International Journal of

Computational Linguistics & Chinese Language Processing

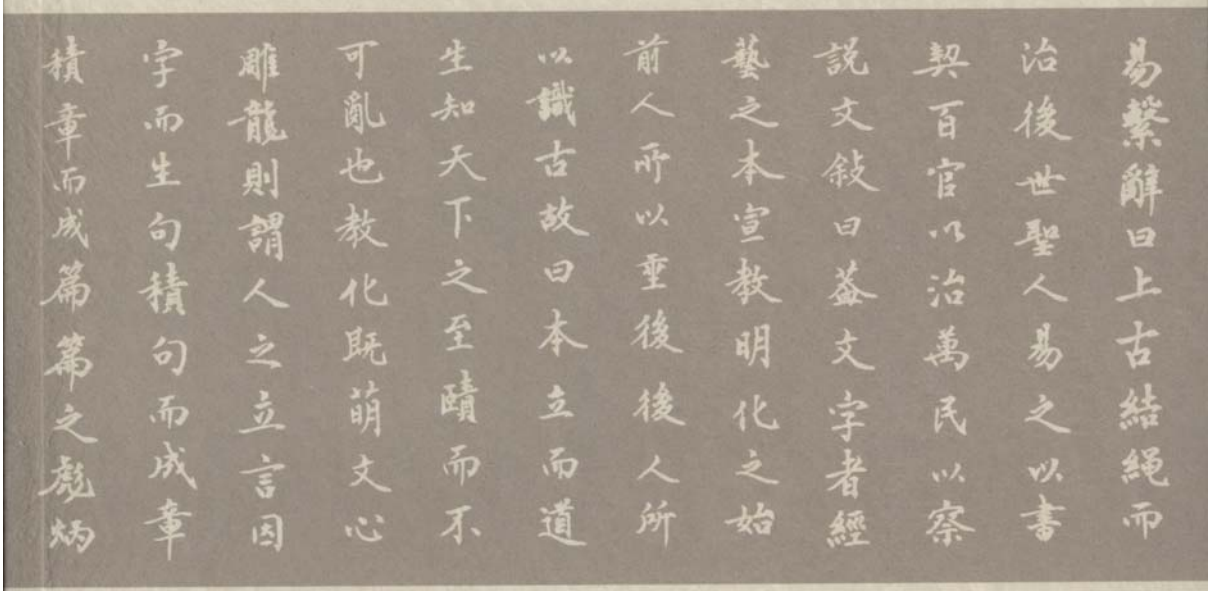
中文計算語言學期刊

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 "Computer Assisted Language Learning"

Guest Editor: Chao-Lin Liu, and Zhao-Ming Gao



Vol.14 No.2 June 2009 ISSN: 1027-376X

International Journal of

Computational Linguistics & Chinese Language Processing

The editing of this journal is subsidized by Center for Humanities Research National Science Council in 2009.

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing 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 from 2005. This journal covers all aspects related to computational linguistics and Chinese speech and language processing. 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

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. 14, no. 2, June 2009

Contents

Special Issue Articles:

Computer Assisted Language Learning

Papers

Short Papers

Regular Issue Articles:

Short Papers

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 133-160

Speech-Based Interactive Games for Language Learning: Reading, Translation, and Question-Answering

Yushi Xu*, and Stephanie Seneff*

Abstract

This paper concerns a framework for building interactive speech-based language learning games. The core of the framework, the "dialogue manager," controls the game procedure via a control script. The control script allows the developers to have easy access to the natural language process capabilities provided by six core building blocks. Using the framework, three games for Mandarin learning were implemented: a reading game, a translation game, and a question-answering game. We verified the effectiveness and usefulness of the framework by evaluating the three games. In the in-lab and public evaluation phases, we collected a total of 4025 utterances from 31 subjects. The evaluation showed that the game systems responded to the users' utterances appropriately about 89% of the time, and assessment of the users' performances correlated well with their human-judged proficiency.

Keywords: Computer Aided Language Learning, Machine Translation, Automatic Question Generation, Automatic Answer Judging

1. Introduction

Computer aids for second language learning have long been a promising yet difficult research topic. Despite much argument about the best way to teach a second language based on pedagogy, the most natural and effective source of second language education is the classroom and human tutors. Statistics, however, have shown a severe shortage of language teachers, compared to the number of language learners. For example, the current estimated number of Chinese language teachers worldwide is around 40,000, while the number of people trying to

* Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory, USA
E-mail: {yushixu, seneff}@csail.mit.edu

[Received July 3, 2009; Revised October 7, 2009; Accepted October 22, 2009]

learn Chinese is about 1,000 times that[1]. The dramatic difference in the numbers not only results in many students not having a chance to find a suitable teacher, but also results in an under-emphasis on spoken communication, which many pedagogists agree to be an important skill, and which cannot be practiced by the student alone.

Given this situation, it is natural to think of replacing a costly human tutor with a computer. Several criteria, however, must be satisfied for such a machine tutor to be interesting to the students. The computer needs to understand the student's speech, and act intelligently enough to avoid being perceived as just an e-textbook. It should be able to offer a variety of activities, and to constantly provide rewards in order to motivate students to invest further effort to improve their skill level.

In an attempt to meet these requirements, we have developed a versatile framework for building speech-based language learning games. The core of the framework is a dialogue manager, which is supported by a set of building blocks, each providing some high-level natural language processing operations. By combining these operations in different ways using a control script, we have implemented three distinct games in two domains. The three games, a reading game, a translation game, and a question-answering game, provide different types of challenges to beginner learners of Mandarin Chinese. The two domains, general travel and flights, expose the students to different sentence patterns and vocabulary. The language processing operations provided by the building blocks are general-purpose, and the control script can be viewed as a high-level programming language. The whole framework thus makes it relatively straightforward to develop other speech-based language learning games, or to export the existing games to other domains of interest with minimal effort.

This paper will be organized as follows. We will first summarize some related work in Section 2. In Section 3, we will give a brief introduction of our three games. Then, in Section 4, the dialogue manager and its core building blocks will be described. Section 5 will describe the implementation of the three games in more detail, followed by their evaluations in Section 6. We will conclude and point to some future work in Section 7.

2. Related Work

There has been a significant amount of previous research in the computer aided language learning (CALL) field. Most of the research has a single focus, for example, vocabulary training (Brown, Frishkoff, & Eskenazi, 2005), or reading comprehension tests (Kunichika, Katayama, Hirashima, & Takeuchi, 2003). Only a few systems have been designed to provide alternative types of activities. Many of these integrated systems have been packaged as a CD-ROM as a delivery mechanism. The software is then installed on a local machine for

[1] Statistics according to China's Ministry of Education, 2006.

deployment. On the other hand, there are some Web-based language learning systems, such as Chengo Chinese (Chengo Chinese, 2004) and Active Chinese (Active Chinese, 2006). Both of these provide online Mandarin learning, which the user can access simply by opening up the web browser. These two systems provide several lessons ranging from easy to hard. In each lesson, a couple of activities and exercises are presented. Typically, the student first watches a conversation between some animated characters. Then, several important sentences are taught along with the vocabulary. After that, the student is expected to complete some pre-designed exercises. Although speech is enabled in both systems, the systems do not go beyond speech recognition. The user interacts with the system mainly via keyboard and mouse.

Examples of language learning systems that use speech as the main input modality are WordWar (McGraw & Seneff, 2008) and Rainbow Rummy (Yoshimoto, McGraw, & Seneff, 2009). In these two systems, the user talks to the system to select and move playing cards. Nevertheless, the systems are designed mainly for vocabulary learning, and do not emphasize other aspects like sentence formation or comprehension.

The work we present in this paper relies on several previously developed language processing systems. TINA (Seneff, 1992), the language understanding system, is a top-down parser, which uses a core context-free grammar augmented with additional rules to enforce long-distance constraints. Special features have been recently added to improve parsing efficiency for Chinese input (Xu, Liu, & Seneff, 2008). The output of TINA is a hierarchical meaning representation that does not explicitly encode word order information. The meaning representation can be converted back into a sentence via a language generation system GENESIS (Baptist & Seneff, 2000). GENESIS uses a context-sensitive lexicon to select appropriate word senses and a set of recursive rules to decide the order of the constituents. Depending on the choice of the rules, GENESIS can produce strings in any format, representing not only natural languages, but also formal languages, such as SQL and HTML.

3. The Games

In this section, we will briefly introduce the three games we have developed for Mandarin learning within the common framework. The games are Web-based and accessible from a shared URL. At the login page, the user chooses the genre of the game, the domain, and the starting level. Figure 1 shows screenshots of the translation game and the question-answering game.

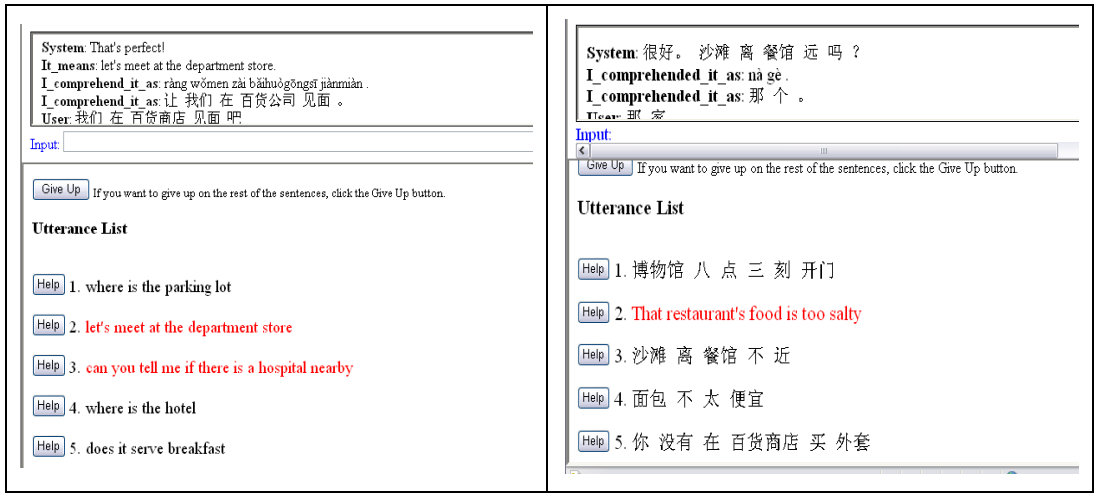


Figure 1. Screenshots of the translation game (left) and the question-answering game (right).

The main goal of the reading game is to help students learn Chinese characters. The student's task is to read out loud a list of Chinese sentences randomly generated by the system. The sentences can be displayed in either Pinyin or Chinese characters, depending on the student's preference. A help button is associated with each of the task sentences to provide a synthesized speech demonstration. To make the game more interesting, the student can read the sentences in any order. When the student records a spoken sentence, the system will not only echo his speech, but also provide an English translation of the sentence, even when it is not in the task list. If the student's speech matches any of the task sentences, the system will congratulate him and mark the sentence as completed. When all of the sentences are cleared, the system assesses the student's performance and reports a score. The game level is then adjusted according to the score.

When the student becomes more familiar with the Chinese characters and accumulates some vocabulary, he can start to play the translation game. In the translation game, instead of a list of Chinese sentences, the student is given a list of random English sentences. The student needs to construct a Chinese sentence of equivalent meaning by himself. Again, he can choose any order to translate the list. The system will echo the student's input, give a Chinese paraphrase and an English translation of it, and judge whether the input sentence is a correct translation of any of the task sentences. The judgment is based on syntax and semantics, so the student is allowed to translate in different ways. If he does not know how to translate a sentence, he can click the help button to hear and see a reference translation, presented in both characters and Pinyin. He can also type an unfamiliar English word or phrase in the input box to get a translation.

After playing the reading game and the translation game, the student should be prepared to try the question-answering game, in which the game scenario is almost completely in Chinese. The system randomly generates a list of Chinese statements, and then poses a question in Chinese based on one of the statements. The student needs to be able to read the displayed statements, to understand the spoken question, and to answer the question correctly in Chinese. Therefore, in this game, all of listening, reading, and speaking abilities can be practiced. Three chances are given for each question. The student can answer the question in various ways, either in short, in full, or somewhere between, as long as it is acceptable in Chinese. If the answer is correct, the corresponding statement will be turned into English. Otherwise, the system will give feedback according to the student's input, and guide her to a desired answer. As in the other two games, the student can ask for help, or ask the system to repeat the question if necessary.

In all three games, the student has an alternate input method. In a noisy environment where speech input is compromised, or if the student is having trouble being understood due to a heavy accent, they can opt to type their sentences into the input box using Pinyin format. The system will propose the character sequence based on the Pinyin input, and will also identify and mark all the characters that the student typed with an incorrect tone.

4. The Framework

The framework of our games is illustrated in Figure 2. The system consists of one or multiple speech recognizers, one or multiple speech synthesizers, the GUI interface, and the dialogue manager with a set of building blocks providing different NLP operations. The recognizers send N-best hypotheses of the student's input to the dialogue manager. After processing, requests are sent to the synthesizers to output the spoken responses. The dialogue manager also communicates with the GUI to receive user information and text input, along with updating the displayed content.

The dialogue manager is the core component of the framework. Together with its building blocks, it provides easy control over the processing steps during a dialogue turn. The control flow is managed by a set of control rules, called a control script. Each rule contains a parameterized operation and an optional trigger condition. The operations are provided by the building blocks. The framework contains six core building blocks. Two blocks, "create frame" and "paraphrase frame," use our pre-existing language understanding and generation systems. In addition to these two most basic NL operations, we have also developed four other core building blocks to handle game creation, management, and evaluation, which are very useful in developing language learning games. Besides the core building blocks, game developers can also provide their own specialized blocks to extend the capabilities of the dialogue manager.

The dialogue manager maintains a shared space representing the dialogue state. Both the dialogue state and the control rules are represented in Galaxy frame format (Seneff, Hurley, Lau, Pao, Schmid, & Zue, 1998). When executing a control script, the dialogue manager examines the conditions of each rule sequentially against the dialogue state. If the conditions are satisfied, the operation specified in the rule is executed. Several control rules can be grouped to form a macro to be reused in the script. An example is shown in Figure 3, in which the sequential operations of "create frame" (parsing) and "paraphrase frame" (language generation) form a Chinese-English translation macro. The macros are an important improvement over our previous design. Not only do they improve the readability of the control script, but they also support disjunction and iteration through recursive macro calls. These extensions provide much better control over program flow. Together with the high level NL operation, the framework provides an easy way for developers to construct different systems through specialized control scripts.

