International Journal of

Computational Linguistics & Chinese Language Processing

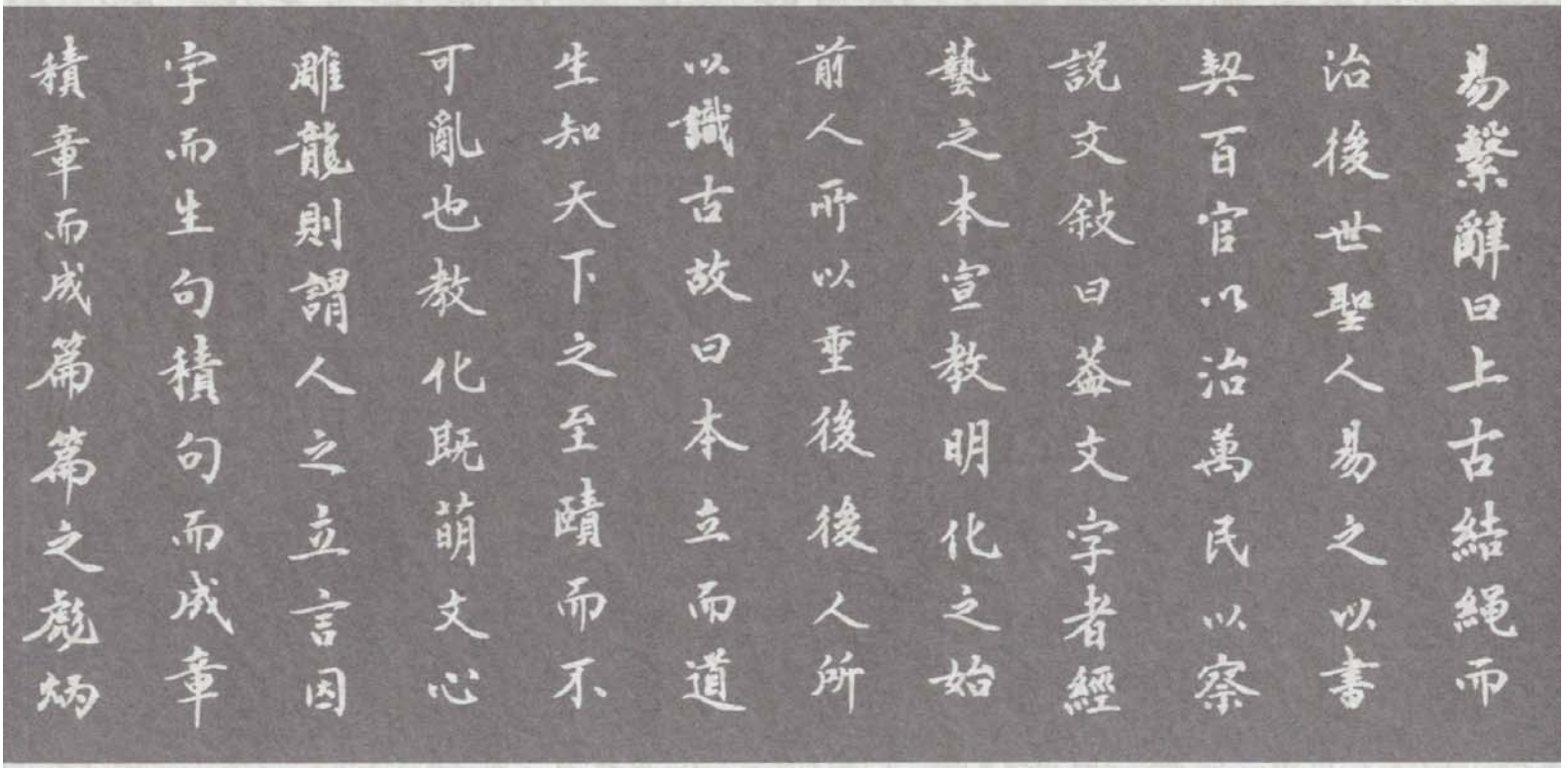
中文計算語言學期刊

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

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

Special Issue on "Selected Papers from ROCLING XXIII"

Guest Editors: Liang-Chih Yu, and Wei-Ho Tsai



Vol.17 No.2 June 2012 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

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

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. 17, no. 2, June 2012

Contents

Special Issue Articles:
Selected Papers from ROCLING XXIII

Papers

Forewords

The 23rd Conference on Computational Linguistics and Speech Processing (ROCLING 2011) was held at the National Taipei University of Technology, on September 8-9, 2011. ROCLING is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants to present their work and discuss recent trends in the field. This special issue presents extended and reviewed versions of four papers meticulously selected from ROCLING 2011.

The first paper "Transitivity of a Chinese Verb-Result Compound and Affected Argument of the Result Verb" proposes a method for predicting the transitive property of Chinese verb-result compounds, and identifying the predication potential between component verbs and arguments. The second paper "Predicting the Semantic Orientation of Terms in E-HowNet" uses a supervised learning method trained using many useful features extracted from multiple sources to predict the semantic orientation of terms taken from E-HowNet. The third paper "Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in Two Conversational Corpora" examines how different dimensions of corpus frequency data may affect the outcome of the statistical modeling of lexical items. Analysis is based on a corpus of elderly speakers, and the target words are temporal expressions. The fourth paper "Using Kohonen Maps of Chinese Morphological Families to Visualize the Interplay of Morphology and Semantics in Chinese" uses Kohonen Maps to analyze and visualize the interplay of morphology and semantics in Chinese. The results could be helpful to linguists in preparing adequate word lists for the behavioral study of Chinese morphological families.

The Guest Editors of this special issue would like to thank all of the authors and reviewers for their contributions. We would also like to thank all the researchers and participants for sharing their knowledge and experience at the conference.

Liang-Chih Yu, and Wei-Ho Tsai

Guest Editors

IJCLCLP Special Issue on "Selected Papers from ROCLING XXIII"

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 1-20

Transitivity of a Chinese Verb-Result Compound and Affected Argument of the Result Verb

You-shan Chung* and Keh-Jiann Chen*

Abstract

The Chinese verb-result compound is productive, but its meaning and syntactic behaviors have posed challenges to theoretical and automatic analyses. Theory-wise, the current study proposes that VRs have inherent affecting direction, which argument mapping principles and selectional restrictions, event structures, or the kinds of semantic/pragmatic principles or real-world knowledge proposed by previous researchers do not seem to account for. Application-wise, we predict the VR's transitivity and whether the result component is predicated of the logical subject or the logical object, based on the transitivity of individual component verbs and the selectional restrictions between component verbs and arguments. Since the transitivity property and selectional restrictions of individual verbs can be annotated in our lexicon, the rules should fare well in automatic processing. Meanwhile, as the rules have been motivated by linguistic theories and have been observed to make correct predictions in most cases, they are worthy of further large-scale testing.

Keywords: Verb-result Compound, Transitivity, Lexical-semantics, Meaning Prediction

1. Introduction

Verb-result compounds (VR[1]) are compounds that are comprised of an event (denoted by the first verb, henceforth V1) and the result of the event (denoted by the second verb, henceforth V2).[2] For example, *da-po* 打破 'broken from being hit,' *chi-bao* 吃飽 'full from eating,' and *tu-hei* 塗黑 'paint something black' are VRs, where *da* 打 'hit,' *chi* 吃 'eat,' and *tu*

* CKIP, Institute of Information Science, Academia Sinica, Taipei, Taiwan
Tel: +886-2-27883799 ext.1502 Fax:+886-2-27881638
E-mail: {yschung; kchen}@iis.sinica.edu.tw

[1] For a complete list of the abbreviations used in this paper, please refer to Appendix I.

[2] Despite slight variations in terminology, VRs roughly correspond to Smith's (1994) resultative verb compounds and Li and Thompson's (1981) Result RVC.

塗 'paint' lead to the results *po* 破 'broken,' *bao* 飽 'full,' and *hei* 黑 'black.' Due to limited space, two variants of VRs, directional verb compounds and phase/completive compounds (Smith, 1994), will not be discussed in this paper.

The high productivity of VRs in Chinese makes them a worthwhile topic of investigation for NLP. Nevertheless, this verbal construction poses challenges to traditional linguistic theories of the syntax-semantics interface aimed at mapping the meaning of verbal constructions to their surface structure. This is because most generative frameworks hold that mappings between event meaning representations and syntactic structure are governed by constraints imposed by the verb (e.g. theta theory, cf. Crystal 1997 for its definition). These constraints, however, do not always seem to be applicable to VRs. Since a VR contains two verbs, it is often argued that one of them is to be identified as the "head," which is supposedly responsible for the mapping. Nevertheless, such an assumption encounters difficulties. For one thing, the head is defined as the component that dominates the resulting meaning and syntactic behaviors, but it is controversial which component it is (i.e. V1 or V2) or even if such a component exists at all in Chinese VRs (Huang, 1998; Li, 2009). Among researchers who do not address head identification, some argue that mappings for VRs are not simply constrained by the argument structures of the component verbs (Yin, 2011; Li, 2007; Huang, 2006).

Various explanations have been proposed to accommodate VRs in a larger framework of syntax-semantics interface, including syntactic operations (Mulkeen, 2011; Shen & Lin, 2005), lexical-semantic properties of the components (Yin, 2011; Li, 2007; Huang, 2006), or a split approach to different VRs (Lin, 1998). Syntactic or lexical-semantic, these accounts all appeal either to extensions of certain underlying syntactic operations, such as movement, or to decisions on thematic role mapping. As far as automatic processing is concerned, the former means adjustments to the general parsing rules and the latter requires the incorporation of real-world knowledge.

In particular, not only do non-lexical accounts present technical challenges, but also a growing consensus in linguistic studies is that the meaning and syntactic behaviors of larger linguistic structures can be explained by the syntactic and semantic properties of the composing words. Along this lexical approach, Levin (1993) presented a heavily-cited framework to explain how the meaning and grammar that are encoded by verbs are mapped onto the syntactic arguments. Therefore, we began the analysis of the VR compound, a kind of verbal compound, with a lexical approach.

For the sake of accuracy and simplicity as a linguistic account and for ease of computer processing, our treatment of VRs is lexically-based. Unlike some lexical accounts, however, we avoid postulations of underlying representations. We address two basic syntactic and

semantic distinctions of VRs by employing information of verbs' transitivity[3] and selectional restrictions encoded in our on-line lexicon. The research questions are how to predict the transitivity of a VR and how to identify V2's affected argument (i.e., the argument predicated by V2).

2. The Prediction of Transitivity and V2's Affected Argument

Transitivity and selectional restrictions are two of the most basic syntactic and semantic distinctions for verbs. A special characteristic of Chinese VRs is that V2's affected argument can either be the logical subject (i.e., subject-controlled, henceforth SC) or the logical object (i.e., object-controlled, henceforth OC) of a VR. Such a contrast can be seen in the VRs *da-po* 打破 'broken from hitting' and *du-dong* 讀懂 'read something and understand it,' which belong to the OC and SC type, respectively. For example, in *Zhangsan da-po-le boli* 張三打破了玻璃 'Zhangsan broke the glass,' the affected argument of *po* 破 'broken' is *boli* 玻璃 'glass,' V1's object, whereas in *Zhangsan du-dong-le naben shu* 張三讀懂了那本書 'Zhangsan (finally) understood the book,' it is V1's subject, *Zhangsan* 張三, that is being predicated of by V2.

Below, we will briefly introduce how two lexical-semantic accounts (Li, 2007, Huang, 2006) and one lexical-syntactic account (Lin, 1998) address transitivity and V2's affected argument.

2.1 Review of Lexical-semantic and Lexical-syntactic Works on Chinese VRs

Lexical-semantic accounts like Li (2007) and Huang (2006) argue that event structure (and argument structure) can explain the meaning and syntactic behaviors of VRs through a set of linking rules that map event participants of various thematic roles (e.g., agent, patient, causer, causee) onto sentence positions (e.g., subject, object).

Li (1997)

Under Li's framework, transitivity of a VR compound and V2's affected argument fall out of the options in expressing the causing factor and the affected entity, assigned the thematic roles "causer" and "causee," respectively; the causee amounts to V2's affected argument in our terms. The roles are assigned by the external and internal arguments of V1 and of V2 as well

[3] A transitive verb is one that has a direct object, although the object does not have to be overt (i.e., occurring in the surface structure), whereas an intransitive verb does not allow a direct object at all (overt or covert). For example, the transitive verb *eat* has a direct object, although both *He ate a sandwich* and *He ate* are possible. In contrast, as an intransitive verb, *go* usually cannot have a direct object. (cf. http://www.glottopedia.de/index.php/Transitive.)

as by the event structure of the V1+V2 combination.

With the causer and causee, as well as their corresponding arguments (i.e., internal/external/neither), identified, a proposed set of linking rules determine the sentence positions for the two thematic roles to be realized. Below is one of them (Li, 2007: 96).

(1) The causer argument is realized in subject position and the causee argument in object position, when both arguments are overtly expressed by different linguistic expressions.

As can be seen, Li's reference point is the causer and the causee. Nevertheless, as far as automatic processing is concerned, the causer and causee have to be identified rather than taken as given. So, Li's rules are not readily applicable for a system that has no access to knowledge of causer and causee.

Besides methodological issues, Li does not explain his criteria for distinguishing between causer and causee, which his linking rules hinge on. As thematic roles, causer and causee should be defined on semantic grounds rather than by sentence positions. Under his analysis, *na-bao yifu* 那包衣服 'the bundle of clothes' in *Na-bao yifu xi-lei-le Zhangsan* 那包衣服洗累了張三 '(Zhangsan washed that bundle of clothes) and the clothes got Zhangsan tired' (Li, 2007: 119) and *shu* 書 'book' in *Zhangsan kan-lei-le shu* 張三看累了書 'Zhangsan read a book and as a result became tired' (Li, 2007: 115) differ in thematic roles. According to Li, in the former sentence, the causer is *na-bao yifu* 那包衣服 whereas in the latter, the causer is *Zhangsan* 張三. In terms of meaning, however, it appears that, in both sentences, it is *Zhangsan* 張三 who is the agent engaging in the activities *xi* 洗 'wash' and *kan* 看 'read,' from which he becomes tired. So, it is not clear why the causer-causee mappings in these two sentences are not the same.

Huang (2006)

There are two components in Huang's framework: event frames and linking rules. The event frames describe the obligatory and optional participants and predicates for transitive and intransitive verbs. These event participants map to certain syntactic categories and sentence positions by default. For example, the subject of a sentence is usually a noun phrase that receives an agent role.

Although Huang's account can explain mappings between thematic roles and sentence positions, his model needs to identify the transitivity of a VR first in order to determine the relevant event participants so as to discover the thematic roles of the subject and the object. Nevertheless, he does not address how to determine the transitivity of a VR compound.

Meanwhile, the issue of the definitions of causer and causee arises again. In *Baozhi kan-hua-le ta de yian* 報紙看花了他的眼 'The newspaper got his eyes blurred from reading it' (Huang, 2006: 27), Huang regards *baozhi* 報紙 'newspaper,' instead of *ta* 他 'he,' who engages in the action that leads to the result state, as the causer. It is not clear why *baozhi* 報

紙 here, *yifu* 衣服 in *Na-bao yifu xi-lei-le Zhangsan* 那包衣服洗累了張三 and *shu* 書 in *Zhangsan kan-lei-le shu* 張三看累了書 from Li (2007) differ in thematic roles. These examples suggest lack of a consistent way to identify causer and causee between researchers and within individual researcher's analyses while such criteria underpin their theories.

Lin (1998)

Lin (1998) predicts a VR's transitivity and V2's affected argument based on V1's transitivity and whether V2 is predicative of animate or inanimate entities. This approach seems promising for automatic processing since transitivity properties and selectional restrictions can be annotated.

Regarding transitivity, she argues that the transitivity value of a VR is the same as that of V1. Her prediction of the affected argument, however, is nondeterministic, as exemplified in her analysis of transitive V1s paired with intransitive V2s predicative of animated noun phrases, e.g., *zhui-lei* 追累 'chase-tired' and *chi-bao* 吃飽 'full from eating.' She argues that, in such cases, the V2 can be predicative of either the subject or the object, or both. For example, she thinks that *Zhangsan zhui-lei-le Lisi* 張三追累了李四 allows three readings, which are, with descending degrees of acceptability, 'Zhangsan chased Lisi and Lisi became tired,' 'Zhangsan chased Lisi and Zhangsan became tired,' and 'Lisi chased Zhangsan and Lisi became tired.' In contrast, *chi-bao* 吃飽 can only have an SC reading. Nevertheless, she does not explain the conditions for a reading to be available, or at least preferred.

In sum, regarding theoretical comprehensiveness, Li's and Huang's identification of causer and causee seems arbitrary. In terms of automatic processing, their predictions depend on already knowing what automatic processing needs to figure out, whereas Lin's account does not address disambiguation.

Despite these suggested theoretical and practical inadequacies, many such lexical-semantic accounts have turned from relying solely on argument structure to what argument structure and selectional restrictions (i.e., the semantic restrictions on what arguments can be taken by a verb) cannot explain. Within the lexical tenets, the general consensus is that, besides the argument structures of individual verbs, the VR construction has its own event structure with thematic roles to assign. The roles from the two structures (i.e., argument structure and event structure) conflate to decide the resulting meaning and syntactic behaviors of the VR. Also frequently considered is the variation in the kinds of composing predicates that exist and whether they are optional/obligatory or general/specific in the event structures of VRs[4] of different event types, consequently taking semantic and syntactic variations between different verbs and within the same verb into account.

[4] Examples of actual representations can be found in Li (2007: 95) and Huang (2006:21).

When these interactions and variations allow ambiguities, it is suggested that semantic and pragmatic, or real-world knowledge[5], motivations screen or rank the remaining readings. Some researchers, like Li (2009) and Huang (2002), only cite the role of real-world knowledge in determining whether a VR is possible or not but leave the question of whether an acceptable compound can occur in a sentence with certain noun phrases in a certain order or not and, if it can, what the predication relation is to syntactic, semantic, or pragmatic constraints. For example, Li proposes the Animate Subject as Agent or Experiencer Strategy (ASAES) as well as the principle of prototypicality, both of which he considers pragmatic principles. The content of the former has been made clear by its name; the latter states that the entity that carries out the V1-denoted act is less likely to be the V2-affected entity. Both principles are meant to rank the possibility of acceptable readings. Other researchers contend that real-world knowledge remains an element in the reading(s) of a legal VR. Huang (2006) incorporates the role of real-world knowledge in the event frame of VRs, which is the interface where linking rules are applied to map semantic configurations to sentence positions.

The current study suggests an even more central role for real-world knowledge in the interpretation of VRs. So far as we are aware, Huang's (2006) constructional framework is one of the few accounts that formally represents real-world knowledge, if not the only account. Nevertheless, even he thinks that real-world knowledge is only an optional participant in the event frame. In spite of this, we not only agree with Huang that real-world knowledge remains at play in deciding the reading(s) of a legal VR, but we also have found real-world knowledge to be central to the meaning and syntactic behaviors of VRs throughout its formation and interpretation. We will support this position with examples in Section 3.

2.2 The Model of the Current Study

We first propose an affecting direction principle to model the transitivity and affected argument of VR compounds. Then, we develop a set of heuristic rules to emulate the principle using only the information of individual transitivity of V1 and V2 and the selectional restrictions of V1 and V2 but not using real-world knowledge. Based on a lexical approach, we encode the semantic and syntactic properties of each word in the lexicon. Transitivity is one of the listed properties, and so are the selectional restrictions. We manually encode such information based on observations of corpus data. For example, *da* 打 'hit' requires an animate agent and a physical theme. This is shown in (2):

[5] Some researchers use the terms "real-world knowledge" and "pragmatic constraints" interchangeably. Li (2007), however, seems to distinguish between them.

(2) *Zhangsan da Lisi* 張三打李四

Zhangsan hit Lisi

'Zhangsan hit Lisi.'

In contrast, the verb *po* 破 'broken' requires an inanimate theme (PERF=perfective marker):

(3) *Boli po le* 玻璃破了

glass broken-PERF

'The glass is broken.'

Thus, the lexical entries of *da* 打 'hit' and *po* 破 'broken' would be partially represented as (4) below (Vt=transitive verb; Vi=intransitive verb; [+ANI]=animate entity; [+PHY]=physical entity; [-ANI]=inanimate entity):

(4) *da* 打 'hit'

Transitivity: Vt

Selectional restrictions: [+ANI] agent, [+PHY] theme

po 破 'broken'

Transitivity: Vi

Selectional restrictions: [-ANI] theme

With the encoded information, when the computer encounters the compound *da-po* 打破 'broken from being hit,' the rules of our model can predict the VR's transitivity and V2's affected argument. Nevertheless, such information cannot always predict the resulting meaning and syntactic behaviors of VRs. We think this is because the fundamental factor that decides a VR's transitivity and V2's affected argument is the affecting direction of the compound.

By "affecting direction" we refer to native speakers' knowledge of the range of entities that can be affected by the state resulting from an action or another state and of whether the VR can take an argument/arguments or other noun phrases or not. For instance, Mandarin Chinese speakers know that the entity affected by of the state denoted by *chi-bao* 吃飽 'full from eating' is an animate being that eats and that the VR cannot be followed by noun phrases inside or outside of V1's argument structure. In formal terms, we say that *chi-bao* 吃飽 'full from eating' is intransitive and SC. Such knowledge is implicit, with linguistic and non-linguistic motivations underlying it. We will discuss the motivations in more detail in Section 3.

