3,812

Table 6. The size of the three classes with the threshold "1.039"

Classes	Reviews	%
Class 1 : Useful Positive	2,712	31.2%
Class 2 : Useful Negative	1,100	12.7%
Class 3 : Not Useful	4,878	56.1%
Total Reviews	8,690	

Table 7. Number of reviews over the threshold "1.5"

Product	Reviews
Book	270
Digital Camera	354
Computer	254
Foods & Drink	189
Movies	341
Shoes	49
Toys	174
Cell Phone	139
Total Reviews	1,770

Table 8. The size of the three classes with the threshold "1.5"

Classes	Reviews	%
Class 1 : Useful Positive	1,265	14.5%
Class 2 : Useful Negative	505	5.8%

Class 3 : Not Useful	6,920	79.6%
Total Reviews	8,690	

Table 9. Number of reviews over the threshold "2.0"

Product	**Reviews**
Book	129
Digital Camera	202
Computer	104
Foods & Drink	72
Movies	160
Shoes	9
Toys	73
Cell Phone	32
Total Reviews	781

Table 10. The size of the three classes with the threshold "2.0"

Classes	**Reviews**	**%**
Class 1 : Useful Positive	604	6.9%
Class 2 : Useful Negative	177	2.0%
Class 3 : Not Useful	7,910	91.0%
Total Reviews	8,690	

We conducted two experiments. The first one was a 10-fold validation on the training set, and the second one was a test on a separated test set.

5.2 Experimental Results

The average accuracy of the 10-fold cross-validation result of each configuration is shown in Table 11. The 7,690 training data were separated into ten folds, and the system used 90% of the data as the training set and the other 10% as the test set. A SVM classifier was trained in each fold and repeated 10 times. The result shows that, with a higher threshold, 1.5 or 2.0, the accuracy of our system is about 72%.

Table 11. The average accuracy result of each data set in the ten-fold cross-validation

Data set	Average Accuracy
LSC threshold 1.039	60.83%
LSC threshold 1.5	72.72%
LSC threshold 2.0	**72.82%**

In the second experiment, we used the 7,690 reviews as a training set and tested the classification on the 1,000 test set, where the number of tests of each class was balanced to 1/3. Note that the actual class of the test was fixed during the test, which corresponds to a threshold 1.039. The classifier was trained with three different class distributions. The confusion matrix of our system is shown in Tables 12 to 14. The precision and the recall of each class are also shown.

Table 12. The confusion matrix (LSC threshold is over 1.039)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	172	75	46	293	59%
Class 2	80	196	24	300	65%
Class 3	81	62	264	407	65%
Total	333	333	334	1,000	
Recall	52%	59%	79%		

Table 13. The confusion matrix (LSC threshold is over 1.5)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	213	47	28	288	74%
Class 2	42	257	14	313	82%
Class 3	78	29	292	399	73%
Total	333	333	334	1,000	
Recall	64%	77%	87%		

Table 14. The confusion matrix (LSC threshold is over 2.0)

Predicted	**Actual**			**Total**	**Precision**
	Class 1	Class 2	Class 3		
Class 1	203	45	27	275	74%
Class 2	46	263	10	319	82%
Class 3	84	25	297	406	73%
Total	333	333	334	1,000	
Recall	61%	79%	89%		

5.3 Feature Analysis Result

To compare which features are more important in the classifier, we conducted a series experiments with one less feature each time. The results are shown in Table 15. We can find that the "detail" feature is the most important. Second, third, and fourth are length, star, and unigram. Since detail is a hybrid feature, this result suggests that a hybrid feature works better than the combination of individual ones.

Table 15. Accuracy with all-minus-one features

Features	**Accuracy**
All-(Detail)	38.569%
All-(Compare)	52.152%
All-(Pros_cons)	49.727%
All-(Length)	39.594%
All-(Star)	39.342%
All-(Unigram)	42.493%
All-(Bigram)	55.339%
All-(Trigram)	49.469%

5.4 Discussion on the Experimental Result

Table 11 shows that the average accuracy numbers of the three data sets are 60.83%, 72.72%, and 72.82%. We find that setting the threshold to 1.5 is expected to prune 79.6% of data; our system can get 72.72% accuracy on the helpful/unhelpful classification. This is a considerable reduction of human labor to find better mining candidates.

From the confusion matrix in Table 13, we find that choosing the threshold as 1.5 enables our system to classify the three classes with precision 74%, 82%, and 73%; while the

system recall for the three classes are 64%, 77%, and 87%. We also can find a similar result in Table 14, where the threshold is 2.0. The precision is almost the same, and the recall is slightly different.

From Table 15, we can find that the “detail” feature is the most important. Without it, the accuracy drops from 60.83% to 38.57%. Nevertheless, each feature helps the performance, so no one feature can be omitted. This result also suggests that more features might be necessary to attain higher performance.

6. Conclusion and Future Works

The paper reports how a system can find helpful online reviews, and the system is tested on a three-class classification problem. The threshold of helpful/unhelpful reviews can be decided according to the amount of data that the users want to prune. The overall accuracy of the three-class problem is about 73%. Helpful negative reviews can be found with 82% precision and 77% recall. Helpful positive reviews can be found with 74% precision and 64% recall. Unhelpful reviews can be filtered out automatically from the consumer reviews with a high recall rate of about 87% with 73% precision. Considering the original data distribution (only 20% as useful), the system performance is quite high.

Currently, our system is based on features observed by humans in previous works, and we only implement some of them. In the future, we will try to implement more features and attempt to extract features from the training corpus automatically.

Acknowledgements

This study was conducted under the "Online and Offline Integrated Smart Commerce Platform (1/4)" of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs of the Republic of China. This study was partially supported by Research Grant NSC 102-2221-E-324 -034 from the Ministry of Science and Technology.

Reference

Connors, L., Mudambi, S. M., & Schuff, D. (2011). Is it the Review or the Reviewer? A Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness. In *Proceedings of the 2011 Hawaii International Conference on Systems Sciences (HICSS)*, January.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence (AAAI'04)*, Anthony G. Cohn (Ed.). AAAI Press 755-760.

Jindal, N., & Liu, B. (2006). Mining comparative sentences and relations. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2* (AAAI'06), Anthony Cohn (Ed.), Vol. 2. AAAI Press 1331-1336.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M**. (**2006). Automatically Assessing Review Helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 423-430.

Li, M., Huang, L., Tan, C., & Wei, K. (2011) Assessing The Helpfulness Of Online Product Review: A Progressive Experimental Approach. In *Proceedings of PACIS*.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. **(**2010). Exploiting Social Context for Review Quality Prediction. In *Proceedings of the 19th international conference on World wide web*, 691-700.

Moghaddam, S., Jamali, M., & Ester, M. (2010). Review Recommendation: Personalized Prediction of the Quality of Online Reviews. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2249-2252.

Mudambi, S. M., & Schuff, D. (2010). What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185-200.

Siersdorfer, S., Chelaru, S., & San Pedro, J. (2010). How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*, 891-900.

Xiong, W., & Litman, D. (2011). Automatically Predicting Peer-Review Helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 502-507.

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 2, June 2014, pp. 33-52
© The Association for Computational Linguistics and Chinese Language Processing

Resolving the Representational Problems of Polarity and Interaction between Process and State Verbs

Shu-Ling Huang[*], Yu-Ming Hsieh[*+], Su-Chu Lin[*], and Keh-Jiann Chen[*]

Abstract

Event classification is one of the crucial tasks in lexical semantic representation. Traditionally, researchers have regarded process and state as two top-level events and discriminated between them by semantic and syntactic characteristics. In this paper, we add cause-result relativity as an auxiliary criterion to discriminate between process and state by structuring about 40,000 Chinese verbs to the two correspondent event hierarchies in E-HowNet. All verbs are classified according to their semantic similarity with the corresponding conceptual types of ontology. As a result, we discover deficiencies of the dichotomy approach and point out that any discrete event classification system is insufficient to make a clear-cut classification for synonyms with slightly different semantic focuses. We then propose a solution to remedy the deficiencies of the dichotomy approach. For the process or state type mismatched verbs, their inherited semantic properties will be adjusted according to their PoS and semantic expressions to preserve their true semantic and syntactic information. Furthermore, cause-result relations will be linked between corresponding processes and states to bridge the gaps of the dichotomy approach.

Keywords: Event Classification, Process and State, Lexical Representation, Cause-result Relativity between Verbs.

1. Introduction

Clarifying the nature of verb classes is a crucial issue in lexical semantic research, being of great interest to both theoretical and computational linguistics. Many classification and representation theories have been presented already, including the widely cited theories

[*] Institute of Information science, Academia Sinica, Taipei, Taiwan
[+] Department of Computer Science, National Tsing-Hua University, Taiwan
E-mail: {josieh,morris,jess,kchen}@iis.sinica.edu.tw

proposed by Vendler (1967), Dowty (1979), Bach (1986), Parsons (1990), Levin (1993), Pustejovsky (1995), and Rosen (2003). Additionally, several online verb classification systems, such as WordNet (Fellbaum, 1998), VerbNet (Kipper-Schuler, 2006), FrameNet (Fillmore *et al*., 2003), and Levin's verb classification also are available. Each approach views events from a different perspective, and each approach clarifies a different part of the overall problem of understanding the linguistic representation of events. Overall, they can be divided into two main schools, one is semantic classification, such as Vendler's approach, and the other is syntactic classification, such as Levin's approach.

Since different event classifications pinpoint the basic features of events that need to be represented, we need to clarify the goal we want to achieve before adopting or proposing an event classification. In this paper, we aim to achieve a better lexical semantic representation framework for E-HowNet (Chen *et al*., 2003), and we adopt the typologies of process and state as the two top-level event types. Since verbs may express different aspects or viewpoints of conceptual events, however, it is difficult to make a clear-cut difference between process and state verbs in some cases. Verb-result compounds, such as 購妥 gou-tuo 'to complete procurement,' are obvious examples of being either pure process or state. Furthermore, semantic interactions of the verbs also need to be clarified. Consider, for example, the synonymous words (strictly speaking, near synonyms and hyponyms) of 記得 ji-de 'remember' in Mandarin Chinese: (a) 想起 xiang-qi 'call to mind,' 記取 ji-qu 'keep in mind,' 背起來 bei-qi-lai 'memorize,' (b) 念念不忘 nian-nian-bu-wang 'memorable,' and 刻骨銘心 ke-gu-ming-xin 'be remembered with deep gratitude'. Although these words are near synonyms, their senses shift slightly according to different semantic focuses, often resulting in different grammatical behavior. If we classify Group (a) as a process type, and Group (b) as a state type by their fine-grained semantic focuses, we may lose the important information that they are actually near synonyms and have the core sense of 記得 ji-de 'remember'. Therefore, in order to design a better semantic and syntactic representational framework for verbs, we try to clarify the polarity and interaction between process and state.

The remainder of this article is organized as follows. In the next section, we begin with a review of past research. Section 3 clarifies the polarity between process and state before addressing difficulties of the dichotomy approach. In Section 4, we describe the interaction between process and state, propose solutions to overcome the difficulties mentioned in the previous section, and discuss other event relations that should be represented in analogy with process state dichotomy. Finally, we conclude our findings and possible future research in Section 5.

2. Background

Over 2300 years ago, Aristotle (in Jonathan Barnes eds., 1984) proposed the first event-based classification of verbs. His main insight was the distinction between states and events (called 'processes' in this paper). Since the late 1960s, a large number of event classifications, variously based on temporal criteria (such as tense, aspect, time point, and time interval), syntactic behavior (such as transitivity, object case, and event structure), or event arguments (such as thematic role mapping, agent type, and verb valence) have been suggested and have aroused heated discussion. These representations can be roughly divided into the two main schools of semantic classification and syntactic classification. In the following discussion, we take Vendler and Levin as representatives for the two schools and we find that both schools treat process and state as two clearly different event types.

2.1 Vendler's Classification

Vendler's classification (1967) is the most influential and representative system in terms of the semantic classification approach. He classified verbs into four categories "to describe the most common time schemata implied by the use of English verbs" (pp. 98-99). The four categories are given in (1).

(1) a. *States*: non-actions that hold for some period of time but lack continuous tenses.

b. *Activities*: events that go on for a time, but do not necessarily terminate at any given point.

c. *Accomplishments*: events that proceed toward a logically necessary terminus.

d. *Achievements*: events that occur at a single moment; therefore, they lack continuous (progressive) tenses.

Distinctly, states denote a non-action condition and are irrelevant to temporal properties, while the other three denote an event process or a time point in an event process. Vendler's successors, such as Verkuyl (1993), Carlson (1981), Moens (1987), and Hoeksema (1983), extended this discussion without changing Vendler's basic framework. According to Rosen (2003), the successors all pointed out that state and process are two major event types. Ter Meulen (1983; 1995) thus suggested a redefinition of Vendler's classes. She defined states as having no internal structure or change, while events, i.e., the processes dealt with in our paper and composing Vendler's other three event types, are defined on the basis of their parts.

2.2 Levin's Classification

Levin (1993) believes that identifying verbs with similar syntactic behavior provides an effective means of distinguishing semantically coherent verb classes. She proposed a coarse-grained classification for verbs based on two observations: the first is that many result verbs lexicalize results that are conventionally associated with particular manners, and vice-versa, many manner verbs lexicalize manners that are conventionally associated with particular results. The examples she gave are listed in (2):

(2) The pervasiveness of the dichotomy (Levin, 2011)

	Manner verbs	**vs.**	**Result verbs**
Verbs of damaging:	*hit*	vs.	*break*
Verbs of putting—2-dim	*smear*	vs.	*cover*
Verbs of putting—3-dim	*pour*	vs.	*fill*
Verbs of removal	*shovel*	vs.	*empty*
Verbs of combining	*shake*	vs.	*combine*
Verbs of killing	*stab*	vs.	*kill*

Levin argued the origin of the dichotomy arises from a lexicalization constraint that restricts the manner and result meaning components to fit in a complementary distribution: a verb lexicalizes only one type and those components of a verb's meaning are specified and entailed in all uses of the verb, regardless of context. Further, not only do manner and result verbs differ systematically in meaning, but they differ in their argument realization options (Rappaport & Levin, 1998; 2005). For example, result verbs show a causative alternation, but manner verbs do not, as shown in Example (3); and, manner verbs show considerably more and different argument realization options than result verbs (Rappaport & Levin, 1998), such as those described in (4).