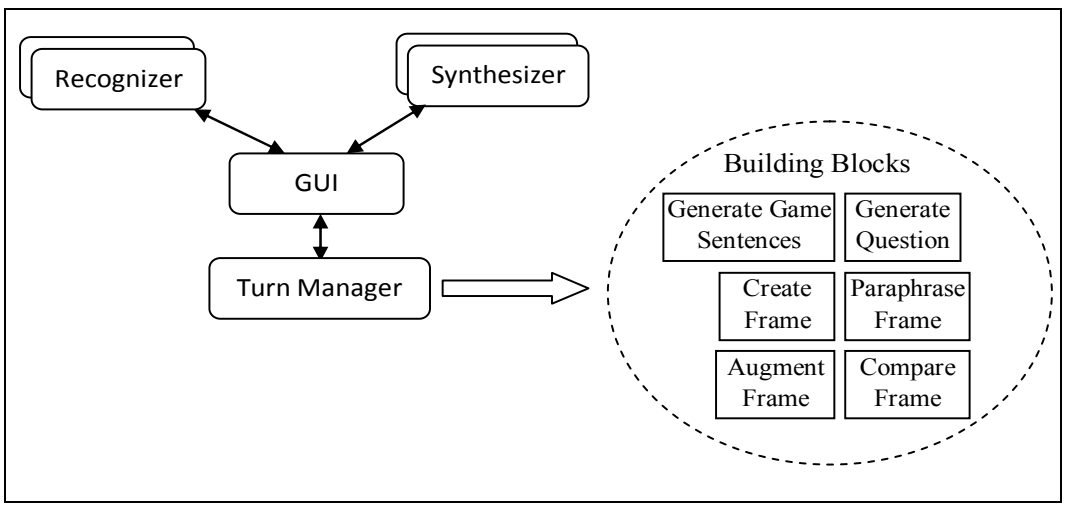


Figure 2. The framework

```
:Chn-Eng-Translate (
        {c rule
            :condition ":input_string"
            :variables {c variables
                            :domain "hanyu" }
            :operation "create_frame" }
        {c rule
            :condition ":parse_frame"
            :variables {c variables
                            :language "English" }
            :operation "paraphrase_frame" } )
```

Figure 3. An example of a Chinese-English translation macro.

In the following subsections, we will describe all of the operations that our core building blocks provide. We will also introduce some of the key macros that are useful in building the three game systems for language learning.

4.1 Generate Game Sentences

This operation controls the game level and generates one or a list of game sentences for the current game level. In the beginning of each game round, a list of task sentences is randomly generated from specified lesson templates. During the round, the operation marks the sentences that the student has completed, and, for games running in system control mode such as the question-answering game, it also chooses another sentence in the list for the next turn. When a round is completed, the operation calculates a performance score for the student and decides how to adjust the game level.

The operation generates the game sentences from a set of templates. The templates are divided into lessons, which can be organized by topics and/or grammar points. Each lesson has a list of sentence patterns and the associated vocabulary, just as in traditional textbooks. The patterns are essentially forest-like nodes, and the vocabulary is contained in the leaf nodes. The patterns and vocabulary introduced in the previous lessons are augmented by later lessons, which makes the templates easy to maintain. In generation, a starting pattern is randomly chosen from the specified lesson, and each non-leaf node is expanded with one of its child nodes based on a random selection process, until every node in the pattern is a leaf node.

In addition to the sentence generation that produces a single sentence in one language, we have also developed a more sophisticated generator that makes use of special non-context-free rules to automatically generate a pair of "synchronized" sentences in two different languages with the same meaning. By "synchronized," we mean that the sentence generator can generate a bilingual pair, guided by a special notation scheme in the templates. As shown in Figure 4, the vocabulary entries of the synchronized templates contain a vertical bar to separate the lexical entries for the two languages. Two special tags, "_L" and "_R", are used to deal with the different word order between the two languages. For instance, English and Chinese demand different positions for the prepositional phrase "from." In the example template, the pattern ":from" can generate a bilingual phrase "from the beach | 离 沙滩". ":from_R" means to take the right part of the output, which is the Chinese string, and put it before the adjective. Likewise, ":from_L" instructs it to take the left part of the output, which is the English string, and put it after the adjective. With this feature, it is easy to provide a generated string and its associated high-quality translation, which is very useful for many aspects of the games.

```
{c lesson
    :templates ( ":place :is :from_R :far :from_L" )
    :place ( "(the hotel | 宾馆)" "(the restaurant | 餐厅)" )
    :is ("(is | )")
    :from ("(from | 离) :attraction")
    :attraction ( "(the beach | 沙滩)" "(the park | 公园)" )
    :far ( ("very far | 很远") ) }
```

Figure 4. An example of the synchronized template. One possible output of this template can be "the hotel is very far from the beach |宾馆离沙滩很远"

4.2 Create Frame and Paraphrase Frame

"Frame" here stands for "linguistic frame," which is a hierarchical meaning representation in the Galaxy frame format. "Create frame" and "paraphrase frame" are a pair of operations which convert between a string and a frame. Going from a string to a frame is the parsing process, and going in the other direction is essentially language generation.

As mentioned in Section 2, we rely on TINA and GENESIS for language understanding and generation in the games. TINA can be used to parse the template-generated game sentences, as well as the N-best list of the student's input. Besides the features mentioned briefly in Section 2, we further implemented a special two-pass parsing scheme in the operation "create frame". In Chinese, the way numbers and proper noun phrases are constructed often causes the parser's theories to grow exponentially when a generic grammar is applied. To avoid this situation, the two-pass parsing scheme first tags out these troublesome phrases using a very small shallow grammar, then creates parse trees for each of them (which we call element trees), and replaces the phrase with a single tag representing each element tree. Then, in the second pass, the parser creates a parse tree for the tagged sentence. Finally, the element trees are inserted at the appropriate locations in the second-pass parse tree to form a complete tree.

The language generation unit, GENESIS, also plays an important role in the system. It can be used to generate a paraphrase in the same language as the input, a translation into another language, a system's response, or an HTML string that can be displayed to the student.

For both TINA and GENESIS, we have developed generic grammar rules and generation rules in both English and Mandarin Chinese. The rules were developed based on the IWSLT[2] corpus, a spoken corpus of telephone quality speech collected from travelers. It covers a wide

[2] Internation Workshop on Spoken Language Translation

range of topics such as weather, flights, navigation, dining, shopping, sports, etc., is quite appropriate for everyday language, and is especially well suited to the needs of a traveler, which fits well with realistic roles for a language learner. With these generic rules, to export an existing game into a new domain of interest only involves adding a new lexicon corresponding to that domain, along with some other minor changes. The form of TINA's output, the linguistic frame, is quite suitable for language portability, especially because it disregards word order information. As most parts of the frame are language independent, we can convert the games for teaching Chinese into teaching English simply by reversing the grammar and generation rules.

For further information about TINA and GENESIS, along with their ability in paraphrasing and translation, we refer you to (Seneff, 1992) and (Baptist & Seneff, 2000).

4.3 Transform Frame

The function of this operation is to alter the elements in the frame. This has many uses, one of which is to convert a frame representing a statement into another frame that represents a question.

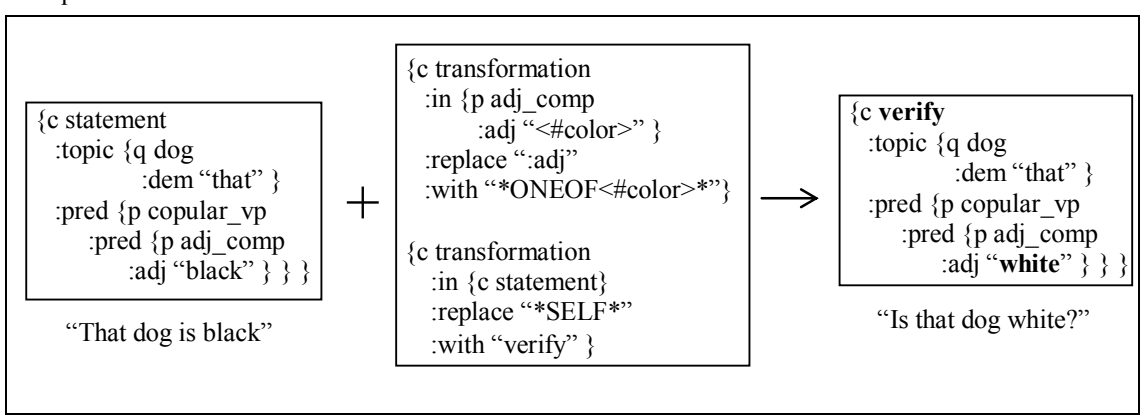


Figure 5. Example of a transformation rule with alternative choices.

The transformations are guided by formal rules. Each rule has three basic clauses, which describe the conditions under which the rule should be triggered, the part to be transformed, and the result after transformation. Wildcard values like ANY, NONE, SELF, etc., are adopted in the syntax to make the rules simple to write but powerful to express all kinds of transformations. A detailed description of the transformation rules can be found in Xu (2008). These transformation rules also support some randomness, allowing alternative outputs depending on a randomly generated outcome. Thus, as exemplified in Figure 5, the color adjective can be replaced with another random color via the rule.

4.4 Augment Frame

For our question-answering game, we need to deal with context resolution, since the answer would oftentimes be a fragment. Also we need to resolve its correctness in terms of both answering the question and not providing additional information that may be inconsistent with the given statement. For this task, we have developed a new building block which provides the "augment frame" operation. The operation does not depend on any domain-specific knowledge. In this algorithm, the frame representation of the previous utterance is aligned with the frame representation of the current utterance. Then, we can determine the omitted information and the pronoun referral in the current utterance, and we can augment the frame to include the complete information.

The alignment algorithm is based on two aspects: the anchor point and the similarity of the aligned frames. Depending on the type of the previous utterance, different anchor points are chosen. For wh-questions, the anchor point is the element that is questioned. For other types of utterances, the top level predicate is chosen. The best alignment is computed based on the constraint that the anchor point should be overlapping, and the similarity score of the two aligned frames is maximized.

Two examples are given in Figure 6. In the first example, the current short utterance "not far" is augmented into "the beach is not far from the hotel." by looking at the previous utterance, which is a yes-no question "is the beach far from the hotel?" In the second example, the previous utterance is a wh-question "where is the beach far from?". After augmentation, "hotel" becomes "the beach is far from the hotel". Note that, in this example, "hotel" is the topic of the current utterance. It, however, becomes the value of the key ":from" after augmentation, so that the anchor point "*question*" is overlapped.

This context resolution by augmentation approach has limited usage. It requires the topic and the basic structures of the utterances to remain unchanged. In semi-dialogue scenarios, like question-answering, this condition holds, and the augmentation algorithm is very effective. A short answer can be augmented into a complete answer by aligning it with the question. Then, if a follow-up question is posed based on the answer, the kv-frame of the second question can be augmented by aligning it with the augmented kv-frame of the previous answer.

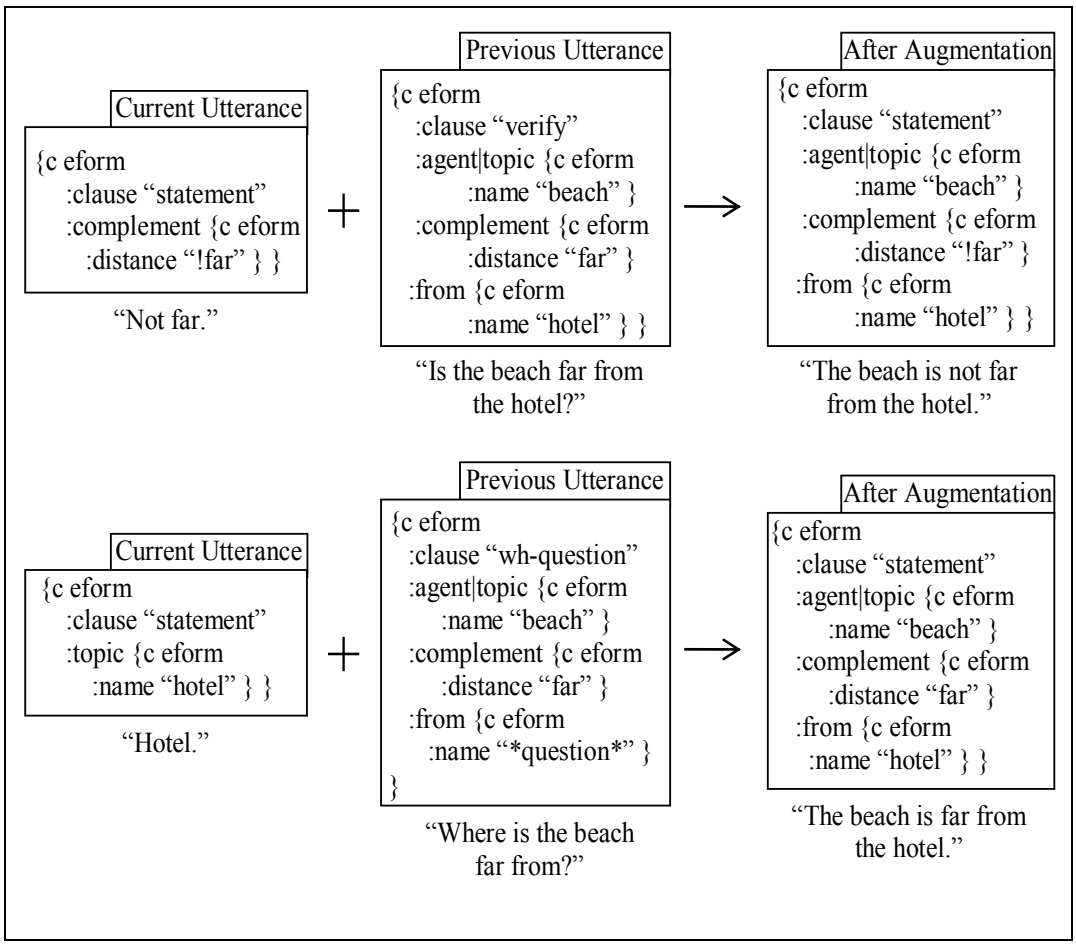


Figure 6. Examples of Context Resolution by Augmentation.

4.5 Compare Frames

This operation provides the ability to examine the differences between two frames. The comparison can be done blindly, *i.e.*, treating each element in the frame as equally important as the others. This setting is suitable when a direct match is desired. For example, in the translation exercise, students are encouraged to follow the way the original sentence is expressed, instead of paraphrasing "not far" into "close". If flexible expression is tolerable, the algorithm can perform a more heuristic comparison according to the parameters sent to it. Thus, it can be instructed to treat "not far" and "quite close" as having equivalent meaning. It can treat head words and modifiers differently, so that a mistake in the name of the patient will result in more deduction than a mistake in its color. It can also make different judgments for binary-value elements and multi-value elements, so that an insertion of the negation "not" will have a different comparison result from an insertion of a degree "quite".

The operation produces a summarization after the comparison, including the substituted elements, the inserted elements, the deleted elements, and an overall score. This output can be used not only to judge the correctness of the student's answer, but also to identify duplication or contradiction in the game sentences that were randomly generated.

4.6 Macros

Macros are formed when several operations are sequentially grouped together. We will introduce four useful macros in this subsection, which are diagrammed in Figure 7.