The affecting direction principle:

Case 1: When V1 is Vt

Vt+R → VRt and SC; if the affecting direction of the VR is toward Vt's subject, e.g., *da-ying* 打贏 'fight and win,' *du-dong* 讀懂 'read something and understand it'

Vt+R → VRt and OC; if the affecting direction of the VR is not toward Vt's subject, e.g., *jiao-hui* 教會 'teach someone and make him/her understand'

Case 2: When V1 is Vi

Vi + Rt → VRt and SC, e.g., *pao-shu* 跑輸 'lose in running'

Vi + Ri → VRi and SC; if the affecting direction of the VR is toward Vi's subject, e.g., *pao-lei* 跑累 'tired from running'

Vi + Ri → VRt and OC; if the affecting direction of the VR is toward Ri's subject, e.g., *pao-po* 跑破 'broken from running'

Two points of clarification are in order here. First, there are some cases where the affecting direction seems to allow both an SC and an OC reading. Two of the oft-cited cases are *qi-lei* 騎累 'ride-tired' and *zhui-lei* 追累 'chase-tired,' as in *Zhangsan qi-lei le ma* 張三騎累了馬 and *Zhangsan zhui-lei-le Lisi* 張三追累了李四, both allowing either the "Zhangsan is tired" reading or the "horse/Lisi is tired" reading. Such VRs are believed to be ambiguous in the sense that they seem to have two readings in the same sentence. Nevertheless, we consider the two readings of these VRs as being separate lexical entries, i.e., polysemous. This is based on Huang *et al*.'s (2005) criterion for distinguishing different words from different meanings of a word, which states that senses that cannot co-exist in the same context have to be treated under different lexical entries. Accordingly, while *qi-lei* 騎累 and *zhui-lei* 追累 are possible with two readings in the above sentences, the ambiguity can be resolved with a moderate number of contextual clues.

(5) 騎累 $_1$

Zhangsan pa qi-lei-le ai-ma, jueding rang ta xiuxi 張三怕騎累了愛馬，決定讓牠休息

Zhangsan fear ride-tired-PERF beloved-horse, decide let it rest

'Fearing that it might be exhausted from extended rides, Zhangsan decided to let his beloved horse take a break.'

騎累 $_2$

Zhangsan qi-lei-le ma, juede haishi kai-che bijiao qingsong 張三騎累了馬，覺得還是開車比較輕鬆

Zhangsan ride-tired-PERF horse, feel nevertheless drive-car relatively relax

'Tired from horse-riding, Zhangsan found driving more relaxing.'

A remaining question is how to decide which *qi-lei* 騎累 or *zhui-lei* 追累 is relevant when the computer encounters such VRs. Indeed, currently, we can only identify them as transitive verbs that are polysemous between an SC and OC reading. Context is required to decide the affected argument of V2.

Second, when the affecting direction of a VR is not toward the Vt's subject, it is not necessarily toward its object.

In a few cases, it is the adjunct of the Vt that is V2's affected object. For example, *qie* 切 'cut' has an agent-denoting subject that is an animate being, e.g., *Zhangsan* 張三, a patient-denoting object that is the physical object being cut, e.g., *rou* 肉 'meat,' and an instrument-denoting adjunct that is the tool for cutting, e.g., *dao* 刀 'knife.' For the VR *qie-dun* 切鈍 'blunt from cutting,' it is the instrument-denoting adjunct instead of the patient-denoting object that is affected by *dun* 鈍 'blunt.' Nevertheless, although the instrument is not the object of the Vt, the prediction that the VR is OC is still borne out since V2's affected argument is indeed the object of the VR *qie-dun* 切鈍 because *dao* 刀 is the receiver of the action[6] described by the compound.

When V1 is Vi, again, the noun phrase that follows the VR is not necessarily V1's object. Actually, it does not even have to be any possible argument, adjunct, or complement in the verb frame. To show this, we may have to first make clear what is meant by being SC and OC. It may be suggested that the affecting direction of the VR *pao-lei* 跑累 'tired from running' is not fixed if sentence pairs like *Zhangsan pao-lei-le* 張三跑累了 'Zhangsan ran and as a result got tired'and *Zhangsan pao-lei-le liang-tiao tui* 張三跑累了兩條腿 'Zhangsan ran, and as a result his legs got tired' are taken into account. In the former sentence, the affecting direction should be toward Zhangsan 張三, but, in the latter sentence, it seems to be either toward *Zhangsan* 張三 or *liang-tiao tui* 兩條腿. As a result, the latter sentence seems to have both an SC and an OC reading.

Our view is that both readings belong to the SC type. First, we find *liang-tiao tui* 兩條腿 'legs' in the sentence to be more like a complement than an object of *pao-lei* 跑累 'tired from running.' The first reason is that, for a noun phrase to be allowed to follow the VR *pao-lei* 跑累, it has to specify a body part of the subject, thus functions like a subject complement, as defined in Crystal (1997). Second, it seems that "Zhangsan's legs being tired" entails "Zhangsan being tired." To the extent that Zhangsan is also tired, the sentence can be labeled as the SC type.

Sometimes, the NP in question is not V1's argument but is still the VR's object. For example, VRs with intransitive V1s like *bing-huang* 病慌 'sick-nervous' and *ji-bing* 急病

[6] cf. Crystal 1997 for the definition of object as "the receiver of the action."

'worry-sick' can have OC readings, as in *Zhangsan bing-huang-le Lisi* 張三病慌了李四 'Zhangsan's illness made Lisi nervous,' *Zhangsan ji- bing-le Lisi* 張三急病了李四 'Zhangsan got Lisi so worried, who became sick as a result.' It is difficult to treat 李四 as a participant in the verb frames of *bing* 病 'sick' and *ji* 急 'worried.' Another probably more drastic example is *ku-zou* 哭走 'to cause someone to leave by crying.' Although *ku* 哭 cannot be subcategorized for objects, the VR obligatorily take objects and is possible with an OC reading only, as in *Zhangsan ku-zou-lie Lisi* 張三哭走了李四 'Zhangsan cried, making Lisi leave.' Note in passing that, in Section 2, we argued for the treatment of the OC and SC readings of *qi-lei* 騎累 and *zhui-lei* 追累 as belonging to different lexical entries. On this account, *bing-huang* 病慌 and *ji-bing* 急病 are also regarded as different VRs.

Since VRs cannot be exhaustively listed, the affecting direction also cannot be exhaustively annotated. Thus, it has to be predicted, too. The affecting direction can generally be predicted by the transitivity properties and selectional restrictions of the two component verbs. Transitivity and selectional restrictions are as exemplified in (4) and are annotated in the lexicon. Therefore, rules based on this information are workable. Where the affecting direction cannot be predicted, however, the VR's transitivity and V2's affected argument become unpredictable as a result. Nevertheless, the results have seemed to be as predicted most of the time so far, suggesting that the affecting direction is usually predictable.

2.3 The Heuristics

With the possibility of unpredictability in mind, we are ready to look at the following heuristics for the automatic prediction of transitivity and V2's affected argument.

Case 1: When V1 is Vt

(a) Vt + Rt → VRt and SC, e.g., *da-ying* 打贏 'fight and win,' *du-dong* 讀懂 'read something and understand it'

(b) Vt + Ri → VRt and OC; if Vt's object or adjunct satisfies the selectional restrictions of Ri's subject, e.g., *da-po* 打破 'broken from hitting,' *qie-dun* 切鈍 'blunt from cutting'

(c) Vt + Ri→ VRt and OC; if Vt's subject and object both satisfy the selectional restrictions of Ri's subject, e.g., *qi-lei*$_1$ 騎累$_1$ 'tired from being ridden,' *zhui-lei*$_1$ 追累$_1$ 'tired from being chased,' *jiao-fan*$_1$ 教煩$_1$ 'vexed from being taught'

(d) Vt + Ri→ VRt and SC; if Vt's subject satisfies the selectional restrictions of Ri's subject, e.g., *qi-lei*$_2$ 騎累$_2$ 'tired from engaging in the activity of riding,' *zhui-lei*$_2$ 追累$_2$ 'tired from engaging in the activity of chasing,' *jiao-fan*$_2$ 教煩$_2$ 'vexed from engaging in the activity of teaching'

The following rules deal with VRs where V1 is an intransitive verb.

Case 2: When V1 is Vi

(e) Vi + Rt → VRt and SC, e.g., *pao-shu* 跑輸 'lose in running'

(f) Vi + Ri → VRi and SC; if Vi's subject satisfies the selectional restrictions of Ri's subject, e.g., *pao-lei* 跑累 'tired from running'

(g) Vi + Ri → VRt and OC; if Vi's subject does not satisfy the selectional restrictions of Ri's subject, e.g., *pao-po* 跑破 'broken from running'

Below, we show how the VR's transitivity and V2's affected argument might be predicted by matching a component verb's transitivity and selectional restrictions with the other's. Again, we *already* know the transitivity and selectional restrictions of the component verbs before the computer encounters a context that contains a VR construction. This means the judgment of the VR's characteristics, except for the cases where both (c) and (d) apply, generally does not depend on on-line context, but we will see exceptions in the following sentences (6), (11), (12), and (15) and will address them in Section 3.

The following sentences from Li (2007) represent some of the most frequent VR patterns, where the subjects and objects are default ones that meet the selectional restrictions of the component verbs. In (6)-(12), V1 is a transitive verb. We will describe the transitivity properties before describing the matching of the selectional restrictions. The relevant rule is given at the end.

(6) *Zhangsan zhui-lei le* 張三追累了

Zhangsan chase-tired-PERF 'Zhangsan chased (someone) and became tired.'

Prediction: *zhui* 追 is Vt and *lei* 累 is Ri. Since *Zhangsan* 張三 can be the subject of *lei* 累, *zhui-lei* 追累 is VRi and SC. (Affecting direction, as predicted by (c) and (d), is ambiguous between VRt and OC and VRi and SC; it is disambiguated by the sentential context[7].)

(7) *Zhangsan ca-liang le boli* 張三擦亮了玻璃

Zhangsan wipe-shiny-PERF glass

'Zhangsan wiped the glass shiny.'

Prediction: *ca* 擦 is Vt and *liang* 亮 is Ri. Since *Zhangsan* 張三 cannot be the subject of *liang* 亮, and the object of *ca* 擦 can be the subject of *liang* 亮, *ca-liang* 擦亮 is VRt and OC. (Rule (b))

[7] As mentioned in Section 2.2, the VR is polysemous between an OC reading (*i.e. zhui-lei*$_1$ 追累 $_1$) and an SC reading (i.e. *zhui-lei*$_2$ 追累 $_2$). Nevertheless, while the heuristics are not sufficient to decide the relevant reading, the sentential context of (6) requires the SC reading. The case of (15) is similar to (6) except that, in the former, the relevant rule (i.e. rule (f)) fails to recognize ambiguities.

(8) *Zhangsan qie-dun-le dao* 張三切鈍了刀

Zhangsan cut-blunt-PERF knife

'Zhangsan cut (something) with the knife, and as a result the knife became blunt.'

Prediction: *qie* 切 is Vt and *dun* 鈍 is Ri. Since *Zhangsan* 張三 cannot be the subject of *dun* 鈍, and the adjunct of *qie* 切 can be the subject of *dun* 鈍, *qie-dun* 切鈍 is VRt and OC. (Rule (b))

(9) *Zhangsan du-dong-le na-shou shi*. 張三讀懂了那首詩

Zhangsan read-understand-PERF that-CL poem

'Zhangsan read and understood that poem.'

Prediction: *du* 讀 is Vt and *dong* 懂 is Rt; *du-dong* 讀懂 is VRt and SC. (Rule (a))

(10) *Zhangsan jiao-fan-le Lisi* 張三教煩了李四

Zhangsan teach-vexed-PERF Lisi

'Zhangsan taught Lisi and as a result Lisi felt vexed.' (教煩 $_1$, OC reading)

'Zhangsan taught Lisi and as a result Zhangsan felt vexed.' (教煩 $_2$, SC reading)

Prediction: *jiao* 教 is Vt and 煩 *fan* is Ri. Since *Zhangsan* 張三 and *Lisi* 李四 do not differ in the possibility of being the arguments of the two verbs, *jiao-fan* 教煩 is predicted to be VRt and either OC (Rule (c)) or SC (Rule (d)). (Context beyond the sentential level is needed for disambiguation.)

(11) *Zhangsan jiao-hui-le Lisi* 張三教會了李四

Zhangsan teach-know-PERF Lisi

'Zhangsan taught Lisi (something), and as a result Lisi learned it.'

Prediction: *jiao* 教 is Vt and *hui* 會 is Rt. Since *Zhangsan* 張三 and *Lisi* 李四 do not differ in the possibilities of being the arguments of the two verbs, *jiao-hui* 教會 is predicted to be VRt and either OC (Rule (c)) or SC (Rule (d)). Nevertheless, as the affecting direction of *jiao-hui* 教會 is toward the object of V1, *jiao-hui* 會 is VRt and OC. (The affecting direction is not predicted by rules.)

(12) *Zhangsan e-bing-le* 張三餓病了

Zhangsan hungry-sick-PERF

'Zhangsan became sick as a result of being hungry.'

Prediction: *e* 餓 is Vt[8] and *bing* 病 is Ri. Since *Zhangsan* 張三 can be the subject of *bing* 病, *e-bing* 餓病 is predicted to be VRi and SC. (The affecting direction, as predicted by (c) and (d), is ambiguous between a VRt and SC and a VRt and OC reading; it is disambiguated by the sentential context.)

---

[8] We think that there are two polysemous *e* 餓. Besides the intransitive sense, sentences like *Baba guyi e xiaohai* 爸爸故意餓小孩 'The father deliberately made/makes his children starve' suggests the existence of a transitive *e* 餓.

In sentences (13)-(15), V1 is an intransitive verb:

(13) *Zhangsan zou-lei-le tui* 張三走累了腿

Zhangsan walk-tired-PERF leg

'Zhang walked until his legs got tired.'

Prediction: *zou* 走 is Vi and *lei* 累 is Ri. Since *Zhangsan* 張三 can be the subject of *lei* 累, *zou-lei* 走累 is predicted to be VRi and SC (Rule (f)).

(14) *Zhangsan ku-ya-le sangzi* 張三哭啞了嗓子

Zhangsan cry-hoarse-PERF throat

'Zhangsan cried his throat hoarse.'

Prediction: *ku* 哭 is Vi and *ya* 啞 is Ri. Since *Zhangsan* 張三 cannot be the subject of *ya* 啞, *ku-ya* 哭啞 is VRt and OC. (Rule (g))

(15) *Zhangsan bing-huang-le Lisi* 張三病慌了李四

Zhangsan sick-nervous-PERF Lisi

'Zhangsan's being sick got Lisi nervous.'

Prediction: *bing* 病 is Vi and *huang* 慌 is Ri. Since *Zhangsan* 張三 can be the subject of both, *bing-huang* 病慌 is predicted to be VRi and SC. Nevertheless, the affecting direction of *bing-huang* 病慌 can be either toward the subject or the object. Therefore, *bing-huang* 病慌 can be either VRi and SC or VRt and OC. (The affecting direction is not predicted by rules; it is disambiguated by the sentential context.)

The predictions of rules (a)-(g) are borne out most of the time. Sometimes, though, these heuristics cannot account for the affecting direction. We will discuss some such cases.

3. Discussion and the Applicability of the Rules

Although language processing can benefit from the heuristics, such rules sometimes fail because they are based on lexical information, to which much of the real-world knowledge, which plays an important role in deciding the affecting direction of VRs, is invisible. As opposed to the views of Li (2007) and Huang (2002), we found real-world knowledge to be no less important in the interpretation of a VR than in its possibility of existence because both a VR's occurrence and its interpretation presume a reasonable cause-effect relationship. Notably, we suggest that the subjectivity involved in such decisions makes it only natural that definite readings are sometimes non-existent. While believing that the affecting direction is largely determined by world knowledge, we also noticed that syntactic constraints regulate how real-world knowledge can be expressed.

Recall that *jiao-hui* 教會 'teach someone and make him/her understand' in (11) is

predicted by (c) and (d) to be either SC or OC, but turns out to be OC only because the affecting direction of the VR is toward the object. It might be asked if one of the verbal components has contributed to the affecting direction of the VR. As far as *jiao-hui* 教會 is concerned, the following sentences show that the affecting direction is a co-product of V1 and V2.

(16) a. *Zhangsan jiao-fan-le Lisi* 張三教煩了李四

Zhangsan teach-vexed-PERF Lisi

'Zhangsan taught Lisi and as a result Lisi felt vexed.' (教煩 $_1$, OC reading)

'Zhangsan taught Lisi and as a result Zhangsan felt vexed.' (教煩 $_2$, SC reading)

b. *Zhangsan jiao-fan-le ying-wen* 張三教煩了英文

Zhangsan teach-vexed-PERF English

'Zhangsan taught English and as a result Zhangsan felt vexed.'

(17) *Zhangsan xue-hui-le gangqin* 張三學會了鋼琴

Zhangsan learn-know-PERF piano

'Zhangsan learned to play the piano.'

That *jiao-hui* 教會 'teach someone and make him/her understand' and *jiao-fan* 教煩 'teach and got (someone) vexed' differ in the affected arguments indicates that the affecting direction at least is not determined by V1 in all instances. Likewise, a comparison between *jiao-hui* 教會 and *xue-hui* 學會 'learned something' shows that the affecting direction of the VR in question is not determined by V2 alone, either. Rather, it appears that real-world knowledge as a result of the *composition* of V1 and V2 deems it unlikely to teach oneself and make oneself understand, hence the inherent OC typing of *jiao-hui* 教會.

Li's principle of prototypicality states that the one who acts usually is not the one that is affected. While he calls it a "pragmatic principle," we think it is still derived from experience living in the world. To the extent that the principle of prototypicality is derived from real-world knowledge is true, it is still an abstraction that competes with other kinds of real-world knowledge. Consider the following four sentences where our heuristics make nondeterministic (i.e., (18) and (20)) or wrong predictions (i.e., (19) and (21)).

(18) *Zhangsan zhui-lei le* 張三追累了

Zhangsan chase-tired-PERF

'Zhangsan chased (someone) and became tired.'

Prediction: *zhui* 追 is Vt and *lei* 累 is Ri. Since *Zhangsan* 張三 can be the subject of *lei* 累, *zhui-lei* 追累 is VRi and SC. (The affecting direction, as predicted by (c) and (d), is ambiguous between VRt and OC and VRt and SC; it is disambiguated by the sentential context.)

(19) *Zhangsan jiao-hui-le Lisi* 張三教會了李四

Zhangsan teach-know-PERF Lisi

'Zhangsan taught Lisi (something), and as a result Lisi learned it.'

Prediction: *jiao* 教 is Vt and *hui* 會 is Rt. Since *Zhangsan* 張三 and *Lisi* 李四 do not differ in the possibility of being the arguments of the two verbs, *jiao-hui* 教會 is predicted to be VRt and either OC (Rule (c)) or SC (Rule (d)). Nevertheless, as the affecting direction of *jiao-hui* 教會 is toward the object of V1, *jiao-hui* 會 is VRt and OC. (The affecting direction is not predicted by rules.)

(20) *Zhangsan e-bing-le* 張三餓病了

Zhangsan hungry-sick-PERF

'Zhangsan became sick as a result of being hungry.'

Prediction: *e* 餓 is Vt and *bing* 病 is Ri. Since *Zhangsan* 張三 can be the subject of *bing* 病, *e-bing* 餓病 is predicted to be VRi and SC. (The affecting direction, as predicted by (c) and (d), is ambiguous between a VRt and SC and a VRt and OC reading; it is disambiguated by the sentential context.)

(21) *Zhangsan bing-huang-le Lisi* 張三病慌了李四

Zhangsan sick-nervous-PERF Lisi

'Zhangsan's being sick got Lisi nervous.'

Prediction: *bing* 病 is Vi and *huang* 慌 is Ri. Since *Zhangsan* 張三 can be the subject of both, *bing-huang* 病慌 is predicted to be VRi and SC. Nevertheless, the affecting direction of *bing-huang* 病慌 can be either toward the subject or the object. Therefore, *bing-huang* 病慌 can be either VRi and SC or VRt and OC. (The affecting direction is not-predicted by rules; it is disambiguated by the sentential context.)

As far as Li's (2007) two principles for ranking possible readings are concerned, (19)-(22) all satisfy ASAES (Animate Subject as Agent or Experiencer Strategy) because this paper only deals with such VR patterns. The other principle, i.e., prototypicality, makes SC readings less likely. Nevertheless, SC readings are not only preferred but are the only possibility in (19) and (21). Therefore, the knowledge that "the one who acts is unlikely to be the one that is affected" is not only insufficient to rank readings but actually allows impossible readings. That (19) and (21) allow SC readings seems to come from real-world knowledge that verifies the possibility for the one who acts (i.e., *Zhangsan* 張三) to be the affected one.

A comparison between (21) and (22) suggests that the ultimate difficulty in pinpointing the most possible reading of a VR lies in the fluidity of real-world knowledge. Such fluidity sometimes comes from subjectivity. Take the readings of *e-bing* 餓病 'hungry-sick' and *bing-huang* 病慌 'sick-nervous' for example. It seems that, for a VR composed of two Vis,

like *bing-huang* 病慌, to have an OC reading, it must have a causative reading. As for the Vt+Vi combination *e-bing* 餓病, the OC reading would be 'someone makes another person starve so that this starving person becomes sick.' While both assuming a causative interpretation under the OC readings, the two VRs contrast in two ways. First, in *e-bing* 餓病, both composing verbs take on causative readings, whereas in *bing-huang* 病慌, only *huang* 慌 has a causative reading. Second, in the absence of context, the SC reading seems more natural than an OC reading for *e-bing* 餓病. Nevertheless, if such a bias is present at all, it is not as pronounced in *bing-huang* 病慌. This is unexpected in structural terms because, while both VRs have V2s that are Vis, the V1 *e* 餓 in *e-bing* 餓病 is a transitive verb, which supposedly is more likely to affect an entity other than the one who acts (thus, have an OC reading) than the intransitive V1 *bing* 病 in *bing-huang* 病慌 does. We are led to the hypothesis that the contrasts could result from judgment of plausibility based on real-world knowledge: *E-bing* 餓病 is less natural with a causative reading because people generally do not intentionally make others suffer; it is unusual to starve people. On the other hand, in the case of *bing-huang* 病慌, *bing* 病 is involuntary and cannot be put in intentional terms, while it is human to find another person's suffering disturbing.