(3) a. Kim broke the window./The window broke.

b. Kim wiped the window./*The window wiped.

(4) a. Terry wiped. (activity)
b. Terry wiped the table. (activity)
c. Terry wiped the crumbs off the table. (removing)
d. Terry wiped the crumbs into the sink. (putting)
e. Terry wiped the slate clean. (change of state)
f. Terry wiped the crumbs into a pile. (creation)

Levin's manner verb and result verb dichotomy characterizes semantic and syntactic interactions between verbs. Specifically, this syntactic dichotomy is caused by the semantic characteristics of the language. We consider a similar semantic relation of cause-result between process verbs and state verbs to show the dichotomy and interactions between them. In fact, Levin's result verbs are verb-result compounds in Chinese, such as the process verb 打破 da-po 'break' in our classification. We regard results of processes to be result states, such as 破裂 po-lie 'broken'. Hence, the aforementioned verb pairs, such as *stab* and *kill* in (2), are both process verbs. By our notion of process and state dichotomy, wounded and die are result states of *stab* and *kill*, respectively.

2.3 E-HowNet's Classification

E-HowNet (Chen *et al*., 2005) is a frame-based entity-relation model that constructs events, objects, and relations in a hierarchically-structured ontology. By following the conventional event classification theories, verbs are partitioned into process and state first, which is a higher priority dichotomous classification criterion than the syntactic classification in E-HowNet, since E-HowNet primarily is a semantic classification system. Furthermore, semantic classification is more intuitive and more in line with the general view of the real world. Based on this criterion, the top-level E-HowNet ontology is established, as depicted in Figure 1, and a snapshot of E-HowNet is given in Appendix A.

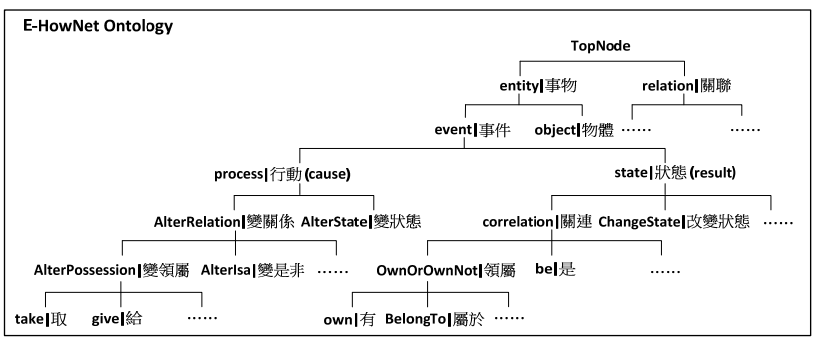


Figure 1. The Architecture of E-HowNet

3. The Polarity and Interaction between Process and State

Process and state have long been treated as two top classes of events. Semantically, their distinctions are evident and intuitive, such as the difference between the process verb 取悅 qu-yue ‘please’ and the state verb 喜悅 xi-yue ‘joyful’. With respect to syntax, process and state verbs also have their own individual characteristics; for example, 取悅 qu-yue ‘please’ must have a patient object but 喜悅 xi-yue ‘joyful’ does not. Differentiating them is considered obvious in theoretical and practical linguistic research areas. Nevertheless, from the perspective of a fine-grained lexical analysis, researchers have also found that it is difficult to make clear-cut differences between process and state. Take the following as examples. The state verb 生氣 sheng-qi ‘angry’ may accept an object goal in Mandarin and is difficult to differentiate from the process verb 發脾氣 fa-pi-qi ‘get angry’ in semantics. In this paper, we do not aim to strictly partition 生氣 sheng-qi ‘angry’ and 發脾氣 fa-pi-qi ‘get angry’ into state and process type. Instead, our objective is to discriminate processes from states with an emphasis on why we encounter difficulties of discriminating them and what better representations may preserve as much semantic and syntactic information as possible. For example, the verb 遇害 yu-hai ‘be murdered’ can be either classified as a process of *kill* or a state of *die*, with neither classification being absolute. A better solution might be that, even if the verb is misclassified into either type, we can still recognize that the experiencer of 遇害 yu-hai ‘be murdered’ is killed and dead. In this section, we emphasize the general distinction between process and state. Then, in the next section, we introduce several approaches we adopted upon encountering difficulties of process-state dichotomy.

The differentiating characteristics between process and state verbs, other than semantic differences, are not obvious. Summarizing the previously mentioned theories in Section 2, the polarities between process and state can be generalized as follows.

(5) The polarities and interactions between process and state

Processes: cause of states, dynamism (*i.e.*, relevant to temporal properties), object domination

States: result of processes, stasis (*i.e.*, irrelevant to temporal properties), object modification

The polarity of dynamism and stasis is a semantic-based distinction, whereas the domination of objects or their modification is a syntax-based distinction. They are both common but coarse-grained event classification criteria, and most verbs can be distinguished by these coarse-grained classification criteria. Nevertheless, some verbs, like 發脾氣 fa-pi-qi ‘get angry’ and 遇害 yu-hai ‘be murdered,’ are not classified easily. In our study, we propose

an interaction between cause and result as an auxiliary criterion, which asserts that *processes* are the cause of states and they denote an event process or a time point on an event process. On the other hand, *states* are the result of processes and they denote a non-action condition and are irrelevant to temporal properties, *i.e*., they have no internal structure or change. Although it would appear that cause-result is a natural differentiation criterion between processes and states, it may not be a one-to-one relation and some verb types may not have obvious cause-result counterparts. For instance, the concept of causative process {earn|賺} may achieve several resultant states, such as {obtain|得到} and {rich|富}, although the process of {swim|游} does not have an obvious result state. Nonetheless, if we can use the characteristics of (5) to differentiate all verbs into process and state types, it may help us achieve the first step towards a lexical semantic classification for verbs. We then use semantic expressions, part-of-speech (PoS) features,[1] and relational links, such as cause-result relationship between process types and state types, to make a better lexical semantic representation. Regarding the verb type classification, the following questions may be raised. Is the process-state dichotomy approach feasible? How are the verbs denoting complex event structures, such as verb-result compounds, classified? Is it true that all states have causing processes and all processes have resulting states? The following observations will provide the answers to these questions.

3.1 Observations and Difficulties of the Process-State Dichotomy in E-HowNet

In order to develop the lexical semantic representation system E-HowNet, we classified all Chinese verbs into a process and state type hierarchy, as illustrated in Figure 1. We use the characteristics (5) of dynamism and stasis as semantic-based distinctions, the domination and modification of objects as syntax-based supporting criteria, and the cause-result relation as a complementary criterion to distinguish process from state. It is interesting that, with the exception of general acts, almost all top-level Chinese verb types, whether of process or state types, necessarily have their cause-result counterpart. Nevertheless, for the fine-grained lower level types or lexical level verbs, there are three different cases of lexical realizations of cause-result dichotomy, which are listed in the following.

Case 1: Process types have result states and *vice-versa*. An example of cause-result mapping between process and state is given in (6).

[1] For simplicity, in this paper, we only tag the top-level PoSs, *i.e*. active PoS and stative PoS, which are adopted from the classification of CKIP group (1993).

(6) Causative process type {brighten|使亮}: *e.g.*, 磨光 mo-guang ‘burnish’, 擦亮 ca-liang ‘polish’ ←→
Resultant state type {bright|明}: *e.g.*, 水亮 shui-liang ‘bright as water’, 光燦 guang-can ‘shining’

For this case, the process and state are two different types and can be differentiated by the fundamental differences between dynamic and static types or by the cause-result relation. Nevertheless, lexemes may shift their senses due to different compounding, resulting in a classification dilemma of semantic similarity first or dichotomy of process and state first. As was mentioned in the above example, the causative process type {kill|殺害}, *e.g.*, 弔死 diao-si ‘hang by the neck,’ has a resultant state type {die|死}, *e.g.*, 往生 wang-sheng ‘pass away’. Then, how about the result-state verb 遇害 yu-hai ‘be murdered’? Should we classify 遇害 yu-hai ‘be murdered’ as a process type {kill|殺害} or as a state type {die|死}? The verb 遇害 yu-hai ‘be murdered’ seems to be the resultant state {die|死} in terms of stativity, but from the perspective of a semantic focus, it is more akin to a causative process {kill|殺害}. This classification difficulty always occurs when we analyze verbs denoting different aspect situations, such as passive or achieved aspects. As a result, near synonyms of the same event type could be separated for denoting different aspectual situations.

In terms of the E-HowNet ontology, the cause-result matching between processes and states almost reaches 100% respecting hypernymy concepts exemplified by corresponding lexical pairs, as shown in Figure 2. Nevertheless, at the hyponym or lexical level, we found that the correspondent rate was not as high as in top-level concepts. This results in Case 2 below.

Case 2: Process types do not have nodes of result states nor do state types have nodes of causing processes in the E-HowNet ontology, which means the result states or causal processes are either vague or they are not lexicalized common concepts. (7), (8) are typical examples.

(7) The causative process type {punish|處罰}, such as 行刑 xing-xing ‘execute’ or 處決 chu-jue ‘put to death,’ have corresponding aspectual resultant states, such as 受刑 shou-xing ‘be put to torture’ and 伏法 fu-fa ‘be executed,’ but no lexicalized concept in common to denote *being punished* or *being tortured* in Chinese. Therefore, there is no proper node of state type to which the above two stative verbs belong in E-HowNet.

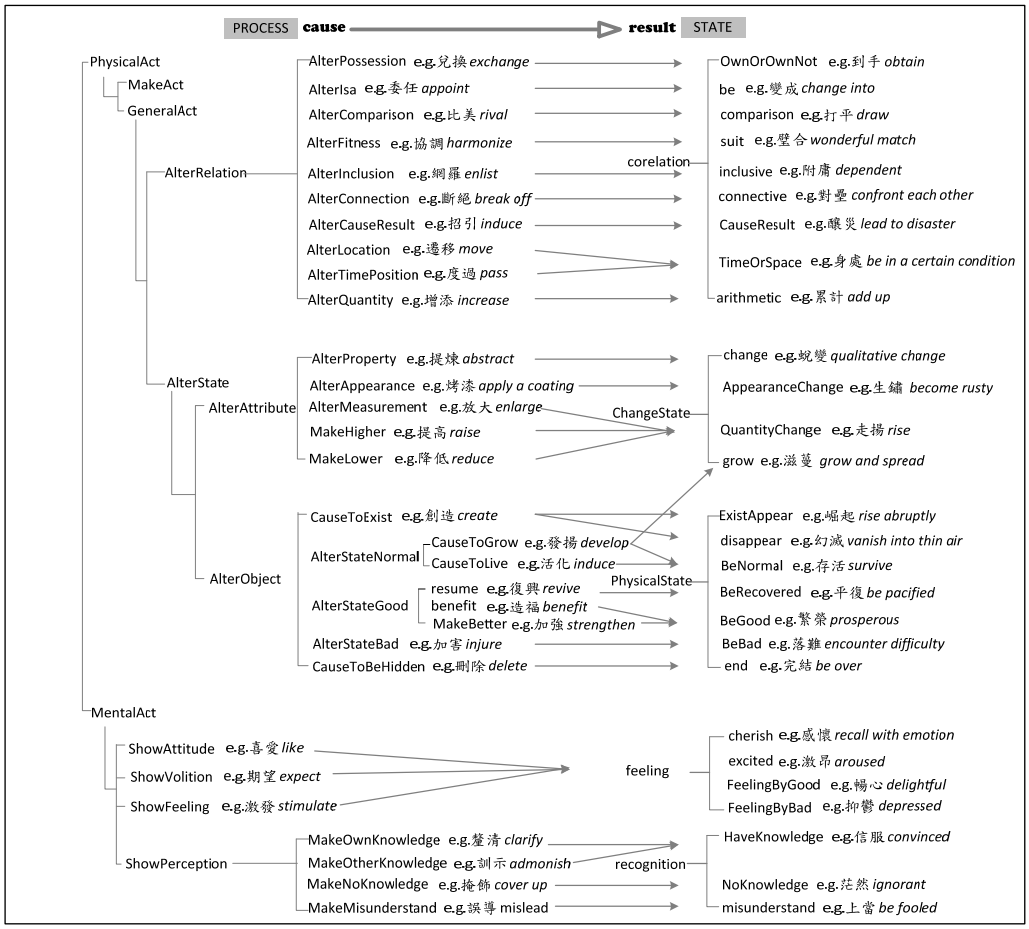


Figure 2. The Matching between Processes and Result States in E-HowNet

(8) There is no lexicalized concept in common to denote causative processes, such as 板起(臉) ban-qu-(lian) 'put on a stern expression' and 正色 zheng-se 'with a stern countenance' in Chinese, which are the cause of the resultant state type {austere|冷峻}, *e.g.*, 凝重 ning-zhong 'serious,' 不苟言笑 bu-gou-yan-xiao 'serious in speech and manner'. That is, there is no proper node of the process type to which the above two process verbs 板起臉 ban-qi-lian 'put on a stern expression' and 正色 zheng-se 'with a stern countenance' belong.

For lexemes of Case 2, the characteristics of process and state of (5) can still differentiate the lexemes on the process and state types, but there are no actual corresponding conceptual nodes in the ontology. This means that some stative verbs must be attached to the process type node and some process verbs should be attached to stative type nodes in the ontology for the

sake of keeping reasonable semantic consistency.

Case 3: Some processes and respective states co-exist concurrently and are not in the cause-result temporal sequence. We call such a concurrent process and state a dual process-state. There are 22 dual process-state type primitives in the E-HowNet ontology (refer to Appendix B), with Example (9) describing one of them.