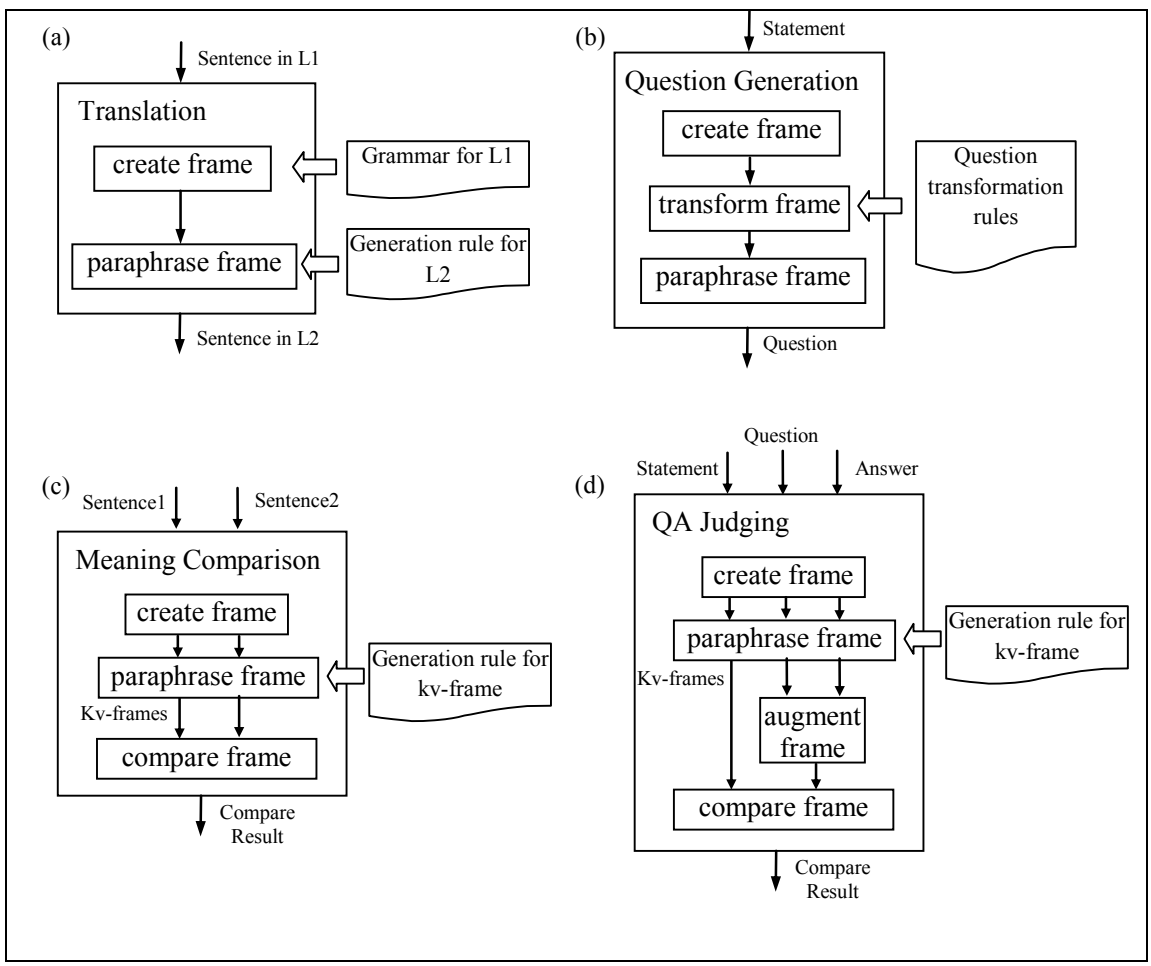


Figure 7. Macros: (a) Translation, (b) Question Generation, (c) Meaning Comparison and (d) QA Judging.

- **Translation.** The translation macro is very simple. The macro parses a string and produces a linguistic frame. Then, the generation rules for the target language are used to convert the linguistic frame into a well-formed text string.

- **Question generation.** This macro produces a question string from a statement string. The statement string is parsed into a linguistic frame. The linguistic frame is then transformed by question transformation rules, and, finally, a question string is generated from the transformed frame.
- **Meaning comparison.** This module compares the meanings of two linguistic frames by first converting them into kv-frames (key:value frames), which provide a more succinct representation of their semantic content. We use the "paraphrase frame" operation to generate the kv-frame. The actual comparison is then performed on the kv-frames instead of the linguistic frames from the parser.
- **QA judging.** This is very similar to the meaning comparison macro, except that, for question-answer judging, an additional step is taken to augment the answer kv-frame into a complete kv-frame by aligning the frame with the question kv-frame. Then, it is compared against the statement kv-frame.

5. The Game Implementation

In this section, we will show how basic operations and the macros described in the last section can be used easily to build different systems. The architectures of the arrangements of the operations and macros will be illustrated, with brief literal descriptions.

5.1 Reading Game

The first game we implemented was the reading game in the travel domain. Although the basic content of the game is very simple, interesting features were added to lessen the possibility of boredom. We wrote the lesson templates in English, rather than in Chinese. Then, we used the translation macro to automatically translate the sentences generated from these English templates into Chinese. This capability allows students to edit and create their own lesson templates without having knowledge of Chinese characters. The system can automatically tell them the corresponding Chinese. We also created an *inverse* translation macro. Whenever the student records an utterance, the system can provide the English meaning of the utterance he just read. When the student mispronounces a word, misrecognition may lead to an amusing English translation, which is more entertaining feedback than simply responding with "please try again". The system can also pronounce the sentence using the synthesizer when the student asks for help.

The framework of the reading exercise after adding these features is shown in Figure 8. The shaded blocks indicate the macros. Although it is a simple game, after utilizing the translation macro, the system already gives the student the impression that it understands what the student is speaking by providing an English translation.

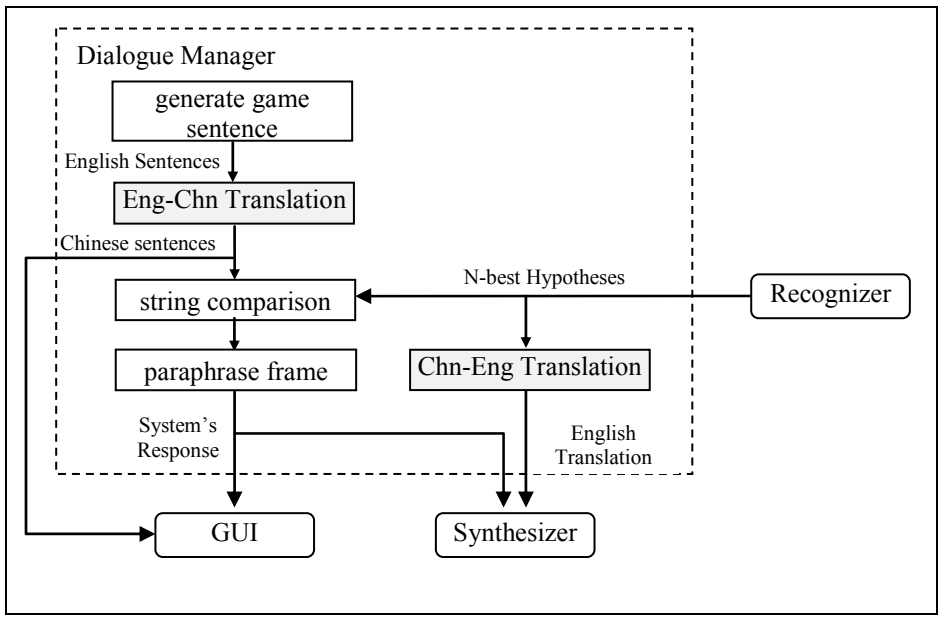


Figure 8. Framework of the reading game.

5.2 Translation Game

It is not difficult to extend the reading game so that it becomes a translation game. Figure 9 shows the framework of the translation game. It is almost exactly the same as the reading game shown above, except that the string comparison is replaced with the meaning comparison macro.

In the translation game, the system generates a list of game sentences from the English lesson templates and translates them into Chinese by the translation macro. This time, however, the English sentences are displayed instead of the Chinese or the Pinyin sentences. Another difference between the two games is that the reading game requires the student to read off the exact characters shown on the screen; in contrast, for translation, there is no unique answer. The student can translate a sentence correctly in multiple ways. So, instead of string comparison, the meaning comparison macro is adopted. To encourage the student to translate as literally as possible, the heuristic frame comparison is not used. Table 1 gives some examples of acceptable and unacceptable translations. The system echoes the student's speech, gives a Chinese paraphrase and an English translation of what the student said, and tells the student if the speech a match.

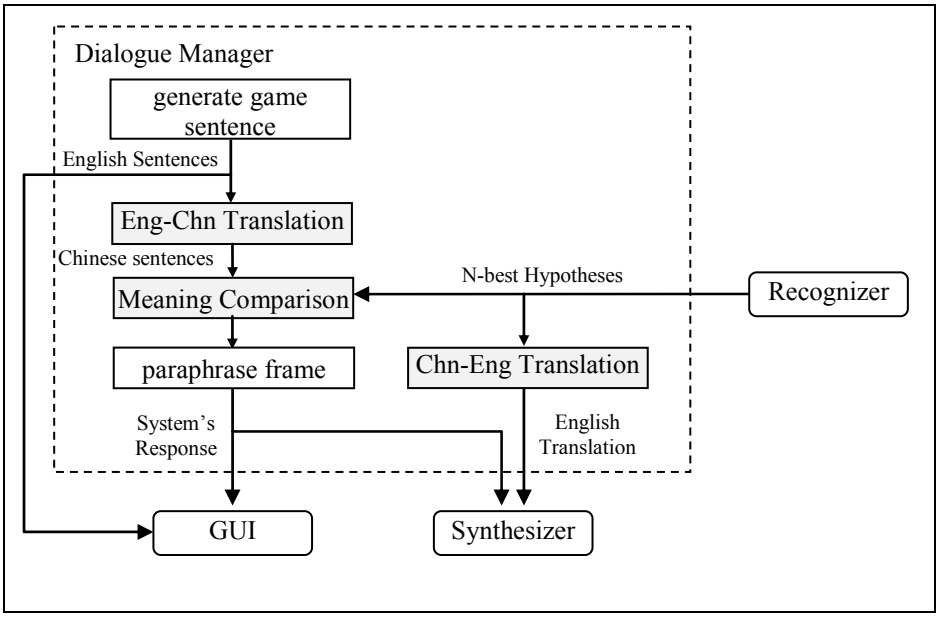


Figure 9. Framework of the translation game.

Table 1. Examples of accepted and rejected translations.

The museum opens at ten thirty.	Let's meet at the stadium.
√ 博物馆十点半开门	√ 让我们在体育馆碰头
√ 博物馆十点三十分开门	√ 咱们在体育馆见面吧
√ 博物馆于十点半开门	X 我们碰头在体育馆吧
√ 十点半博物馆开门	X 在体育馆见面
X 博物馆开门十点半	

Both the reading game and the translation game were developed in the travel domain first, and then exported to the more specific flight domain. The whole process of exporting, including writing new lesson templates, adding flight domain specific lexical and semantic information into the grammar and generation rules, training a new recognizer, and testing the system, took less than three weeks.

5.3 Question-Answering Game

The third game we built is a question-answering game. With the experience of the previous two games, this game was developed within two months, including the time spent developing the frame transformation rules for generating questions from statements. In this game, the student reads a list of statements on the screen, listens to the question posed by the system, and speaks the answer. When the answer is correct, the statement will be marked and turned into the English equivalent.

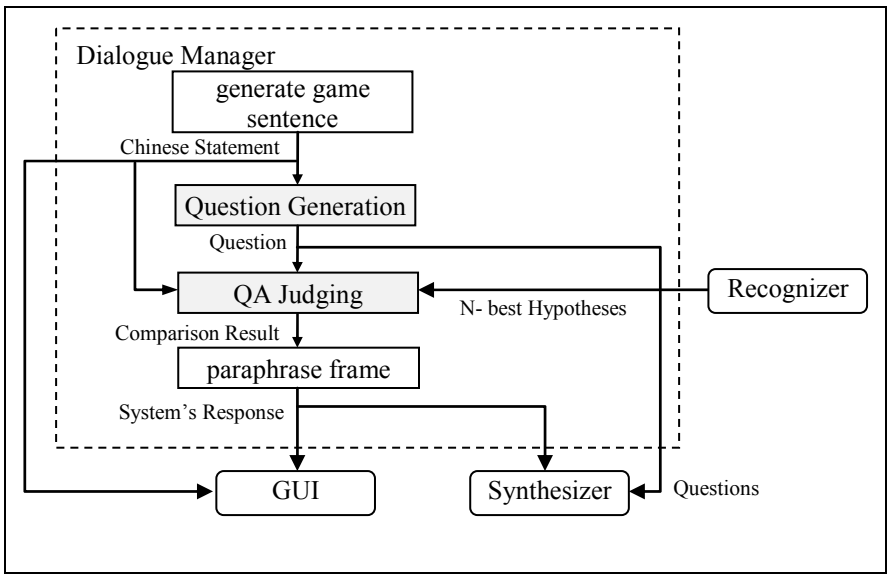


Figure 10. Framework of the question-answering game.

On the screen: 那只狗是黑色的 (That dog is black.)
你很喜欢吃蔬菜 (You like to eat vegetables very much)
System: 那只狗是白的吗？ (Is that dog white?)
Student: 不对，不是白的 (No, not white.)
System: 那么是什么颜色？ (So what color is it?)
Student: 是黑的 (It's black)
System: 很好。你喜欢吃蔬菜吗？(Good job. Do you like to eat vegetables?)
Student: 喜欢 (I do.)
System: 再具体一点 (Please be more specific.)
Student: 我很喜欢吃蔬菜 (I like to eat vegetable very much)

Figure 11. An example conversation between the system and the student in the question-answering game.

The framework of this game, shown in Figure 10, is a little different from the previous two, but not significantly. The translation macro is removed. Instead, we use the synchronized templates to generate Chinese statements and their English meanings at the same time. The Chinese sentences are processed into questions through the question generation macro. Transformation rules are written to include all kinds of possible questions. The question generation macro determines which rules apply to the current game statement, and randomly selects one to apply. After the student has spoken the answer, the QA judging macro takes in the N-best hypotheses from the recognizer, the questioned statement, and the question, and judges the correctness of the answer. Based on the comparison result, the system might give some advice or pose a follow-up question to guide the student to include the desired content in

the answer. An example of a conversation between the system and a student is given in Figure 11. From the conversation, we can see that the game is actually a simplified dialogue game, only that the dialogue is strictly limited in the scope of the game statements.

6. Evaluation of the Games

The most straightforward way of evaluating the framework and the dialogue manager is to evaluate the three games we implemented. As the games utilized different operations and macros, along with being connected in different ways, the effectiveness and flexibility of the framework can be proved by the successfulness of the three games.