As can be seen, the above judgment is a reflection of the perception of the world. Thus, it is expected that the perceived probability of OC and SC readings will be as varied as people's perception of the world. A definite OC or SC judgment sometimes could be non-existent.

Finally, we will discuss two cases that seem to be consistent with real-world knowledge but are either downright unacceptable or odd. Sentence (22) presents a reading that is a possible real-world scenario but linguistically unacceptable:

(22) *Zhangsan zhui-lei-le Lisi* 張三追累了李四

Zhangsan chase-tired-PERF Lisi

'Zhangsan chased Lisi and as a result Zhangsan got tired.'

In terms of real-world knowledge, it is possible that Zhangsan chased Lisi and as a result Zhangsan got tired, but such a reading is not as natural as one where Lisi is tired.

Another example is the intransitive-and-SC-only reading of *chi-bao* 吃飽 'full from eating.' Consider the following sentence:

(23) ?*Zhangsan chi-bao-le ji* 張三吃飽了雞

Zhangsan eat-full-PERF chicken

'*Zhangsan ate (something) and the chicken became full.' (OC reading)

'?Zhangsan ate the chicken and became full.' (SC reading)

Since both *chi* 吃 'eat' and *bao* 飽 'full (from eating)' can take animate subjects, the VR is predicted to be transitive and either SC or OC. Nevertheless, the impossibility of

becoming full from someone else's eating rules out the transitive-and-OC reading. As for why the transitive-and-SC reading is also impossible, however, real-world knowledge does not seem to be a factor, as there is nothing unusual about eating something and becoming full as a result.

For now, we have not been able to explain the underlying motivations for every (im)possible reading. Nevertheless, the proposed rules so far have made the correct prediction most of the time in the VRs we encountered.

4. Conclusion

In the current study, we explain how we predict a VR's transitivity and identify the argument predicated by the second verb. We propose a set of rules that are motivated by a lexical-semantic analysis. Although these rules have not been tested against a corpus, they are testable and are worth testing. This is because linguistic analyses suggest their theoretical credibility and because these rules have a uniform formalism and are modular in the sense that they use two kinds of formally-represented information, i.e., 1) the transitivity of V1 and V2 and 2) whether the arguments can be the subject and object of the two component verbs or not. If these rules prove to achieve high prediction rates, they can be readily applied.

References

Chen, K.-J., Huang, S.-L., Shih, Y.-Y., & Chen, Y.-J. (2005). Extended-HowNet-A representational framework for concepts. In *Proceedings of OntoLex 2005*, 1-6.

Crystal, D. (1997). *Dictionary of phonetics and linguistics*. Oxford: Wiley-Blackwell.

Huang, C.-T. J. (2002). Resultatives and unaccusatives: a parametric view. *Bulletin of the Chinese Linguistic Society of Japan*, 253, 1-43.

Huang, H.-C. (2006). A constructional approach to argument realization of Chinese resultatives. *UST Working Papers in Linguistics*, 2, 13-31.

Huang, S. (1998). Chinese as a headless language in compounding, In J. L. Packard (Ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, 261-284. Berlin: Mouton De Gruyter.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Li, C. (2009). On the headedness of Mandarin resultative verb compounds. *Concentric: Studies in Linguistics*, 35(1), 27-63.

Li, C. (2007). *Mandarin resultative verb compounds: Where syntax, semantics, and pragmatics meet*. Ph. D [Dissertation]. Yale University, New Haven, USA. [Online]. Available: ProQuest. [Accessed: 2011/11/13].

Li, C. N. & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Taipei: Crane.

Lin, H.-L. (1998). *The syntax-morphology interface of verb-complement compounds in Mandarin Chinese*. Ph. D [Dissertation]. University of Illinois at Urbana-Champaign, Urbana-Champaign, Illinois, USA. [Online]. Available: ProQuest. [Accessed: 2011/7/13].

Mulkeen, S. (2011). A conflation account of mono-clausal resultatives in English and Chinese. MA thesis. National Cheng-Kung University. [Online]. Available: http://www.airitilibrary.com/searchdetail.aspx?DocIDs=U0026-0802201119204700 [Accessed:2012/4/3]

Shen, L., & Lin, T.-H. J. (2005). Agentivity agreement and lexicalization in Resultative Verbal Compounding, Paper Space, 2005. [Online]. Available: http://ling.nthu.edu.tw/faculty/thlin/pdf/Shen_Lin.pdf. [Accessed: 2011/7/9].

Smith, C. C. (1994). *The parameter of aspect*. New York: Kluwer Academic Press.

Yin, H. (2011). The semantic structures of Chinese verb-resultatives. *International Journal of English Linguistics*, 1(2), 126-133.

黃居仁（2005）主編。《意義與詞義》系列。中文的意義與詞義。中央研究院語言所文獻語料庫與資訊所中文詞知識庫小組技術報告 05-01。

詞庫小組。廣義知網知識本體架構線上瀏覽系統。[Online]. Available: http://ehownet.iis.sinica.edu.tw [Accessed: 2012/4/3]

Appendix I

Abbreviations

[+ANI]=animate entity

[-ANI]=inanimate entity

CL=classifier

LOC=locative preposition

[+PHY]=physical entity

OC=object-controlled

PERF=perfective marker

R=result

Rt= transitive verb that denotes a result

SC=subject-controlled

Vi=intransitive verb

VRi= VR is an intransitive verb

VRt＝VR is a transitive verb

Vt=transitive verb

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 21-36

廣義知網詞彙意見極性的預測

Predicting the Semantic Orientation of Terms in E-HowNet

李政儒*、游基鑫、陳信希

Cheng-Ru Li, Chi-Hsin Yu, and Hsin-Hsi Chen

摘要

詞彙的意見極性是句子及文件層次意見分析的重要基礎，雖然目前已經存在一些人工標記的中文情緒字典，但如何自動標記詞彙的意見極性，仍是一個重要的工作。這篇論文的目的是爲廣義知網的詞彙自動標記意見極性。我們運用監督式機器學習的方法，抽取不同來源的各種有用特徵並加以整合，來預測詞彙的意見極性。實驗結果顯示，廣義知網詞彙意見極性預測的準確率可到達92.33%。

關鍵字: 廣義知網，情緒分析，情緒字典, 語義傾向, 向量支援機

Abstract

The semantic orientation of terms is fundamental for sentiment analysis in sentence and document levels. Although some Chinese sentiment dictionaries are available, how to predict the orientation of terms automatically is still important. In this paper, we predict the semantic orientation of terms of E-HowNet. We extract many useful features from different sources to represent a Chinese term in E-HowNet, and use a supervised machine learning algorithm to predict its orientation. Our experimental results showed that the proposed approach can achieve 92.33% accuracy.

Keywords: E-NowNet, Sentiment Analysis, Sentiment dictionary, Semantic orientation, SVM

* 國立台灣大學資訊工程系
Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {crlee, jsyu}@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

1. 緒論

情緒分析（Sentiment Analysis）在現今的網路世界中，有許多實際且重要的運用，例如從網路的評論文章中分析消費者對產品的評價，或分析消費者對產品性能的關注焦點等等。不管對句子或文件層次的情緒分析，意見詞詞典都是一個重要的資源。通常意見詞詞典是用人工來收集詞彙，並用人工標記詞彙的各種情緒屬性，包括主客觀（subjective or objective）、極性（orientation/polarity)、）及極性的強度（strength）(Esuli & Sebastiani, 2005)。這些情緒屬性對不同的應用有不同的重要性，標記難度也各不相同，通常詞彙的極性是最容易進行標記的屬性。

標記情緒屬性時，研究者可以從零開始收集詞彙以建立意見詞詞典，如台大意見詞詞典 NTUSD(Ku & Chen, 2007)。在另一方面，也有研究者嘗試爲自然語言處理中的許多現存的資源，添加情緒屬性，如 SentiWordNet(Esuli & Sebastiani, 2006a)。但現有資源的語彙量通常很大，如 WordNet 3.0 就包括 206,941 個不同的英文字義（word-sense pair），要全部用人工進行標記之成本太高。因此，通常的作法是少量標記一些詞彙，再用機器學習方法，爲剩下的詞彙進行自動標記，雖然自動標記的準確率不如人工標記，但對一般應用有某種程度的幫助。

在中文自然語言處理，NTUSD 是一部重要的意見詞詞典，但此詞典只包括詞彙及極性的資訊。另一方面，董振東先生和陳克健教授所建立的知網和廣義知網(Z. Dong & Dong, 2006; 陳克健, 黃, 施, & 陳, 2004)，是重要的語意資源。對於每個詞彙，都用有限的義原給予精確的定義，但這些定義卻缺乏情緒的語意標記。因此，如何自動爲廣義知網加上情緒標記，成爲一個重要的課題，也是本研究的目的。

本研究提出爲廣義知網加上情緒標記的方法，首先利用 NTUSD 跟廣義知網詞彙的交集建立標準答案集，再由標準答案集訓練出分類器，爲其他廣義知網詞彙進行標記。如何有效的運用監督式機器學習演算法，如何爲詞彙抽取出有用的特徵，是主要的挑戰議題。在此研究中，我們有系統的嘗試抽取各種不同的詞彙特徵，最後得到高準確率的二元分類器（binary classifiers）用以自動標記正負面情緒標記。

第二節介紹廣義知網、及英文和中文相關的情緒屬性標記研究，第三節介紹從 E-HowNet 及 Google Chinese Web 5-gram 抽取特徵的方法，第四節呈現各種實驗的結果及分析，包括跟 NTUSD 人工標記的比較，最後總結論文的成果。

2. 相關研究

董振東先生於 1998 年創建知網（HowNet），並在 2003 年，跟中央研究院資訊所詞庫小組在 2003 年，將中研院詞庫小組詞典（CKIP Chinese Lexical Knowledge Base）的詞條跟知網連結，並作了一些修改，最後形成廣義知網（Extended-HowNet, E-HowNet[1]）。詞庫小組修改並擴展知網原先的語義義原角色知識本體，建構出廣義知網知識本體

[1] http://ehownet.iis.sinica.edu.tw/

（Extended-HowNet Ontology），並用這些新的語義義原，以結構化的形式來定義詞條，詞條定義式的例子如圖 1。

有關情緒屬性標記的研究，我們分爲英文及中文來討論。在英文方面，最早是由 Hatzivassiloglou & McKeown(1997) 在 1997 年針對形容詞所做的研究，他們所用的形容詞分別有正面詞 657 個及負面詞 679 個，該論文依據不同的實驗設定，監督式機器學習的準確率（Accuracy）由 82% 到 90%。之後陸續有不同的研究，所用多爲半監督式機器學習的演算法(Esuli & Sebastiani, 2006b; Kamps, Marx, Mokken, & De Rijke, 2004; Turney & Littman, 2003)，效能從 67%到 88%不等，但因爲這些演算法所用的資料集並不相同，實驗過程及評估標準也不一樣，（有用 Accuracy、Precision、或 F-Measure），所以效能沒有辦法直接比較。

```
<Word item = "汽油">
   <WordFreq>15</WordFreq>
   <WordSense id="1">
      <English>gasoline</English> <Phone>ㄑㄧˋ ㄧㄡˊ</Phone> <PinYin>qi4 you2</PinYin>
      <SyntacticFunction> <POS>Naa</POS> <Freq>15</Freq> </SyntacticFunction>
      <TopLevelDefinition>{material|材料:attribute={StateLiquid|液態},telic={burn|焚燒}}</TopLevelDefinition>
      <BottomLevelExpansion>{material|材料:attribute={StateLiquid|液態},telic={burn|焚燒}}</BottomLevelExpansion>
   </WordSense></Word>
```

圖 1.「汽油」的廣義知網定義式

在中文的情緒屬性標記相關研究，Yuen et al.(2004) 2004 年利用 Turney & Littman(2003) 的半監督式機器學習演算法，在正面詞 604 個及負面詞 645 個的資料集上做實驗，得到最高的成績是 80.23%的精確度及 85.03%的召回率。之後從 2006 到 2010 年，陸續的研究使用不同的資料集，用不同類型的機器學習演算法來處理這個問題(Han, Mo, Zuo, & Duan, 2010; Li, Ma, & Guo, 2009; Lu, Song, Zhang, & Tsou, 2010; Yao, Wu, Liu, & Zheng, 2006)，所得到的效能在不同的指標（Accuracy、Precision、或 F-Measure）下，從 89%到 96%不等。因爲基準不同，這些效能一樣沒有辦法直接比較，但相較於英文，成績則明顯提高。

3. 特徵抽取及機器學習演算法

由於我們運用監督式機器學習演算法來訓練二元分類器（binary classifier），最重要的問題是爲詞彙抽取出有用的特徵。在此論文中，我們分別從 E-HowNet 及 Google Chinese Web 5-gram 這兩個來源抽取兩大類的特徵，接著將這兩個來源的特徵組合訓練分類器。此外，我們也嘗試使用組合式的監督式機器學習演算法（ensemble approach），來更進一步得到更高的效能，以下我們分別詳細介紹。

3.1 基礎義原特徵

從 E-HowNet 抽取的特徵稱之爲基礎義原特徵，也就是對每一個 E-HowNet 的詞彙 i，用一向量 $V_i = (w_{i,j}) = (w_{i,1}, w_{i,2}, \cdots, w_{i,n})$ 表示，其中 n 爲向量的維度。

由於每一詞彙的每一個語意（sense）都有一個結構化的定義式，而且定義式中都用義原來進行定義，公式 (1) 定義 V_i 中每個特徵的權重：

$$w_{i,j} = \begin{cases} 1, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \quad (1)$$

以圖 1「*汽油*」這個詞彙爲例，其定義式中出現了義原 ***material***，所以它的值 $w_{\text{汽油, material}}$ 就會是 1，其他沒出現的義原，值就會是 0。我們共使用了 2567 個義原來當特徵。

廣義知網的詞彙有歧異性，也就是每個詞彙可能有許多語意。而詞彙的第一個語意，是出現頻率最高的語意（除了四個詞彙例外），所以我們用詞彙的第一個語意來抽取特徵。只從詞彙的一個語意抽取特徵，而不把該詞彙所有的語意放在一起，代表這種方法可爲不同的語意給出不同的極性預測。只是由於目前 NTUSD 極性標記只到詞彙的層級，所以無法對語意的層級進行極性預測。但只要有語意層級的極性標記，我們這種做法可馬上套用。

3.1.1 基礎義原特徵加權值

除了公式 (1) 的方式外，我們可以利用更多 E-HowNet 的特性，來抽取出有用的資訊。一個可能的方式是定義式中的結構，如果把定義式展開，會得到如圖 2 的樹狀結構。在這樹狀結構中，義原所在的深度是一個有用的資訊，因此我們仿照劉群&李素建(刘 & 李, 2002)的公式，將深度的資訊當作權重引入公式 (1)，得到公式 (2)。

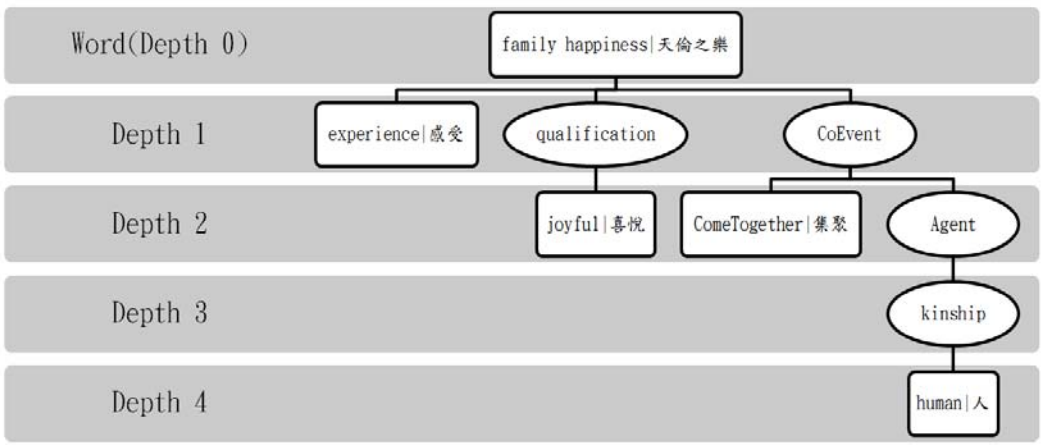


圖 2.「天倫之樂」定義式的樹狀表示

$$w_{i,j} = \begin{cases} \dfrac{1}{1+\alpha \times d_{i,j}}, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \quad (2)$$

公式 (2) 中，α 是可調的參數，$d_{i,j}$ 是詞彙 i 跟義原 j 的距離，這可用義原 j 的深度表示。調整公式 (2) 中的 α ，讓我們可以實驗那一種方式，才應給較高的權重：

（可能一）$\alpha < 0$: 深度越深，表示該義原有較多資訊，應給較高權重。

（可能二）$\alpha > 0$: 深度越深，表示該義原有較少資訊，應給較少權重。

由於 $\alpha < 0$ 時，$w_{i,j}$ 可能變爲負值，所以最小的 α 設爲 −0.05。另外，當 α = 0，公式 (2) 會等於公式 (1)，所以我們在做實驗時，只要使用公式 (2) 即可。

3.1.2 加入否定關係調整特徵的加權值

在計算義原深度時，可能會經過帶有否定意義的關係，例如「*一事無成*」定義式中有「{*not*({*succeed*|*成功*})}」，可以發現 *succeed* 被 *not* 所修飾。這時，義原 *succeed* 的權重用負值來表示可能會更好，因此我們將否定的概念引入公式 (3) 如下：

$$w_{i,j} = \begin{cases} \dfrac{Neg_{i,j}}{1+\alpha \times d_{i,j}}, & \text{如果定義式 i 中出現義原 j} \\ 0, & \text{不出現義原 j} \end{cases} \quad (3)$$

其中，$Neg_{i,j}$ 表示義原 j 是否有被否定意義的關係所修飾，若有則 $Neg_{i,j}$ 爲 −1，若無則 $Neg_{i,j}$ 爲 +1。另外，如果樹狀結構上面的義原被否定意義的關係所修飾，這否定意義會傳遞到下面的義原。

3.2 語篇（context）特徵

廣義知網雖然有嚴謹的定義式可用以表示詞彙，但是有四個缺點，造成只用義原當特徵無法正確獲得詞彙的極性。

第一個缺點是詞彙所標的義原量太少，因爲詞彙是用人工標示義原，所以無法給予很多標示。這表示詞彙擁有的資訊量有限，會造成分類器無法有效學習。第二個缺點是義原數量太少，這會造成語義的劃分不夠精確，無法顯示出真實的語義差別，例如「*明哲保身*」跟「*見風轉舵*」的定義式都是「*{sly|狡}*」，但「*明哲保身*」是正面意見，「*見風轉舵*」卻是負面意見。第三個缺點是廣義知網定義標準的差異，例如，專有名詞在廣義知網中會用客觀的義原來定義，但該專有名詞經過使用，卻可能會引起人的正反情緒（如「莫札特」是專有名詞，但卻常用來當正面意見），這種差異會引入程度不等的雜訊到分類器中。第四個缺點是廣義知網尚未對所有詞彙標上定義式，例如「*乾淨俐落*」在廣義知網及 NTUSD 中都出現，但廣義知網卻沒有標上定義式。

因此我們引入語篇的特性，從該詞彙在語言中的實際使用情況，抽取出詞彙的特徵，來補償這些缺點。我們使用 Liu et al. (F. Liu, Yang, & Lin, 2010) 所建立的 Google Web

5-gram Version 1，來抽取語篇特徵。Google Web 5-gram 是 Google 從網路中收集大量的簡體中文網頁，並經過處理所建立的資源。他們收集了 882,996,532,572 個字符（token），共 102,048,435,515 個句子，經過斷詞後建成 n-gram。n-gram 的 n 從 1 到 5，並且只保留頻率大於 40 的 n-gram。Google Web 5-gram 的例子如圖 3 所示。

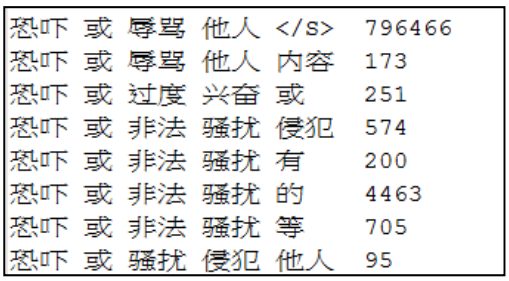


圖 3. Google Web 5-gram 資料範例

上圖中，表示「*恐吓 或 非法 骚扰 的*」這一 5-gram 共出現了 4463 次。從圖中我們也可看到，Google Web 5-gram 是簡體中文，但廣義知網爲繁體中文，所以我們先將廣義知網用 Microsoft Word 翻譯爲簡體中文，之後才使用 Google Web 5-gram 這一語料庫。語料庫在使用時，只用 5-gram 的部分來抽取特徵。

3.2.1 Google Web 5-gram特徵抽取

我們使用特徵跟詞彙的同出現（co-occurrence）次數做爲特徵值，以圖 3 爲例，如果詞彙是「*恐吓*」，以「*非法*」當特徵值，則同出現次數會將所有「*恐吓*」及「*非法*」一同出現的 5-gram 次數相加。在上面的例子中，「*恐吓*」及「*非法*」的同出現次數爲 574+200+4463 + 705=5942 次。