> (9) The dual process-state {living|生活} includes: (a) 求生 qiu-sheng ‘seek to survive,’ 度日 du-ri ‘subsist,’ and (b) 生存 sheng-cun ‘exist,’ 在世 zai-shi ‘be living,’ 一息尚存 yi-xi-shang-cun ‘be still alive’. The semantic focus of (a) indicates a process of *making a living* or *to live*, while (b) indicates the state of *being alive* or *be living*. The two types of process and state coexist and they do not have cause-result relation.

For the dual process-state type, we encounter the similar dilemma of the previous two cases. If we choose the bipartite process and state approach, near synonyms will belong to two nodes far apart in the ontology. If we adopt the approach of a unified conceptual node for each dual process-state type, the result will be the same problem as in Case 2, *i.e.*, stative verbs and process verbs are of the same type.

Furthermore, in Mandarin Chinese we have many verb-result compounds (VR), such as 累病 lei-bing ‘sick from overwork,’ 驚退 jing-tui ‘frighten off,’ and 購妥 gou-tuo ‘to complete procurement’. Since the causative process and resultant state are contained in the same verb, how should we classify them?

4. Knowledge Representation for Process and State Verbs

The difficulties of the dichotomous approach are caused by the semantic interaction between state and process. We thus propose the classification criterion (5) and a representational scheme according to the above observations, and we try to solve the difficulties without changing the framework of the dichotomy structure. The idea is that all verbs are classified into the most similar conceptual types, according to their respective sense. The process or state type mismatched verbs will have their types adjusted by their PoS or semantic expressions. Such an approach is functional, like using the feature of ‘*don’t fly*’ to adjust the flying property for penguins as bird type and still maintaining the inherent properties. Furthermore, cause-result relations will be established between corresponding processes and states to bridge the gaps of the dichotomy approach.

4.1 Lexical Semantic Representation for Process and Stative Verbs

For the Case 1 verbs, every process has a corresponding result state, and every state has a corresponding causal processes. For synonymous verbs with a process and state dichotomy, each verb is placed under its corresponding conceptual node. In addition, the cause-result relationship will be established between corresponding process types and state types, as exemplified in Figures 2 and 4. In real implementation, there are 310 corresponding cause-result pairs established. Nevertheless, from a practical point of view, all semantic representation systems are discrete systems. Given that they use a limited number of primitive concepts to express complex concepts, the result is that some words are forced to be classified to the most similar concept node but with a mismatched major semantic type, such as 遇害 yu-hai 'be murdered' possibly being classified as the process type {kill|殺害} instead of the state type {die|死}. We will resolve such problems by following the same method for Case 2 verbs.

As shown in the observation of Case 2, some of the cause-result corresponding concepts are vague and some are not lexicalized, neither of which occurred as conceptual nodes in the ontology. As a result, for verbs whose potential hypernyms are missing, we will classify these verbs to their cause-result counterpart conceptual nodes instead. After that, we use the part-of-speech to recover the correct semantic type of state or process, as exemplified in (10).

(10) Causative process: {FondOf|喜歡} ←→ there is no corresponding resultant state

The typical examples of semantic type of {FondOf|喜歡} are 看中 kan-zhong 'take fancy to,' 喜愛 xi-ai 'love,' 酷愛 ku-ai 'ardently love,' and 熱衷 re-zhong 'be addicted to'. They are tagged with an active PoS. The verbs 癡情 chi-qing 'be infatuated' and 興致盎然 xing-zhi-ang-ran 'full of interest,' however, are stative verbs, but there is no lexicalized state primitive to place these verbs. Hence, they are classified to the most similar hypernym concept node, *i.e.*, {FondOf|喜歡}.

With part-of-speech tags, we have no problem discriminating state verbs that are attached to a process primitive. In fact, we can define state verbs in {result({process})} format; or process verbs in {cause({state})} format in order to make both semantic distinctions and link relations.

(11) 看中 kan-zhong 'take fancy to,' 喜愛 xi-ai 'love,' 酷愛 ku-ai 'ardently love,' 熱衷 re-zhong 'be addicted to' are defined as {FondOf|喜歡};

癡情 chi-qing 'be infatuated,' 興致盎然 xing-zhi-ang-ran 'full of interest' are defined as {result({FondOf|喜歡})} and have stative PoS.

Moreover, fine-grained part-of-speech tags also provide syntactic information for each verb; this solves the difficulty of Case 2 and effectively expresses fine-grained semantic and syntactic distinctions for near-synonyms.

4.2 Lexical Representation for Dual Process-State Verbs

For Case 3 dual process-state verbs, the bipartite classification for state and process no longer exists for two reasons. First, it is difficult to make a distinction between process and state for the dual types. Second, state and process are just two different viewpoints of the same events. A single dual process-state conceptual type may contain both process and stative verbs of the same event type but different viewpoints. We use part-of-speech tags to tell the difference between semantic focus and the syntactic behavior of each verb. In addition, the dual process-state type also indicates that the process and state coexist at the same event duration. For instance, both verbs 度日 du-ri ‘subsist’ and 在世 zai-shi ‘be living’ are belong to the same conceptual type of {living|生活}, but have the active PoS and stative PoS, respectively.

4.3 Lexical Semantic Representation for Verb-Result Compounds

In addition to the verbs belonging to Cases 1-3, we also wanted to address the solution for classification difficulty of VR compounds. A VR compound may be a composition of a process event followed by an event of result state, such as 打破 da-po ‘break’. The verb in (12.a) is more process-like, but the same verb in (12.b) is more state-like. It is a dilemma to classify the verb into either process or state.

(12.a) 張三打破花瓶 Zhang-san da-po hua-ping ‘Zhang San broke the vase.’

(12.b) 花瓶打破了 hua-ping da-po le ‘The vase is broken.’

Nevertheless, if its semantic expression provides sufficient information to clarify the accurate word meaning and relation between V1 and V2, as well as a suitable PoS classification, there is no difference in the event type where it was classified. Although it is controversial to recognize the semantic focus of these verbs, *i.e.*, to determine whether they are more state-like or more process-like, it is not an important issue in making a semantic and syntactic distinction in lexical representation. We built explicit links of cause-result relations between sub-events in the LESRE framework of E-HowNet (Chen *et al.*, 2013). For example, the sense of VR verb 驚退 jing-tui ‘frighten off’ is expressed as in (13). We also encoded the co-indexed arguments for all related event pairs, *i.e*. the patient of {frighten|嚇唬} is the agent of {leave|離開} in (13).

(13) 驚退 jing-tui ‘frighten off’ def:{frighten|嚇唬: patient={x}, result={leave|離開:agent={x}}}

In order to maintain fluency and legibility of the article, the PoS features and semantic expressions of all of our examples are listed in Appendix C.

4.4 Lexical Representation for Linking Semantic Related Concepts

The connection of semantic relations between concepts is almost as important as the classification of events in a hierarchical framework. Since the construction of a hierarchical taxonomy is primarily by hypernym-hyponym relations, many semantically related concepts may be far apart in the taxonomy, such as cause process and result state. Therefore, we must also take semantic connection and fine-grained lexical representation into account when classifying events into groups. There are 11 types of explicit relations in HowNet identified by Dong & Dong (2006), also adopted by E-HowNet, to link the related concepts. They are synonym, synclass, antonym, converse, hypernym, hyponym, cognate role-frame, part-to-whole, value-to-attribute, attribute-to-host, and semantic-roles-to-event. In fact, the supplement linking relations between two semantically related but hierarchically far apart concepts in E-HowNet are more than the aforementioned relations. We use E-HowNet expressions to express semantic equivalence and link the two concepts. For instance, for the related concepts of *able* and *ability*, *bad* and *good* their relations can be expressed as ability=degree({able|能}) and {bad|壞}=not({nice|良好}). In this paper, we find processes and states exist with a cause-result relation that can be expressed in a function form as result({act|行動})={state|狀態} or cause({state|狀態})={act|行動}, such as result({CauseToAppear|顯現})={appear|出現} or cause({appear|出現})={CauseToAppear|顯現}. In the future, important relations regarding entailment and precondition between two concepts will be further explored.

5. Discussion and Conclusion

Levin (2010) pointed out that different studies support positing verb classes of varying grain-sizes, including (a) coarse-grained classification discriminating *manner verb*, *result verb*; (b) medium-grained classification discriminating *motion verbs*, *speaking verbs*, *etc*., with Fillmore’s verb classification being regarded as a representative of medium-grained classification; and (c) fine-grained classification discriminating *run*, which lexicalizes a manner of motion that causes directed displacement towards a goal. Nevertheless, while these classifications are different in grain-size, they are not contradictory for the classification criteria.

In E-HowNet, we carry this viewpoint through the whole construction by first classifying events into causative processes and their corresponding resultant states, *i.e.*, the top two levels of events we mainly discussed in this paper. We then further subdivided these into more than 1200 generic events (*i.e.*, primitives) into a semantic hierarchy framework as a medium-grained event classification. Finally, the near synonyms were attached to each primitive and discriminated by fine-grained features that were integrated in the lexical event structure representation of E-HowNet (abbreviated as LESRE; see Chen *et al.*, 2013). The content and formation of LESRE is shown in Figure 3.

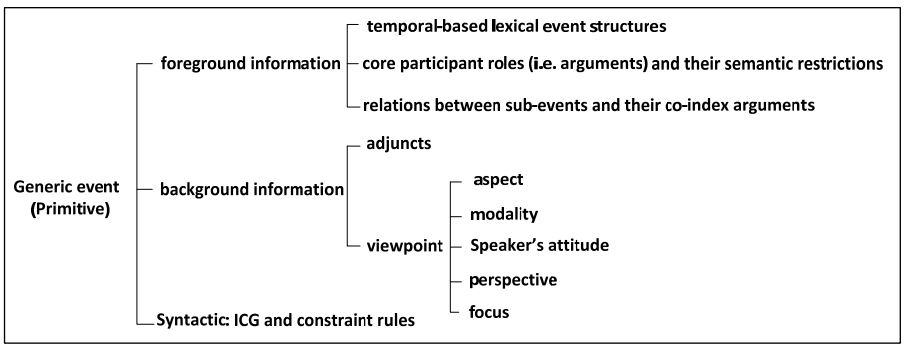


Figure 3. The Content and Formation of LESRE

We believe the varying grain-size classifications provide different semantic and syntactic realization options, such as the coarse-grained classification in which process verbs show considerably more and different argument options than state verbs; further, the idiosyncrasy of each grain-size classification, as well as their interaction, will provide us with advanced knowledge in lexical representation. We will, therefore, continue to complete the LESRE theory in the near future, with the ultimate objective being to establish a completed event classification system that can be applied to both theoretical and computational linguistics. The sketch of different grain-sized event classifications in the E-HowNet construction is shown in Figure 4.

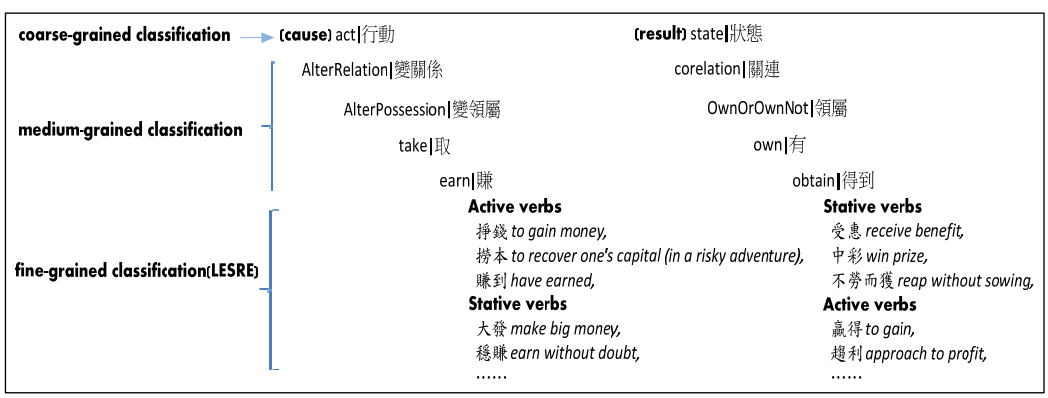


Figure 4. Three Grain-sizes of Event Classification in E-HowNet Construction

Event classification is one of the crucial tasks in lexical semantic representation. Traditionally, researchers have regarded process and state as the two top level events and defined them by counter temporal features and syntactic rules. In this paper, we added cause-result relativity as an auxiliary criterion to discriminate between process and state, and structured about 40,000 Chinese verbs to the two correspondent event classes. All verbs were classified according to their semantic similarity with the conceptual types of the ontology. The process or state type mismatched verbs would have their types adjusted by their PoS or semantic expressions. Furthermore cause-result relations would be linked between corresponding processes and states to bridge the gaps of the dichotomy approach.

We not only aimed to claim the deficiency of dichotomy approach, but also to point out that any discrete event classification system is insufficient to make a clear-cut classification for all verbs, such as synonyms with slightly different semantic focuses. Although misclassification maybe unavoidable, under our framework of event classification, we proposed the remedy of using fine-grained feature expressions to recover erroneous information inherited from the mismatched classification and differentiated the fine-grained semantic differences for near synonyms. The E-HowNet feature expression system is an incremental system, *i.e*., fine-grain features can be added gradually without side effects. Currently, we have resolved the medium-grained classification among 1200 generic event types for about 40,000 Chinese verbs. In the future, we will improve their fine-grained feature expressions to achieve better lexical semantic and syntactic representations.

Acknowledgments

This research was supported in part by National Science Council under Grant NSC 99-2221-E-001-014-MY3.

Reference

Aristotle, (1984), Metaphysics. In Jonathan Barnes (eds.), The Complete Works of Aristotle: The Revised Oxford Translation, Volume 2. Princeton, NJ: Princeton University Press.

Chen, K.-J., *et al*., (2003), *E-HowNet*, CKIP Group, Academia Sinica, http://ehownet.iis.sinica.edu.tw /ehownet.php.