We conducted the evaluation of the three games in two phases. In the first phase, we recruited several subjects to come to our lab, and gave them detailed instructions. In the second phase, we advertised our games to a list of users who are interested in Mandarin learning games and asked them to play the games by accessing a public URL via the Internet. We offered them gift certificates based on the amount of data they provided. They were less instructed on the games, and they might play the game in various environments. Due to the different settings of the two phases, we provide separate analyses for the two data sets. In both phases, we focused our evaluation on the system's performance, rather than proving pedagogical effectiveness. The reading game was not evaluated, because of its similarity to the translation game in terms of the architecture and the game procedure.

In all three games, we used SUMMIT, a landmark-based recognizer (Glass, 2003). The recognition output is constrained by an *n*-gram language model, that was trained using data automatically generated from our game templates. We developed an N-best selection process to score and select the hypotheses, choosing the one that best matched the dialogue context, if such an utterance existed.

We use two separate off-the-shelf synthesizers for synthesizing English and Chinese, respectively. Dectalk is used to synthesize English, and, for Chinese, a synthesizer provided by the Chinese Academy of Sciences is used.

6.1 In-Lab Evaluation Phase

6.1.1 The Translation Game

We implemented the translation game in two domains: travel and flights, which we did not distinguish during the evaluation. The lesson templates include twelve lessons for the travel domain and ten lessons for the flight domain. A single recognizer was used for both domains. The acoustics were trained from native speakers' data. An *n*-gram language model trained on the template-generated sentences augmented with IWSLT 2006 data was used to constrain the

recognition output. The vocabulary size was about 8.6K.

We recruited 5 subjects, 3 females and 2 males, to come to the lab. Each subject started at the first level, and was given five randomly generated utterances to translate in each round. We recorded the waveforms and the system's activity, as well as watching their behavior throughout their play. Advice was provided when they got stuck. Altogether, 615 utterances were collected from these five subjects.

We calculated the false rejection and false acceptance rate based on manual judgment. The false rejection rate was 8.6%, with almost all of the cases being caused by recognition errors. We listened to all of these waveforms and determined that most of the mis-recognized utterances were pronounced poorly or disfluently by the learners. The false acceptance rate was 0.9%. All of the false acceptances occurred when there was a minor syntactic problem in the sentence that was not identified by the system. For example, the user used an incorrect measure word for the noun. Encouragingly, we found that in the Chinese paraphrase the system gave back to the student, the syntactic problem had been automatically fixed, and we observed that the subjects did notice the implicit correction.

We calculated the average number of utterances the users spoke to complete one round, the average number of rounds they took to advance one level, and the average number of times per utterance they asked for help. The results are shown in Figure 12. The users are sorted on the horizontal axis to indicate their human-judged Chinese proficiency. The leftmost user is a native Chinese speaker. We can see that there is a good correlation between their real proficiency and the three values we measured. The users with lower proficiency tend to produce more utterances in one round, and tend to ask for help more frequently. The two numbers are the major factors for the system to assess the student's performance and to decide whether to adjust the game level. The result is that the poorer students tend to stay longer in the same level, as illustrated in the figure.

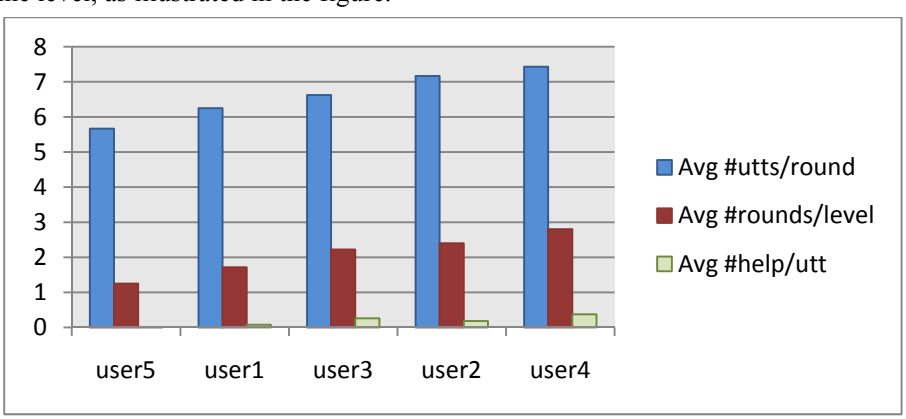


Figure 12. Performances of the users in the translation game. Users are arranged left-to-right in order of decreasing proficiency.

The game received positive feedback from the users. The users liked the feature that the system praised them, and they also appreciated the gradual introduction of new vocabulary and sentence patterns.

6.1.2 The Question-Answering Game

The question answering game was evaluated in a similar way as the translation game, but, a simulation phase was conducted as well to evaluate the quality of the questions and the coverage of the question types. The lesson templates are composed of seven lessons. Forty frame transformation rules were written to create 17 types of questions. We simulated 42 game rounds, 6 for each lesson. In each round, 5 statements and questions were generated. We determined manually that all the questions were well-formed. The distribution of the question types is illustrated in Figure 13. A fair percentage of yes-no questions and wh-questions were generated in the 210 questions, and within the wh-questions, the different types of questions were distributed reasonably.

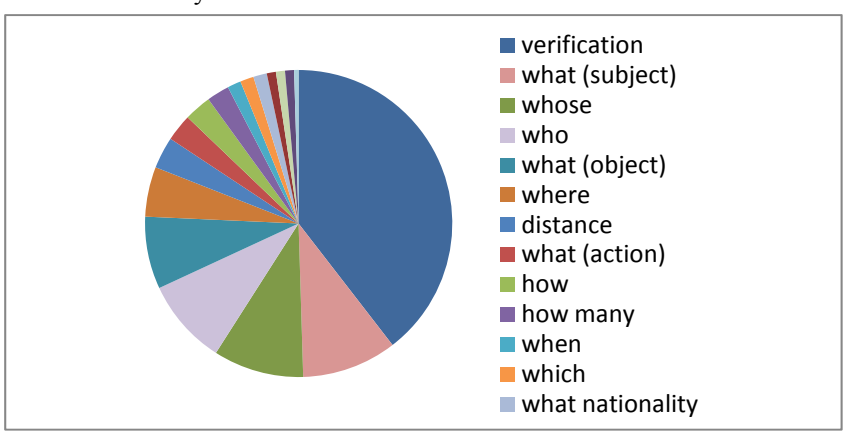


Figure 13. Distribution of the question types.

For the game system evaluation, we retrained the recognizer on an augmented synthetic corpus of utterances to model the statistics of both the translation game and the question answering game. The vocabulary size of the language model was enlarged to around 9K. Seven subjects, 3 males and 4 females, participated in the in-lab evaluation. Three of them were native speakers. Although the participants accessed the game from different computers, we ensured that they all used a high-quality microphone in a quiet environment. 732 utterances were collected from these subjects.

We categorized the utterances into three types of answers: blank-filling style short answers, such as a single yes/no or a single noun; full answers which essentially are a repetition of the statement in the list that answers the question; and other answers that are somewhere between the short answers and the full answers. The distribution of the three types,

shown in Figure 14, is quite balanced.

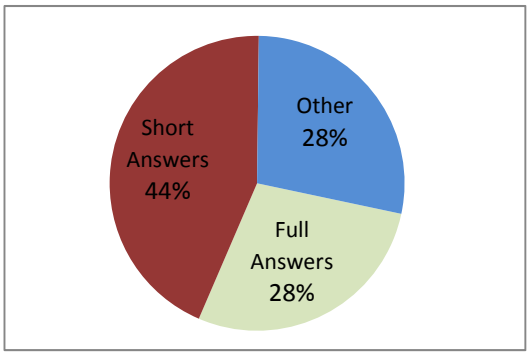


Figure 14. Types of answers

In the question-answering game, the system has several different responses instead of binary choices. Due to this, we calculated the accuracy of the responses instead of the FA/FR rates. The accuracy was 91.7%, with 57 out of 61 incorrect responses caused by recognition errors. The rest of the errors were caused by ill-formed kv-frames, which were fixed before the public evaluation phase.

6.2 Public Evaluation Phase

In this phase, we opened our games to the Internet users. An email message containing the URL and some game instructions was sent to a list of possibly interested users worldwide. We provided awards for the users who completed a certain number of game rounds. The users were free to choose to play any of the three games they liked, as well as to select their own initial game level. The number of utterances in each round was fixed at five.

In ten days, 23 users accessed our games, including three users whose data we discarded in the analysis due to quality issues: User 3 only provided two utterances in the middle of two game rounds of User 2; User 11 recorded his almost inaudible speech in an extremely noisy background; and User 12 used a poor-quality microphone which output highly saturated waveforms and resulted in a very high recognition error that was not comparable to that of any of the other users. All of the remaining 20 users tried the translation games; 9 also played the question-answering game; and 1 also tried the reading game. The 20 users include 7 females and 13 males. We manually judged their Chinese proficiency on a 5-point scale based on their pronunciation and intonation. Five points indicates a native speaker, and one stands for really poor pronunciation. The average proficiency score was 3.1, with four of the users judged to be native speakers.

From the 20 users, we successfully collected 1754 utterances for the reading/translation game, and 924 utterances for the question-answering game. We discarded 151 empty

utterances and 26 utterances that the dialogue manager did not receive due to communication problems. We also discarded utterances related to one problematic game sentence pattern, which produced an incorrect reference translation and led to confusion. This problem was fixed after the first two days of the experiment. After pruning, we were left with 1530 utterances for the reading/translation game, and 875 utterances for the question-answering game.

The overall sentence recognition error rate for all three games was 29.6%. Although this number is quite high, two factors played a critical role. Nearly a third (30.4%) of the mis-recognized sentences were either not a Chinese sentence, an ungrammatical Chinese sentence, or contained a totally mispronounced word. The other factor is that there were many repeated errors. When an utterance was not recognized correctly, the user usually spoke it again, essentially repeated verbatim, and it was very likely that the second utterance would not be recognized correctly as well. To verify this theory, we calculated the rate of repeated recognition errors. We define the rate of repetition to be the total number of mis-recognized utterances divided by the unique number of mis-recognized utterances. The unique number of mis-recognized utterance with recognition errors were counted independently within each game round, so that two identical misrecognized utterances in two different game rounds are distinguished. The rate of repetition of the three games was 1.77, which means that each unique recognition error is repeated almost twice. If the repeated errors are excluded, the sentence error rate for recognition goes down to 19.2%.

The recognition error rate also varies greatly among users, as shown in Figure 15. The users in the plot are sorted by their human-judged proficiency. It is clear from the plot that the recognition error is influenced greatly by factors other than their nativeness, which are likely to be microphone quality and environmental noise.

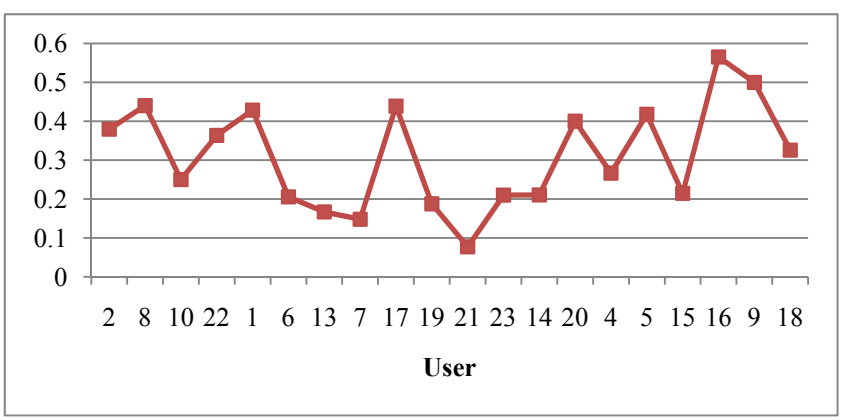


Figure 15. Sentence recognition error rate by users. Users are arranged left-to-right in order of decreasing proficiency.

Table 2 shows the error rates of the system responses. As in the in-lab evaluation, we calculated the false acceptance rate and the false rejection rate for the reading/translation game, and we did not distinguish the detailed error type for the question-answering game. We can see that the error rates were similar to those in the in-lab evaluation. Most of the errors were still caused by recognition errors. Others were mainly due to incorrect or missing information in our meaning representations. For example, "饭店" can mean either restaurant or hotel, but our linguistic frame only contains one of these interpretations. Also, we did not handle verb reduplication appropriately, so that in the utterance "请帮帮我" (please help me), we treated the two occurrences of the verb "帮" as two different verbs, and falsely rejected the utterance.

Table 2. Error rates of the system responses in the public evaluation phase

Game Genre	Error Type	Error Rate	% Caused by Recognition Error
Reading/Translation	False Acceptance	2.0%	90.3%
	False Rejection	11.6%	89.8%
Question-Answering	Incorrect Responses	9.8%	88.3%

In the public evaluation, it is more difficult to determine whether the users with poorer Chinese got more practice from simple statistics like average number of utterances they took per round. The problem is that the number of utterances per round is also dependent on environmental factors such as microphone quality and background noise level. We also notice that some users inexplicably repeated an already matched utterance, and thus had more utterances in each round. To take these two factors into consideration, we define a normalized average number of utterances per match as in Equations (1) and (2). In the equations, SER is the sentence recognition error rate, SER_{user} is the sentence recognition error rate attributed to users' mistakes. SER - SER_{user} gives the recognition error rate caused by other factors like background, channel, and acoustic models. Thus, a high c_{norm} means the user recorded in a quiet environment with a high-quality microphone. On the other hand, a low c_{norm} means the user probably used a poor recording device or played the game in a noisy environment.

$$\bar{u} = c_{norm} \times \frac{\#Total\ utterances}{\#Total\ matches} \tag{1}$$

$$c_{norm} = 1 - (SER - SER_{user}) \tag{2}$$

Figure 16 shows a plot of $\bar{u}$ for the users who completed at least one round of the reading/translation games. The users are sorted by decreasing Chinese proficiency. The logarithmic trend line illustrates that it took more effort for the lower proficiency user to complete a match. Two anomalously low points for User 7 and 16 result from their frequent actions of asking for help. They clicked help every two utterances on average, so their translations were mostly our reference translations, which were mistake-free and easy to

recognize. The high value of User 17 is due to his multiple repetition of two wrong translations which he probably thought to be correct.

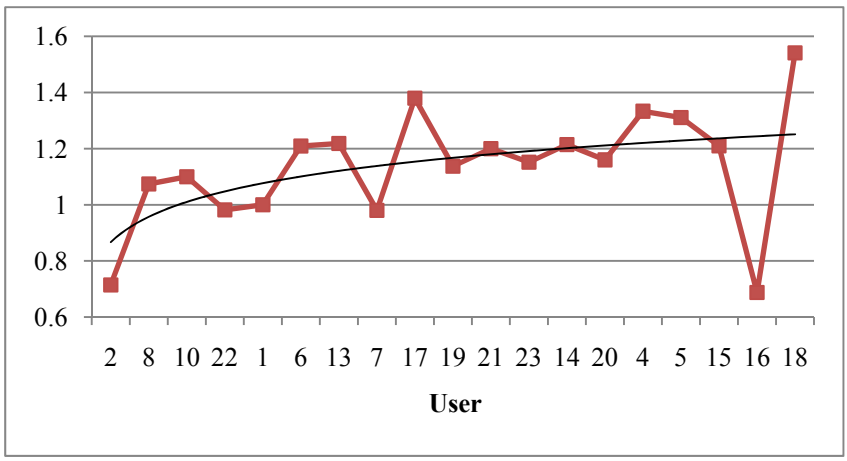


Figure 16. Normalized average number of utterances per match with the logarithmic trend line for the reading/translation game. Users are arranged left-to-right in order of decreasing proficiency.