另外，由於廣義知網跟 Google Web 5-gram 的斷詞標準並不一致，所以在處理時把 Google Web 5-gram 的空白去掉，直接找出「*詞彙*」跟「*特徵*」這兩字串是否同時出現，來計算次數，這樣可以避免斷詞標準不一所產生的問題。例如「*一事無成*」在 Google Web 5-gram 中被斷成四個獨立的詞，將空白去掉就可以正確比對到。

因爲這裡的詞彙集合就是等待標示極性的詞，所以我們只要指定特徵的集合包括那些詞，就可算出表示詞彙 i 的向量 $V_i = (c_{i,j}) = (c_{i,1}, c_{i,2},\ldots, c_{i,m})$。其中，m 是特徵集合的大小，$c_{i,j}$ 是「*詞彙 i*」跟「*特徵 j*」這兩字串同出現的次數。在我們的實驗中，共嘗試了十種不同的特徵集合，分別是廣義知網的名詞、廣義知網的動詞、廣義知網的副詞、廣義知網的形容詞、廣義知網所有詞彙、Google Web 5-gram 最常出現的 5000 詞、Google Web 5-gram 最常出現的 5000 詞（但詞彙長度最少爲 2）、Google Web 5-gram 最常出現的 10000 詞、Google Web 5-gram 最常出現的 10000 詞（但詞彙長度最少爲 2）、以及 NTUSD 完整版。

3.2.2 Google Web 5-gram特徵值處理

用 $V_i = (c_{i,1}, c_{i,2},\ldots, c_{i,m})$ 的方式來表示的缺點，是 $c_{i,j}$ 的值變化的範圍會非常大，最小爲40，最大會到上千萬。這在機器學習中，通常需要做進一步的處理才會有比較好的結果。我們實驗了兩個不同的方法來處理這一問題：第一種是一般的餘弦標準化（cosine-normalization），將原本的向量 V_i 用公式 (4) 處理；第二種是 Esuli & Sebastiani (2005) 所提的餘弦標準化 TFIDF （cosine-normalized TF-IDF），他們用該方法來處理 WordNet 中的詞彙的權重，如公式 (5) 所述。

$$CosNorm(V_i) = \frac{V_i}{\sqrt{\sum_{1\le k\le m} c_{i,k}^2}} \quad \in \Re^m \tag{4}$$

$$CosNorm(TFIDF_i) = \frac{TFIDF_i}{\sqrt{\sum_{1\le k\le m} tfidf_{i,k}^2}} \quad \in \Re^m \tag{5}$$

$$TFIDF_i = (tfidf_{i,1}, tfidf_{i,2}, ..., tfidf_{i,m})$$

$$tfidf_{i,j} = tf_{i,j} * idf_j$$

$$tf_{i,j} = \frac{c_{i,j}}{\text{特徵 } j \text{ 總出現次數}} = \frac{c_{i,j}}{\sum_{k\in D} c_{k,j}}$$

$$idf_j = \log(df_j)^{-1} = \log\frac{|D|}{|\{i : c_{i,j} > 0, \forall i \in D\}|}$$

公式 (5)中 D 表示文件的集合，此處把詞彙 i 當成文件，特徵 j 當成 term。

公式 (4) 的標準化可以讓所有詞彙的向量等長，消掉次數變化過大的缺點。公式 (5) 的想法則認爲特徵 j 的權重，應該先跨詞彙進行標準化（normalization），所以 $tf_{i,j}$ 會除以特徵 j 的總出現次數，另外再考慮特徵 j 的稀有度，所以乘上 idf_j，最後再讓所有詞彙的向量等長。我們會在後面的實驗中，比較這兩種不同權重處理方式的效能。

3.3 不同特徵的組合

我們用了基礎義原特徵 $(w_{i,1}, w_{i,2},\ldots, w_{i,n}) = (w_{i,j})$ ，及語篇特徵 $(c_{i,1}, c_{i,2},\ldots, c_{i,m}) = (c_{i,j})$ 來表示詞彙 i。如果想同時使用這兩種特徵中的資訊，一種直觀的方式，是將兩種特徵表示方式混合，用 $V_i = (w_{i,1}, w_{i,2},\ldots, w_{i,n}, c_{i,1}, c_{i,2},\ldots, c_{i,m})$ 來表示。由於基礎義原特徵及語篇特徵都有許多不同的變形，我們無法一一嘗試所有可能的組合，所以會先分別用實驗找出最好的基礎義原特徵 $(w_{i,j})$ 及語篇特徵 $(c_{i,j})$，再把兩種特徵混合來進行實驗。我們沒有對混合後的向量做其它的處理，只是直接串接成爲 n+m 維向量。

3.4 組合式的監督式機器學習演算法（ensemble approach）

由於廣義知網詞彙的每一個意義（sense）都標有詞性，而且我們用了很多不同的特徵集

合，這表示我們會有很多不同的分類器。如果依不同詞性選擇做得最好的分類器，則可以有更好的效能。例如，如果分類器 A 的總體效能不是最好，但如果它對名詞做的效能是最好的，也許拿它來預測名詞的極性會更準確，依此類推。我們把廣義知網的詞性，分爲名詞、動詞、副詞、形容詞及其他共五類，分別選在該類別預測最好的分類器來預測。這作法是一種常見的組合不同分類器的策略（ensemble approach），我們也會對此進行實驗，來觀察效能。

4. 實驗與分析

4.1 實驗資料與實驗設定

本研究使用國立台灣大學意見詞詞典完整版（NTUSD）、與廣義知網的交集，作爲實驗資料，這兩個資料集的詞彙數如表 1。資料集 E-HowNet∩NTUSD 會作爲標準答案集（只包含正負面意見詞彙，不包含中立詞彙），在我們所看的相關論文中，這個答案集的大小是最大的一個。實驗使用標準答案集其中的 80% 爲訓練資料集，其餘 20%爲測試資料集，並依照實驗資料的詞性分布以及語意極性分布作分層抽樣（stratified sampling）。

表1. 廣義知網、NTUSD、以及交集的資料筆數

資料集	正面	負面	總數
E-HowNet	N/A	N/A	88,127
NTUSD	21,056	22,750	43,806
E-HowNet∩NTUSD	5,346	6,256	11,602

分層抽樣詳細的作法如下：先將資料依照五種詞性以及兩種極性分成十個子集合，再針對每個子集合取其中 80%作爲訓練資料，另外 20%作爲測試資料。由於我們的資料量夠多，所以可以使用這種抽樣。這種抽樣主要的好處在於我們更容易對測試結果進行更多的分析，我們把分層抽樣的結果列於表 2。

表2. 訓練資料的詞性以及傾向分布

詞性		全部資料集		正面傾向百分比	訓練資料集		測試資料集	
		正面	負面		正面	負面	正面	負面
名詞	2,040	931	1,109	45.64%	745	887	186	222
動詞	9,056	4,134	4922	45.65%	3,307	3,938	827	984
副詞	383	206	177	53.79%	165	142	41	35
形容詞	74	45	29	60.81%	36	23	9	6
其他	49	30	19	61.22%	24	15	6	4
總數	11,602	5,346	6,256	46.08%	4,277	5,005	1,069	1,251

本研究使用 Chang & Lin(2001) 所發布的 LIBSVM 支援向量機，來當監督式機器學習演算法，使用 radial basis function (RBF) kernel function， RBF 的兩個手動參數 cost c 與 gamma g 以網格搜尋（Grid Search）的方式尋找最佳參數值 (c, g)，搜尋範圍 $c \in \{2^{-5}, 2^{-3}, 2^{-1}, ..., 2^{15}\}$、$g \in \{2^{-15}, 2^{-13}, 2^{-11}, ..., 2^{-3}\}$，總共 110 組參數，取五疊交叉驗證（5-fold cross validation）中平均準確率最高的參數。

我們使用預測準確率（accuracy）來比較分類器間的效能，這是看訓練出的分類器在測試資料集中的成績，而分類器會對測試資料集中的所有詞彙都進行極性的預測。另外，使用 McNemar 檢定(Dietterich, 1998)來測試分類器的效能差距是否爲顯著，顯著水準設定爲 0.95。

McNemar 檢定將測試資料依照兩個分類器(以下稱爲分類器 A 與分類器 B)的標記，分成四組並計數。其中測試樣本數即爲下面 $n_{1,1}$、$n_{0,1}$、$n_{1,0}$、$n_{0,0}$ 四個數字的總合，在虛無假設（null hypothesis）中，兩個分類器應具有相同的錯誤率，也就是 $n_{0,1}=n_{1,0}$ 。

$n_{1,1}$：	分類器 A 與分類器 B 皆正確標記的樣本數	$n_{0,1}$：	分類器 A 標記錯誤，但分類器 B 標記正確的樣本數
$n_{1,0}$：	分類器 B 標記錯誤，但分類器 A 標記正確的樣本數	$n_{0,0}$：	分類器 A 與分類器 B 皆錯誤標記的樣本數

McNemar 檢定建構在卡方適合度檢定（$\chi 2$ test goodness of fit）上，整理而得的檢定值爲 $\frac{(|n_{0,1}-n_{1,0}|-1)^2}{n_{0,1}+n_{1,0}}$ ，此檢定值在 $n_{0,1}+n_{1,0}$ 夠大的時候會趨近於自由度爲 1 的卡方分配，因此在顯著水準（significant level）爲 0.95 時，此值若大於 $\chi^2_{1,0.95} = 3.8415$，則拒絕虛無假設。我們用 (McNemar 檢定結果, p-value) 來顯示我們的檢定結果，例如檢定結果 (1.50, 0.22) 表示，McNemar 檢定結果爲 1.50 < 3.84，所以沒有通過 McNemar 檢定，p-value 爲 0.22。

4.2 基礎義原特徵的效能

圖 4 爲基礎義原方法在不同 α 值所得到的預測準確率，其中公式 (2) 的結果是 PBF（Prime-Based Feature）那條折線，最佳的 α 值爲 −0.02，準確率爲 89.4397%。當 PBF 中 α = 0，該結果即爲公式 (1) 的結果。公式 (3) 的結果是 PBFN（Prime-Based Feature with Negation）那條折線，最佳的 α 值爲 −0.02 及 −0.03，準確率爲 89.6121%。

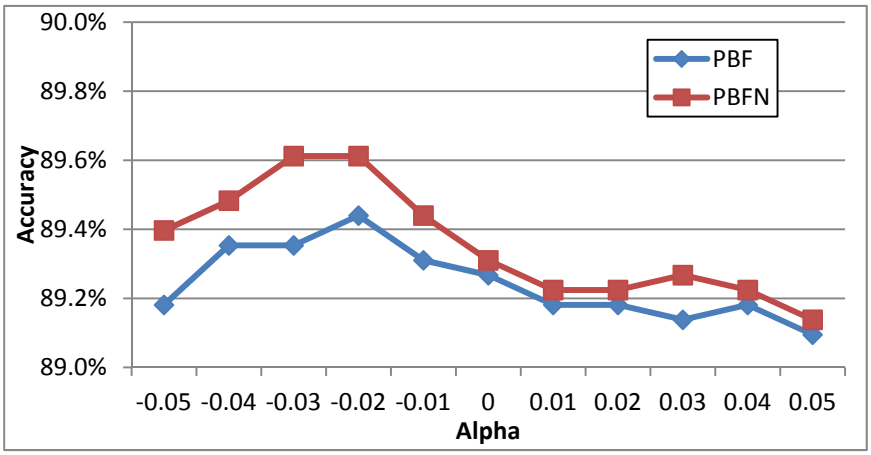


圖 4. 廣義知網特徵於不同 α 值的效能比較

我們從圖 4 可以看出，描述 PBFN 的折線在所有的 α 值下，準確率皆略高於 PBF，但是兩個最大值($\alpha = -0.02$)的差距僅 0.1724%，此差距爲不顯著，檢定結果 (1.50, 0.22)。由於 $\alpha < 0$ 有最佳效能，這表示深度較深給較高權重，該義原有較好的特徵，可以給分類器學習。

4.3 語篇特徵的效能

語篇特徵使用十組特徵集的名稱，以及特徵數量，如表 3 所示。在表中，我們使用特徵集代號來代表該特徵集。十組特徵集中，最少的是 *Adj* 的特徵集，只有 948 個詞，最多的是 *All* 的特徵集，有 86,712 個詞。

表 3. 語篇特徵所使用的特徵集與其特徵數

特徵集	特徵集代號	特徵數
廣義知網名詞	*Noun*	46,807
廣義知網動詞	*Verb*	37,109
廣義知網副詞	*Adv.*	2,364
廣義知網形容詞	*Adj.*	948
廣義知網所有詞彙	*All*	86,712
最常出現 5000 詞	*F5000-1*	5,000
最常出現 5000 詞（長度≧2）	*F5000-2*	5,000
最常出現 10000 詞	*F10000-1*	10,000
最常出現 10000 詞（長度≧2）	*F10000-2*	10,000
NTUSD（完整版）	*NTUSD*	42,614

我們使用三種不同的加權方式得到的預測準確率如圖 5，圖中我們也把特徵集的特徵數由左至右由小到大排列。

從圖 5 可以看出，沒有標準化的原始頻率的最佳準確率爲 59.70%，使用的特徵集爲「廣義知網名詞」，其效能最差且差距很大。餘弦標準化 TFIDF 的效能排在中間，最佳準確率爲 83.41%，使用的特徵集爲「最常出現 10000 詞」。而經過餘弦標準化的特徵值則可以得到最佳效能，其最佳準確率爲 88.23%，此時使用的特徵集爲「廣義知網動詞」，此效能跟其他兩者的差距爲顯著，檢定結果 (4.61, 0.03)。

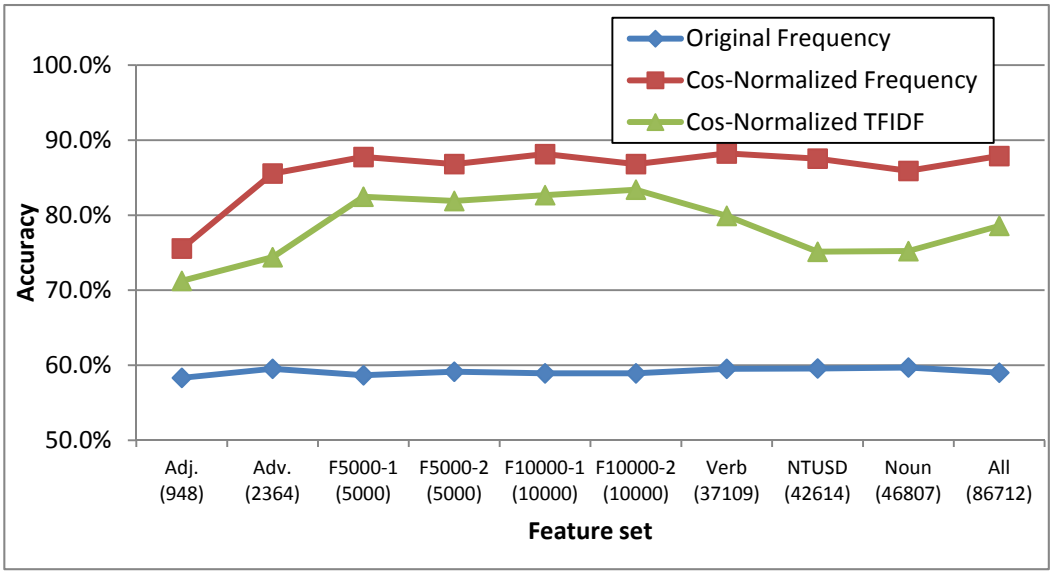


圖 5. 使用語篇特徵時的預測效能

圖 5 中特徵集的個數，並沒有絕對的影響，但若個數太少，如特徵個數小於 2364 個，則效能會明顯變差。圖 4 中的最佳值 PBFN（$\alpha = -0.02$）爲 89.61%，特徵個數爲 2,567 個，這個值比圖 5 中的最佳值 88.23%還要大，這表示廣義知網中的特徵比較準確，但這差距爲不顯著，檢定結果 (2.49, 0.11)。

4.4 組合不同特徵的效能

組合特徵時，因爲餘弦標準化有最好的效能，所以語篇特徵選擇餘弦標準化後的十組特徵集，分別與廣義知網特徵效能最好的 $\text{PBFN}_{\alpha=-0.03}$ 組合，來訓練分類器，分類器預測準確率如圖 6。其中廣義知網特徵的特徵集效能爲固定，因此以水平直線表示（gloss 那條折線）。組合而成的特徵集，以「*語篇特徵集代碼*+$\text{PBFN}_{\alpha=-0.03}$」加以命名，例如「*F10000-2*+$\text{PBFN}_{\alpha=-0.03}$」表示「最常出現 10000 詞（長度≧2）」跟「$\text{PBFN}_{\alpha=-0.03}$」兩個特徵集的組合。

我們從圖 6 可以看出，將廣義知網特徵與外部語料特徵組合之後，準確率都有顯著

提升，提升後的最高準確率爲 92. 3276%，使用「廣義知網所有詞彙 *All*+PBFN$_{\alpha = -0.03}$」和「最常出現 10000 詞（長度≧2） *F10000-2*+PBFN$_{\alpha = -0.03}$」爲特徵集時皆有相同的準確率。上圖中，「廣義知網所有詞彙 *All*」準確率從 88.23%提升至 92.33%時，此差距爲顯著，檢定結果 (32.14, $1.4*10^{-8}$)。

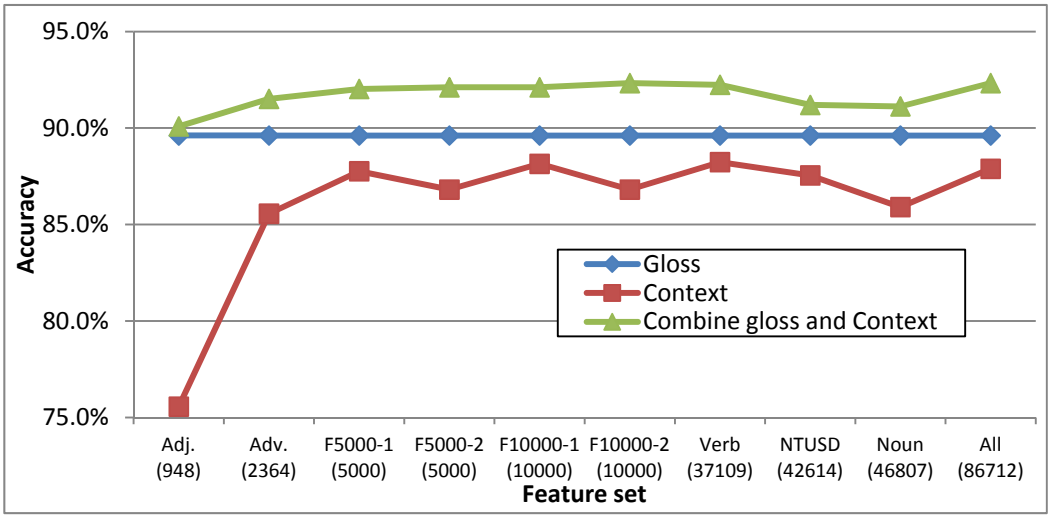


圖 6. 廣義知網、語篇特徵、與組合特徵的準確率比較

4.5 組合式的監督式機器學習演算法效能

在圖 6 中，組合出的特徵集有十個，所以共有十個分類器，每個分類器在訓練時，對不同詞性有不同的效能，我們將這十個分類器對於每個詞性標記的效能（內部測試）整理成表 4。

表 4. 訓練資料集中，組合特徵對不同詞性的標記準確率

特徵集代號	總體效能	訓練資料集中，依詞性分別計算的準確率				
		名詞	動詞	副詞	形容詞	其他
Adj.	94.3223%	95.9559%	94.2167%	89.9023%	93.2203%	82.0513%
Adv.	95.3243%	96.5074%	95.2795%	92.1824%	91.5254%	84.6154%
F5000-1	96.1000%	97.3039%	96.0110%	92.8339%	94.9153%	89.7436%
F5000-2	97.2635%	98.0392%	97.1705%	**96.0912%**	94.9153%	**94.8718%**
F10000-1	96.2400%	97.3652%	96.1767%	92.8339%	94.9153%	89.7436%
F10000-2	**97.5005%**	**98.2843%**	**97.4189%**	**96.0912%**	94.9153%	**94.8718%**
Verb	96.5632%	97.5490%	96.5079%	94.4625%	91.5254%	89.7436%
NTUSD	96.8218%	97.3039%	96.8254%	95.1140%	93.2203%	**94.8718%**
Noun	96.8541%	98.1005%	96.6460%	**96.0912%**	**96.6102%**	89.7436%
All	96.4124%	97.4265%	96.3699%	93.1596%	94.9153%	89.7436%

表 4 中的特徵集代號是「*語篇特徵集代碼*+PBFN$_{\alpha\,=\,-0.03}$」的簡寫，因爲使用相同的 PBFN$_{\alpha\,=\,-0.03}$，所以將其忽略。「總體效能」是指分類器訓練時的整體效能。表中，一欄中最佳的標記效能以**粗體字**表示。

表 4 中我們可以發現，訓練時，*F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 有最高的總體效能，其各詞性效能除了形容詞外，多是最好；考量到資料集中形容詞的數量並不多，這表示組合多個分類器後，效能的提昇空間可能有限。表 4 中另一個值得注意的一點是訓練資料集的內部測試效能（inside test）*F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 的 97.5005% 跟實際測試效能 92.3276%相比，降低了 5.31%，這降低幅度並不大，顯示這分類器的 generalization 能力不錯，這也是使用 Google Web 5-gram 的優點，它可產生較強健（robust）的分類器(Bergsma, Pitler, & Lin, 2010)。