Chen, K.-J., Huang, S.-L., Shih, Y.-Y., & Chen, Y.-J. (2005). Extended-HowNet- A Representational Framework for Concepts. *OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop*, Jeju Island, South Korea.

Chen, K.-J., Huang, S.-L., & Lin, S. (2013), *The Lexical Event Structure Representation of E-HowNet*, technical report, CKIP Group, Academia Sinica. (In preparation).

CKIP group. (1993). *Technical report no.93-05*, Academia Sinica.

Dong, Z., & Dong, Q. (2006). *HowNet and the Computation of Meaning*. World Scientific Publishing Co. Pte. Ltd.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press. http://www-personal.umich.edu/~jlawler/levin.verbs.

Levin, B., (2010). What is the best grain-size for defining verb classes? *Conference on Word Classes: Nature, Typology, Computational Representations, Second TRIPLE International Conference*, Università Roma Tre, Rome, pp. 24-26.

Levin, B., (2011). Verb Classes Within and Across Languages. *Valency Classes Conference, Max Planck Institute for Evolutionary Anthropology*, Leipzig, pp. 14-17.

Meulen, A. G.B. ter, (1983). *The Representation of Time in Natural Language*. In A.G.B. ter Meulen (ed.), Studies in Modaltheoretic Semantics. Dordrecht: Foris.

Meulen, A. G.B. ter, (1995). *Representing Time in Natural Language: The Dynamic Interpretation of Tense and Aspect*. Cambridge, MA: MIT Press.

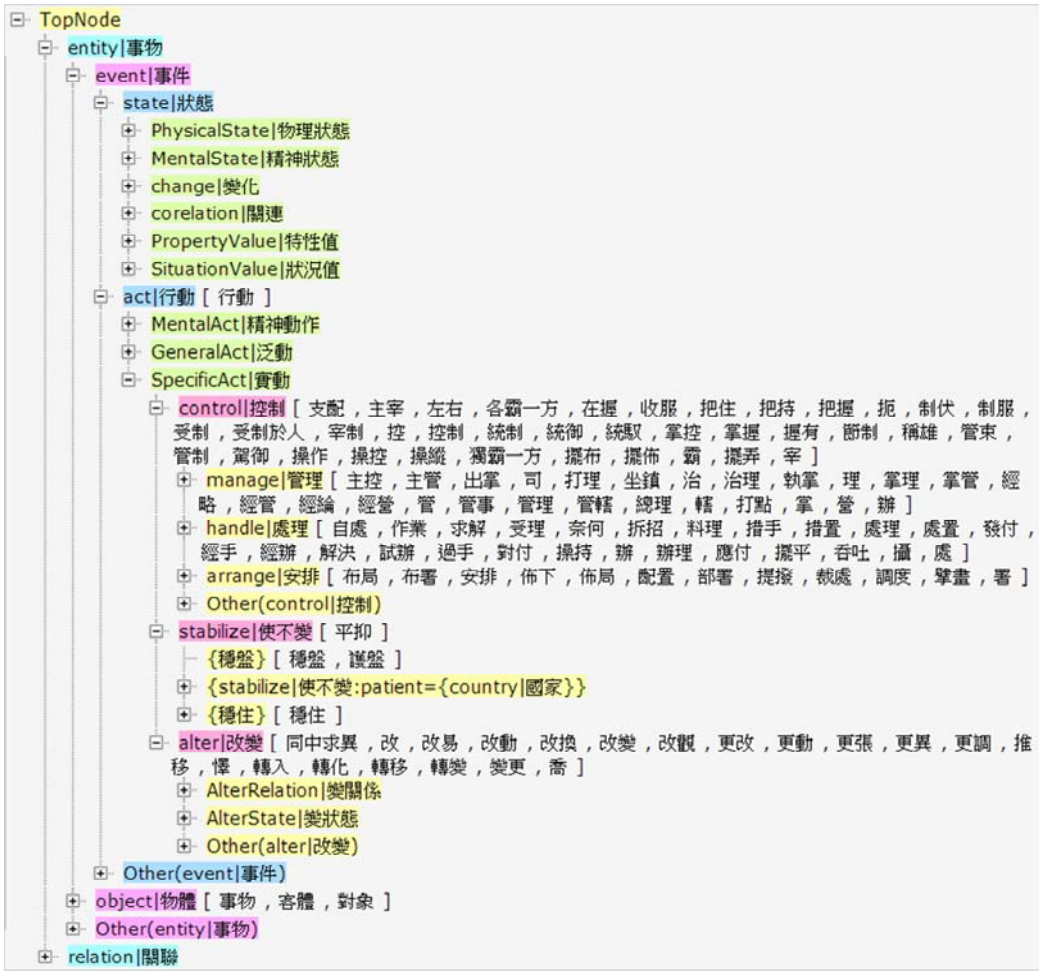
Rappaport, H. M., & Levin, B. (1998). *Building Verb Meanings*. in M. Butt and W. Geuder, eds., The Projection of Arguments, CSLI Publications, Stanford, CA, pp. 97-134.

Rappaport, H. M., & Levin, B. (2005). Change of State Verbs: Implications for Theories of Argument Projection, in N. Erteschik-Shir and T. Rapoport, eds., *The Syntax of Aspect*. Oxford: Oxford University Press, pp. 274-286.

Rosen, S. T., (2003). The Syntactic Representation of Linguistic Events. In L. Cheng & R. Sybesma (eds.), *The 2nd State of the Article Book*. Mouton de Gruyter, Berlin, pp. 323-365. Reprinted from Glot International, 4, 3-10.

Vendler, Z., (1967). *Linguistics in Philosophy*. Ithaca, New York: Cornell University Press, pp. 97-121.

Appendix A: A Snapshot of E-HowNet Ontology



中央研究院 中文詞知識庫小組 版權所有©

Appendix B: Dual Process-State Type Primitives in the E-HowNet Ontology

	Process	State	→	Dual Process-State	Example
1.	reproduce|生殖	ComeToWorld|問世		ComeToWorld|問世	降生 jiang-sheng ‘born’
2.	CauseToLive|使活	animated|有生命		living|生活	維生 wei-sheng ‘earn a living’
3.	MakeLiving|謀生	alive|活著			
4.	KeepOn|使繼續	GoOn|繼續		KeepOn|使繼續	待續 dai-xu ‘continued’
5.	keep|保持	withstand|抗住		keep|保持	顧全 gu-quan ‘keep in mind’
6.	resume|恢復	BeRecovered|復原		resume|恢復	復甦 fu-shu ‘resuscitate’
7.	stay|停留	situated|處於		stay|停留	獨處 du-chu ‘solitary’
8.	AimAt|定向	facing|朝向		AimAt|定向	迎向 yi-xiang ‘face to’
9.	attract|吸引	attractive|誘人		attract|吸引	迷住 mi-zhu ‘preoccupy’
10.	economize|節省	thrifty|儉		economize|節省	節儉 jie-jian ‘scrimp’
11.	lavish|浪費	extravagant|奢		lavish|浪費	虛靡 xu-mi ‘waste’
12.	ExpressAnger|示怒	angry|生氣		ExpressAnger|示怒	生氣 sheng-qi ‘angry’
13.	forgive|原諒	lenient|寬大		forgive|原諒	寬容 kuan-rong ‘broadminded’
14.	slack|偷懶	lazy|懶		slack|偷懶	混 hun ‘drift along’
15.	pity|憐憫	benevolent|仁		pity|憐憫	心軟 xin-ruan ‘softhearted’
16.	betray|背叛	treacherous|逆		betray|背叛	變節 bian-jie ‘defect’
17.	recreation|娛樂	enjoy|享受		enjoy|享受	自娛 zi-yu ‘amuse oneself’
18.	SeekPleasure|尋歡	enjoy|享受			
19.	None	ill|病態		ill|病態	生病 sheng-bing ‘sick’
20.	None	err|出錯		err|出錯	失誤 shi-wu ‘mistake’
21.	None	lack|缺少		lack|缺少	缺欠 que-fa ‘lack’
22.	None	ServeAsFoil|陪襯		ServeAsFoil|陪襯	相襯 xiang-chen ‘match’

Appendix C: The PoS Features and Semantic Expressions of Examples in the Paper

Examples	PoS	Semantic Expression
購妥 gou-tuo ‘to complete procurement’	active	{buy|買:aspect={Vachieve|達成}}
想起 xiang-qi ‘call to mind’	active	{remember|記得}
記取 ji-qu ‘keep in mind’	active	{remember|記得}
背起來 bei-qi-lai ‘memorize’	active	{remember|記得:aspect={Vachieve|達成}}
念念不忘 nian-nian-bu-wang ‘memorable’	stative	{remember|記得:manner={continuous|連續}}
刻骨銘心 ke-gu-ming-xin ‘be remembered with deep gratitude’	stative	{remember|記得:degree={extreme|極}}
打破 da-po ‘break’	active	{beat|打:patient={x},result={split|破開:patient={x}}}
破裂 po-lie ‘broken’	stative	{FormChange|形變:StateFin={incomplete|缺}}
取悅 qu-yue ‘please’	active	{please|取悅}
喜悅 xi-yue ‘joyful’	stative	{joyful|喜悅}
生氣 sheng-qi ‘angry’	stative	{ExpressAnger|示怒}
發脾氣 fa-pi-qi ‘get angry’	active	{ExpressAnger|示怒}
遇害 yu-hai ‘be murdered’	stative	{kill|殺害}
磨光 mo-guang ‘burnish’	active	{brighten|使亮:means={rub|摩擦}}
擦亮 ca-liang ‘polish’	active	{brighten|使亮:means={wipe|擦拭}}
水亮 shui-liang ‘bright as water’	stative	{bright|明}
光燦 guang-can ‘shining’	stative	{bright|明}
弔死 diao-si ‘hang by the neck’	active	{kill|殺害:means={coil|纏繞}}
往生 wang-sheng ‘pass away’	stative	{die|死}
行刑 xing-xing ‘execute’	active	{punish|處罰:means={kill|殺害}}
處決 chu-jue ‘put to death’	active	{punish|處罰:means={kill|殺害}}
受刑 shou-xing ‘be put to torture’	stative	{punish|處罰:domain={police|警}}
伏法 fu-fa ‘be executed’	stative	{punish|處罰:means={kill|殺害}}

板起(臉) ban-qu-(lian) ‘put on a stern expression’	active	{austere|冷峻}
正色 zheng-se ‘with a stern countenance’	active	{austere|冷峻}
凝重 ning-zhong ‘serious’	stative	{austere|冷峻}
不苟言笑 bu-gou-yan-xiao ‘serious in speech and manner’	stative	{austere|冷峻}
求生 qiu-sheng ‘seek to survive’	active	{living|生活}
度日 du-ri ‘subsist’	active	{living|生活}
生存 sheng-cun ‘exist’	stative	{living|生活}
在世 zai-shi ‘be living’	stative	{living|生活}
一息尚存 yi-xi-shang-cun ‘be still alive’	stative	{living|生活}
累病 lei-bing ‘sick from overwork’	stative	{ill|病態:cause={tired|疲乏}}
驚退 jing-tui ‘frighten off’	active	{frighten|嚇唬: patient={x}, result={leave|離開:agent={x}}}
看中 kan-zhong ‘take fancy to’	active	{FondOf|喜歡}
喜愛 xi-ai ‘love’	active	{FondOf|喜歡}
酷愛 ku-ai ‘ardently love’	active	{FondOf|喜歡:degree={extreme|極}}
熱衷 re-zhong ‘be addicted to’	active	{FondOf|喜歡}
癡情 chi-qing ‘be infatuated’	stative	{FondOf|喜歡:manner={mad|瘋痴}}
興致盎然 xing-zhi-ang-ran ‘full of interest’	stative	{FondOf|喜歡:cause={interesting|趣}}

Computational Linguistics and Chinese Language Processing
Vol. 19, No. 2, June 2014, pp. 53-72
© The Association for Computational Linguistics and Chinese Language Processing

不同母語背景華語學習者的用詞特徵：以語料庫為本的研究*

Salient Linguistic Features of Chinese Learners with Different L1s: A Corpus-based Study

張莉萍+

Li-ping Chang

摘要

本文採用語料庫語言學方法，對比不同母語背景學習者語料庫，試圖尋找不同母語背景學習者的詞語特徵。首先透過主題關鍵性方法快速找出不尋常高頻、不尋常低頻關鍵詞語，進一步從語言類型、詞類概念、文化因素等面向來解釋這些關鍵詞語突出的原因。這些特徵包括母語為英語學習者代名詞和能願動詞「會、可以」的使用、相對少用句尾語氣助詞、多用量詞「個」；日、韓學習者在假設句使用上優先選用「…的話」形式、因果句優先選用「所以」等等。

關鍵詞：華語、學習者語料庫、主題關鍵性、中介語對比、語言類型、母語遷移

Abstract

The study aims to explore the salient linguistic features of Chinese lexical items from different L1s learners. The research method is corpus-based, including comparing the learner corpus and the native-speaker corpus, as well as sub-corpora for different L1s. The learner corpus which consists of more than 1.14 million Chinese words from novice proficiency to advanced learners' texts is mainly from the computer-based writing Test of Chinese as a Foreign Language (TOCFL). The

*本研究得到科技部計畫(NSC 101-2631-S-003 -003)、科技部跨國頂尖研究中心計畫（NSC 103-2911-I-003-301）、教育部邁向頂尖大學計畫以及國立台灣師範大學華語文與科技研究中心部分經費補助，特此感謝。並感謝本期刊兩位匿名審查人給予之寶貴意見。

+ 國立台灣師範大學國語教學中心, Mandarin Training Center, National Taiwan Normal University
E-mail: lchang@ntnu.edu.tw

sub-corpora of Japanese, English, Korean, Vietnamese, Indonesia and Thai are observed. Japanese corpus is top 1, which occupies twenty four percent of the total data, followed by English, Korean, and etc. And the native corpus is from the Academia Sinica balanced corpus. Through the overuse or underuse linguistic forms and keyword-keyness analysis, some salient features are discovered. For examples, comparative to Chinese learners with other L1s, English language background learners show the unusual high frequency on pronouns and unusual low frequency on sentential final particles in Chinese writing. And Japanese as well as Korean background learners tend to overuse the post form '*de hua*' instead of '*ruguo*' when expressing the '*if*' sentence, and overuse '*suoyi*' instead of '*yinwei*' when expressing the cause-effect relation. The article also provides possible explanations for these results from the aspects of learners' native language typology, linguistic structure, syntactic category and culture.