For the question-answering game, we did not find a good correlation between ū and proficiency. In examining the log files, we determined that many users were confused with the pronoun reference of "you" and "I". Many users did not catch the conversational design of the game, and answered "your dad is Mike" when the system asked "who is your dad?". This confusion added much noise to ū, which resulted in it not being representative of the proficiency level.

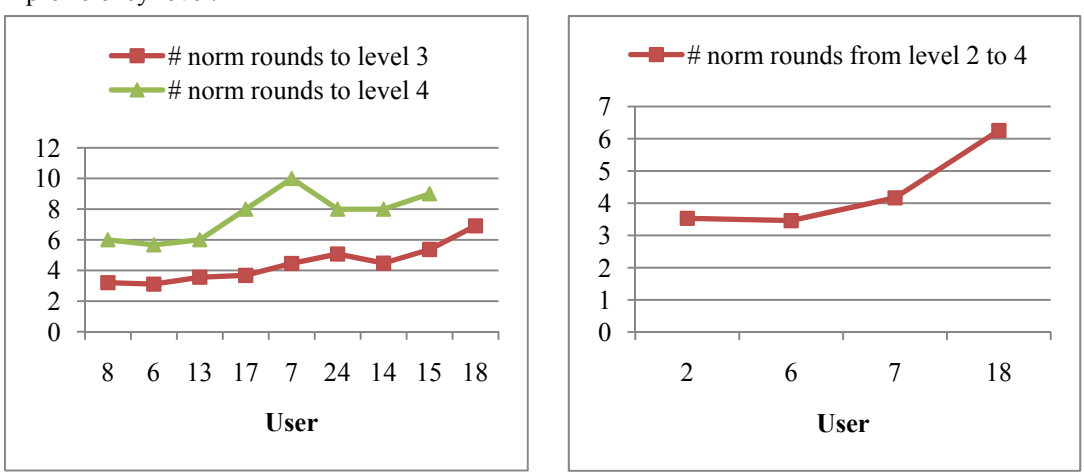


Figure 17. Normalized number of rounds to reach Level 3 and Level 4 for reading/translation game (left), and from Level 2 to 4 for the question-answering game (right). The users are sorted by decreasing human-judged proficiency.

We also analyzed how closely the system's assessment is related to the user's Chinese proficiency. Since many users did not play enough rounds, and often quit the last round in a session without completing it, it is not meaningful to calculate the average number of rounds per level. Instead, we counted how many rounds they took in one game session to reach Level 3 and Level 4 from Level 1 for the translation game. For the question-answering game, we noticed that it took the users one or two rounds to understand how to play the game, as well as the pronominal reference, so we discarded the information in Level 1 and counted the number of rounds they took from Level 2 to Level 4. The numbers of rounds are normalized by coefficient c_{norm} to reduce the differences in the recording conditions. The result is plotted in figure 17. It can be observed from the plots that as a whole, to reach the same level, users with lower proficiency spent more rounds, which means that our game has a reasonable assessment algorithm. The exceptional high number for User 7 to reach Level 4 resulted from an incomplete round at Level 3 which dropped him back to Level 2.

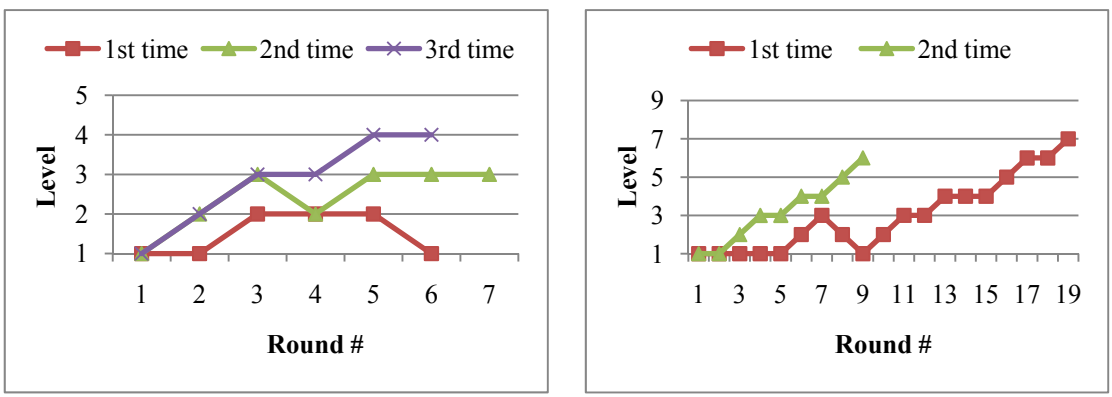


Figure 18. Levels User 18 achieved in different game sessions for translation game (left) and question-answering game (right).

Several users accessed our system multiple times. Among them, we noticed a low-proficiency user who played a total of 70 rounds. We found her making a lot of progress during these game plays. Figure 18 illustrated the levels she achieved in different game sessions. We can see that for the same number of rounds she reached a higher level when she repeated the game for a second and third time. The progress can be attributed to both increased acquaintance with the game and improvement in Chinese proficiency. For example, she had trouble with the syllable "chi" which she pronounced as "qi" causing much misrecognition. After several rounds, she realized the problem and tried hard to correct it. Finally, she learned the correct pronunciation and had it recognized correctly.

The users gave us considerable feedback on the games. In most of the feedback, the users showed their fondness for the games. Figure 19 shows some of the comments we received from the users. Most of the users found the games to be fun and helpful. They would like to

play again and recommend them to their friends. Some of the users also advised that the interface should be improved to become easier for first-time users. Some of the users were very careful and pointed out mistakes in the synthesized replies. Several users tried to explore the space that our system is able to handle by speaking their own utterances. Their feedback was very helpful for our future development.

"It's a confidence booster for one. When practicing speaking, it's nice to have it repeat back what I said and to know I said it right. You can't really get that with a human, it would probably drive them nuts."

"The hardest part of learning Chinese to me is finding someone to practice with. I haven't used any tool thus far that had such a great amount of feedback."

"It's a good way to learn new words."

"I think this is just good. Besides you already have other games focusing on vocabulary. Though for me building my vocabulary is important, making proper sentences in Chinese is even (more) important and compelling."

" (The game helps) Recalling different ways of saying the same thing."

Figure 19. Some of the comments from the users.

7. Conclusions and Future Work

We have developed a framework for building interactive speech-enabled language learning games. We introduced the Galaxy frame representation based dialogue manager, which operates according to a control script to enable the game developers to access natural language process capabilities in an easy way. Several generic building blocks have been newly developed, or adapted into the framework to provide different natural language operations, including game sentence generation, parsing, language generation, frame transformation, frame augmentation and frame comparison.

Three games have been built using the framework: a reading game, a translation game, and a question answering game. From the subject-based evaluation, we verified that the game systems were successful. The system responded to the users appropriately about 89% of the time. The assessment of users' performance correlated well with the users' true proficiency. The users were generally positive towards the systems. The success of the three games showed that the framework is useful. The dialogue manager handles the different game procedures correctly according to the control scripts, and the building blocks performed the desired functions correctly.

The complexity of the three games increases gradually. Starting from the simple reading game to the question answering game, more language processing units were utilized. As stated

in Section 5, the question answering game can be viewed as a semi-dialogue game, so the next step is to build a real dialogue game on the framework. In the question answering game, the approach to context resolution is simply by augmentation. This approach is simple, but it also limits the complexity of the dialogue. For a real dialogue game, a more generic approach would be needed. Also, more sophisticated dialogue management is required. Our group has developed dialogue systems in specific domains, and we believe that, with the help of these existing technologies, it would not be too hard to build a dialogue game for language learning purposes with domain and language portability. This will be the main focus of our future research.

Acknowledgements

This research was funded by ITRI and a grant from the Delta Electronics Environmental and Educational Foundation. We would like to thank Ian McGraw for his help in providing the WAMI toolkit that made development of the Web interface relatively easy. We would also like to acknowledge Anna Goldie for her help in making up the lessons for the question-answering game.

References

Active Chinese. (2006). Retrieved 2008, from Active Chinese: http://www.activechinese.com/.

Baptist, L., & Seneff, S. (2000). Genesis-II: A Versatile System for Language Generation in Conversational System Applications. *Proc. ICSLP*, (pp. 271-274). Beijing, China.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proc. HLT*, (pp. 819-826). Morristown, USA.

Chan, M. K. (2003). The Digital Age and Speech Technology For Chinese Language Teaching and Learning. *Journal of the Chinese Language Teachers Association* , 38(2), 49-86.

CHANG, J. S., & CHANG, Y.-C. (2004). Computer Assisted Language Learning Based on Corpora and Natural Language Processing: The Experience of Project CANDLE. *An Interactive Workshop on Language e-Learning*, (pp. 15-23). Tokyo, Japan.

Chengo Chinese. (2004). Retrieved 2007, from Chengo Chinese: http://www.elanguage.cn/

Ehsani, F., & Knodt, E. (1998). Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. *Language Learning & Technology* , 2(1), 54-73.

Glass, J. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech and Language* , 17(2-3), 137-152.

Kunichika, H., Katayama, T., Hirashima, T., & Takeuchi, A. (2003). Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation. *Proc. ICCE2004.*

McGraw, I., & Seneff, S. (2008). Speech-enabled Card Games for Language Learners. *Proc. AAAI.* Chicago, USA.

Seneff, S. (1992). TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics ,* 18(1), 61 - 86.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). GALAXY-II: A Reference Architecture for Conversational System Development. *Proc. ICSLP.* Sydney, Australia.

Xu, Y. (2008). *Combining Linguistics and Statistics for High-Quality Limited Domain English-Chinese Machine Translation.* Master's Thesis, MIT, Cambridge, Massachusetts.

Xu, Y., & Seneff, S. (2008). Two-Stage Translation: A Combined Linguistic and Statistical Machine Translation Framework. *Proc. AMTA.* Honolulu, USA.

Xu, Y., Liu, J., & Seneff, S. (2008). Mandarin Language Understanding in Dialogue Context. *Proc. ISCSLP*, (pp. 113-116). Kunming, China.

Yoshimoto, B., McGraw, I., & Seneff, S. (2009). Rainbow Rummy: A Web-based Game for Vocabulary Acquisition using Computer-directed Speech. *Proc. SIGSLaTE 2009.* Warwickshire, UK.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 161-180

Evaluating Two Web-based Grammar Checkers - Microsoft ESL Assistant and NTNU Statistical Grammar Checker

Hao-Jan Howard Chen*

Abstract

Many ESL students need to improve writing skills to pass various language tests; thus, writing teachers need to read many compositions and provide feedback. To help ESL teachers reduce their teaching load and to give students faster feedback, various English grammar checkers have been developed. Few of these PC-based grammar checkers, however, are widely available to ESL learners. As the Internet has become an important tool for language education, web-based grammar checkers have begun to emerge. In this paper, we first introduce two new web-based grammar checkers (Microsoft ESL Assistant and NTNU statistical grammar checker) and then compare their performance. Ten common EFL errors selected from a large Chinese EFL learner corpus were used to test these two grammar checkers. The test results showed that the NTNU statistical checker was far more sensitive to various learner errors, and it could detect eight types of selected errors. Microsoft ESL Assistant could only deal with five types of errors. Moreover, these two checkers both could not deal with fragments and run-on sentences errors. It seems clear that both checkers have room for improvement before they can provide satisfactory service to ESL learners. The Microsoft ESL Assistant should expand its coverage to detect more learner errors. NTNU checker should reduce false alarms and indicate the locations of errors more accurately. Learner errors are indeed complicated for developers of grammar checkers, but the strong need for a functional grammar checker deserves CALL researchers' special attention.

Keywords: ESL Writing, Errors, Grammar Checker, Ngrams, Rules.

* English Department, National Taiwan Normal University
E-mail: hjchen@ntnu.edu.tw

[Received August 9, 2009; Revised November 15, 2009; November 14, 2009]

1. Introduction

It is challenging for second language learners to become proficient writers of the target language. Learners need considerable writing practices before they can write accurately and fluently. In addition to providing more writing opportunities to students, many teachers and researchers believe that learners need to receive proper corrective feedback on their writings (Ferris, 1999, 2003, 2006). If learners only keep on writing and do not receive any corrective feedback, they will not be able to make progress quickly. Even though the role of corrective feedback in second language learning remains controversial, many teachers and learners firmly believe that feedback plays an important role in second language writing (*cf.* Ferris, 1999, 2003, 2006; Truscott, 1996, 1999, 2007; Goldstein, 2006; Guenette, 2007).

In many ESL/EFL settings, writing teachers need to work with 40 or 50 students in their classes. Reading and correcting students' essays is a great burden for many writing teachers. To reduce teachers' loads in correcting common errors and to help students enhance their writing accuracy, various grammar checking tools have been developed in different countries. Many CALL researchers consider the development of grammar checkers to be part of the Intelligent CALL research (*cf.* Holland *et al*., 1995). The development of a useful grammar checker to identify and correct learners' errors has been considered a very important research direction in CALL. However, most of the English grammar checkers developed by academic institutes could only deal with a limited set of grammar errors and were not made available to ESL students (*e.g*., Liou, 1991, 1992, 1993; More, 2006; Naber, 2003; Park, Palmer, & Washburn, 1997). Some commercial PC-based grammar checkers (*e.g*. Whitesmoke) are available, but their error detecting capacities are still limited according to some recent evaluation studies (Chiu, 2008).

With the rapid development of artificial intelligence and natural language processing technologies, several automated essay scoring programs (*e.g.*, Vantage My Access and ETS Criterion) have appeared recently in the ESL market (Attali, 2004; Burstein, Chodorow & Leacock, 2004; Elliot, 2001; Elliot & Mikulas, 2004; Han, Chodorow & Leacock, 2006; Higgins, Burstein & Attali, 2006). Grammar checkers are also included in these two leading writing tools. By subscribing to these commercial programs, students can choose from a wide range of practice essays topics to write multiple drafts and receive immediate corrective feedback in the form of both holistic scores and diagnostic comments on grammar, organization, style, and usage.

These commercial software packages, however, are expensive, and ESL students who subscribe to these services can only use these programs for 3-6 months during the subscription period. Because of these limitations and high prices, many ESL/EFL students still do not have access to any of these programs. The better way of helping a large number of students is to

look for similar grammar checking programs on the Internet. Although the Internet has become one of the most important resources for second/foreign language education worldwide, very few web sites offer grammar checking tools for ESL learners.