我們用內部測試效能來挑選分類器，以便用在組合式的監督式機器學習演算法中。我們在表 4 中選不同詞性做得最好的分類器來組合，如果效能相同，則選特徵數量較少的那一個分類器，因爲特徵數較少通常在未看過的資料集會做得較好。組合出的分類器我們稱爲 *EnsembleClassifier*，其結果列在表 5，其中 *F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 於各詞性的標記效能也列出來比較。

表 5. 組合分類器於各詞性的標記效能及比較

分類器 / 詞性	*F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 分類器於各詞性的標記效能			組合分類器 *EnsembleClassifier* 於各詞性的標記效能				
	正確個數	錯誤個數	準確率	使用的分類器	正確個數	增減	錯誤個數	準確率
名詞	371	37	90.9314%	*F10000-2*	371	(+0)	37	90.9314%
動詞	1,681	130	92.8216%	*F10000-2*	1,681	(+0)	130	92.8216%
副詞	67	9	88.1579%	*F5000-2*	69	(+2)	7	90.7895%
形容詞	14	1	93.3333%	*Noun*	12	(-2)	3	80.0000%
其他	9	1	90.0000%	*F5000-2*	9	(+0)	1	90.0000%
總數	2,142	178	92.3276%		2142	(+0)	178	92.3276%

表 5 中，我們也列出每種詞性做錯與做對的個數，並以 *F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 分類器爲基準，看組合後的分類器，在各詞性中做對做錯的次數的增減，用括號來標出增減的數量。

EnsembleClassifier 所得成績跟 *F10000-2*+PBFN$_{\alpha\,=\,-0.03}$ 相同，這表示目前的分類器組合方式，無法提升效能。

4.6 相關研究效能比較

我們總結前面各種不同的實驗結果，畫成圖 7，來方便我們比較效能。其中，gloss 表基礎義原特徵 $PBFN_{\alpha=-0.03}$，最好的效能到 92.3276%。

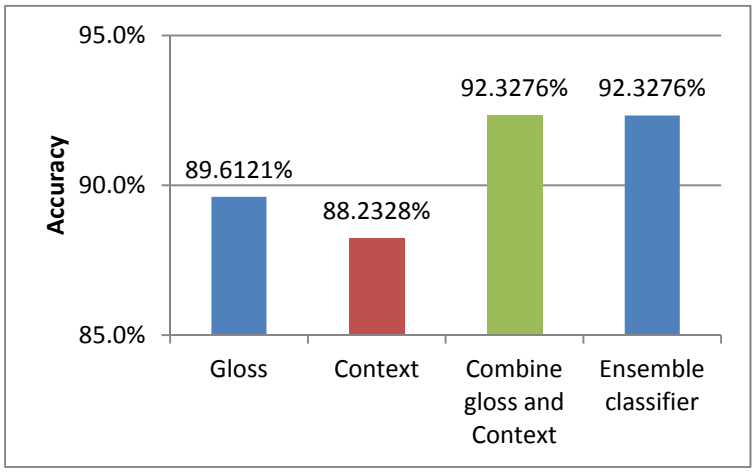


圖 7. 四種方法效能比較

由於我們使用 NTUSD，我們想看看 NTUSD 人類標記的效能跟我們分類器的效能有何差異。在 Ku & Chen (2007)的研究中，對詞有分四類標記，分別是正面、負面、中立、及非意見詞，並聘請標記者對 NTUSD 進行標記，我們將該研究中標記者的最佳標記效能與本研究的比較如表 6。由於人類標記者是將詞分成四類，但我們的系統只分兩類，所以這數據沒有辦法跟我們的結果直接相比較；但我們仍可從表 6 中看出，本研究所產生的正負面詞彙自動標記演算法，已達到了很高的效能。

表 6. NTUSD 標記者與本研究標記效能比較

分類器	Recall	Precision	F-Measure
F10000-2+$PBFN_{\alpha=-0.03}$	92.36%	92.20%	92.27%
三人中最佳的人類標記者	96.58%	88.87%	92.56%

表 6 中，人類標記者的 Recall 及 Precision 取自 Ku & Chen (2007)。*F10000-2*+$PBFN_{\alpha=-0.03}$ 的預測結果爲 (True Positive, False Positive, True Negative, False Negative) = (TP, FP, TN, FN) = (968, 77, 1174, 101)，其中 Positive 表正面極性。我們分別對正負面極性計算 Recall、Precision 及 F-Measure (R^+、P^+、F^+、R^-、P^-、F^-)，其中，P^+=TP/(TP+FP)、R^+=TP/(TP+FN)、F^+= $2P^+R^+/(P^++R^+)$、P^-=TN/(TN+FN)、R^-=TN/(TN+FP)、F^-= $2P^-R^-/(P^-+R^-)$，最後系統的 Recall=$(R^++R^-)/2$、Precision=$(P^++P^-)/2$ 及 F-Measure = $(F^++F^-)/2$ = (91.58% + 92.95%)/2 = 92.27%。由計算中我們可以看到，我們的系統對負面極性做得較好，而且因資料集有較多的負面詞彙，所以最後的準確率 92.33% 比 F^+ 高。

5. 結論

本研究使用了 Google Web 5-gram Version 1 來抽取語篇特徵，並加上來自 E-HowNet 的基礎義原特徵，用監督式機器學習的方法，來預測 E-HowNet 詞彙的意見極性。雖然單獨使用不同的特徵已經可以接近 90% 的準確率，但如果把兩種特徵都加以使用，分類器的極性預測的準確率可到達 92.33% 的高準確率；以這種方式建立的分類器，可用來自動標記 E-HowNet 詞彙的意見極性。

我們希望在未來能把這種方式，往不同的方向擴展，來給予 E-HowNet 詞彙更多意見的屬性，這包括對詞彙標記主客觀的屬性及正負面傾向的強度等。除此之外，因爲 E-HowNet 詞彙有許多不同的詞性，我們也希望能把我們的方法，運用詞性的層次來進行標記。藉由提供更精確的字彙意見標記資訊，來支援句子及文件層次的意見分析。

致謝

Research of this paper was partially supported by National Science Council (Taiwan) under the contract NSC 98-2221-E-002-175-MY3.

參考文獻

Bergsma, S., Pitler, E., & Lin, D. (2010). Creating robust supervised classifiers via web-scale N-gram data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 865-874.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. In *Neural computation*, 10(7), 1895-1923.

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

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification, In *Proceedings of CIKM-05*, 617-624.

Esuli, A., & Sebastiani, F. (2006a). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06)* , 417-422.

Esuli, A., & Sebastiani, F. (2006b). Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 193-200.

Han, Z., Mo, Q., Zuo, M., & Duan, D. (2010). Efficiently identifying semantic orientation algorithm for Chinese words. In *International Conference on Computer Application and System Modeling*, Vol. 2, 260-264.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of ACL-97*, 174-181.

Kamps, J., Marx, M., Mokken, R. J., & De Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Vol. IV, 1115-1118.

Ku, L.-W., & Chen, H.-H. (2007). Mining opinions from the Web: Beyond relevance retrieval. In *Journal of the American Society for Information Science and Technology*, 58(12), 1838-1850.

Li, D., Ma, Y.-tao, & Guo, J.-li. (2009). Words semantic orientation classification based on HowNet. In *The Journal of China Universities of Posts and Telecommunications*, 16(1), 106-110.

Liu, F., Yang, M., & Lin, D. (2010). *Chinese Web 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

Lu, B., Song, Y., Zhang, X., & Tsou, B. (2010). Learning Chinese polarity lexicons by integration of graph models and morphological features. In *Information Retrieval Technology: 6th Asia Information Retrieval Societies Conference, AIRS 2010*, 466-477.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactions on Information Systems*, 21, 315-346.

Yao, J., Wu, G., Liu, J., & Zheng, Y. (2006). Using bilingual lexicon to judge sentiment orientation of Chinese words, In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology*, p. 38.

Yuen, R. W., Chan, T. Y. ., Lai, T. B. ., Kwong, O. Y., & T'sou, B. K. . (2004). Morpheme-based derivation of bipolar semantic orientation of Chinese words, In *Proceedings of the 20th international conference on Computational Linguistics*, 1008-1014.

陳克健、黃淑齡、施悅音、陳怡君(2004)。多層次概念定義與複雜關係表達－繁體字知網的新增架構。漢語詞彙語義研究的現狀與發展趨勢國際學術研討會，北京大學。

刘群、李素建(2002)。基于《知网》的词汇语义相似度计算。第三届汉语词汇语义学研讨会。

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 37-54

Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in Two Conversational Corpora[1]

Sheng-Fu Wang*, Jing-Chen Yang*, Yu-Yun Chang*,

Yu-Wen Liu+, and Shu-Kai Hsieh*

Abstract

This study examines how different dimensions of corpus frequency data may affect the outcome of statistical modeling of lexical items. Our analysis mainly focuses on a recently constructed elderly speaker corpus that is used to reveal patterns of aging people's language use. A conversational corpus contributed by speakers in their 20s serves as complementary material. The target words examined are temporal expressions, which might reveal how the speech produced by the elderly is organized. We conduct divisive hierarchical clustering analyses based on two different dimensions of corporal data, namely raw frequency distribution and collocation-based vectors. When different dimensions of data were used as the input, results showed that the target terms were clustered in different ways. Analyses based on frequency distributions and collocational patterns are distinct from each other. Specifically, statistically-based collocational analysis generally produces more distinct clustering results that differentiate temporal terms more delicately than do the ones based on raw frequency.

[1] Acknowledgement: Thanks Wang Chun-Chieh, Liu Chun-Jui, Anna Lofstrand, and Hsu Chan-Chia for their involvement in the construction of the elderly speakers' corpus and the early development of this paper.

* Graduate Institute of Linguistics, National Taiwan University, 3F, Le-Xue Building, No. 1, Sec. 4, Roosevelt Rd., Taipei Taiwan, 106
E-mail: {sftwang0416; flower75828; june06029}@gmail.com; shukaihsieh@ntu.edu.tw

+ Department of English, National Taiwan Normal University, No. 162, He-ping East Road, Section 1, Taipei, Taiwan, 106
E-mail: Yw_L7@hotmail.com

Keywords: Clustering, Collocation, Corpus Linguistics, Temporal Expression, Gerontology

1. Introduction

The study of gerontology is gaining wide attention, as the aging population is becoming a major issue in society. Research has noted that aging causes not only physiological changes for elderly people, but also affects their language production (Burk & Shafto, 2004), cognitive load (Wilson, 2008), context processing speed (Rush, Barch, & Braver 2006), language performance patterns compared to younger individuals (Veliz, Riffo, & Arancibia, 2010), *etc*. To study gerontology from a linguistic point of view, Green (1993) proposed that the phenomenon of gerontology could be studied through discourse analysis. Therefore, we use conversations the elderly and from younger speakers as our speech corpora and take these corpora as an input to exemplify the procedures and usage of lexical modeling.

As for the examination of temporal terms, we postulated that people's social roles, especially the elderly's, might be embedded in conversations where speakers share personal experiences or judgment of the past (Kuo, 2008) and the present. Thus, we presume that temporal expressions might serve as the anchoring points in a conversation-based corpus and might reveal a certain aspect of the speech behavior patterns of the elderly.

Statistical modeling can serve to describe a given set of data, be it diachronic subsets, register, or lexical units. Statistical models often take the so-called "bottom-up" approach that suits most corpus linguists' empirical state of mind. Moreover, nice and neat visualization is often a feat in such modeling techniques, to an extent that some of the models are called "graph models" (Widdows & Dorow, 2002). When the proper behavior of lexical units and the structure of the lexicon are applied, statistical modeling may help us develop NLP-oriented lexicographic modules in the form of dictionaries, thesauri, and ontologies (Mitrofanova, Mukhin, Panicheva, & Savitsky, 2007).

A glimpse at relevant studies would reveal that the most prominent kind of data input is related to the distributional patterns of the lexical items in corpora. The distributional data could be in the form of word frequencies and variability of frequencies (Gries & Hilpert, to appear) or the distribution of n-grams as a whole (Gries, Newman, & Shaoul, 2011). The distributional pattern or dependency with syntactic patterns is also a prominent source of data input (Cimiano, Hotho, & Staab, 2004a, 2004b; Lin, 1998; Pereira & Tishby 1992). Target lexical items' dependency and co-occurrence with particular word types also has been taken as the basis of lexical modeling in some studies (Redington, Chater, & Finch, 1993). Moreover, statistically-based collocational patterns are used for modeling similarities among lexical units of interest (Chen, 2009; Gries & Stefanowitsch, to appear).

The above-mentioned different methods, or rather, different data inputs, are considered to fall somewhere between raw distributional data and relational data, or between lexical items and syntactic patterns. In our study, we aim to compare the two endpoints of this methodological continuum, namely the "frequency distributional data" input and "collocation data," in order to see how these different types of inputs may result in different data in lexical modeling. The former is the analysis based on the raw distributional data of lexical items, and the latter is based on more fine-grained examination of the relationship between lexical items, such as a particular item's collocational pattern with other lexical units in a corpus. By comparing the two kinds of analyses, we hope to reveal the extent to which these two methods demonstrate different patterns, thereby making contributions to the evaluation and selection of research methods in modeling lexical items.

This paper is organized as follows: Section 2 introduces the corpora used in this study, including data collection, guidelines for transcription, and annotation standards. Section 3 reports basic corporal information and preliminary analysis of six selected temporal expressions from the corpus. Section 4 demonstrates the methods and results of statistical modeling of temporal expressions, as well as a meta-analysis on different models. Section 5 and Section 6 summarize our findings and suggestions for further research respectively.

2. Corpora

2.1 The Elderly Speaker Corpus

Speech data were collected from four pairs of elderly people. Each pair consisted of one male and one female speaker. All subjects are native speakers of Mandarin and Taiwanese Southern Min. One pair is from Changhua while the others are from Taipei. The mean age of the subjects is 65.75 years old (SD = 6.16). Each pair of speakers was asked to do a face-to-face conversation in Mandarin for 30 to 40 minutes. The designated conversational topic was the speakers' life experience in the past and the present. During the recording, other participants, such as the subject's relatives or the observer, might also be involved in the talk. All files were recorded by a digital recorder in the WAV format. The total length of the speech samples is 145 minutes.

Speech samples collected from the elderly people's conversations were then transcribed into Chinese characters, following Du Bois *et al.*'s transcription standards for discourse analysis (Du Bois, Schuetze-Coburn, Cumming, & Paolino, 1993). Prosodic features and vocal qualities of the intonation units (IUs) were excluded from the transcription since they are not the main interest in this study. A short guideline of transcription standards is provided below.

Conversation samples were manually processed into several IUs. Each IU was labeled with a number on the left, as shown in Example (1).

(1)

34 SM: a 你 看 這 個 做工 的
P. you see this CL. do.work DE
35 ...(1.3) 那 個 有--
that CL. have
36 有夠 重
have.enough heavy

Speech overlap occurring during the conversation was indicated by square brackets, as shown in Example (2). To indicate overlap, brackets were vertically aligned where the overlaps began. Double square brackets were used for numerous overlaps occurring within a short stretch of speech, with their left brackets displaying temporal alignment.

(2)

70 SF: ...都 [送 人家]
all give others
71 SM: [送 人家] [[撫養 la]]
give others to raise P.
72 SF: [[撫養]]
to raise

Sometimes, the subjects switched between Mandarin and other languages when speaking. Such cases of code-switching were enclosed in square brackets and labeled with *L2* as well as the code for that non-Mandarin language. Example (3) demonstrates a case where a speaker switched between Mandarin and Taiwanese Southern Min (TSM). It should be noted that, since this study targeted elderly people's Mandarin speech performance, code-switching to languages other than Mandarin was excluded from our analysis.

(3)

268 SF: [L2 TSM 單輪車 TSM L2]
single wheeler

Laughter was also identified in the transcription. Each syllable of laughter was labeled with one token of the symbol @ (see Example 4a). Longer laughter was indicated by a single symbol @ with the duration in the parentheses (see Example 4b). Two @ symbols were placed at each end of an IU to show that the subject spoke while laughing (see Example 4c).

(4)

a. 163 F1: @@@@@
b. 200 SM: @(3.3)
c. 828 O: @沒 那麼 嚴重 la@
not that serious P.

The occurrence and duration of a pause in discourse were transcribed. Pauses were represented by dots: two dots for short pauses of less than 0.3 seconds, three dots for medium pauses between 0.3 and 0.6 seconds, and three dots with the duration specified in parentheses for pauses longer than 0.7 seconds. Example (5) below is the instance for pauses.

(5)

40 SF: ..以前 o..是--
before P. is

41 SF: ...eh ..都 是..父母...(0.9)做 X
P. all is parents do X

Particles were transcribed in phonetic transcription to avoid disagreement on the employment of homophonic Mandarin characters, as Example (6) shows. Phonetic transcriptions for the particles included *la, hoNh, a, o, le, haNh, hioh,* and *ma.*

(6)

26 SM: hoNh.. a 我們 二十 幾 歲 結婚
P. P. we twenty more age get.married

The recorded utterances were not always audible or clear enough for the transcribers to identify what was being said. Each syllable of uncertain hearing was labeled with a capital X, as shown in Example (5) above. Last but not least, truncated words or IUs were represented by double hyphens --, as shown in Examples (1) and (5).

The transcription was automatically segmented and POS (part of speech) tagged through the CKIP Chinese Word Segmentation System provided by the Chinese Knowledge Information Processing (CKIP) group at Academia Sinica (2004). The segmentation and POS standards were based on the Sinica Corpus guidelines (1998). The annotated language samples then were checked manually. The procedure is described below.

First, every segmentation result derived from CKIP was examined and corrected if wrong, as in the following examples. Example (7a) is the original IU before segmentation and tagging. Through CKIP, we attain the result in Example (7b), which is falsely processed. Example (7c) shows the proper segmentation after manual correction.

(7)

a. 我爸爸是他媽媽的哥哥
"My father is his mother's brother."

b. *我 爸爸 是 他媽 媽的 哥哥
I father is he.mom mom.DE brother

c. 我 爸爸 是 他 媽媽 的 哥哥
I father is he mom DE brother

Second, POS tags were viewed as correct if the main word classes were correct, while the details of their sub-classes were not of primary concern. For instance, in Example (8), the main word class of each POS tag (in this case, *N*, *DE*, *V*, or *D*) was examined, but not the sub-class tagging, as we give less consideration for the subcategories they belong to.

(8)

他(**N**h)	的(**DE**)	腦筋(**N**a)	動(**V**AC)	得(**DE**)	比較(**D**fa)	快(**V**H)
he	DE	brains	act	DE	more	fast

"He gets new ideas faster."

Third, particles' tags were manually corrected to *I* for IU-initial particles[2], and *T* for IU-final particles. If an IU contained nothing but particles, then the particles were tagged as *I*.

Finally, POS tags were removed for truncations (*e.g.* 這--), uncertain hearing (*i.e.* X), and code-switching. Given that truncations are not generally viewed as lexical items, they were not suitable for analysis at the lexical level.

2.2 The NTU Conversational Corpus

This corpus contains speech data collected by Master's students of the Graduate Institute of Linguistics at the National Taiwan University. The transcription follows Du Bois *et al.*'s standards for discourse analysis (1993). The data was collected by graduate students in their early 20s, and most of the recruited speakers were similar in age to the data collectors. In other words, this corpus contains mostly "youth speech," which is suitable as a complementary corpus to the elderly speakers' corpus for our analysis.

Although constant effort has been made in data collection of the NTU conversational corpus for more than ten years, little effort has been devoted to data organization and preprocessing. For this study, we selected a subset of face-to-face conversations between speakers (mostly students) less than 30 years old. The topics of these conversations were mostly everyday life experiences. The chosen subset, containing around 66,000 words, was tokenized and annotated the same way as the elderly speakers' corpus for further analysis.

3. Corpus Information & Preliminary Analysis

The elderly speaker corpus contains 4,982 IUs of Mandarin utterances and 22,090 word tokens produced by all speakers. Elderly people's production in Mandarin contains 3,739 IUs, and there are 18,076 word tokens in total. The subset of the NTU conversational corpus used in this study contains 15,863 IUs and 65,742 word tokens.

---

[2] According to the standards provided by Sinica Corpus, *I* represents "interjections" which usually occur in the IU-initial position.

The corpus processing tool used here is R (R. D. C. Team, 2010), which allows us to perform tasks, including preprocessing, word frequency, KWIC (KeyWord In Context) extraction, and statistical modeling.

We assume that time-related words may provide some vital clues to the elderly's and the youth's speech patterns, so the following analyses will focus on the subjects' use of temporal expressions. We are interested in how elderly people use 現在 (now) and 以前 (before), as well as other temporal expressions (tagged as Nd), to frame the present- and the past-related concepts, and how their use differs from the younger generations' employment of temporal expressions. Thus, the six most frequent temporal expressions from each corpus were selected for the analysis. Table 1 lists the frequency of the six target temporal expressions. As shown in the rankings, "now" is the most frequent temporal expression in both corpora.

Table 1. The frequency of the most frequent temporal expressions in the two corpora.

Elderly Speaker Corpus			NTU (Youth) Corpus		
Rank.	**Term**	**Freq.**	**Rank.**	**Term**	**Freq.**
1	現在(now)	169	1	現在(now)	137
2	以前(before)	169	2	後來(later)	76
3	小時候(in one's childhood)	12	3	今天(today)	52
4	民國(R.O.C. year)	11	4	以前(before)	52
5	當初(back then)	9	5	昨天(yesterday)	34
6	最近(recently)	6	6	今年(this year)	31

4. Statistical Modeling of Temporal Expressions

In this section, we will present quantitative analyses with the help of hierarchical clustering, a data-driven approach, to see how the temporal terms of interest are grouped together with the frequency data extracted from our corpus.