Keywords：Mandarin Chinese, Learner Corpus, Contrastive Inter-language Analysis , Keyword-keyness, CEFR, Language Transfer

1. 前言

語料庫語言學興起於 20 世紀 80 年代，藉由語料的大量蒐集及標記工作，提供客觀、真實的語料給語言學研究者分析以及做為教學用例。同時間，第二語言學習者或外語學習者產出的語料也受到重視，所建置的語料庫稱為學習者語料庫（learner corpora）或中介語語料庫（interlanguage corpora）。最早的學習者語料庫是 1980 年代晚期由朗文出版集團所建立的朗文學習者語料庫（Longman Learners' Corpus），約一千萬詞規模，語料來源為各國英語教師所提供的學生作文或考試語料。之後所建立的學習者語料庫多為語料加註了偏誤標記（Díaz-Negrillo & Fernández-Domínguez, 2006），並為不同母語背景的學習者建立子語料庫，例如，ICLE (International Corpus of Learner English) 為英語學習者建立了 14 個子語料庫，目前子庫還在不斷增加中。學習者語料庫可以讓語言研究者或教學者觀察學習者的實際運用情況，對學習者的語言特徵和語言發展進行全面而有系統的描述和對比分析[1]（Granger, 1998; Granger, Hung & Petch-Tyson, 2002; Hawkins & Buttery, 2009)。

關於漢語學習者語料庫的建置相對晚些，以 2009 年在大陸正式公開的「HSK 動態作文語料庫」為代表[2]，該語料庫是母語非漢語的外國人參加高等漢語水平考試（HSK）紙筆作文考試的語料，於 2003 年開始建置，蒐集了 1992-2006 年的部分外國考生的作文答卷，共計 11569 篇，424 萬字。雖然有四百多萬字的規模，但僅限當初參加 HSK 高

[1] 本文使用「特徵」一詞，是指廣義的語言使用特徵，只要是足以描繪出學習者使用語言的個別特色，都可稱為特徵，非指理論語言學中的抽象特徵。

[2] 可參見 http://202.112.195.192:8060/hsk/login.asp

等考試的文本[3]，也就是這些語料多是高等程度學習者的寫作語料。台灣的「TOCFL 學習者語料庫」也是類似性質，蒐集 2006 年至今參加華語文能力測驗（TOCFL）考生的作文，不同的是，這個測驗是電腦考試，也就是考生直接於線上輸入文字。現階段約 114 萬詞（174 萬字左右），語料涵蓋 42 種不同母語背景、不同能力考生所寫的作文，共 5092 篇，130 個主題，語料量仍在持續擴充中，關於語料庫的建置與語料分布請參考張莉萍（2013a）。

在對外漢語方面，以語料庫為本的語言特徵研究相對地少，在詞彙特徵方面，張莉萍（2012）利用語料庫文本覆蓋率的概念，驗證能力越高的學習者所需要的詞彙量越大，並從教學與學習的角度，建議對應於 CEFR（Common European Framework of Reference for Languages）(COE, 2001)不同等級學習者所需要的詞彙量，A2 為 1000 個詞、B1 為 2300-3000 個詞、B2 為 4500-5000 個詞、C 級為 8000-10000 個詞[4]。張莉萍（2013a）也嘗試分析了不同能力學習者使用高頻動詞的情況，與英語學習一樣，呈現同樣的趨勢（Hawkins & Buttery, 2009:164），即，隨著學習者語言水平的提高，使用高頻動詞的頻率越趨近本國人（p.147）。在語法特徵方面，張莉萍（2013b）則是利用本國人語料庫與學習者語料庫的對比，企圖找出不同能力階段學習者的句式表現特徵，例如，把字句、被字句。這些研究結果可以直接或間接應用在教學與評量上，本研究的目標則是企圖找出不同母語背景學習者的用詞特徵。

2. 研究方法與步驟

這個研究採用中介語對比分析方法（contrastive interlanguage analysis, 以下簡稱 CIA），CIA 是語料庫語言學與第二語言習得研究結合後的新領域，主要是針對兩類語料進行對比，一是本國人語料庫和學習者語料庫的比較；一是不同母語背景學習者語料庫之間的比較（Granger, 1998: 12）。應用前者對比分析可以找出學習者少用 (underuse) 或過度使用 (overuse) 的語言特徵；應用後者對比分析，則可能找出針對不同母語背景學習者的語言特徵。本研究著重在後者的描述與討論，所選用的學習者語料庫即是前述的 TOCFL 語料庫，主要是這個語料庫建置時，即載有學習者的母語背景資料；HSK 則僅有國籍資料。

為了觀察不同母語背景學習者的用詞特徵，我們先將語料庫依學習者不同母語分成六個子語料庫(sub-corpora)—英、日、韓、越、印尼、泰，從表 1 可以看到，母語為日語的語料量最多，佔全體語料的 24%左右。而這六個母語背景的語料在 TOCFL 語料庫中排名前六名，約佔全體語料量的百分之七十左右，以下是每個子語料庫的總詞數以及佔全體語料的百分比統計訊息。

[3] 這裡是指舊版的 HSK 考試，當初設計雖有初、中、高等級，但僅高等考試需要寫作文。

[4] CEFR 的能力架構為三等六級，由初級到精通等級依序是 A1/A2, B1/B2, C1/C2；TOCFL 現階段語料庫中 A2 到 C1 等級語料的分布比例分別是 13%、48%、33%、6%。

表 1. 六個子語料庫詞數及百分比訊息

	日語	英語	韓語	越南語	印尼語	泰語
詞數 (token)	271869	160400	110396	108999	87989	40071
佔全部語料 %	23.92	14.11	9.71	9.59	7.74	3.53

本研究應用語料庫檢索工具中常用的主題關鍵性（keyword-keyness）這個統計方法嘗試快速有效地找出特徵[5]。這個方法最早是應用在尋找文本中的主題詞，通過詞語的關鍵性分析 (keyness) 找出某一主題文本的詞語特徵 (William, 1976)。雖然早期多應用於文類或風格分析，但筆者假設利用主題關鍵性分析的統計概念—不尋常高頻 (unusually high) 或不尋常低頻 (unusually low) (Scott, 1997: 236)，找出不同子語料庫中相對高頻或低頻的關鍵性詞語 (Baker, 2006: 139)，應該可以呈現不同母語背景學習者的用詞特徵。

目前運用這關鍵性方法做的研究主要有兩種類型 (Rayson & Garside, 2000)，一是拿一個語料庫與一個較大（標準）語料庫的比較，目的是藉由顯著不同用法（頻率）來發現被觀察的語料庫的特徵；第二種是拿兩個差不多大小語料庫來比較，目的是想要找出可以區辨兩者（語料庫）的特徵。基本的原理都是利用統計的方法，比對兩個語料庫衍生出來的詞表，看看哪些詞是屬於不尋常高頻或不尋常低頻，進而從這些主題詞中來分析文本。不論哪個類型，原則上都要有一個參照語料庫 (reference corpus)，一個受觀察語料庫 (observed corpus)，參照語料庫要大於受觀察語料庫 (Scott & Tribble, 2006: 58)。因此第三節所呈現的結果，是使用兩種比對的結果，一是拿 114 萬詞學習者語料庫為參照語料庫，每個子語料庫為受觀察語料庫；一是分別拿日語背景語料庫、英語語料庫為參照語料庫[6]，其他子語料庫為受觀察語料庫。

分析的步驟是，先將不同母語（英、日、韓、越、印尼、泰）的語料庫製作成子語料庫（受觀察語料庫），進行斷詞工作後[7]，放入 AntConc（3.2.4w 版）這個語料庫工具產出屬於各子語料庫的詞表，然後利用工具上 Keyword list 功能，計算出 Keyness，觀察這六個不同母語背景學習者的用詞特徵。圖 1 為進行計算日語學習者語料庫 keyness 的 AntConc 畫面[8]。

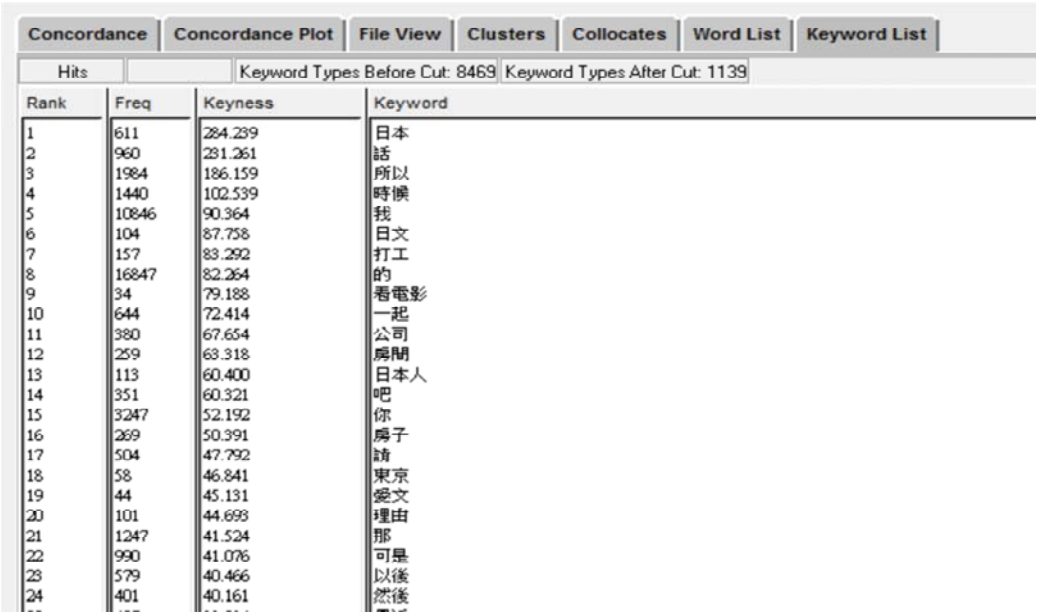
---

[5] 在語料庫工具 *WordSmith* (Scott ,1997, 2008) 或 *AntConc* (Anthony, 2009)中，都可見 keyness 這個功能。

[6] 分別選定日語、英語為參照語料庫，主要是因為這兩者語料量是前二名，其他子語料庫的量則相對地小，不確定統計數據的可信度。

[7] 寫作文本斷詞是採用中央研究院的自動斷詞與詞性標註程式進行自動處理。參考網址：http://ckipsvr.iis.sinica.edu.tw/

[8] 畫面中第九個詞語「看電影」是自動斷詞的結果，其實語料庫中大多數都斷成「看」和「電影」兩個詞，這個斷詞結果不知為何原因和其他語料庫不同。雖然如此，並不影響我們後面的討論。



Rank	Freq	Keyness	Keyword
1	611	284.239	日本
2	960	231.261	話
3	1984	186.159	所以
4	1440	102.539	時候
5	10846	90.364	我
6	104	87.758	日文
7	157	83.292	打工
8	16847	82.264	的
9	34	79.188	看電影
10	644	72.414	一起
11	380	67.654	公司
12	259	63.318	房間
13	113	60.400	日本人
14	351	60.321	吧
15	3247	52.192	你
16	269	50.391	房子
17	504	47.792	請
18	58	46.841	東京
19	44	45.131	愛文
20	101	44.693	理由
21	1247	41.524	那
22	990	41.076	可是
23	579	40.466	以後
24	401	40.161	然後

圖 1. AntConc 3.2.4w 計算 Keyness 畫面

3. 不同母語背景的用詞特徵

基本上，這六個子語料庫都是來自同一個大語料庫，寫作的主題都在一樣的範圍內，如果在不同子語料庫中可以看到不尋常高頻或不尋常低頻的現象，應該可以推測是受學習者母語這個因素的影響。

3.1 關鍵性詞表結果

下表是依第 2 節研究方法中的陳述，為每個子語料庫做出關鍵詞表，然後依序列出前 20 高關鍵詞的資訊以便觀察[9]。從表 2 可以看到一個一致的現象，即這些不同母語背景的子語料庫顯現出來的高關鍵性詞都有代表該母語國的專有名詞，例如，日語語料庫中的「日本、日本人」，泰語語料庫中的「泰國、泰文、泰國人」等等。這個與整體語料庫的性質有關，因為考試作文題目中包括了介紹自己、說明喜好等與個人切身相關的主題，自然這些代表寫作者個人資料的訊息就會進入到寫作文本中。從這點，也可以初步確定這個關鍵性的方法，的確可以讓我們很快速輕易地看出屬於某個特定語料庫中的用詞特徵。

[9] 如圖 1 所示，第一個欄位顯示的是排名高低，可以找出前 20 名高關鍵詞。

表 2. 六個子語料庫與整體學習者語料庫對比的前 20 高關鍵性詞

子語料庫	前 20 高關鍵詞
日	日本 話 所以 時候 我 日文 打工 的 看電影 一起 公司 房間 日本人 吧 你 房子 請 東京 愛文 理由
英	我 會 她 你 個 美國 幫 感謝 台北 我們 相片兒 如果 王宜家 夢 李吉他 見面 他們 可以 穿
韓	韓國 李廠長 不一樣 韓國人 時候 話 我 辣 的 錢包 泡菜 吧 以後 韓國菜 可是 中國菜 所以 還 寵物 容易
越	家長 你 您 小孩 工廠 都 電腦 攝影機 越南 影響到 上網 學習 生活 大家 處理 噪音 臭味 讓 每 很
印	印尼 學校 會 寵物 您 也 工廠 要 因為 在 們 或是 他 阿德 欣賞 可以 老師 畢業 狗 來
泰	泰國 他 她 就 書 今天 李孟 本 老師 平平 泰文 貴方 我 筆者 喬丹 去 幫 馬老師 泰國人 讓