There are some reasons for the limited availability of grammar checkers. The most important reason is that it is very difficult to develop a reliable grammar checker. Daniel Kies, on his web site (http://papyr.com/hypertextbooks/grammar/gramchek.htm), compared several popular grammar checking tools. These PC tools are Microsoft's Word (both the Windows and the Mac versions of the program), Corel's WordPerfect (Windows only), Grammarian Pro X (Mac only), and Open Office Writer with the Language Tool extension added (platform independent). The best program for Windows is Corel's WordPerfect, but it can only reach about 40 percent accuracy for about twenty types of errors.

According to Naber (2003), there are basically three different ways to implement a grammar checker. Each approach has its strengths and weakness.

1. Syntax-based checkers. In this approach, a text is completely parsed, *i.e*. the sentences are first analyzed and each sentence is assigned a tree structure. The text is considered incorrect if the parsing does not succeed. A robust syntactic parser plays a very important role in this approach (*cf*. Jensen, 1993).

2. Statistics-based checkers. In this approach, a language model is trained from a large training corpus (*e.g*., a native corpus), which contains many short phrases (*cf*. Atwell & Elliott, 1987; Chodorow & Leacock, 2000). It can be used for detecting and correcting certain types of grammar errors, where local information is sufficient to make decision. Sequences which occur often in the corpus can be considered correct in other texts, and uncommon sequences might be errors. Some recent developments of grammar checkers based on corpus linguistics and statistics seem to be quite useful (Chodorow & Leacock, 2000; Sjobergh, 2005; Wu, Su, Jiang, & Hsu, 2006). A popular example is the grammar and style checker developed by ETS.

3. Rule-based checkers. In this approach, a set of rules is matched against a text which has at least been POS tagged. This approach is similar to the statistical approach, but all the grammar rules are developed manually. Park *et al*. (1997) and Naber (2003) are studies using this approach.

Although it is time-consuming and difficult to develop a robust grammar checker that can provide specific and clear grammar feedback, many ESL/EFL teachers and students worldwide desperately need a good grammar checker to improve teaching and learning. With funding from the National Science Council of Taiwan, a research team at National Taiwan Normal University (NTNU) developed a statistical grammar checker based on large English corpora. At the same time, another research team in Microsoft also developed their new grammar checker called Microsoft ESL Assistant. It seems that a grammar checking service has begun to emerge on the Web. These two projects are briefly introduced in the sections below.

1.1 Microsoft ESL Assistant

The Microsoft Research ESL Assistant is a web service that provides correction suggestions for ESL writing errors. The web service also provides suggestions for word choice. It consists of three parts: a set of modules that identify possible corrections, a large language model that evaluates the possible suggestions, and a module that produces search results using Live Search. The individual error modules each target specific errors. Some of these models are based on heuristics, while others use machine learned classifiers. Information that the modules take into account includes the presence of specific words as well as the sequence of part-of-speech tags that are automatically assigned. The language model of ESL Assistant is trained on the Gigaword corpus. The English Gigaword Corpus is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania.

The ESL Assistant has recently been updated with a new user interface that uses the Microsoft Silverlight™ browser plug-in. As shown below in Figure 1, text to be checked is entered in the box at the top. When the user clicks the check button, potential errors are identified and highlighted. Users can proceed from one highlighted segment to the next, reviewing the possible errors and the suggestions presented in the box below. In addition, a pie chart allows the user to compare the approximate frequency distributions of their own input and the suggestions, as found by the Microsoft Bing™ search engine. When the user hover their mouse over any of the options available, a dropdown panel shows selected usage examples found on the Web. Users can explore additional examples by clicking the link at the bottom of the dropdown panel.

Figure 1. The Microsoft ESL assistant.

1.2 NTNU Statistical Grammar Checker

The ngram-based statistical grammar checker was developed at NTNU. The grammar checker was based on the statistical approach. This is the approach currently used by ETS and several other research teams (*e.g.*, Chodorow & Leacock, 2000; Sjobergh, 2005). The advantages of using an ngram-based statistical approach follow. First, there is no need to write grammar rules manually. Second, it can detect many errors which cannot be detected by traditional grammar checkers and parsers. The bigrams and trigrams can detect more errors. Third, the lexical errors are more likely to be detected by this type of grammar checker.

The procedures used to detect the violations of English grammar in a statistical grammar checker are not complicated. The grammar checker system is first trained on a large corpus of edited texts, from which it extracts and counts *bigrams* that consist of sequences of adjacent words. British National Corpus (BNC) was used as the reference corpus for the NTNU checker. The British National Corpus is a 100-million-word collection of samples of written and spoken English from a wide range of sources. The corpus covers British English of the late twentieth century from a wide variety of genres with the intention that it be a representative sample of spoken and written British English of that time. Then SRI (Stanford Research Institute) Language Modeling Toolkit (SRILM) was used to develop an ngram language model based on the BNC corpus. Various bigrams and trigrams were extracted from the BNC corpus and stored in a large database. A web-based grammar checker interface linked to the ngram database was also developed. The web interface of the ngram-based grammar checker is shown in Figure 2.

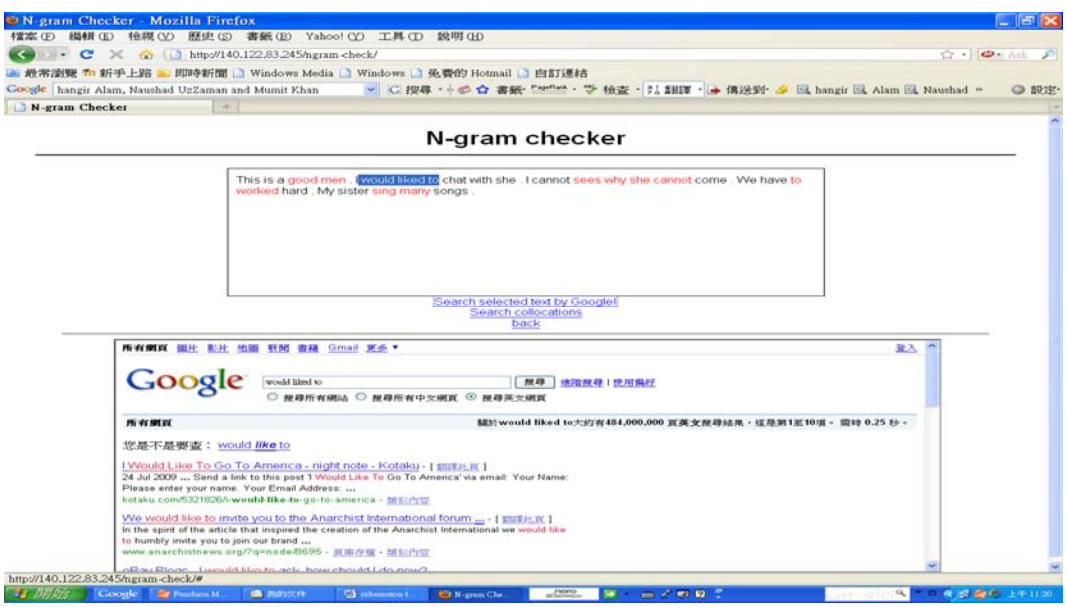


Figure 2. The Web interface of NTNU Ngram-based statistical grammar checker.

ESL students can submit their writings to the web-based grammar checker, and the system searches the students' writings for bigrams and trigrams. The bigrams and trigrams which never or rarely show up in BNC corpus were highlighted by the system. Within a few seconds, this checker can quickly check the article and highlight the problematic word strings (clusters) not found in the native corpora. With the help of the bigram and trigram information from the very large native corpora, the system can efficiently detect and highlight the problematic usage in students' essays.

After students input their essays into the grammar checker, the potential errors are automatically highlighted in red. The Google search engine can be used to search for the better usage when users use the mouse to highlight any word string in their writing. Google either directly recommends some more commonly used strings or shows various sentences which include these words.

Since the Microsoft ESL Assistant and NTNU ngram-based statistical checker are web-based services, they have the potential to allow many ESL students to use them. It is not clear, however, if these two new checkers can reach the accuracy level of some of the leading products in detecting or correcting ESL student errors. To determine their usability, it is essential to further evaluate their performance in dealing with ESL learner errors.

2. Methodology

To assess if these two grammar checkers can provide similar services as the leading products in the market, it is necessary to compare the performance of these two checkers with some of the leading commercial products in the market. One of the best grammar checking tools available is the ETS Criterion. It can detect many ESL learner errors and can provide appropriate feedback. Based on the results of a previous research evaluation study (Chen, 2009), ten types of common errors, as shown below in Table 1, which can be correctly identified by ETS Criterion were used to test the performance of the Microsoft ESL Assistant and NTNU grammar checker. Five sentences in each error category were randomly selected from a large set of incorrect sentences which were identified by ETS criterion.

Table 1. Ten types of common ESL errors detected by ETS Criterion.

Error Types of ETS Criterion
1. Missing or Extra Article
2. Spelling
3. Fragment or Missing Comma
4. Run-on Sentences
5. Subject-Verb Agreement
6. Confused Words
7. Ill-formed Verbs
8. Proofread This!
9. Wrong Article
10. Compound Words

Three research questions were proposed in this paper.

1. What types of errors detected by ETS Criterion can be detected by these two web-based grammar checkers?
2. What types of errors cannot be identified by these two grammar checkers?
3. What are the accuracy rates of these two grammar checkers? Which checker has the higher accuracy rate? Which checker can find more errors?

3. Evaluation Result

3.1 The Performances of Two Grammar Checkers in Each Type of Error

In the following section, the performances of each grammar checker in each different error category are shown in tables. The errors detected by each grammar checker are underlined. The first part of the table shows the errors marked by ETS Criterion, the second part shows the errors marked by Microsoft ESL Assistant, and the third part shows the errors marked by the NTNU ngram-based grammar checker.

1. Missing or Extra Article: As shown in Table 2, Microsoft ESL Assistant could detect 4 out of 5 errors for missing or extra articles. The NTNU grammar checker could also detect 4 errors, but the highlighting of errors was not accurate and there were some false alarms. Although both checkers could detect this type of errors, it was clear that ESL Assistant was more effective.

Table 2. Comparison of checker performance on missing or extra articles.

ETS Criterion	1. From going to <u>lab</u> to do experiments, I learned a lot. 2. Here is an old lady who speaks some words she sees on <u>street</u> randomly. 3. In <u>first</u> grad of high school, I did not know anyone. 4. It is <u>honor</u> that I am one of her best friends. 5. Bear could get <u>good</u> relationship with everyone in our class.
Microsoft ESL Assistant	1. From going to lab to do experiments, I learned a lot. 2. Here is an old lady who speaks some words she sees on <u>street</u> randomly. 3. In <u>first</u> grad of high school, I did not know anyone. 4. It is <u>honor</u> that I am one of her best friends. 5. Bear could get <u>good</u> relationship with everyone in our class.
NTNU Grammar Checker	1. From going <u>to lab to do experiments</u>, I learned a lot. 2. Here is an old lady who <u>speaks some words she sees on street</u> randomly. 3. In first <u>grad of</u> high school, I did not know <u>anyone</u>. 4. It is <u>honor</u> that I am one of her best friends. 5. Bear could <u>get</u> good relationship with everyone in <u>our class</u>.

2. Spelling errors: As shown in Table 3, Microsoft ESL Assistant could not detect any spelling errors. The NTNU checker, however, could detect all the 5 spelling errors. It was clear that the NTNU checker was more effective in this category. It is odd that why Microsoft ESL Assistant did not incorporate the very effective Microsoft spelling checker into the web system.

Table 3. Comparison of checker performance on spelling errors.

ETS Criterion	1. I am a deligent student. 2. She hated to study the textbook and even frquently skipped the class. 3. but it seemes everyone is so busy. 4. She is also humurous. 5. He is very hansome.
Microsoft ESL Assistant	1. I am a deligent student. 2. She hated to study the textbook and even frquently skipped the class. 3. But it seemes everyone is so busy. 4. She is also humurous. 5. He is very hansome.
NTNU Grammar Checker	1. I am a deligent student. 2. She hated to study the textbook and even frquently skipped the class. 3. But it seemes everyone is so busy. 4. She is also humurous. 5. He is very hansome.

3. Fragment or Missing Comma: As shown in Table 4, Microsoft ESL Assistant again could not detect any spelling errors. Similarly, the NTNU checker could not detect any of the fragment errors but it highlighted some words. It was clear that both types of grammar checkers could not deal with fragment errors.

Table 4. Comparison of checker performance on fragment or missing comma.

ETS Criterion	1. Although now we are in various areas over this island. 2. Because his or her friends are someone who relate to him or her and know about him. 3. When our friends' or family's birthday is coming. 4. In a word, because people analyze the question in different ways, such as the TV station and audience. 5. If she can be more considerate.
Microsoft ESL Assistant	1. Although now we are in various areas over this island. 2. Because his or her friends are someone who relate to him or her and know about him. 3. When our friends' or family's birthday is coming. 4. In a word, because people analyze the question in different ways, such as the TV station and audience. 5. If she can be more considerate.

NTNU Grammar Checker	1. Although now we are in various areas over this island. 2. Because his or her friends are someone who relate to him or her and know about him. 3. When our friends' or family's birthday is coming. 4. In a word, because people analyze the question in different ways, such as the TV station and audience. 5. If she can be more considerate.

4. Run-on Sentences: As shown in Table 5, Microsoft ESL Assistant could not detect any run-on sentence errors. Similarly, the NTNU checker again could not detect any of these run-on sentence errors and it also generated some false alarms. It was clear that both types of grammar checkers could not deal with run-on sentence errors.

Table 5. Comparison of checker performance on run-on sentences.

ETS Criterion	1. Like I would play strong and tell her firmly that she should go to Purdue no matter what and that's a chance of a life time, though in reality, and actually the first instinct, I forbid her to go. 2. Then the day came, Bob received the grading and bounced to Charlie. 3. Bob was turned down and felt disappointed and, however, felt angry at Charlie. 4. We have people and place, the rest is food. 5. It was like a dream, I was flying!
Microsoft ESL Assistant	1. Like I would play strong and tell her firmly that she should go to Purdue no matter what and that's a chance of a life time, though in reality, and actually the first instinct, I forbid her to go. 2. Then the day came, Bob received the grading and bounced to Charlie. 3. Bob was turned down and felt disappointed and, however, felt angry at Charlie. 4. We have people and place, the rest is food. 5. It was like a dream, I was flying!
NTNU Grammar Checker	1. Like I would play strong and tell her firmly that she should go to Purdue no matter what and that's a chance of a life time, though in reality, and actually the first instinct, I forbid her to go. 2. Then the day came, Bob received the grading and bounced to Charlie. 3. Bob was turned down and felt disappointed and, however, felt angry at Charlie. 4. We have people and place, the rest is food. 5. It was like a dream, I was flying!