The clustering method employed here is divisive hierarchical clustering. This differs from agglomerative hierarchical clustering in that a group of entities is first divided into large groups before smaller groups are classified. Such a method is useful for finding a few clusters that are large in size (Rush, Barch, & Braver, 2006). We would like to find out whether the terms for "the present" and "the past" can really be grouped into clusters that are different in temporality. Thus, divisive hierarchical clustering serves our need. In our current study, the clustering was conducted with the help of the *dist()* function in R (2010), which takes a table of correlations between the vectors containing different temporal expressions' frequency data or collocational data in the corpus and transforms it into a "table of distances" based on the square distance between these vectors.

4.1 Modeling Results from the Elderly Speaker Corpus

We executed a series of hierarchical clustering with different data input. The first analysis was run with the frequencies of the temporal terms across different files/texts in our corpus. Such an input was expected to capture the co-occurrence pattern of these temporal terms affected by individual speaker's style or idiolect, as well as by differences in the topic of conversation. The output is presented in Figure 1, where 現在 (now) is separate from 以前 (before) under a major cluster on the left. Also, 最近 (recently) stands independently from the other expressions, suggesting that temporal terms within a particular time domain are more likely to occur in the same text, which is really a conversational event in our corpora.

Next, four clustering analyses were made based on the frequency data across subsets of different sizes. The sizes chosen for producing subsets were 10, 50, 200, and 500 words. Smaller subsets may reflect linguistic patterns in a few clauses, and larger subsets may reflect patterns in a larger unit, such as major or minor topics in the flow of conversation. As we can see in Figure 2, 現在 (now) and 以前 (before) are classified in the same small cluster.

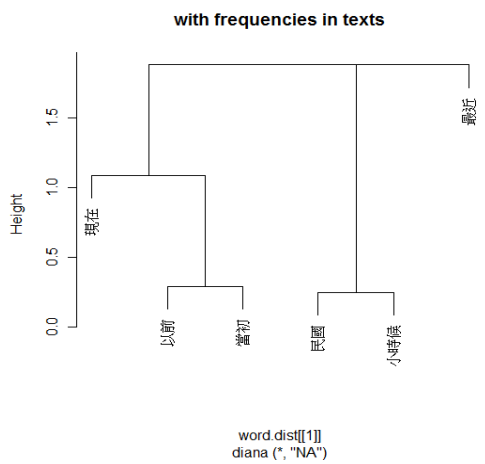


Figure 1. Clustering based on frequencies in texts in the elderly speaker corpus. "Height" in the y-axis represents the furthest distance between the entities under a particular node in terms of the distance, based on the correlation of data vectors of these expressions. Thus, it is sensitive to how far apart the entities in question are.

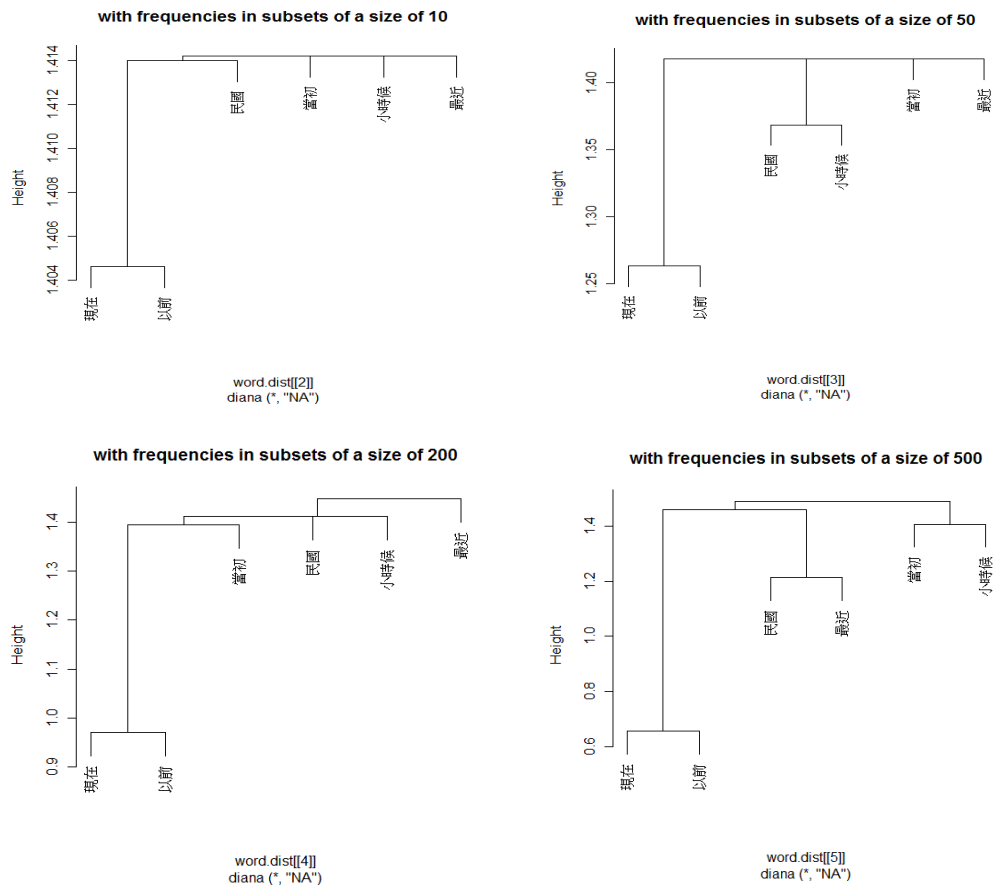


Figure 2. Clustering based on frequencies across subsets in the elderly speaker corpus. Upper left, with subsets of a size of 10 words. Upper right, of 50 words. Lower left, of 200 words. Lower right, of 500 words.

We can also conduct clustering analysis according to how these terms collocate with other words in the corpus, on the premise that collocational patterns should reveal some characteristics of lexical items. Thus, two more analyses based on this assumption were given. The first analysis was done using each word type's collocational pattern (span = 3) with the six temporal terms as input. The second analysis was achieved through the dependency patterns of sentential particles (*i.e.* lah, hoNh, ah, oh, le, haNh, hioh, mah, as described by Li, 1999), taking the temporal terms as input. There are two reasons for the inclusion of particle collocation. First, with regard to methodology, running more than one collocational test allows one to see whether collocational analyses with different approaches generate similar results. Second, sentential particles' dependency patterns might help us understand how the "referent" of each temporal expression is conceived and presented in discourse. The outcome is

illustrated in Figure 3. Again, 現在 (now) and 以前 (before) are clustered closely, showing that their collocational patterns might be similar.

There is a potential problem in using raw frequencies in studying collocates. Collocates with high frequencies might simply be high frequency words, rather than being "exclusively close" to the terms of interest. Thus, we bring forth collexeme analysis (Gries, Hampe, & Schönefeld, 2005; Gries, 2007), a statistical method developed for finding "true collocates," that is, collocates with strong collocational strength (coll.strength hereafter). The coll.strength of each word type and particle was calculated and used as input for clustering analysis. The output is shown in Figure 4.

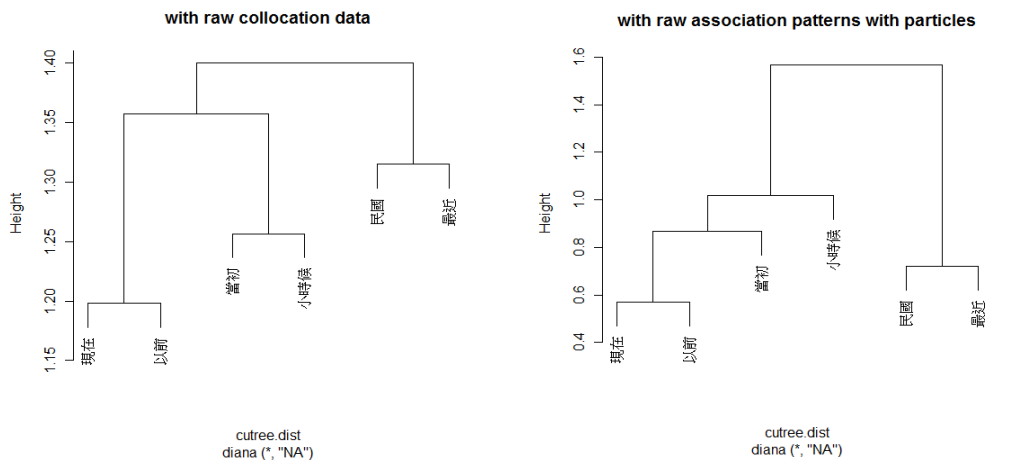


Figure 3. Clustering based on association/collocation frequencies. Left, with all word types in the elderly speaker corpus. Right, with particles.

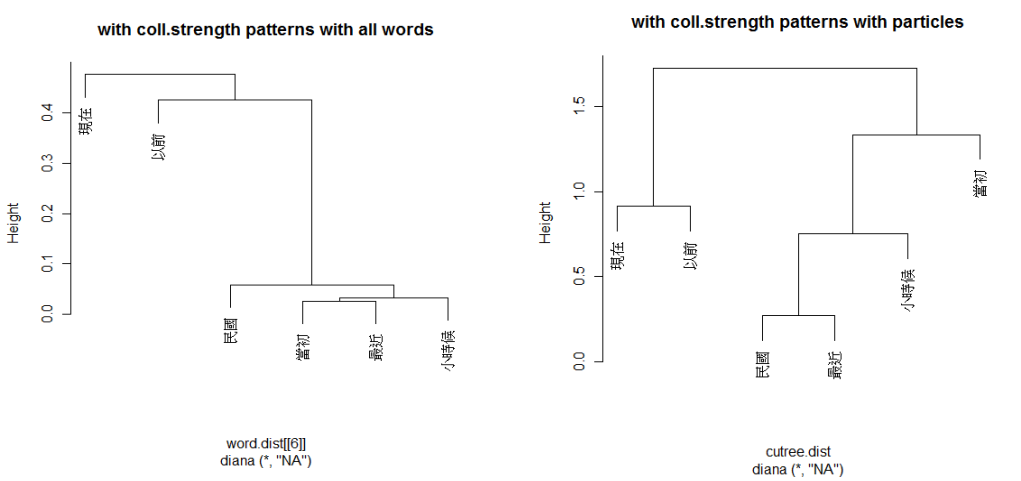


Figure 4. Clustering based on coll.strength patterns. Left, with all word types in the elderly speaker corpus. Right, with particles.

To sum up, 現在 (now) and 以前 (before) seem to intertwine, concerning their occurrences in different subsets of the corpus. This may suggest that, when elderly speakers talk about the past, the present follows as a contrast in time regarding the same subject matter, and *vice versa*. Only the by-text analysis shows a difference between the terms for the present and those for the past, suggesting that there are some conversations featuring more present than the past, and *vice versa*. Collocational strength analysis is another approach revealing a difference between 現在 (now) and 以前 (before), showing that, although the two terms usually are used closely, they still attract different words with different strengths. It should be noted that association patterns with particles do not distinguish between the present and the past. A possible explanation for this is that such a difference in pragmatic and discourse meaning is too fine-grained to be shown with information based on quantitative data. In other words, it shows that a quantitative method with corpus data has its limitation, especially when the annotation only functions at the basic POS level. Such findings of the temporal terms may in turn suggest that modeling lexical items is not a simple matter of looking for any types of analyzable data input. In addition to surface frequencies, taking collocational patterns into account, especially those based on statistical analyses, seems to be a requirement to capture the nuances among lexical items.

4.2 Modeling Results from the NTU (Youth) Conversational Corpus

We performed similar analyses for the selected portion of the NTU conversational corpus, with a few modifications. First, the analysis based on the frequencies of temporal expressions in corpus subsets was only conducted with subset sizes of 10, 50, and 200 words, since the individual texts (conversations) in the NTU corpus are mostly not as long as those in our elderly speakers' corpus. For certain conversations, a 500-word window can cover almost all of the words in the text, making the analysis too similar to the by-text analysis we conducted. Second, the analyses of temporal expressions' raw collocation and collocational strength with particles were not conducted for the NTU corpus because the manual annotation on collocation between particles and temporal terms has not yet been completed.

Figure 5 shows the results of clustering analysis based on raw frequencies of the temporal expressions in the NTU conversational corpus. Temporal scope seems to be the factor determining the clustering patterns. Expressions that have a less definite time frame, such as 後來 (later) and 以前 (before), are grouped together, whereas expressions denoting a specific temporal scope, such as 今天 (today) and 昨天 (yesterday), are grouped under the same node.

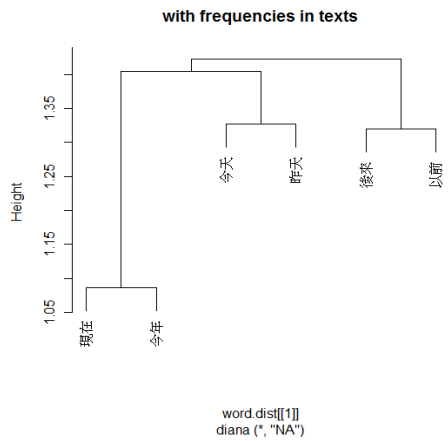


Figure 5. Clustering based on frequencies in texts in the NTU conversational corpus.

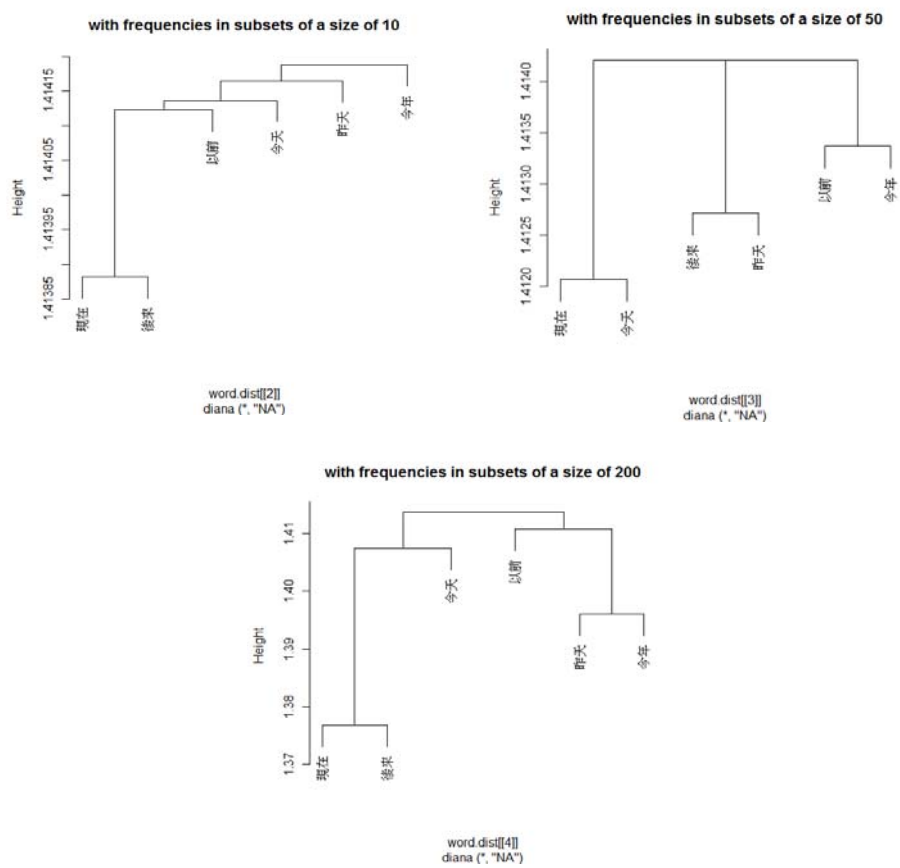


Figure 6. Clustering based on frequencies across subsets in the NTU conversational corpus. Left, with subsets of 10 words. Middle, of 50 words. Right, of 200 words.

Figure 6 shows clustering results based on the target expressions' frequencies in the subsets of the data. The results are not clear-cut, lacking a consistent pattern across the outputs from different inputs. One of the more consistent patterns may be that 現在 (now) and 後來 (later) are clustered together in two of the graphs above and are rather closely clustered in one of the two. This might imply that, within a context with a size smaller than the whole text, these two temporal expressions are commonly used together to form conversations.

As for the result of the target temporal expressions' collocates in raw frequencies, clustering patterns differentiate expressions for the past, the present, and those with different time scopes. In Figure 7, the expression 今年 (this year), which denotes the present within a bigger scope, is clustered away from other expressions. The expressions 現在 (now) and 今天 (today), both denoting the present within a smaller scope, are clustered together. Two expressions for the past, 以前 (before) and 昨天 (yesterday), are grouped under the same node. Finally, 後來 (later), an expression for denoting time sequences, is clustered alone. This indicates that different types of temporal expressions collocate with other words in the corpus in different manners.

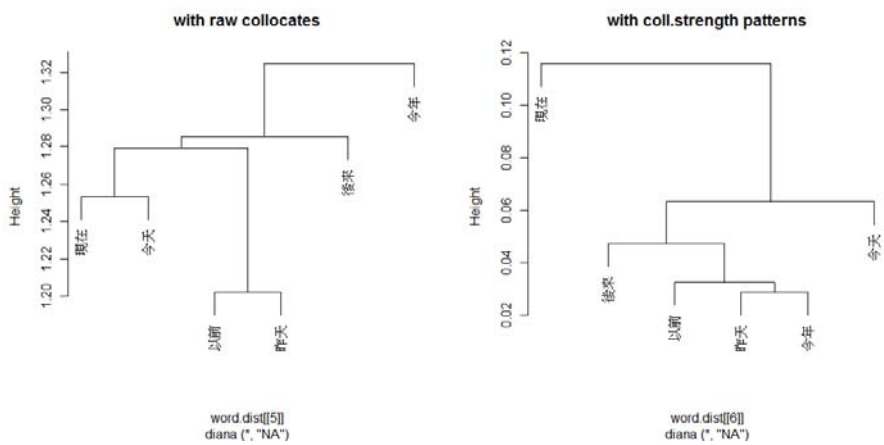


Figure 7. Left, clustering based on collocational frequencies with all the word types in the NTU conversational corpus. Right, clustering based on collocational strength patterns.

Yet, when the input data are collocational strength patterns, which should reveal these expressions' true collocates, the results of the clustering change. 現在 (now) and 今天 (today), two expressions about the present, stand out in two distinct clusters, while 後來 (later) also stands out to a certain extent. This reveals that expressions about the present are strongly and uniquely collocated with a certain group of words in the youth corpus, compared to other temporal terms. Such a result of modeling may imply that the youth use some particular patterns in structuring events or topics about the present.

4.3 Evaluation on Modeling Results

How do we evaluate all of these different results? The answer may not be surprising: We can do it with clustering analysis. The "clustering" package for R offers a function "cutree" for a simple quantification of different clustering, where each "tree" is quantified in terms of which cluster an item is clustered to. We collect the data for all of the trees shown above and execute clustering as meta-analysis. The outcome for the elderly speaker corpus is shown in Figure 8, and the one for the NTU conversational corpus is shown in Figure 9.

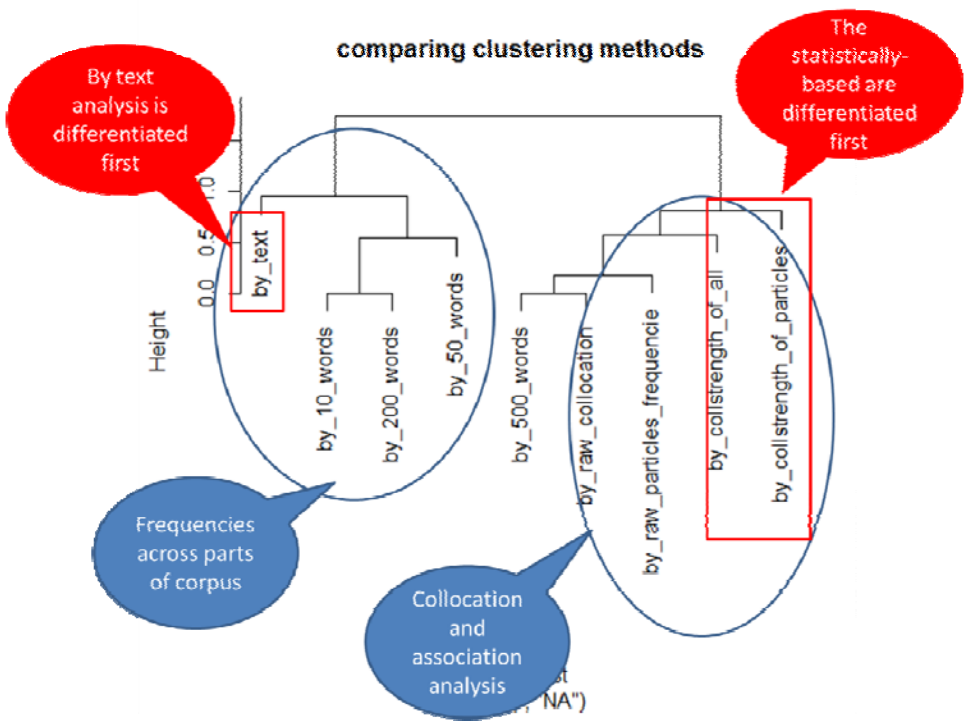


Figure 8. Clustering of various results from the elderly speaker corpus.

An interesting pattern shows up in Figure 8. There are two major clusters. The left one is based on frequency patterns of temporal terms, and the right one basically contains analyses of how these terms collocate or associate with other words or particles. Despite the curious occurrence of the "by-500-words" analysis in the right major cluster, the result of this meta-analysis seems to be able to characterize the major differences in terms of data input. More specifically, in the left major cluster, the "by-text" analysis is the first one singled out. This conforms to our impression that temporal terms are clustered differently, with 現在 (now) and 最近 (recently) placed relatively far away from other past-related expressions. Moreover, in the right major cluster, the analyses with coll.strength are the first ones differentiated from the others. Again, this reflects that statistically-based analyses produce

different patterns from the ones based on simple frequency values. What can be inferred from the patterns in Figure 8 is that different types of data input certainly influence the outcome of clustering analysis.