下面表 3、表 4 的結果則是，我們分別將日語和英語做為參照語料庫，得出其他子語料庫與它們之間的高關鍵性詞表（一樣取前二十個）。由於這個研究方法的先決條件是參照語料庫要比受觀察語料庫大，所以下面表 3（以日語語料庫為參照）可以看到其他五個子語料的關鍵性詞，表 4 是以英語語料庫為參照，只能看到其他四個子語料庫的關鍵性，因為日語語料庫比英語大，無法作關鍵性分析。

表 3. 與日語語料庫對比的前 20 高關鍵性詞

子語料庫	前 20 高關鍵詞（類）
英	會 Nh 張愛文 和 美國 個 可以 Neu 他們 讓 Caa 我 一 就 幫 D 快 Nb VL 國家 她 [話、所以、吧、的]
韓	韓國 韓國人 李廠長 辣 錢包 泡菜 韓國菜 不一樣 媽 皮帶 膩 學生 D 主人 更 自然 種 熱 全 皮鞋 [個、然後、不過]
越	都 D 家長 越南 讓 COLONCATEGORY 就 發展 影響到 學習 可以 生活 Da 而 小孩 您 現象 很 所 電腦 [的、話、時候、我、所以]
印	印尼 D 會 在 人民 可以 此 要 能 他們 也 讓 或是 因為 Caa 學校 來 P 家長 牠們 [的、話、所以]
泰	泰國 就 他 讓 Nh Nb D 她 張愛文 VL 平平 李孟 泰文 書 幫 本 叫 喬丹 曼谷 影片 [的、話、時候、所以、吧]

表 4. 與英語語料庫對比的前 20 高關鍵性詞

子語料庫	前 20 高關鍵詞（類）
韓	韓國 話 T 還 韓國人 寵物 吧 PAUSECATEGORY Ng 所以 李廠長 臺灣 你們 的 時候 辣 韓國菜 FW 台灣 以後 [個、會、代名詞(Nh)]
越	PAUSECATEGORY T 越南 您 家長 大家 學習 都 生活 處理 喔 電腦 還 而 這樣 Da 攝影機 影響到 網路 孩子 [的、代名詞(Nh)、個、會]
印	印尼 D 寵物 T 也 ETCCATEGORY 您 b 學校 你們 狗 畢業 來 們 喔 養 或是 阿德 事 出租 [代名詞(Nh)、個]
泰	泰國 T COMMACATEGORY 今天 他 就 你們 筆者 平平 李孟 泰文 月 泰國人 書 喬丹 曼谷 PAUSECATEGORY V 女生 有 [的、會、個]

在與日語語料庫和英語語料庫對比的資料表中，為了比較宏觀地來看彼此差異，我們在做關鍵性分析時，特別保留了詞類訊息（包括標點符號標記），因此在上面兩個表格中可以看到詞類訊息[10]。另外，也為了容易比對分析，將不尋常低頻的高關鍵性詞也一併選取前幾個具語法意義的詞彙用方括弧表示在表 3、表 4 中。例如，和日語學習者比較，英語學習者相對少用[所以、吧、…的話][11]。

以下是我們歸納出不同母語背景學習者的用詞特徵。在歸納過程中，我們忽略表 3、4 中專有名詞（例如，「日本」）或具有實詞意義的詞（例如，「書」），而選擇具有語法或語用功能的詞語，以進一步分析並解釋可能的原因。

1. 母語為英語者的用詞特徵

(1) 與整體對比，使用代名詞「我、她、你、我們、他們」的頻率相對突出。與日語對比，也可以看到代名詞（Nh）的高關鍵性。

(2) 與整體對比，能願動詞「會、可以」的頻率相對突出。與日、韓、越、泰對比，「會」的使用頻率相對地高。

(3) 與整體對比，使用量詞「個」的頻率相對突出。與日、韓、印尼、越、泰對比，也高。

(4) 與整體對比，使用連接詞「如果」的頻率相對突出。與日語對比時，「所以」和「…的話」相對少用。

(5) 與韓、越、印尼、泰對比，語助詞(T)相對少用。與日語對比，相對少用語助詞「吧」。

2. 母語為日語者的用詞特徵

(1) 與整體對比，使用代名詞「我、你」的頻率相對突出。

(2) 與整體對比，「的」的使用相對地突出。

---

[10] 這些詞類訊息標記代表的意義請參考詞庫小組（1995）。

[11] 表中方括弧內「的、話」是因為自動斷詞規則所致，筆者進語料庫實際觀察，發現「的、話」的高關鍵性主要是學習者使用「（如果）…的話」而來。

(3) 與整體對比，連接詞語「所以、...的話」的使用頻率相對突出。與英、越、印尼、泰對比，也都高。

(4) 與整體對比，語助詞「吧」的使用頻率相對地高。與英、泰對比，也都高。

3. 母語為韓語者的用詞特徵

(1) 與整體對比，代名詞「我」的頻率相對突出。

(2) 與整體對比，「的」的使用相對地突出。

(3) 與整體對比，連接詞「所以、可是、...的話」的使用頻率相對突出。與英對比時，「所以」相對地高。

(4) 與整體對比，語助詞「吧」的使用頻率相對地高。與英對比時，也高。

(5) 與整體對比，副詞「還」的使用頻率相對突出。與日語對比時，副詞「更」相對地高；與英語對比時，副詞「還」相對地高。

4. 母語為越語者的用詞特徵

(1) 與整體對比，代名詞「你、您」的頻率相對突出。

(2) 與日語對比時，能願動詞「可以」的使用頻率相對突出。

(3) 與英語對比時，語助詞「喔」的使用頻率相對地高。

(4) 與整體對比，副詞「都」的使用頻率相對突出。與日語對比時，副詞使用頻率相對地高；與英語對比時，副詞「都、還」相對地高。

5. 母語為印尼語者的用詞特徵

(1) 與整體對比，代名詞「您、他」的頻率相對突出。

(2) 與整體對比，能願動詞「會、可以」的頻率相對突出。與日語對比時，「會、可以、要、能」相對地高。

(3) 與整體對比，使用連接詞「或是」的頻率相對突出。與日語對比時，「或是、因為」相對地多。與英語對比時，「或是」相對地高。

(4) 與整體對比，副詞「也」相對突出。與英、日對比時，也高。

(5) 與整體對比，後綴詞「們」相對突出。與英語對比，也高。

(6) 與英語對比時，語助詞「喔」相對地高。

6. 母語為泰語者的用詞特徵

(1) 與整體對比，代名詞「他、她、我、筆者」的頻率相對突出。與日語對比時，代名詞使用頻率相對地高。與英語對比時，「他、你們、筆者」相對地高[12]。

(2) 與整體對比，量詞「本」的頻率相對突出。與日語對比時，也高。

---

[12] 我們實際進泰語子語料庫查詢，使用「筆者」都是同一位學習者（B1 等級），屬個別現象。

(3) 與整體對比，副詞「就」相對突出。與英、日對比時，也高。

(4) 與英語對比時，語助詞（T）的使用頻率突出。

3.2 分析討論

根據上述的歸納，為了較有系統地來討論不同母語背景學習者用詞的特徵與可能原因，下面以詞類來分項敘述。

3.2.1 代名詞

從關鍵性分析結果來看，英語代名詞(表中的 Nh) 的使用相對於其他五個語言背景學習者特別顯著，除了高關鍵性的代名詞種類最多，與日語相比，英語學習者的代名詞這個詞類的關鍵性也是突出。韓、越、印尼學習者則沒有這個特色。

代名詞在篇章中主要起銜接作用。不少研究者或語言教師都會指出初階學習者的特徵之一就是不會使用零代詞(zero anaphora)，使得篇章結構鬆散、連貫性差(肖奚強, 2001: 54），而多數實證研究的對象都是歐美或西方學習者，例如，Jin (1994) 從語言類型的角度，探討母語為英語的學習者，在學習屬於主題顯著的語言（華語），有將英語的主語結構特徵遷移到第二語言的現象；周曉芳（2011）則是針對歐美學生敘述語篇中的回指習得研究指出學習者呈現代詞多用、零代詞嚴重少用的情況。研究者對歐美學習者代詞習得感興趣的原因，主要應該是從英漢語言類型差異的角度出發。研究者會從主題顯著語言/主語顯著語言（Li & Thompson, 1976）、代詞脫落 (pro-drop) 語言/非代詞脫落語言來解釋學習者過渡語言的現象。

我們從語料分析數據中得出整體學習者有代名詞多用的情況，但如果以日語做為參照語料庫，則只有英語和泰語的代名詞類出現關鍵性數值。表示英語與泰語學習者代名詞多用的情況是比其他四個語言背景的學習者來得顯著。由於我們所觀察的這六個語言背景中，只有英語屬於非代詞脫落/主語顯著語言，其他五個亞洲語言都屬於代詞脫落/主題顯著語言 (Chomsky, 1981: 284, footnote 47; Li and Thompson, 1976)。這裡泰語母語者在代名詞部分的突出表現則顯得與本研究中其他亞洲語言的表現不一致，我們嘗試從前人的研究中找線索，但只有關於泰語學習者在回指偏誤這方面的統計結果，例如，楊帆（2011: 14）、林雪鳳（2008: 62）指出指稱偏誤中，零形式的偏誤情形最多，也就是學習者不善於使用零形式，而使得篇章行文不流暢；並沒有從語言類型或對比的角度來分析偏誤的原因，前人分析這方面偏誤的原因，都偏重在教學上，普遍認為教材或教學沒有著重在這方面做有系統的安排。

不過，泰語的代名詞數量相當驚人，會因說話者與聽話者的性別和相對關係等有所不同，光是「我、你」的說法就各有十餘種[13]。或許泰國學習者對於代名詞的使用特別

[13] 引自網路上文件：thai.world68.com/upfiles/泰語人稱代詞詳解.doc

有意識，也就是說，母語者的詞類系統對於二語習得也是一個影響因素[14]，當然這需要未來更多的研究來驗證。

另外我們也發現無論與整體語料庫或日語、英語語料對比時，越南母語者代名詞「您」都具關鍵性。前人的文獻顯示，越南語的人稱代詞系統很複雜，而且多數帶有情感色彩，例如不同形式可以表示親密、尊敬、中性、輕蔑等態度（阮氏懷芳, 2008: 46），漢語則大不相同，除了第二人稱「您」帶有尊敬色彩，其他人稱都是不帶感情色彩的中性詞。推測越南學習者在使用人稱代詞時，特別有意識地，會根據場合、身份，選用第二人稱「您」。我們實際進入越南語子語料庫驗證了我們的推測，凡寫作主題中牽涉到給老師、父母、工廠廠長、政府公務人員等便條或信件的內容，需使用第二人稱代詞時，越南學習者幾乎都選用敬語形式，包括「您、您倆、您們」。

3.2.2 連接詞

在連接詞的部分，本研究先就「因為/所以、如果/…的話」這兩組呈現高關鍵性數值進一步分析。

3.2.2.1 「因為/所以」

張莉萍（2013a）觀察本國人與學習者連接詞使用分佈時，即注意到學習者使用「所以」的頻率較「因為」高，與本國人不同。現在關鍵性的分析呈現日語和韓語者使用「所以」的高關鍵性。由於語料庫中絕大多數初階學習者是在台灣接受華語課程，我們可以排除教學、教材這些變數對學習者的影響，加上在華語教學中，表因果關係的「因為、所以」都是同時成對授予學生的，因此，我們推測日語、韓語學習者的這個使用傾向受母語影響的可能性不低。

下表是針對「因為、所以」這兩個連接詞，英、日、韓這三個母語背景學習者各等級的使用狀況。表 5 顯示，三類學習者都有一個共同趨勢，即這兩個詞的使用率隨著能力的提高而降低。如果參照本國人的使用率[15]，「所以」是 0.13-0.56，「因為」是 0.16-0.85，可以看出學習者使用「因為」的頻率和本國人較接近，「所以」的使用率則要到 B2（以上）能力比較接近本國人。

---

[14] Jarvis, Gastaneda-Jimenez and Nielsen (2012:61) 研究中指出不少學習者母語影響外語學習的例證，例如，芬蘭語學習者明顯較其他母語者少用英語的 a, the，是因為芬蘭語中沒有冠詞(articles)這一個語法系統。另一例是，芬蘭語中沒有區分性別的代名詞系統，從語料庫數據中也可以看到以芬蘭語為母語的學習者在英語 he, she 的使用上，相對較其他母語者來得少。

[15] 為免頻率訊息受書面或正式語體的因素影響，這裡的使用頻率是同時引自中研院平衡語料庫詞頻頻率（http://elearning.ling.sinica.edu.tw/CWordfreq.html）以及對話語料庫頻率訊息（http://mmc.sinica.edu.tw）。對話語料庫是由 30 個自由對話，26 個地圖導引對話與 29 個主題對話的內容整理而成，約計 40 萬詞，該語料庫詞表於 2012 年 4 月公開。平衡語料庫則是台灣第一個對外開放的大型語料庫，以書面語料居多，1995 年開放 1.0 版語料庫是 200 萬詞，本研究使用的是 3.0 版 500 萬詞（http://app.sinica.edu.tw/kiwi/mkiwi/）。

表 5. 「所以、因為」英、日、韓語學習者使用情況

		A2 (%)	B1 (%)	B2 (%)	C1 (%)
所以	日　だから	1.4609[16]	1.0139	0.5603	0.4166
	韓	1.4859	1.0445	0.3855	0.1543
	英	0.9783	0.5923	0.2912	0.2419
因為	日　ので	0.6598	0.5564	0.4854	0.2406
	韓	0.7676	0.6017	0.3905	0.3305
	英	0.8145	0.6282	0.5670	0.3892

這其中又以日語學習者使用「所以」的頻率最為突出。英語學習者使用「因為、所以」這組連接詞的傾向和目標語人士較接近，即「因為」的頻率高於「所以」；韓語學習者則到 B2 等級才漸漸有這個傾向(0.3905:0.3855)；日語學習者則一直呈現「所以」高於「因為」的結果。

周剛（2001）在〈漢、英、日語連詞語序對比研究及其語言類型學意義〉一文中舉了因果目的關係連詞的例子如下[17]：

(1) 因為他的汽車在路上出了事，所以來晚了。

(1') He came late because his car broke down on the way.