5. Subject verb agreement: As shown in Table 6, Microsoft ESL Assistant could only detect one agreement error among the five errors. The NTNU checker, however, could detect five agreement errors. It was clear that the NTNU grammar checker performed much better in this category.

Table 6. Comparison of checker performance on subject-verb agreement.

ETS Criterion	1. There are many advertisement. 2. My teacher tell me a story. 3. No matter what I plans after graduation. 4. The advertisement appear in TV. 5. She sing very loudly.
Microsoft ESL Assistant	1. There are many advertisement. 2. My teacher tell me a story. 3. No matter what I plans after graduation 4. The advertisement appear in TV 5. She sing very loudly.
NTNU Grammar Checker	1. There are many advertisement. 2. My teacher tell me a story. 3. No matter what I plans after graduation. 4. The advertisement appear in TV. 5. She sing very loudly.

6. Confused Words: As shown in Table 7, Microsoft ESL Assistant could only detect three errors which were related to the confusion between *a* and *an*. Also, it could not detect any confused word pairs (quite vs. quiet; effect vs. affect). The NTNU checker could detect the five confused words errors, although the error highlighting is not very accurate. It was clear that NTNU ngram grammar checker could find more errors in this category.

Table 7. Comparison of checker performance on confused words.

ETS Criterion	1. It should be an perfect idea. 2. I was a ambassador sent by the Chin dynasty. 3. She is more independent as a individual. 4. He is a very quite student. 5. The new drug might have a better affect.
Microsoft ESL Assistant	1. It should be an perfect idea. 2. I was a ambassador sent by the Chin dynasty. 3. She is more independent as a individual. 4. He is a very quite student. 5. The new drug might have a better affect.
NTNU Grammar Checker	1. It should be an perfect idea. 2. I was a ambassador sent by the Chin dynasty. 3. She is more independent as a individual. 4. He is a very quite student. 5. The new drug might have a better affect.

7. Ill-formed Verbs: As shown in Table 8, Microsoft ESL Assistant could only detect one of the five verb form errors. The NTNU checker could detect three errors correctly but the error highlighting does not indicate the error location very accurately. It seems that NTNU ngram grammar checker performed better in this category.

Table 8. Comparison of checker performance on ill-formed verbs.

ETS Criterion	1. I like the lessons they have gave us. 2. I should have react like Washington when I face the situation. 3. We are excel at different subject. 4. Please be sure that they are satisfy with it. 5. She and me are so happy to had a friend like this.
Microsoft ESL Assistant	1. I like the lessons they have gave us. 2. I should have react like Washington when I face the situation. 3. We are excel at different subject. 4. Please be sure that they are satisfy with it. 5. She and me are so happy to had a friend like this.
NTNU grammar Checker	1. I like the lessons they have gave us. 2. I should have react like Washington when I face the situation. 3. We are excel at different subject. 4. Please be sure that they are satisfy with it. 5. She and me are so happy to had a friend like this.

8. Proofread This!: As shown in Table 9, Microsoft ESL Assistant could detect three out of the five "proofread this" errors. The NTNU checker could detect all the five errors correctly though it also provided one false alarm, "player". It was clear that the NTNU ngram grammar checker still performed better in this category.

Table 9. Comparison of checker performances on proofread this!

ETS Criterion	1. She had a warn family. 2. I mayor in engineering at school. 3. I'm a fourth years student of National Taiwan Normal University. 4. I find a job that I desirable. 5. The famous player kicked the ball at the time they bended!
Microsoft ESL Assistant	1. She had a warn family. 2. I mayor in engineering at school. 3. I'm a fourth years student of National Taiwan Normal University. 4. I find a job that I desirable. 5. The famous player kicked the ball at the time they bended!
NTNU grammar Checker	1. She had a warn family. 2. I mayor in engineering at school. 3. I'm a fourth years student of National Taiwan Normal University. 4. I find a job that I desirable. 5. The famous player kicked the ball at the time they bended!

9. Wrong Article: As shown in Table 10, Microsoft ESL Assistant could detect three out of the five article errors and generated one false alarm. The NTNU checker also could identify four errors correctly, but it also provided more false alarms. It was clear that the NTNU ngram grammar checker still performed better in this category.

Table 10. Comparison of checker performance on wrong article

ETS Criterion	1. I turned these experience into my own energy for next challenge. 2. Those stuff in high school days is absolutely insufficient. 3. People always like to be a popular guys. 4. Finding a best friend can make your life more special. 5. If you do not have a slightest idea.
Microsoft ESL Assistant	1. I turned these experience into my own energy for next challenge. 2. Those stuff in high school days is absolutely insufficient. 3. People always like to be a popular guys. 4. Finding a best friend can make your life more special. 5. If you do not have a slightest idea.
NTNU Grammar Checker	1. I turned these experience into my own energy for next challenge. 2. Those stuff in high school days is absolutely insufficient. 3. People always like to be a popular guys. 4. Finding a best friend can make your life more special. 5. If you do not have a slightest idea.

10. Compound Words: As shown in Table 11, Microsoft ESL Assistant could not detect any of the five compound word errors. The NTNU grammar checker could identify four compound word errors correctly, but it also generated a few false alarms. Although the error highlighting is not very accurate, users can highlight any word strings and get useful feedback from the Google search engine. It was clear that the NTNU ngram grammar checker still performed better in this category.

Table 11. Comparison of checker performance on compound words

ETS Criterion	1. When ever I feel lonely. 2. Do no say any thing when you don't now what you are saying. 3. Every one should have the experience of making friends with others. 4. I try to find any one to substitute but it seems everyone is so busy. 5. At last, no body will get benefits from cheating.
Microsoft ESL Assistant	1. When ever I feel lonely. 2. Do no say any thing when you don't now what you are saying. 3. Every one should have the experience of making friends with others. 4. I try to find any one to substitute but it seems everyone is so busy. 5. At last, no body will get benefits from cheating.

NTNU Grammar Checker	1. When ever I feel lonely. 2. Do no say any thing when you don't now what you are saying. 3. Every one should have the experience of making friends with others. 4. I try to find any one to substitute but it seems everyone is so busy. 5. At last, no body will get benefits from cheating.

3.2 General Comments on the Performance of Grammar Checkers

Based on the detailed analysis of the ten common ESL errors, it was obvious that the NTNU ngram grammar checker performed better than the Microsoft ESL Assistant. The NTNU checker performed better in 7 error categories. The Microsoft checker only performed better in the "missing or extra article" category. In addition, both grammar checkers failed to detect any fragment and run-on sentence errors. Table 12 summarizes the results of comparing these two grammar checkers on the error types they can detect.

Table 12. Summary of the performance of two grammar checkers

Error types	Microsoft	NTNU
1. Missing or Extra Article	O (Better)	O
2. Spelling	X	O (Better)
3. Fragment or Missing Comma	X	X
4. Run-on Sentences	X	X
5. Subject-Verb Agreement	O	O (Better)
6. Confused Words	O	O (Better)
7. Ill-formed Verbs	O	O (Better)
8. Proofread This!	O	O (Better)
9. Wrong Article	O	O
10. Compound Words	X	O (Better)

Notes: O means that the checker can detect the errors; X means that the checker fails to detect the errors.

Another possible way to compare the performances of these two grammar checkers is to compare the precision and the recall rates of these two checkers. Precision and recall are two widely used measures for evaluating the performance of information retrieval systems. In information retrieval, a perfect Precision score of 1.0 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved) whereas a perfect Recall score of 1.0 means that all relevant documents were retrieved by the search (but says nothing about how many irrelevant documents were also retrieved). In evaluating the performances of grammar checkers, a perfect precision score means that all the

errors found by a checker are indeed real errors. A perfect recall score means that a checker would find all the errors made by language learners.

Table 13 summarizes the precision and the recall rates of these two grammar checkers. It seems that Microsoft ESL Assistant and NTNU grammar checker have similar precision rates (50% vs. 61%). However, the recall rate of NTNU grammar checker is clearly higher than that of Microsoft ESL Assistant (72% vs. 30%). The results indicate that NTNU can detect more errors in ESL learners' writing.

Table 13. Summary of the precision and recall rate of two grammar checkers

	Microsoft		**NTNU**	
	Precision	Recall	Precision	Recall
1. Missing or Extra Article	4/4 (100%)	4/5 (80%)	4/7 (57%)	4/5 (80%)
2. Spelling	0 (0%)	0/5 (0%)	5/5 (100%)	5/5 (100%)
3. Fragment or Missing Comma	0 (0%)	0/5 (0%)	0 (0%)	0/5 (0%)
4. Run-on Sentences	0 (0%)	0/5 (0%)	0 (0%)	0/5 (0%)
5. Subject-Verb Agreement	1/1 (100%)	1/5 (20%)	5/5 (100%)	5/5 (100%)
6. Confused Words	3/3 (100%)	3/5 (60%)	5/5 (100%)	5/5 (100%)
7. Ill-formed Verbs	1/2 (50%)	1/5 (20%)	4/6 (67%)	4/5 (80%)
8. Proofread This!	3/4 (75%)	3/5 (60%)	5/6 (83%)	5/5 (100%)
9. Wrong Article	3/4 (75%)	3/5 (60%)	4/8 (50%)	4/5 (80%)
10. Compound Words	0 (0%)	0/5 (0%)	4/7 (57%)	4/5 (80%)
Average	50%	30%	61 %	72%

4. Discussion

Based on the test results, eight types of errors identified by ETS Criterion can also be detected by the NTNU grammar checker. Microsoft ESL Assistant can only detect five types of learner errors. It seemed obvious that the NTNU grammar checker is more sensitive to ESL learner errors. The Microsoft ESL Assistant seems less sensitive in reporting learners' errors. Some common ESL errors, such as spelling and subject-verb agreement, are not treated by the current version of Microsoft ESL Assistant. It is unclear why Microsoft did not target any of these common errors in their new web-based grammar checker.

In addition, these two grammar checkers cannot deal with two types of errors, sentence fragments and run-on sentences. The reason Microsoft ESL Assistant cannot deal with these two types of errors is not clear because the Microsoft checker uses both a rule-based approach and a statistical approach. A rule-based approach might help to resolve these two common ESL errors.

Nevertheless, the failure in detecting these two types of learner errors highlights one of the major weaknesses of the statistical grammar checker. The ngram-based statistical grammar checker is good at detecting errors in the "local" domains or "narrow "domains because the language models used by a statistic-based grammar checker are bigrams and trigrams. Therefore, it can thus find many errors based on inappropriate word combinations. Nevertheless, the statistical grammar checker has great difficulty in detecting errors that cannot be inferred based on word combinations. For instance, the errors like fragments and run-on sentences cannot be detected by a statistical checker because these errors are hard to detect solely based on the information of word combination. Similarly, errors like pronoun errors and tense errors cannot be detected effectively by the statistical grammar checker. To sum up, the statistical grammar checker will fail to capture errors if the errors are not word combination problems or they involve problems of non-adjacent word strings or conflicts across different clause boundaries.

When these two grammar checkers are examined in terms of precision and recall, Microsoft ESL Assistant evidently has high precision rate in several error categories (Missing or Extra Article, Subject-Verb Agreement, and Confused Words) but overall its average precision rate is only about 50%. In addition, ESL Assistant has much lower recall rate (only about 30%). On the other hand, the NTNU statistical grammar checker clearly has higher average recall rate (72%), but its precision rate (61%) is similar to Microsoft system because it often generated false alarms.

4.1 Possible Directions for Improvement

Even the state-of-the art grammar checkers like ETS Criterion and Vantage MyAccess have some limitations in dealing with ESL learners' errors. It was found that they could not deal with errors like word order, tenses, and collocations (*cf.* Chen, 2006, 2009). It is clear that these two web-based grammar checkers have much room for improvement. The coverage and capacities of Microsoft ESL Assistant are still quite limited. This could be the major problem for ESL users around the world who use this service via Internet. Many types of ESL learner errors (spelling, subject-verb agreement) are not treated in the current version. The recent incorporation of the Bing search engine into the system is successful, and this change has made the system more user-friendly. Users now can quickly check their usage with the help of Bing. Microsoft has a very strong research team behind the system, and it should not be too difficult to deal with some common errors like spelling and subject-verb agreement. It is expected that the new version of Microsoft ESL Assistant would significantly enlarge their coverage.

Although the Microsoft Bing search engine can help users in some cases, the help it can provide is still limited. For some errors, Bing engine can search the native corpus and find

similar expressions and show them to learners. Making correct recommendations for ESL learners, however, is not an easy task in many cases. For example, it would be difficult if the errors are collocation errors. If a grammar checker needs to provide automatic feedback on a sentence like "I would like to increase my life," it would be very difficult since there are so many English verbs that collocate with the noun "life". Even if the system tries to find all the possible verbs, right suggestions cannot be provided easily. It seems obvious that Microsoft ESL Assistant might also need to improve its feedback mechanisms and provide more specific suggestions.

For the NTNU grammar checker, there are at least four major directions for improvement. First, the number of false alarms should be reduced. Although the NTNU grammar checker can be effective in detecting various learner errors, the sensitivity to learner errors can cause some false alarms. This noise can be rather annoying for ESL learners since they might not have the ability to judge if these messages are false alarms or real errors. Some possible ways of improving the ngram-based statistical grammar checker would involve a large reference corpus and perhaps the proper use of a POS-tagged reference corpus. The new data set, Web 1T 5-gram Version 1 contributed by Google Inc., contains English word n-grams and their observed frequency counts. The length of the n-grams ranges from unigrams (single words) to five-grams. The new huge corpus will be useful for statistical language modeling. The NTNU team is now developing another ngram-based checker based on the combination of Google data and BNC data (http://140.122.83.245/gchecker/index2.html), and we hope this new version with a much larger corpus will reduce some false alarms. Incorporating information from tagged corpora would be another way of improving the statistic-based checker. A better statistical checker might use the tag information to detect errors. Similar to raw text data, a part of speech (POS)-annotated corpus can be used to build a list of POS tag sequences. Some sequences will be very common (for example, determiner, adjective, noun as in "the old man"), others will probably not occur at all (for example, determiner, determiner, adjective). The powerful tagger can help us to tag large text files. The clusters of various POS tags can be further extracted and used to help to identify learners' errors. The tagged corpus might also help to provide better grammatical explanations.

Second, the feedback mechanism can be further improved. The external link to the Google search engine can only solve some problems like word forms and misspellings. Google can only provide correct examples in some cases. If a student writes "…prepare the exam," the student can easily find the correct answer "prepare for the exam" from Google. Nevertheless, Google cannot always find useful information based on the word strings provided by ESL learners. In addition, learners might need some metalinguistic feedback. It seems that a more robust feedback mechanism should be built.

Third, the highlighting techniques were not accurate enough. Sometimes the NTNU

grammar checker highlighted the adjacent words but not the keyword. Inaccurate highlighting can confuse ESL learners if they do not receive some explanations about using the grammar checker. A possible way of resolving this problem is to show the types of errors made by learners. If the checker can also indicate the types of errors that learners make, then the information can give learners more direct help.