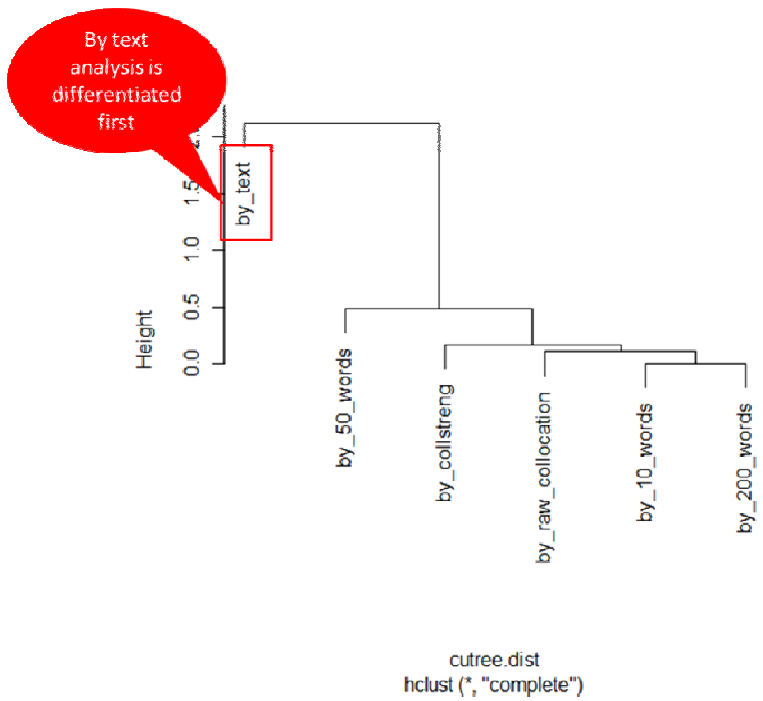


Figure 9. Clustering of various results from the NTU conversational corpus.

The evaluation of the clustering results of the NTU conversational corpus, shown in Figure 9, clearly is different from the one of the elderly speakers' corpus. The only similar pattern is that the by-text analysis still stands out as a particular kind of clustering. Other clustering analyses are lumped together with no clearly emerging pattern. There are several possible reasons for such a result. First, one may argue that the relationship between clustering methods and their results does not hold a consistent pattern. On the other hand, it could mean that there are some differences in the use of the temporal items between these two corpora, hence, between speakers different in age. At this stage, the two corpora are not equal in size, so any claims about how elderly speakers' linguistic patterns actually differ from those of younger speakers' would appear to be unsound. To sum up, the evaluation method we propose here is a technique that can possibly reveal differences among methods of modeling or even differences between various corpora, as shown in our preliminary research results. For further development, similar investigations on better-controlled and comparable corpora must be conducted.

5. Conclusion

Statistical modeling based on different types of data input displays different patterns, with modeling derived from frequencies and collocational patterns forming two major clusters (as revealed in the meta-analysis, which visualizes the difference between models based on quantitative data). In the "frequency" cluster, analysis based on distributional patterns is differentiated from the ones based on arbitrarily divided subsets. In the "collocation" cluster, statistically-oriented (*i.e.* collocational strength) analyses are distinguished from those based on surface collocational frequencies. For our present study, these findings are not overwhelmingly surprising because it is not hard to imagine the impact of the difference in texts and subsets on research, as well as the impact of surface frequencies and statistically-calculated relational patterns. Yet, when it comes to evaluating more types of modeling methods or inputs, meta-analysis of this kind provides a valuable means of choosing adequate methods. For instance, when researchers try to model different aspects of the lexical structure, the hierarchical modeling proposed here may help avoiding utilizing methods that are in fact very similar.

According to our analysis on temporal terms, the core expressions of the present and the past have very similar distributional patterns, showing that elderly speakers in the corpus tend to compare the present with the past in the same textual domains. The difference between these terms is disclosed only in models based on by-text frequency and statistical collocational analysis. The former shows that different speakers or conversation events may have their own preferred usage of temporal expressions. The latter indicates that these terms are still different in terms of their collocations, yet the difference can only be revealed through statistical tests on "true collocates" proposed by Gries (2007). These findings can be seen as a pilot result of the linguistic patterns of aging people and young people in comparison.

6. Future work

The primary purpose of this study was to attempt to highlight certain methodologies applicable to an elderly speaker corpus through several statistical approaches, rather than recklessly leaping to a conclusion that some universal elderly speech patterns are found in our corpus. The inclusion of the analysis on a younger speaker corpus helps us become more cautious with the claims we can make about our statistical approaches. Yet, to further explore the issue and confirm the validity of potential general linguistic patterns discovered in the current research, we must carefully conduct qualitative analyses of each temporal expression and interpret the results with the evidence from the quantitative methods we adopted previously. For example, the expansion of the size of the elderly speaker corpus does may alter the outcome of our statistical modeling.

Also, the inequality in terms of the size of our two corpora and the different ways of data collection make the direct comparison between the elderly speaker corpus and the NTU (youth) conversational corpus difficult. In the future, it might be advisable to collect speech materials from a small number of younger speakers by asking them to narrate personal experiences and stories just like what we asked the elderly speakers to do. By doing so, we can directly compare this small corpus contributed by younger speakers with our elderly corpus, to see whether we can prompt even similar high-frequency temporal expressions for comparing two corpora of two generations of speakers.

References

Baayen, R. H. (2008). Analyzing linguistic data. *A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bruke, D. M. & Shafto, M. A. (2004). Aging and language production. *Current directions in psychological science,* 13, 21-24.

Chen, C.-H. (2009). Corpus, lexicon, and construction: A quantitative corpus approach to Mandarin possessive construction. *International Journal of Computational Linguigstics and Chinese Language Processing,* 14, 305-340.

Cimiano, P., Hotho, A., & Staab, S. (2004a). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. in *Proceedings of the European Conference of Artificial Intelligence*, 435-439.

Cimiano, P., Hotho, A., & Staab, S. (2004b). Clustering concept hierarchies from text. in *Proceedings of LREC 2004*, 1-4.

CKIP. (2004). CKIP Chinese word segmentation system. Retrieved June 2, 2011 from http://ckipsvr.iis.sinica.edu.tw/

CKIP. (1998). Introduction to Sinica Corpus: A tagged balance corpus for Mandarin Chinese. Taipei, Taiwan (R.O.C.): Academia Sinica.

Du Bois J. W., Schuetze-Coburn, S., Cumming, S. & Paolino, D. (1993). Outline of discourse transcription. In Edwards J. A. & Lampert M. D. (Eds.), *Talking data: Transcription and coding in discourse research.* Hillsdale: New Jersey, Lawrence Erlbau.

Green, B. S. (1993). *Gerontolgy and the social construction of old age.* New York: Aldine De Gruyter.

Gries, S. T. (2007). Collostructional analysis: Computing the degree of association between words and words/constructions. Retrived May 30, 2011 from: http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r

Gries, S. T., Hampe, B., & Schönefeld, D., (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. C*ognitive Linguistics*, 16, 635-676.

Gries, S. T. & Hilpert, M. (to appear). Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Nevalainen, T. & Traugott, E. C.,

(Eds.), *Handbook on the history of English: Rethinking approaches to the history of English.* Oxford: Oxford University Press.

Gries, S. T., Newman, J. & Shaoul, C. (2011). N-grams and the clustering of registers. *Empirical language research*. Retrived May 28, 2011 from: http://ejournals.org.uk/ELR/article/2011/1

Gries, S. T. & Stefanowitsch, A. (to appear). Cluster analysis and the identification of collexeme classes. In S. Rice and J. Newman (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI.

Kuo, S.-H. (2008). Discourse and aging: A sociolinguistic analysis of elderly speech in Taiwan (NSC96-2411-H007-024). Taipei, Taiwan (R.O.C.): National Science Council.

Li, C. I. (1999). Utterance-final particles in Taiwanese: A discourse-pragmatic analysis. Taipei, Taiwan (R.O.C.): Crane Publishing Co.

Lin, D. (1998). Automatic retrieval and clustering of similar words. in *Proceedings of the 17th International Conference on Computational Linguistics*, 768-774.

Mitrofanova, O., Mukhin, A., Panicheva, P., & Savitsky, V. (2007). Automatic word clustering in Russian texts. *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, 85-91.

Pereira, F. & Tishby, N. (1992). Distributional similarity, phase transitions and hierarchical clustering. *Probabilistic approaches to natural language, papers from 1992 AAAI Fall Symposium*, 108-112.

R. D. C. Team. (2010). R: A language and environment for statistical computing. Retrieved June 1, 2010 from: http://www.R-project.org

Redington, M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. in *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 848-853.

Rush, B. K., Barch, D. M., & Braver, T. S. (2006). Accounting for cognitive aging: Context processing, inhibition or processing speed? *Aging, Neuropsychology and Cognition*, 13, 588-610.

Veliz, M., Riffo, B., & Arancibia, B. (2010). Cognitive aging and language processing: Relevant issues. *Revista de Lingüística Teórica y Aplicada*, 75-103.

Widdows, D. & Dorow, B. (2002). A graph model for unsupervised lexical acquisition. in *Proceedings of the 19th International Conference on Computational Linguistics*, 1093-1099.

Wilson, K. R. (2008). *The effects of cognitive load on gait in older adults* (Doctoral dissertation). Florida State University, Tallahassee, FL.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 2, June 2012, pp. 55-68

Using Kohonen Maps of Chinese Morphological Families to Visualize the Interplay of Morphology and Semantics in Chinese

Bruno GALMAR*

Abstract

A morphological family in Chinese is the set of compound words embedding a common morpheme, and Self-organizing maps (SOM) of these Chinese morphological families can be built. Computation of the unified-distance matrices for the SOMs allows us to perform semantic clustering of the members of the morphological families. Such semantic clustering sheds light on the interplay between morphology and semantics in Chinese. We studied how the word lists used in a lexical decision task (LDT) (Chen, Galmar, & Su, 2009) are mapped onto the clusters of the SOMs. We showed that this mapping is helpful to predict whether repetitive processing of members of a morphological family would elicit a satiation in an LDT - habituation - of both morphological and semantic units of the shared morpheme. In their LDT experiment, Chen, Galmar, and Su (2009) found evidence for morphological satiation but not for semantic satiation. Conclusions drawn from our computational experiments and calculations are in accordance with the behavioral experimental results in Chen *et al*. (2009). Finally, we showed that our work could be helpful to linguists in preparing adequate word lists for behavioral study of Chinese morphological families.

Keywords: Self-Organizing Maps, Computational Morphology and Semantics.

1. Introduction

In this paper, we call a morphological family the set of compound words embedding a common morpheme. Hence, the compound words in Table 1, which all contain the morpheme ‘明’ (míng) as a first character, belong to the morphological family of ‘明’.

* Institute of Education, National Cheng Kung University, Tainan, Taiwan
E-mail: hsuyueshan@gmail.com

Table 1.Some examples of the明 morphological family (Chen, Galmar, & Su, 2009).

明朝	明天	明白	明確	明星	明亮
Ming Dynasty	tomorrow	to understand / clear	explicit	star	bright

In Chinese, the meaning of a morpheme can be either transparent or opaque to the meaning of the compound word embedding it. For example, the common morpheme in Table 1 "明" can mean (*clear)* or (*bright)* and is transparent to the meaning of "明星" (*star*) but rather opaque to the meaning of "明天" (*tomorrow*). If some members of a morphological family are semantically similar, one could advance as a reason for such a similarity that these members are transparent to the same meaning of the shared morpheme. Most Chinese morphemes are polysemous (Chen & Chen, 2000). Hence, in theory, *transparent members* of a morphological family could belong to different semantic clusters whose centers would be the different meanings of the shared polysemous morpheme.

This paper is aimed primarily at using computational linguistics methods to perform semantic clustering of the members of the morphological families. Such a clustering is used to predict the results of a behavioral Lexical Decision Task[1] (LDT) designed by Chen, Galmar, & Su (2009) to study the phenomenon of morphological satiation in Chinese.

In visual word recognition, morphological satiation is an impairment of morphological processing induced by repetitive exposure to the same morpheme embedded in different Chinese compound words (Chen *et al*., 2009; Cheng & Lan, 2009). Chen, Galmar, and Su (2009) posited that morphological satiation is due to habituation of the morphological unit of the repeated morpheme. This is represented in Figure 1 by Diagram (a).

As a morpheme is thought to be a meaningful unit, it is logical to consider whether a semantic satiation (Kounios, Kotz, & Holcomb, 2000; Smith & Klein, 1990; Tian & Huber, 2010), an impairment of semantic processing causing a temporary loss of the meaning of the common morpheme, would occur concomitantly with morphological satiation.[2] In other words, the satiation observed by Chen *et al*. (2009) could have two loci: a morphological locus and a semantic locus, as represented in Figure 1 by Diagram (d).

A morphological satiation could also have its loci of satiation on the links between the morphological, lexical, and semantic units, as represented in Figure 1 by Diagrams (b) and (c). We quickly can rule out the possibility of a locus on the link between morphological and

[1] An LDT is a behavioral task for which subjects have to identify whether presented visual stimuli are words or non-words.

[2] If most of the members of a morphological family used in an experimental task are transparent to the same meaning of the shared morpheme, the same semantic units of the shared morpheme are repeatedly accessed and finally habituate - satiation diagram (d) -. Therefore, there could be a semantic satiation in addition to morphological satiation.

lexical units, as represented by Diagram (b). The reason is that, in a LDT, this link is changing at each presentation of a new two-character word. The morphological unit of the repeated morpheme constitutes one fixed endpoint of the morphological/lexical link but the over endpoint is always changing.

The present work of semantic clustering focuses on clarifying by computational means whether morphological satiation would probably have a sole morphological locus (Diagram (a)) or whether it would have both a morphological and semantic locus (Diagram (d)). The behavioral LDT experiment results in Chen *et al*. (2009) point to the existence of a sole morphological locus.

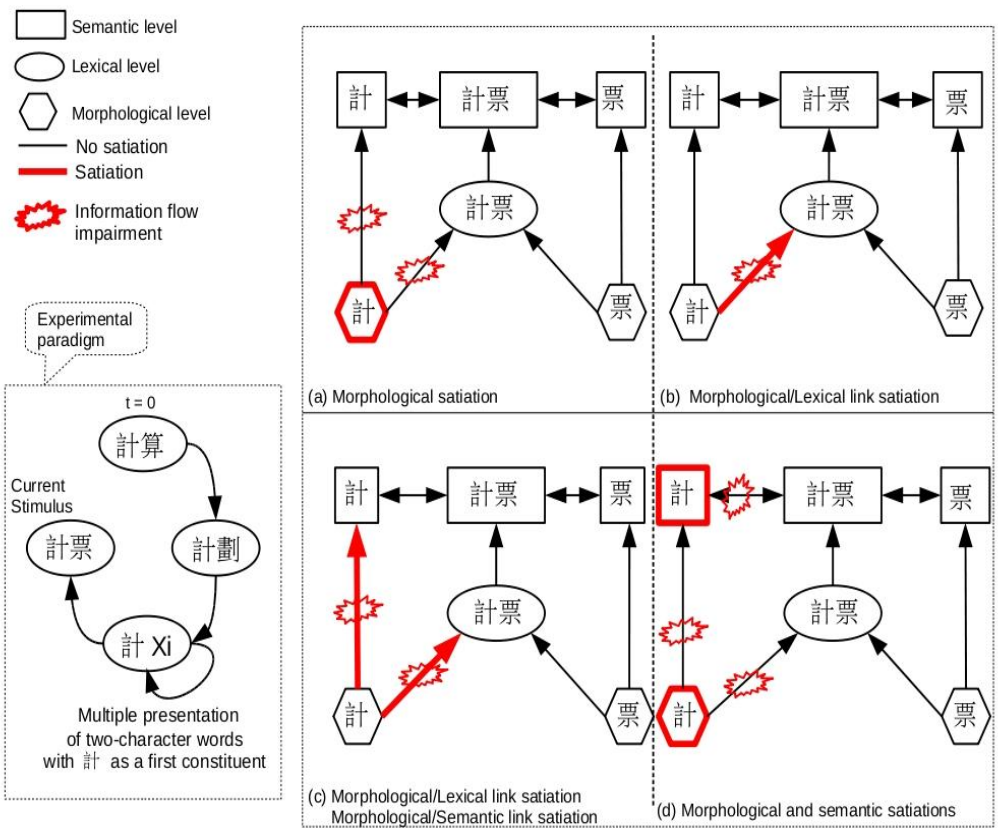


Figure 1. Different possible loci of satiation for morphological satiation.

2. Rationale of our Approach

As human subject agreement for semantic clustering tasks is low (Jorgensen, 1990), computational corpus-based semantic clustering was thought to be a valuable and complementary experimental approach compared to a behavioral one with human subjects.

A corpus of written texts is a human artifact, its content is relevant to the human reader and, from a cognitive psychology standpoint, a corpus does embed a subset of organized human semantic knowledge and is worthy to be studied in computer simulations as a pure abstract semantic memory stripped out of sensory and motor representations.

In natural language processing, proponents of the "bag of words" approach simplify each document internal structure to a set of words and use a whole corpus to build a matrix of co-occurrence of the words corpus (Landauer & Dumais, 1997). Computational methods, such as Latent Semantic Analysis (LSA), take as input such a high dimensional matrix and reduce its dimensionality to form a vector space of the documents and words (Landauer, McNamara, Dennis, & Kintsch, 2007). This space embeds only an associative kind of semantic information[3]: words that co-occur in the same documents or that have common co-occurrents are close associates.

For a news corpus, the association can often be of the situational type. For example, "Father Christmas" will be a close associate of "department store," as there are many news reports around Christmas about the bustling agitation in department stores full of "Father Christmas".[4] In cognitive science and AI, it is said that the two terms "Father Christmas" and "department store" belong to a common memory frame, a frame being defined by Minsky as "*a data-structure for representing a stereotyped situation*" (Minsky, 1974).

In the present work, we follow a "bag of words" approach by first building a term document matrix (TDM). Then, Self-Organizing Maps (SOMs) and associated unified-distanced matrices (called U-matrix thereafter) are built from the TDM. The SOMs and U-matrices serve to visualize semantic clusters in a morphological family on a 2D hexagonal grid of bins (Kohonen, 2001).

On the SOMs, a semantic cluster is made of members of a morphological family that have been fitted into the same bin of the grid and into contiguous bins which are close neighbors - according to the U-matrix information - in the original high dimensional space. SOMs have been used successfully to capture associative semantic relationships between words in corpora. Closer to the present approach, SOMs have been used to study the developmental aspect of vocabulary acquisition in Chinese (Li, 2001, 2009; Li, Farkas, &

[3] Semantic information also can be, for example, of the categorical or featural types.

[4] This example is borrowed from Galmar & Chen (2010b).

MacWhinney, 2004; Zhao & Li, 2008). Zhao, Li, and Kohonen (2010) studied the clustering of subsets of the most common Chinese words along both linguistic and semantic dimensions. Kohonen and Xing (2011) computed the SOMs of different linguistic classes for Chinese and studied how word frequency modulates the clusters on the SOMs. Our study is the first one to use SOMs to study the interplay between morphology and semantics in Chinese compounds words sharing a common morpheme, *i.e.* to study the semantics of morphological families. Previous studies on the semantics of morphological families (Galmar & Chen, 2010a; Galmar & Chen, 2010b) followed a supervised approach that relied upon etymological knowledge of Chinese morphemes to identify the meaning of a morpheme in a given compound word. The present work followed an unsupervised approach, and the goal is meaning discrimination through clustering rather than meaning identification. We chose the SOM algorithm to ensure that our work can be replicated thanks to the widespread availability of SOM packages and toolboxes and because it has not yet been applied to our specific research topic.

3. The Corpus and the Term Document Matrix (TDM)

3.1 The Academia Sinica Balanced Corpus

We used the Academia Sinica Balanced Corpus (ASBC), a five million word annotated corpus based on Chinese materials from Taiwan, mostly newspapers articles. The corpus is made of roughly 10000 documents of unequal length.

We removed the foreign alphabet words and most of the Chinese functional words from the corpus. We kept POS tag information to allow differentiation between different grammatical instances of the same word[5] (Galmar & Chen, 2010).

3.2 The Term Document Matrix (TDM)

The TDM was built using the *TermDocumentMatrix* function of the R package *tm* (Feinerer, 2008) with a self-customized Chinese tokenizer. The TDM is a 136570 terms * 9179 documents matrix.

The TDM was weighted:

1. Using the classical term frequency-inverse document frequency (TfIdf) weighting scheme for both local and global weighting of the terms in the TDM (Landauer & Dumais, 1997). We used the function *weightTfIdf* of the package *tm* (Feinerer, 2008).
2. Using a weighting scheme at the document level to reduce the effect of the size difference between documents:

[5] Some of the Chinese words can have up to 5 different POS tags [10].

$$\log_2\left(\frac{Max_{Document\_size}}{Document\_size}+1\right) \quad (1)$$

Each document of the TDM is a genuine article of the ASBC corpus and is considered as a semantic unit. More weight is given to small documents of the ASBC corpus. A complete justification for such a decision is given in Galmar & Chen (2010). Briefly, one can say that the gist of a news article is easier to extract from a very short article than from a very long one for a human reader due to attention capacity limitations.