(1'')a. 彼　は 自動車　が 途中 事故 に　あったので、来る の　が　遅れた。

Kare wa zidousha ga tochu jiko ni attanode, kiru no ga okureta

b. 彼 は 自動車　が 途中 事故 に　あった。だから　来る の が 遅れた。

Kare wa zidousha ga tochu jiko ni atta dakara kiru no ga okureta

（引自周剛, 2001: 46 例句 20）

說明日語表因果目的的連詞可以用接續助詞，如(1")a 的ので（因為），也可以用接續詞，如(1")b 的だから（所以），接續助詞都是後置先行的[18]；接續詞都是前置後續的。

---

[16] 數據的算法是將日語者使用「所以」詞數除以日語者總詞數(e.g. 403/271869)，表示母語為日語學習者，使用「所以」的比例。

[17] 引用之原文例子沒有附逐字翻譯之羅馬拼音，為利閱讀，我們補充於此。後面例句(2)同此處理方式。

[18] 該文所探討的語序敘述如右：「連詞的語序應該包括兩種，一種是指連詞處於所連接的成分的前後位置，我們稱之為前置和後置；另一種是指連詞處於話語中的語序，有固定和不固定之分，我們稱之為定序和不定序，定序連詞中又可以分為先行和後續兩種。先行連詞中有些詞由於語用因素，可與所連接的成分移動至後續小句的位置，有些則不能，我們把它們稱之為可後移與

而且表示因果關係的連詞，只能用單個詞，不能用兩個詞搭配的關聯形式（像「因為…所以…」）。

而漢、日表示因果關係的語序，都是原因在前，結果在後，如果要強調結果或是做為事後補充語，有時表示結果的子句，也會出現在前（Chao, 1968: 132）[19]。我們進一步查詢ので（因為）和だから（所以）這兩個詞在日語的頻率訊息，發現だから的頻率遠高於ので。所以我們推測日語學習者使用「所以」頻率較高的原因有兩個，一是因為日語在這類連接句中，只能選擇單一連接詞語，而だから（所以）的頻率又高於ので（因為）許多，這個習慣遷移到漢語中[20]，使用「所以」的頻率相較於其他不同母語背景學習者，才會那麼突出；二是語言類型的影響，我們假設學習者傾向選擇和目標語一致的前置連詞來使用，だから位於（小）句首位置，同漢語連詞的典型位置相同。不論是哪一個因素，這個現象一直持續，在高階學習者身上仍可見。韓語表達因果關係的方法與形式與日語大同小異，大多也都使用一個連詞，不用成對的形式，呈現在漢語學習上，用詞特徵與日語學習者大致相同。

3.2.2.2 「如果…的話」

漢語表示假設條件關係的連接詞語，最常見的就是「如果…（的話）」。其中「如果」的頻率不論在平衡語料庫或對話語料庫中都高於「的話」，「…的話」並不是漢語典型的連接形式[21]，漢語典型的連接詞語多放在句首位置，不論是主要句子或從屬句子。但關鍵性分析中，日語、韓語都呈現「的話」的高關鍵性。於是我們進一步篩選統計了英、日、韓三種語言單用「如果」、單用「的話」以及使用「如果…的話」這三種不同形式，結果如下表：

表6. 英、日、韓學習者「如果...的話」使用比例統計

	如果…的話	如果…	…的話	小計
英	15.51%	71.62%	12.87%	100%
日	45.45%	18.18%	36.36%	100%
韓	29.60%	25.37%	45.03%	100%
本國人	6.51%	80.72%	12.77%	100%

不可後移連詞。反之，後續連詞中有些詞由於語用因素，可與所連接的成分移動至先行小句的位置，有些則不能，我們把它們稱之為可前移與不可前移連詞。」（引自周剛, 2001: 42）

[19] Tai (1985) 解釋「先說因，後說果」是漢語的自然語序(natural order)；「先說果，後說因」則是凸顯語序(salient order)。

[20] 關於 L1 頻率的遷移，在 Jarvis and Crossley (2012)主編的書中，也有不少討論，例如，瑞典語與芬蘭語學習者在使用英語 come 這個詞彙的頻率高於葡萄牙語和西班牙語學習者(p.62)，是受到他們母語中相對應詞彙頻率的影響。葡語和西語母語中使用 *llegar* (arrive)較高，反映在使用英語時，arrive 的使用率高於 come。

[21]「的話」形式出現於元朝（余志鴻, 1992; 祖生利, 2002），是受阿爾泰語言影響（感謝曹逢甫教授提供此資訊）。

從上表可以發現，對於這三種形式的使用，英語母語者選擇單用「如果」這個連接詞的比例最高（71.62%），也最接近本國人使用情況（80.72%）[22]；韓語者選擇單用「的話」比例最高，日語者選擇「如果…的話」比例最高。而日韓共同點是單用「如果」的情況最少，選擇含有「的話」的表達形式佔了 75%以上，日語更高達 82%。我們從前人文獻知道，日語中一般表達假設條件關係的連詞只用接續助詞，如下面(2”)例句中的“たら”。

(2) 如果她拒絕了，那麼你怎麼辦？

(2’) In case she refuses, what will you do?

(2”)も し　彼女　が　断ったら、　あなた は　どうする の　ですか。

Moshi kanojo ga kotowaruttara anata wa dosuruno desu ka.

（引自周剛, 2001: 49 例句 36）

推測日語母語者受日語條件關係表達中，都用接續助詞（後置黏著詞）的影響，傾向使用漢語中性質接近的「的話」來表達假設語義。韓語在連接詞語的表現和日語相近，大都使用詞尾成分，接在詞幹或時間詞尾後[23]。因此日韓學習者的表現較一致，與英語學習者選用連接成分（形式）差異較大。

筆者推測日韓學習者傾向於使用後置成份，因此另外觀察「除了…以外/之外/外」這句式，統計數據顯示日韓學習者保留使用後置成份「以外/之外/外」的比例遠高於英語背景者；日語有 93.65%、韓語有 89.23%、英語有 69.23%，雖然保留後置成份的比例都很高，不過，英語背景者不用後置成份的情形還是明顯高於日韓學習者。

對表達同一個語義的不同形式，日韓學習者似乎有志一同的選用和自己母語相似的表達形式，而非選用目標語人士較高頻使用的形式。從這兩組連接詞的表現，似乎也可以看出日韓語言類型接近，呈現在學習漢語的表現形式也接近[24]。

---

[22] 這個統計結果是根據中央研究院漢語平衡語料庫 4.0 版（http://140.109.19.114/）檢索得來。

[23] 郭麗娟（2011: 26）指出漢語假設連詞「的話」和韓語「으면（的話）」 都放在動詞的後面，位置相同。

[24] 從構詞類型的角度來看（Lehmann, 1978: 219），日語、韓語偏向黏著型語言 (agglutinative language)，是透過詞綴加入詞幹以增加意義或改變語法功能的語言。英語偏向屈折型語言 (inflectional language)，透過詞形變化表示意義和語法功能的改變。漢語偏向孤立型語言 (isolating language)，透過語序和功能詞表達語法關係。

3.2.3 能願動詞

在能願動詞的部分，英語母語者的「會、可以」具高關鍵性；印尼母語者的「會、要、可以」具高關鍵性。如果以日語為參照語料庫，則印尼和越南學習者對能願動詞的使用頻率相對突出。能願動詞這一類無論在語言學或華語教學中，都有相當多的討論，在華語學習上，也是一個公認的難點（鄧守信, 2009: 113）。一般教師或研究者多從能願動詞混用的角度來分析原因（賴鵬, 2006），葉信鴻（2009）則從 HSK 動態作文語料庫中，實際統計「會、能、可以、想、要」等五個主要能願動詞的偏誤類型發現，缺漏 (omission) 與誤加 (addition) 這兩類遠較錯用 (mis-selection) 與錯序 (mis-ordering) 等類型比例來得高，一般文獻中所關心的混用或誤用情況反而不是偏誤較高的類型。

由於筆者不熟悉印尼語，也沒找到相關探討印尼學習者學習漢語能願動詞的研究，因此在這裡，僅能就英語學習者來討論。從語言類型的角度來解釋，似乎可以推測因為英語助動詞與漢語能願動詞類有相近的語義和語法範疇，幾乎也都有可以對應的詞語，雖然不一定是一對一，可能是一對多，但存在著相應的類，對英語學習者來說應該比母語沒有獨立的這類學習者較不會有缺漏或誤加的現象。我們從英語語料庫「會、可以」相較於其他子語料庫的高關鍵性，可以大膽推測其他語種學習者缺漏情況較英語學習者顯著。由於缺漏，需要分析者一筆一筆觀察標記，無法由機器判讀，因此限於時間、人力，在這個研究裡，僅觀察 B2 等級（中高級，相當於 HSK 語料等級）子語料庫中「會」的使用情況，結果如表 7 所示：

表 7. TOCFL 語料庫 B2 等級子語料庫「會」的使用情況

	篇數	使用次數	偏誤小計	缺漏次數	%	誤加次數	%	錯用次數	%
日	260	396 (0.53)	305	204	66.89	81	26.56	20	6.56
英	122	418 (1.07)	127	54	42.52	37	29.13	36	28.35
韓	130	268 (0.68)	42	16	38.10	9	21.43	17	40.48
泰	41	96 (0.89)	28	10	35.71	4	14.29	14	50.00
印尼	112	335 (1.05)	38	2	5.26	3	7.89	33	86.84
越	188	401 (0.73)	28	5	15.15	12	36.36	16	48.48

日語學習者缺漏「會」的比例最高，英語學習者次之，並不如筆者先前推測，因此英語這個高關鍵性的表現，未來還需要進一步探討。

從表 7 也可以看出，不同母語者「會」的偏誤類型表現不一，但值得我們注意的是，除了傳統教學者或研究者關注的混用（錯用）情況，學習者在該用卻不用的情形也相當顯著，而且這情況是存在已趨近本國人表現的 B2 學習者身上，可見「會」或其他能願動詞的使用問題，包括該用不用或誤加等等，也值得研究者或教學者深入發掘探討。以下僅舉幾例語料庫中缺漏偏誤供參（方括弧內是筆者加入的）。

(3) *希望李先生能夠了解一個小小的消費者的情境，並相信政府當局［會］妥善處理。（B2 程度學習者，以下 2 例皆是）

(4) *如果製造「有錢就可以解決問題」這樣的世界，人的心［會］越來越空虛。

(5) *而且長期電腦使用［會］對身體帶來很多不好的影響，例如視力不良，肩膀酸痛等。

3.2.4 語氣詞

這裡的語氣詞指的是放在句尾的語氣助詞，如，「啊、吧、嘛」等等。在高關鍵性的分析結果中，可以清楚看出英語和其他五個亞洲語言在這方面明顯的差異。也就是，英語學習者跟整體學習者比較下，相對少用語助詞。日、韓學習者則是相對多用語助詞「吧」；越南語、印尼語、泰語則是和英語對比時，才呈現出語助詞的關鍵性。從類型學的觀點來看，英語沒有句尾語氣助詞這類詞語，多用助動詞或疑問句型來表達這樣的情態；而日、韓等東亞語言幾乎都有表達情態的助詞，推測這是英語背景學習者少用的主要原因。徐晶凝（2003）從日漢對譯的語料中來分析「吧」的語法分佈情形，指出漢語「吧」基本上的出發點是在於說話者對命題真值的關心，可以從日語對譯的某種形式「ね 疑問句」得出，這是一種預料聽話者對詢問內容能給以肯定性確認的問句；然而日語說話者在交際時，頻繁使用語氣助詞，例如“ね”，並不是為了讓對方針對他提供的命題信息進行確認，而是一種潤滑雙方關係的手段。這個觀點，我們從日語學習者語料中也得到證實，日語母語者的語言習慣讓「吧」的使用頻率較其他母語背景者高，而且多用在表示禮貌的委婉語氣，而非想得到訊息的確認。例如，下面這段擷取自日籍學習者的考試作文語料，可以看出使不使用「吧」並沒有合法性的問題，使用了語氣助詞具有緩和語氣、促進雙方情感交流的效果。

> 我已經詳細地看過了妳的信。關於妳同學的錢包不見了這事，我可以理解妳的心情，同時我也替你感到憤怒。無論你的位子在不在他的旁邊，總不該把同學當做小偷來對待*吧*！...後來，他知道你背包裡沒有他的錢包，他向妳道了歉嗎？...我不知道平常妳跟他好不好，但我想，他是一時焦急才這麼做的*吧*。下次碰到他時，妳可以問一下他的錢包找到了沒有。並且可以告訴他，妳對他的對待多麼的傷心。...聽到妳講的話，他有什麼道歉，那妳就原諒他*吧*。...妳剩下的留學期間也不多了*吧*，好好學習，好好玩啊！
>
> （資料來源：TOCFL 語料庫日籍 B2 學習者）

韓語的語氣表現方式與日語差不多，主要是以終結詞尾來表達說話者的態度和意圖（謝瑩 2011: 26）[25]，與漢語的句末語氣相似。謝瑩（2011: 27）認為韓國學生出於對於禮貌策略的敏感性，會主動選擇在句末添加語氣詞以達到表示禮貌的目的，於是出現語氣詞多用的偏誤，其中包括了「吧」的使用。本研究尚未考察語氣詞偏誤情況，不過，從上述日語學習者的這段語料，可以知道如果多用不至於構成偏誤，可以視為日語學習者，甚至是東亞學習者的語體風格。至於多用的因素，那可能是文化或語用因素導致學習者遷移而多用句尾語氣詞。至少從關鍵性分析結果來看，東亞語系背景學生在語氣詞使用的表現上，是較英語背景學生突出許多。

3.2.5 量詞

在量詞部分，英語學習者「個」的高關鍵性也是相當突出的[26]。相較於其他五個亞洲語言，英語雖也有度量衡詞或以容器表達的數量詞語，但對於歐美學生來說，漢語的量詞在英語並沒有相對應的形式，如，一「本」書、一「匹」馬。那麼，為什麼英語學習者反而在量詞「個」呈現不尋常高頻的現象，我們推測是教學時，這個語法上的結構「數詞＋量詞＋名詞」並不難，學習者知道要使用量詞，只是不一定能掌握量詞和名詞的搭配，量詞不一定用得正確，而「個」是漢語中最通用的量詞，如果不確定使用哪個量詞或學習者能力還不及，多會選擇「個」，因此歐美學生使用「個」的頻率可能還高於母語人士，這一點我們從張博等人（2008: 91）從中介語語料庫中得到的調查結果也可印證，他們從 1636 個歐美學生使用量詞的用例中，得出「個」的用例有 1076 句，佔總數的 66％，其中在名量搭配這類偏誤中，「個」的泛化，佔了 52.53%，也是比例最高的。

另外，我們從探討日、韓、泰學習者學習漢語量詞的文獻中發現，雖然這些語言有量詞這一類，但是名量的搭配、量詞的位置或使用量詞的條件與漢語不盡相同（宮下研介, 2011；金龍勛, 2011；宋帆, 2008），對這些學習者而言，難度並不比歐美學習者來得低，我們從張博等人（2008: 79-97）統計韓國學生和歐美學生的正誤用例，前者偏誤的比例（11%）還比後者（7.3%）來得高，也可以看出量詞不只對歐美學習者而言，是難點；對整體學習者而言，也是一樣的。

表 8 是 TOCFL 語料庫中六個子語料庫量詞的使用情況，可以看出英語學習者使用「個」這個量詞佔所有量詞用例的 46%左右，較其他母語學習者高了三至九個百分比左右。這個結果與張博等人利用英、韓中介語語料庫所調查出來的結果一致，不過，他們並沒有觀察到英語學習者較其他語言背景學習者在使用「個」的頻率突出表現。

---

[25] 終結詞尾是指位於一句話終了時，處於最後一個謂語的詞尾。除了表示這句話已經結束，還表達說話者的態度與意圖。

[26] 泰國學習者「本」也具有高關鍵性，不過，從表 2 中，可以知道是受「書」高頻率的影響，屬個別情況，這裡不做討論。

表 8. 不同母語背景學習者量詞「個」使用情況

	英	日	韓	越	泰	印尼
使用量詞種類	112	135	84	97	64	84
「個」使用次數	2405	3164	1099	1289	486	1086
使用量詞次數	5256	8162	2973	3410	1342	2541
「個」所佔比例	45.76%	38.77%	36.97%	37.80%	36.21%	42.74%

究竟是泛化（誤用）或是誤加？抑是英語學習者較日、韓語學習者更有意識地使用量詞？則需要進一步觀察分析。初步我們利用偏誤檢索工具觀察「個」誤加的情況[27]，發現在 33 個誤加用例中，西方學生用例佔了 60%；東方學生用例佔了 40%。似乎西方學生誤加的情況較東方學生嚴重，不過，這個誤加的情況不只是量詞「個」的問題，而是英語學習者會在不需要加數量結構時，使用「一個」，推測與英語冠詞 “a/the”的表達形式有關，所以會造出「我看見一個漂亮的風景」這樣的句子。

4. 結語與研究限制

本研究利用主題關鍵性方法觀察不同母語背景學習者在選詞、用詞上的特徵。這個方法讓我們快速、有目標地進一步分析不同母語者的用詞特徵，這些特徵包括英語學習者代名詞和能願動詞「會、可以」的使用突出性、相對少用句尾語氣助詞、多用量詞「個」；日韓學習者在假設句使用上優先選用「的話」形式、因果句優先選用「所以」等等。

這些關鍵性特徵大致上都可以從語言類型、語言結構、詞類、語用（文化）等面向上給予解釋，包括：（1）在選詞上，可以看出學習者傾向於選擇與自己母語類型接近的語言形式，例如，日語者在表達漢語因果關係時，優先選擇了「所以」，因為日語裡沒有像漢語中「因為…所以…」這樣成對的形式，只能用單一詞語，而日語中的だから，非黏著性而且位於句首的位置，最接近漢語的「所以」；相對於「因為」的ので，則是黏著語素，在日語的使用頻率上也遠低於だから。另外，還有假設條件句中「的話」形式的高關鍵性。（2）母語中是否有與漢語相對應的詞類，似乎也是影響選詞、用詞的因素之一，例如英語背景學習者有助動詞這類，以及沒有句尾語助詞這類所呈現的關鍵性表現。（3）母語中的文化因素，也會遷移到漢語的使用，例如，越南學習者使用漢語第二人稱敬語形式的高關鍵性。這些資訊都可以經過轉化做為提升教學成效的基礎。

這個研究原來只是個嘗試，幸運的是看到應用主題關鍵性方法的可行性，可以有效地找出不同母語背景學習者的語言特徵，也可以透過這個方法，讓二語習得研究者有針對性地探討分析。除此之外，主題關鍵性分析方法未來也可以嘗試應用在機器自動辨識文本寫作者的國籍訊息上（Jarvis & Crossley, 2012），進一步偵測出可能的偏誤（尤其是學習者誤加或遺漏成分）。

[27] 檢索網址為 http://kitty.2y.idv.tw/~hjchen/cwrite-error/

當然這個研究由於現階段語料量不大（一百七十多萬字），因此無法將不同母語者的語料再依不同能力（等級）區分來觀察，只能等待語料庫日漸茁壯後，可作更精緻的分析，例如，觀察學習者這些顯著特徵是否隨著能力不同而產生變化。最後，本研究限於時間與人力，僅整理出前 20 高關鍵性詞語，實際上可能還有更多待發現的關鍵特徵。即使是已經整理出的關鍵特徵，部分也因為限於筆者自身對這些外語的知識不足或找不到相關對比研究文獻而無法進一步解釋，或做深入的分析探討，例如，能願動詞「可以」和一些高頻副詞，未來需要更多研究者投入開發。

參考文獻

Anthony, L. (2009). Issues in the design and development of software tools for corpus studies: The case for collaboration. *Contemporary corpus linguistics*, ed. by P. Baker, 87-104. London, UK: Continuum Press.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Chao, Y.-R. (1968). *A grammar of spoken Chinese*. Berkeley/Los Angeles: University of California Press.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Holland: Foris Publications. Reprint. 7th Edition. Berlin and New York: Mouton de Gruyter, 1993.

Council of Europe [COE]. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Resla, 19*, 83-102.

Granger, S. (Ed.). (1998). *Learner English on computer*. London & New York: Longman.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition and foreign teaching.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hawkins, J. A., & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English profile programme. *Language testing matters: Investigating the wider social and educational impact of assessment, Studies in Language Testing*, ed. by L. Taylor, and C. J. Weir, 31, 158-175. Cambridge: UCLES/Cambridge University Press.

Jarvis, S., & Crossley, S. A. (Eds.) (2012). Approaching language transfer through text classification: Explorations in the detection-based approach. Bristol, UK: Multilingual Matters.

Jarvis, S., Gastaneda-Jimenez, G., & Nielsen, R.. (2012). Detecting L2 Writers' L1s on the Basis of Their Lexical Styles. *Approaching language transfer through text classification: Explorations in the detection-based approach*, eds. by Jarvis and Crossley, 34-70. Bristol, UK: Multilingual Matters.

Jin, H. G. (1994). Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer. *Language learning* , *44*, 101-122.

Lehmann, W. P. (Ed.). (1978). *Syntactic Typology: Studies in the phenomenology of language*. Hassocks: Harvester Press.

Li, C. N., & Thompson, S. A. (1976). Subject and topic: A new typology of language. *Subject and topic*, ed. by Charles N. Li, 457-489. New York: Academic Press.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *proceedings of the workshop on comparing corpora*, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), 1-6. Hongkong.

Scott, M. (1997). PC analysis of key words - and key key words. *System*, *25*(2), 233-245.

Scott, M. (2008). *WordSmith tools. Version 5*. Liverpool: Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Keywords and corpus analysis in language education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins.

Williams, R. (1976). *Keywords: a vocabulary of culture and society*. New York: Oxford University Press.

余志鴻（1992）。元代漢語的後置詞系統。*民族語文*，3，1-10。

宋帆（2008）。*漢泰語量詞比較研究和泰語量詞教學*。上海市：上海外國語大學語言學及應用語言學專業碩士論文（未出版）。

肖奚強（2001）。外國學生照應偏誤分析。*漢語學習*，2001 年第 1 期，50-54。

阮氏懷芳（2008）。*漢越人稱代詞對比研究*。桂林：廣西師範大學語言學及應用語言學專業碩士論文（未出版）。

林雪鳳（2008）。*泰國初級漢語學習者敘述體語篇銜接之研究*。廈門：廈門大學語言學及應用語言學專業碩士論文（未出版）。

周剛（2001）。漢、英、日語連詞語序對比研究及其語言類型學意義。*語言教學與研究*，2001 年第 5 期，42-54。

周曉芳（2011）。歐美學生敘述語篇中的回指習得研究過程。*世界漢語教學*，*25*(3)，422-432。

宮下研介（2011）。*漢日量詞對比研究—對日漢語中的量詞教學*。黑龍江：黑龍江大學漢語國際教育專業碩士論文（未出版）。

金龍勛（2011）。*中韓量詞對比研究*。蘇州：蘇州大學漢語國際教育專業碩士論文（未出版）。

徐晶凝（2003）。語氣助詞"吧"的情態解釋。*北京大學學報（哲學社會科學版）*，*40*(4)，143-148。

祖生利（2002）。元代白話碑文中助詞的特殊用法。*中國語文*，*5*，459-480。

張莉萍（2012）。對應於歐洲共同架構的華語詞彙量。*華語文教學研究*，*9*(2)，77-96。

張莉萍（2013a）。TOCFL 作文語料庫的建置與應用，載於崔希亮、張寶林（主編），*第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集*（頁 141-152）。北京：北京語言大學出版社。

張莉萍（2013b）。華語學習者句式使用情況分析。發表於 2013 台灣華語文教學學會年會暨國際學術研討會（2013.12），高雄：文藻外語大學。

張博等人（2008）。*基於中介語料庫的漢語詞彙與專題研究*。北京市:北京大學出版社。

詞庫小組（1995）。中央研究院漢語語料庫的內容與說明。*技術報告第 95-02/98-04 號*，中央研究院。

葉信鴻（2009）。*現代漢語助動詞的界定與教學應用*。台北：國立台灣師範大學華語文教學研究所碩士論文（未出版）。

郭麗娟（2011）。*中高級水平韓國留學生漢語連詞使用情況研究*。南京：南京師範大學漢語國際文化教育學院碩士論文（未出版）。

楊帆（2011）。*泰國學生漢語語篇銜接手段偏誤分析及教學研究*。重慶市：西南大學語言學及應用語言學專業碩士論文（未出版）。

鄧守信（2009）。*對外漢語教學語法*（修訂二版）。臺北市：文鶴出版社。

賴鵬（2006）。漢語能願動詞語際遷移偏誤生成原因初探。*語言教學與研究*，2006(5)，67-74。

謝瑩（2011）。*高級水平韓國留學生漢語語氣詞“吧、嗎、呢”偏誤分析*。上海：華東師範大學漢語國際教育碩士專業碩士論文（未出版）。

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

<table>
<tr><td>會員類別</td><td colspan="2">□終身 □個人 □學生</td><td>會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td>學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td colspan="2"></td><td>E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　（簽章）

中　華　民　國　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

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

2. 連絡處：

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

3. 本表可自行影印

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

Name: _____(Please print) Date: _____

Please debit my credit card as follows: US$ _____

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

Card No.: _____ -_____-_____ -_____ Exp. Date:_____(M/Y)

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

CARD HOLDER SIGNATURE: _____

Phone No.: _____E-mail: _____

Address: _____

PAYMENT FOR

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

Quantity Wanted: _____

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

Quantity Wanted: _____

US$ ____ ❑ Publications:_____

US$ ____ ❑ Text Corpora: _____

US$ ____ ❑ Speech Corpora:_____

US$ ____ ❑ Others: _____

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

US$ _____ = Total

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

ACLCLP
% IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　　日期:：_____

卡別：❑ VISA CARD　❑ MASTER CARD ❑ JCB CARD　發卡銀行：_____

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

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____ ❑ 中文計算語言學期刊(IJCLCLP) _____

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

NT$_____ ❑ 中研院詞庫小組技術報告_____

NT$_____ ❑ 文字語料庫 _____

NT$_____ ❑ 語音資料庫 _____

NT$_____ ❑ 光華雜誌語料庫1976~2010

NT$_____ ❑ 中文資訊檢索標竿測試集/文件集

NT$_____ ❑ 會員年費：❑續會　❑新會員　❑終身會員

NT$_____ ❑ 其他: _____

NT$_____ = 合計

填妥後請傳真至 02-27881638 或郵寄至:

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

Publications of the Association for Computational Linguistics and Chinese Language Processing

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

10% member discount: _____Total Due:_____

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

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address：_____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲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.19 no.2 June 2014

Contents

Papers

Social Metaphor Detection via Topical Analysis..... 1
Ting-Hao (Kenneth) Huang

Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem..... 17
Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen

Resolving the Representational Problems of Polarity and Interaction between Process and State Verbs..... 33
Shu-Ling Huang, Yu-Ming Hsieh, Su-Chu Lin, and Keh-Jiann Chen

不同母語背景華語學習者的用詞特徵：以語料庫爲本的研究... 53
張莉萍

ISSN: 1027-376X
The Association for Computational Linguistics and Chinese Language Processing