4. The Self-Organizing Maps

For a given morphological family, the rows corresponding to the members of the family in the TDM were extracted. The extracted rows constitute a submatrix of the TDM. From this submatrix, an SOM is built using the *Batch map algorithm* (Kohonen, 2001). The U-matrix (Ultsch & Siemon, 1990) is computed to assess how close members fitted to contiguous bins (bins are hereafter called units) on the SOM are in the original high-dimensional space (hereafter called input data space).

4.1 The batch version of the SOM algorithm

As all of the data - the TDM - can be presented to the SOM algorithm from the beginning of learning, the batch version of the SOM algorithm (called "Batch Map") is used instead of the incremental learning SOM algorithm. The batch SOM is very similar to the k-means (Linde-Buzo-Gray) algorithm (Kohonen, 2001).

Our SOM defines a mapping from the input data space $\Re^n$ of observation samples onto a hexagonal two-dimensional grid of N_u units. Every unit *i* is associated with a *reference vector* $m_i \in \Re^n$. The set of units located inside a given radius from unit *i* is termed *neighborhood set* N_i.

The Batch Map algorithm can be described as follows (Kohonen, 2001, p. 139-140; Kohonen, Oja, Visa, & Kangas, 2002, p. 1360).

1. Initialize the N_u reference vectors by taking the first N_u observation samples.
2. For each unit i, collect a list L_i of copies of all those observation samples whose nearest reference vector belongs to N_i.
3. Update the value of each reference vector m_i with the mean over L_i.
4. Repeat from Step 2 a few times.

The Batch Map presents a main advantage over the incremental learning version of the SOM algorithm (Kohonen, 2001; Fort, Letremy, & Cottrell, 2002) in that no learning rate

parameter has to be specified. To double-check the computed batch SOM's representativeness of the input data space, we followed the recommendation of both Fort, Letremy, and Cottrell (2002) and Kohonen (2001) to compare organization in the Batch Map and in the incremental learning SOM.

We used the code in the R package *class* (Venables & Ripley, 2002) for the batch SOM to build the SOMs. For the morphological family 計 (jí), used as an illustrative example in Section 5, we built the SOMs on a 7*8 hexagonal grid of 56 bins. The size of the SOM was determined by experimental testing while ensuring a lower quantization error (Kohonen, 2001).

4.2 The Unified-Distance Matrix

We reused and modified the code in the R package *kohonen* (Wehrens & Buydens, 2007) to build the U-matrix for the Batch Map and to plot a grey-level map superimposed on the SOM map. The U-matrix is the distance matrix between the reference vectors of contiguous units. On the grayscale SOMs, contiguous units in a light shade on the SOM are representative of existing clusters in the input data space. Contiguous units in a dark shade draw boundaries between existing clusters in the input data space (Ultsch & Siemon, 1990).

5. Results

We present the results for the study of the 計 (jí) morphological family.[6] This Chinese morpheme has two main meanings: (1) to count, to calculate or (2) to plan, to scheme. The study was limited to the members in the ASBC corpus embedding 計 as a first character. The SOM map of these members is noted SOM_{93} and is shown in Figure 2.

At the first level, the map is divided in two zones: a dark shaded one – the upper part of the map - and a light shaded one. Most of the words belong to the light shaded zone. Among the diverse existing clusters, we note that:

- Cluster C_1 mainly gathers word sharing and other words related to the first meaning of 計.
- Cluster C_2 gathers three words related to the frame *taxi* in the same unit.
- Cluster C_3 includes many words belonging to two contiguous units in a light shade. We decided to recompute a Batch Map SOM for the members in these two units to zoom in and have a clearer map of these members. The map is shown in Figure 3.

[6] Others examples are also given in the script file – available upon request - to create and plot the SOMs presented in the present paper.

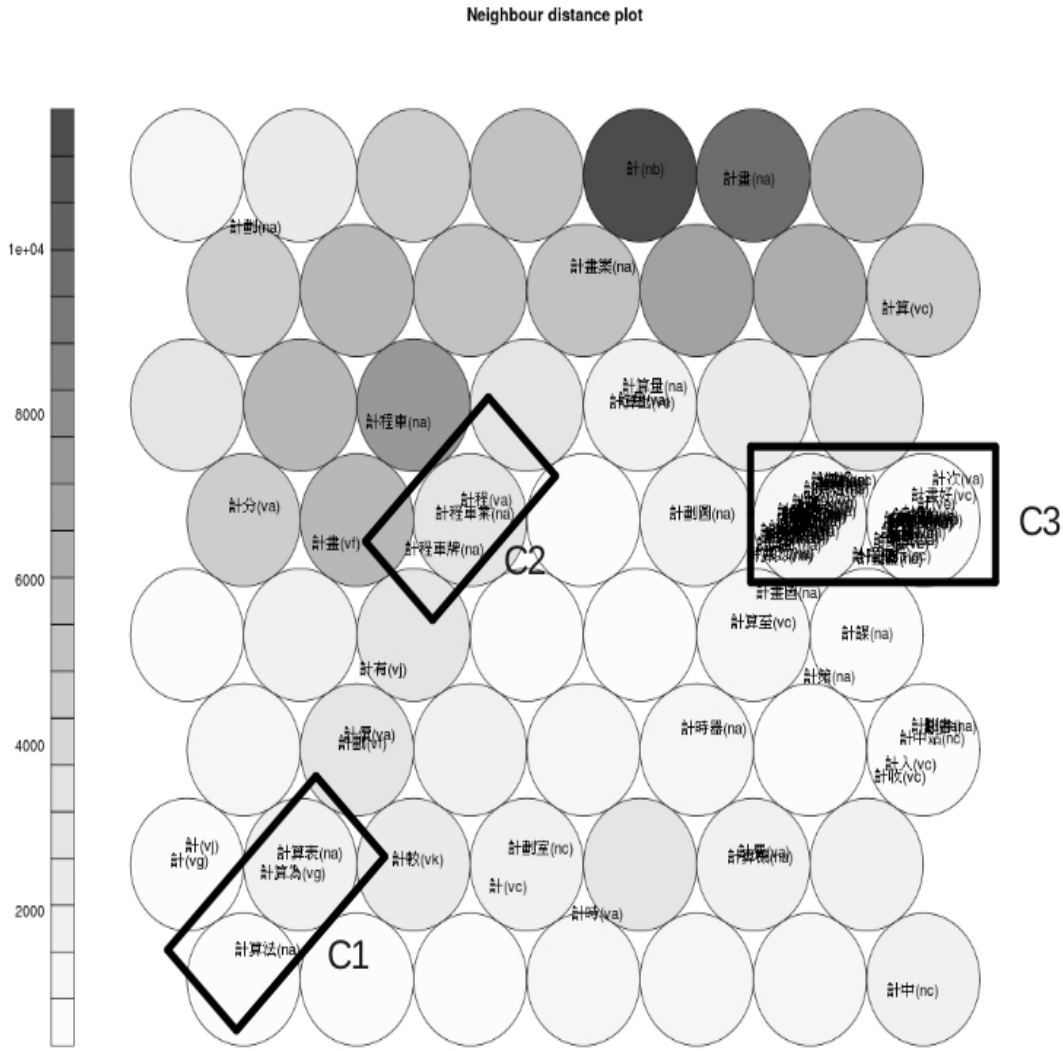


Figure 2. SOM$_{93}$ of the 計 (jì) morphological family.

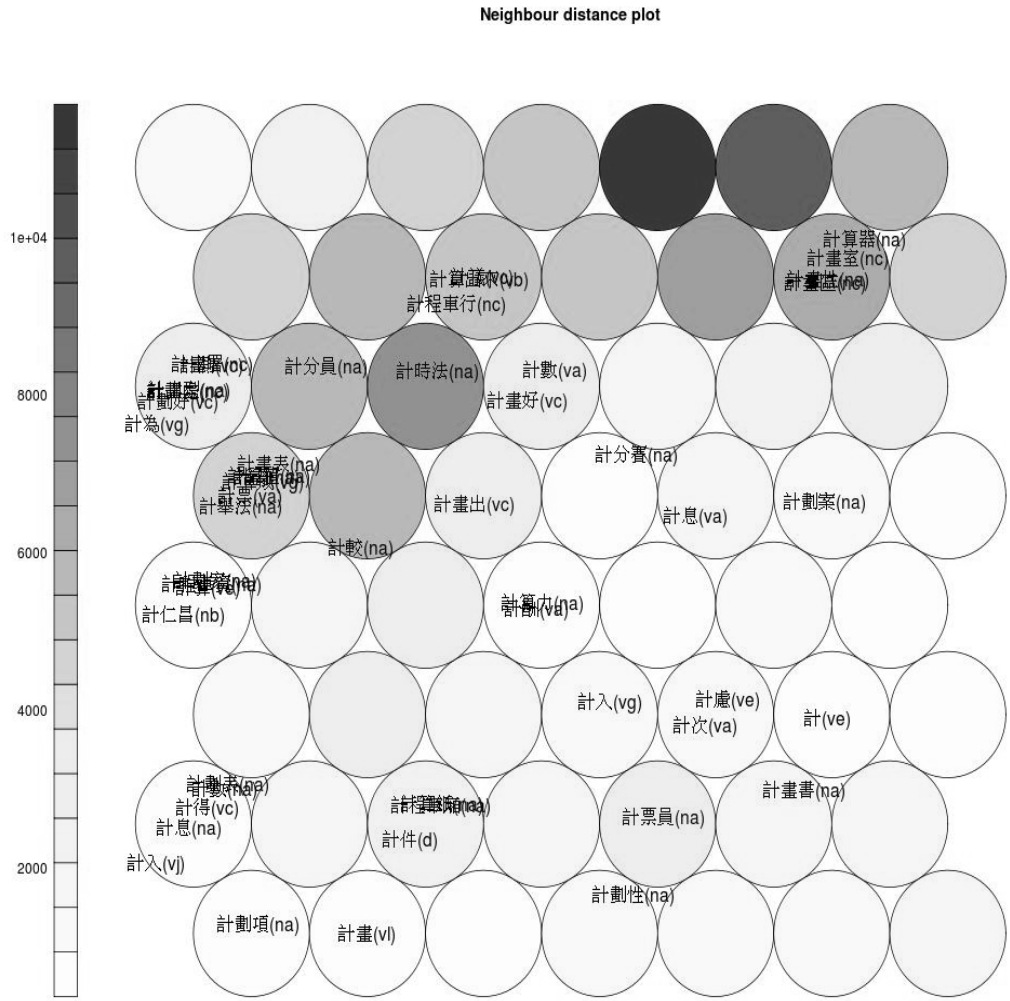


Figure 3. SOM for Cluster C3 in Figure 2.

Figure 4 shows only the 13 words used by Chen *et al.* (2009) in one block of their LDT experiment.[7] Some of the words have two POS tags, so the total number of data points represented in Figure 4 is 17.

[7] One word of the original experiment not existing in the ASBC corpus is missing here.

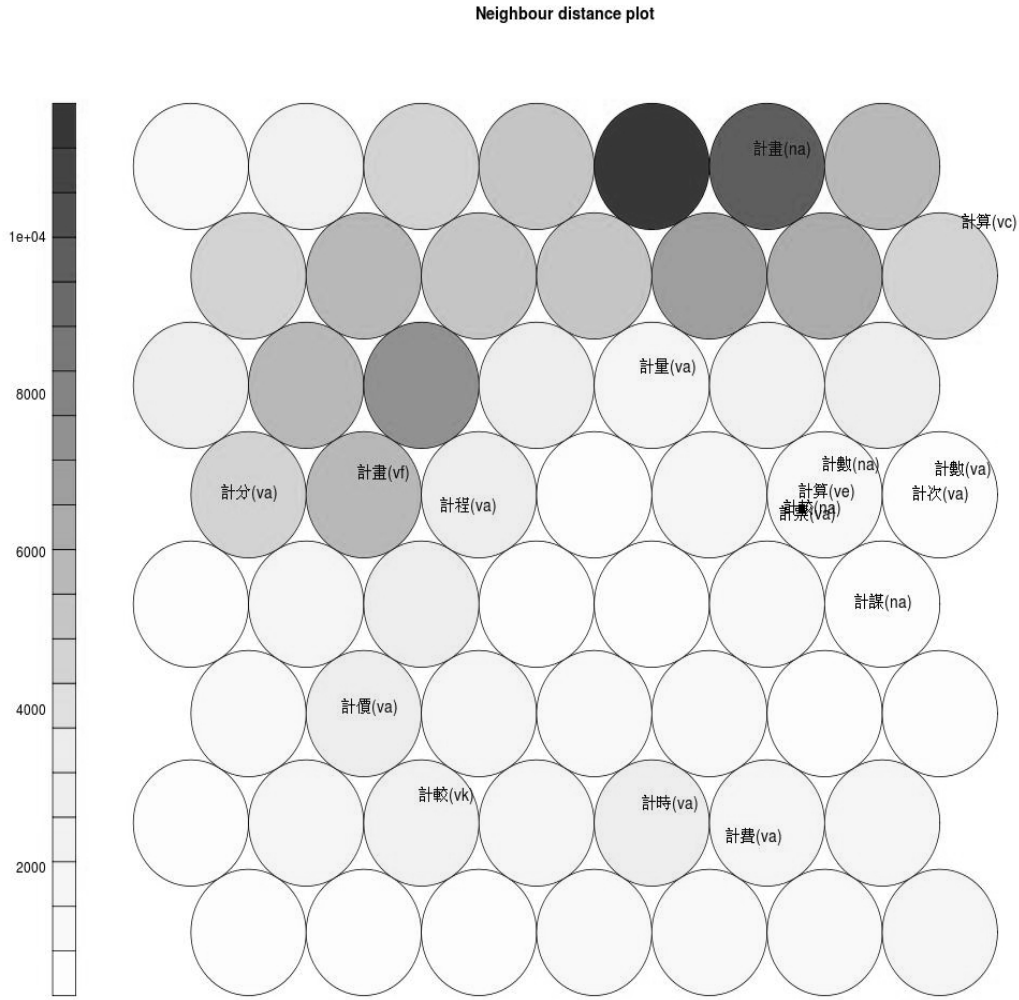


Figure 4. SOM of the members of the 計 morphological family used in Chen et al. (2009).

Clustering is observed easily with such a few words. Three contiguous units in a light shade form the unique big cluster with a total of six different words. In the latest experimental research on semantic satiation, Tian and Huber (2010) found that after five or seven repetitions of a given word, the word's meaning starts to be satiated. From two to five repetitions, there is semantic priming - behavioral enhancement in semantic tasks - and more repetitions are the realm of semantic satiation.

If, in Chen *et al.* (2009)'s lexical decision task (LDT), these six[8] words occur successively, there should be semantic satiation. In Chen *et al.*'s LDT, the 13 words in Figure 3 were randomly mixed with 13 non-words. Non-words, being meaningless, should not contribute significantly to satiate the semantic units of the different meanings of the shared morpheme. Therefore, from the analysis of our SOM, we predict that there could be a preliminary sign of semantic satiation only in the case where the 6 members of the big cluster occur successively in the 26-word list - we call this the best case.

To compute the probability of this best case, we need to calculate two numbers:

1. N_a, the number of distinguishable arrangements of n=26 words of which 6 - belonging to our big cluster - constitute a first set S1 and the 20 remaining ones constitute another set S2. The order of occurrence of the 6 words of S1 does not matter; therefore, the words of S1 are considered to be of a same type T1. For the same reason, words of S2 are of a same type T2, different from type T1.

$$N_a = \frac{26!}{6!20!} = 230230 \tag{2}$$

2. The number of distinguishable arrangements of 6 successive occurrences of S1 words[9] in a 26-word list: 21.

The probability *p* of the best case is given by dividing the number of distinguishable arrangements of 6 successive occurrences of S1 words by the number of distinguishable arrangements of n=26 words made of the two types T1 and T2.

$$p = \frac{21}{230230} \approx 9*10^{-5} \tag{3}$$

This best case has a very low probability, so subjects in Chen *et al.* would almost always be given a 26-word list that do not warrant - according to our analysis - elicitation of semantic satiation.

Hence, we agree with Chen *et al.* that there was no semantic locus of satiation in their experiment. On the other hand, we refine Chen *et al.*'s conclusions by advancing that one could prepare specific experimental word lists that would maximize the probability of observing semantic satiation.

6. General Conclusion

By visualizing the SOMs augmented with neighboring distance information from the U-matrix, one can observe whether semantic clusters exist in a morphological family and how the

[8] Tian and Huber (2010) found satiation effects after six repetitive accesses to a word's meaning.

[9] Order of occurrence of the S1 words does not matter.

experimental data in Chen *et al.* (2009) is mapped to these clusters.

Conclusions drawn from our computational experimental results are in accordance with Chen *et al.*'s behavioral experimental results revealing the absence of a semantic satiation while morphological satiation occurs. Nevertheless, we propose that semantic satiation theoretically could be elicited with specifically arranged word lists for Chen *et al.*'s experiment. Such lists have a very low probability of occurrence when a random assignment of words is used to prepare experimental word lists. Therefore, the present work shows the necessity of preparing adequate experimental word lists based on computational semantic clustering - as shown here - or human norms of semantic similarity if available.

7. Future Directions

Alternatives to SOMs, such as GTM (Bishop, Svensen, & Williams, 1998), exist and could be used for comparison purposes with the present results.

8. Code to generate the SOMs from the ASBC corpus

The source code and R command lines are available upon request in a script file. In order to run the whole script file from the very beginning, one needs the Academia Sinica Balanced Corpus (ASBC). The ASBC has to be purchased.[10]

Acknowledgements

This work was supported by a doctoral fellowship grant (NSC100-2420-H-006-007-DR) awarded to Bruno Galmar.

References

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural computation*, 10(1), 215-234.

Chen, J. Y., Galmar, B., & Su, H. J. (2009). *Semantic Satiation of Chinese Characters in a Continuous Lexical Decision Task.*

Chen, K. J., & Chen, C. (2000). *Automatic semantic classification for Chinese unknown compound nouns.*

Cheng, C. M., & Lan, Y. H. (2011). An implicit test of Chinese orthographic satiation. *Reading and Writing*, 24(1), 55-90.

Cottrel, M., Fort, J., & Letremy, P. (2005). *Advantages and drawbacks of the batch Kohonen Algorithm.*

[10] Contact the Academia Sinica (中央研究院語言所).

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software,* 25(5), 1-54.

Galmar, B., & Chen, J. (2010). Identifying Different Meanings of a Chinese Morpheme through Latent Semantic Analysis and Minimum Spanning Tree Analysis. *International Journal of Computational Linguistics and Applications,* 1(1-2), 153-168.

Galmar, B., & Chen, J. Y. (2010). Identifying different meanings of a Chinese morpheme through semantic pattern matching in augmented minimum spanning trees. *The Prague Bulletin of Mathematical Linguistics,* 94(-1), 15-34.

Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research,* 19(3), 167-190.

Kohonen, T. (2001). *Self-Organizing Maps*: Springer.

Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE,* 84(10), 1358-1384.

Kohonen, T., & Xing, H. (2011). Contextually self-organized maps of chinese words. *Advances in Self-Organizing Maps,* 16-29.

Kounios, J., Kotz, S. A., & Holcomb, P. J. (2000). On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory & cognition,* 28(8), 1366-1377.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review,* 104(2), 211.

Landauer, T. K., McNamara, D. S., Dennis, S. E., & Kintsch, W. E. (2007). *Handbook of latent semantic analysis*: Lawrence Erlbaum Associates Publishers.

Li, P. (2001). *A self-organizing neural network model of the acquisition of word meaning.*

Li, P. (2009). Lexical organization and competition in first and second languages: Computational and neural mechanisms. *Cognitive science,* 33(4), 629-664.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks,* 17(8-9), 1345-1362.

Minsky, M. (1974). A framework for representing knowledge. *AIM-306*.

Smith, L., & Klein, R. (1990). Evidence for semantic satiation: Repeating a category slows subsequent semantic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 16(5), 852.

Tian, X., & Huber, D. E. (2010). Testing an associative account of semantic satiation. *Cognitive psychology,* 60(4), 267-290.

Ultsch, A., & Siemon, H. P. (1990). *Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis.* Paper presented at the Proceedings of the International Neural Network Conference (INNC' 90).

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*: Springer verlag.

Wehrens, R., & Buydens, L. M. C. (2007). Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software,* 21(5), 19.

Zhao, X., & Li, P. (2008). *Vocabulary development in English and Chinese: A comparative study with self-organizing neural networks*.

Zhao, X., Li, P., & Kohonen, T. (2010). Contextual self-organizing map: software for constructing semantic representations. *Behavior Research Methods*, 1-12.

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:_____

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

CARD HOLDER SIGNATURE : _____

Tel.: _____ E-mail: _____

Add: _____

PAYMENT FOR

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

Quantity Wanted: _____

US$ _____ ❑ Publications:_____

US$ _____ ❑ Text Corpora: _____

US$ _____ ❑ Speech Corpora:_____

US$ _____ ❑ Others: _____

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

US$ _____ = Total

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

ACLCLP

% Institute of Information Science, Academia Sinica

R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　日期:： _____

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

卡號:_____-_____-_____- _____　有效日期：_____

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：　_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____　E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____❑ 中文計算語言學期刊(IJCLCLP)

NT$_____❑ 中研院詞庫小組技術報告

NT$_____❑ 中文（新聞）語料庫

NT$_____❑ 平衡語料庫

NT$_____❑ 中文詞庫八萬目

NT$_____❑ 中文句結構樹資料庫

NT$_____❑ 平衡語料庫詞集及詞頻統計

NT$_____❑ 中英雙語詞網

NT$_____❑ 中英雙語知識庫

NT$_____❑ 語音資料庫_____

NT$_____❑ 會員年費　❑續會　❑新會員　❑終身會員

NT$_____❑ 其他: _____

NT$_____= 合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		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.17 no.2 June 2012

Contents

Special Issue Articles: Selected Papers from ROCLING XXIII

Papers

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