Finally, as mentioned earlier, the limitation of statistical grammar checker is obvious. The NTNU checker would need to integrate the strength of the rule-based checker. The strength of a rule-based checker is that it can allow programmers to specify what needs to found and what needs to be replaced. In the future, when a tagger or a chunker is integrated with the existing system, more grammar rules based on POS tags and structural relations can be added in to deal with more grammar problems like run-on sentences, fragments, conjunction errors, and tense errors. If the two different grammar checkers can be integrated into one system, then the integrated system should be able to detect more learner errors. (*cf.* Chen, 2006, 2009). If these four problems can be solved, the NTNU grammar checker can be more useful for ESL learners.

4.2 Limitations and Implications

This study is a preliminary study on assessing the performances of two new web-based grammar checkers. Several limitations should be noted. There were only 10 types of errors tested. More types of errors from different L1 backgrounds should be included in future tests. In addition, it would be also interesting to compare human error corrections and machine corrections in future research. The strengths and limitations of machine feedback can be further identified.

It is indeed very challenging to develop a robust grammar checker. Researchers have identified some possible reasons. First, the errors made by second/foreign language learners are complex and diverse. There are so many different types of errors in learners' writing. A short sentence written by ESL learners might contain several errors. Because of the complexity, it is not easy to provide satisfactory feedback on these grammar errors made by learners. Second, the NLP technologies which are used to support these grammar checking tools are not in a mature stage. English parsers are still not perfect in detecting various errors and computers still cannot understand the meaning that learners try to convey. Many ESL errors cannot be detected, and appropriate suggestions for corrections cannot be provided.

For developers and researchers of Computer-Assisted Language Learning systems, despite the richer technology and corpus resources available now, the task for developing a robust grammar checker remains arduous and complicated. As indicated earlier, feedback plays a central role in the second language writing process. If specific and clear feedback messages can be provided to learners in the writing process, it is very likely learners can

incorporate the feedback to further improve their writing performance. A robust grammar checker can significantly reduce writing teachers' load and enhance students' writing performance. More research effort should be made to develop a robust grammar checker that can provide specific and clear feedback to ESL writers.

Acknowledgments

This research was supported in part by National Science Council under Grant 96-2411-H-003-042-MY2.

Reference

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion.* Paper presented at the National Council on Measurement in Education (NCME) in San Diego, CA.

Atwell, E. & Elliott, S. (1987). Dealing with Ill-formed English Text. In: R. Garside, G. Leech and G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine,* 25(3), 27-36.

Chen, H. J. (2006). Examining the Scoring Mechanism and Feedback Quality of My Access. *Proceedings of Tamkang University Conference on Second Language Writing.*

Chen, H. J. (2009). Developing Statistic-based and Rule-based Grammar Checkers for Chinese ESL Learners. *Proceedings of the 2009 LTTC International Conference on English Language Teaching and Testing*. Taipei: Taiwan.

Chiu, T. L. (2008). An Evaluation of Whitesmoke: A Grammar-checking and Text-enrichment Software. *Proceedings of the Seventeenth ETA/ROC International Symposium on English Teaching*. Taipei: Taiwan.

Chodorow, M. & Leacock, C. (2000). *An unsupervised method for detecting grammatical errors.* Paper presented at the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics.

Elliot, S. M. (2001). *IntelliMetric: From here to validity*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Elliot, S. M., & Mikulas, C. (2004). *The impact of MY Access! use on student writing performance: A technology overview and four studies.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Ferris, D. (1999). The Case for Grammar Correction in L2 Writing Classes: A Response to Truscott. *Journal of Second Language Writing,* 8(1), 1-11.

Ferris, D. (2003). *Treatment of Error in Second Language Student Writing*. Ann Arbor, MI: University of Michigan Press.

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In Hyland, K. & Hyland, F. (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81-104). Cambridge: Cambridge University Press.

Goldstein, L. (2006). Feedback and revision in second language writing: Contextual, teacher, and student variables. In Hyland, K. & Hyland, F. (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 185-205). Cambridge: Cambridge University Press.

Guenette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing*, 16, 40-53.

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

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145-159.

Holland, M. V., Kaplan, J.D., & Sama, M.R., eds. (1995). *Intelligent Language Tutors: Theory Shaping Technology*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jensen, K. (1993). PEG: the PLNLP English Grammar, in Jensen K., Heidorn G.E., & Richardson S.D., (Eds.), *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers, Boston, 29-45.

Liou, H. C. (1991) Development of an English grammar checker: A progress report. *CALICO Journal*, 9(1), 57-70.

Liou, H. C. (1992). An automatic text-analysis project for EFL writing revision. *System*, 20(4), 481-492.

Liou, H. C. (1993). Integrating text-analysis programs into classroom writing revision. *CAELL Journal*, 4(1), 21-27.

Moré, J. (2006). A Grammar Checker Based on Web Searching. Digithum [online article]. Retrieved October 14, 2009, from http://www.uoc.edu/digithum/8/dt/eng/more.pdf.

Naber, D. (2003). *A Rule-Based Style and Grammar Checker*. Unpublished doctoral dissertation, University of Bielefeld, Germany.

Park, J. C., Palmer, M., & Washburn, G. (1997). An English grammar checker as a writing aid for students of English as a second language. *Proceedings of the Conference of Applied Natural Language Processing (ANLP)*. Washington, DC.

Sjöbergh, J. (2005). Chunking: an unsupervised method to find errors in text. *Proceedings of the 15th Nordic Conference of Computational Linguistics, NODALIDA 2005*.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46, 327-369.

Truscott, J. (1999). The case for "The case against grammar correction in L2 writing classes": A response to Ferris. *Journal of Second Language Writing*, 8(2), 111-122.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, 16, 255-272.

Wu, S.-H., Su, C.-Y., Jiang, T.-J., & Hsu, W.-L. (2006). An Evaluation of Adopting Language Model as the Checker of Preposition Usage. *Proceedings from ROCLING 2006.*

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 181-204

An Exploratory Application of Rhetorical Structure Theory to Detect Coherence Errors in L2 English Writing: Possible Implications for Automated Writing Evaluation Software

Sophia Skoufaki*

Abstract

This paper presents an initial attempt to examine whether Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) can be fruitfully applied to the detection of the coherence errors made by Taiwanese low-intermediate learners of English. This investigation is considered warranted for three reasons. First, other methods for bottom-up coherence analysis have proved ineffective (e.g., Watson Todd *et al.*, 2007). Second, this research provides a preliminary categorization of the coherence errors made by first language (L1) Chinese learners of English. Third, second language discourse errors in general have received little attention in applied linguistic research. The data are 45 written samples from the LTTC English Learner Corpus, a Taiwanese learner corpus of English currently under construction. The rationale of this study is that diagrams which violate some of the rules of RST diagram formation will point to coherence errors. No reliability test has been conducted since this work is at an initial stage. Therefore, this study is exploratory and results are preliminary. Results are discussed in terms of the practicality of using this method to detect coherence errors, their possible consequences about claims for a typical inductive content order in the writing of L1 Chinese learners of English, and their potential implications for Automated Writing Evaluation (AWE) software, since discourse organization is one of the essay characteristics assessed by this software. In particular, the extent to which the kinds of errors detected through the RST analysis match those located by *Criterion* (Burstein, Chodorow, & Leachock, 2004), a well-known AWE software by Educational Testing Service (ETS), is discussed.

* Graduate Institute of Linguistics, National Taiwan University
E-mail: sophiaskoufaki@ntu.edu.tw

[Received August 10, 2009; Revised December 6, 2009; Accepted December 22, 2009]

Keywords: Automated Writing Evaluation, Discourse Organization, Coherence Errors, Rhetorical Structure Theory.

1. Introduction

Research findings indicate that English language learners produce various kinds of discourse errors in their writing, such as inductive patterns (e.g., Kaplan, 1966) and inappropriate coordination (e.g., Soter, 1988). However, the discourse errors of second language (L2) learners of English have not been examined in detail partly because at least some of them are more difficult to detect than other kinds of errors (e.g., syntactic, spelling). This paper describes an initial attempt to examine whether Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) can be fruitfully applied to the detection of the coherence errors made by Taiwanese low-intermediate learners of English. In particular, this paper reports on a pilot study where 45 written samples from the LTTC English Learner Corpus, a Taiwanese learner corpus of English currently under construction, were analysed according to RST. It is hoped that this pilot study will provide some preliminary indication of the viability of this approach to coherence error detection.

The results of this analysis will also serve as a preliminary list of coherence errors which may prove typical or not in further large-scale studies of this kind. A categorization of second language (L2) English coherence errors in general and of the coherence errors of particular learner populations has not been provided yet by applied linguists. Therefore, this pilot study is warranted because of its possible utility for research on English L2 discourse and the instruction of writing in English as an L2.

Another aim of this study is to examine whether the most frequent of the errors detected through the RST analysis can be located by *Criterion*, a well-known AWE software by the Educational Testing Service (ETS). Automated Writing Evaluation (AWE) software such as *Criterion* (e.g., Burstein, Chodorow, & Leachock, 2004) and *My Access!* (e.g., Vantage Learning, 2007) locate and give diagnostic feedback only for a limited number of discourse errors. This issue has been pointed out by the computational linguists involved in the creation of AWE software (e.g., Higgins, Burstein, Marcu, & Gentile, 2004), but no study has been conducted with specific English learner populations to examine what discourse errors should be added to the inventory of discourse errors currently located via AWE software. Being a pilot study, the study reported here does not purport to fill this research gap but only to provide an initial step towards this goal.

In the following two sections, this paper will offer further information on the motivation of this study. Then, it will offer some background information on RST. Third, it will provide an overview of the LTTC English Learner Corpus and will describe the data and method of the study. Fourth, it will describe findings from a qualitative and quantitative perspective. Fifth,

these results will be discussed in relation to a) whether RST analysis seems a viable method for coherence error detection, b) which factors seem to affect the coherence errors located in the data, c) whether results indicate inductive order patterns and d) how much they overlap with the coherence errors that can be located via *Criterion*. The paper will end with a summary of conclusions and directions for future research.

2. RST and Discourse Coherence Error Detection

It is difficult to reliably identify coherence errors because readers of the same text may form different interpretations of the coherence relations among elements of the text (Mann & Thompson, 1988). Therefore, a bottom-up method of coherence error detection should be used so that coherence errors will be identified as reliably and objectively as possible.

RST was chosen first because the output of other methods of locating coherence breaks, such as topical structure analysis and genre analysis in Watson Todd *et al.* (2007), has been shown to have little relationship with English teachers' judgments. Second, strong correlations have been found between RST analyses which show that a text is coherent and subjective judgments that a text is coherent (Taboada & Mann, 2006a). Finally, RST has not been applied to the location of coherence errors (Higgins, Burstein, Marcu, & Gentile, 2004: 185), so an evaluation of its application for this purpose is interesting from a methodological perspective.

3. Discourse Coherence and L1 Chinese Learners of English

Given the paucity of discourse error tagging in learner corpora (Díaz-Negrillo & Fernández-Domínguez, 2006) and the sparse research on discourse errors by learners of English, this pilot study aims to provide a preliminary categorization of discourse errors in the writing of low-intermediate Taiwanese learners of English. This list of errors will be supplemented and refined through further research.

L1 Chinese learners of English make similar discourse errors to learners with other native tongues, but there have also been claims for typical L1 Chinese errors. However, these claims have not been examined sufficiently through quantitative methods. Therefore, the pilot study reported in this paper also partly functions as a preliminary quantitative test for one of these claims. This claim is that the paragraphs and essays of L1 Chinese learners of L2 English have an inductive rather than deductive order. It has been claimed that these learners present the main point of their writing only at the end of a paragraph or essay, whereas in L1 English writing the main point is presented first (e.g., Kaplan, 1966; Matalene, 1985).

The claim for the use of an inductive order only by L1 Chinese learners of English (and not by native speakers of English) has been challenged. For example, Scollon and Scollon (1995) used ethnomethodology to show that inductive and deductive patterns both exist in the speech of

both native speakers of English and native speakers of Chinese. The only difference between the two languages is that these patterns are used for different pragmatic purposes. However, their analysis relates only to spoken discourse, so one cannot draw any conclusions about the existence of inductive patterns in written native English. This research gap is filled by Chen (2008). In a quantitative study, he found, among other things, that the minority of the native speakers of English preferred essays written with an inductive rather than deductive pattern and nearly half of them preferred paragraphs written in an inductive rather than a deductive order. This finding indicates that inductive patters can be used in written English but they are more acceptable in paragraphs rather than in essays. Finally Mohan and Lo (1985) review Chinese writing textbooks and analyse Classical Chinese texts to show that the deductive pattern is the most usual and prescribed essay writing pattern in Chinese[1].

From a theoretical perspective, if the RST analysis of the texts in the pilot study can point to instances of inductive order, the controversial issue of whether the English discourse of L1 Chinese learners is characterized by inductive order will be able to be examined in more detail in later research. Moreover, if the present study indicates that inductive-order errors occur frequently in the data, this may be seen as a preliminary indication that AWE software should try to detect and categorize as errors cases of inductive content order.

4. Discourse Errors and *Criterion*

The pilot study reported here is also motivated by one of the criticisms made about AWE software, that is, that the effectiveness of AWE software should not be tested only through "*a posteriori* statistical validation" but also through an "*a priori* investigation of what *should* be elicited by the test before its actual administration" (Weir, 2005: 17). In other words, high levels of agreement in the grades assigned to essays between human judges and software should not be the only criterion for software evaluation; the kinds of errors which are located by software should also match those located by human judges. Such concerns are warranted for practical reasons as well, since it has been shown that learners can fool AWE software, that is, they can get high scores although the content of their essays is inadequate (Herrington & Moran, 2001; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Ware, 2005). Therefore, if AWE software is designed so as to locate the errors that a human judge would locate, wrong essay

[1] Controversy also exists over the cause of inductive patterns whenever they are found in the writing of L1 Chinese learners of English. For example, one possible reason is the influence from L1 rhetorical structure, as contrastive rhetoric theorists claim (eg., Chen, 2001; Kaplan, 1966; Matalene, 1985). Another is the lack of relevant or useful feedback and instruction from teachers (e.g., Gonzales, Chen, & Sanchez, 2001; Mohan & Lo, 1985). Yet another possible reason is the inability to properly structure an essay not only in the L2 but also in the L1 because one has not reached the right developmental stage in his/her writing ability (e.g., Mohan & Lo, 1985).

evaluations will be prevented.

To examine this issue, this section of the paper will summarize the kinds of discourse errors located by *Criterion*[2]. Then, section 9.3 will compare them with the discourse errors located through the RST analysis in the present study. The rationale is that any discrepancies between the two lists of errors should warrant large-scale empirical work testing whether these discrepancies really exist.

The main discourse errors which are located by *Criterion* are those of absence or insufficient number of discourse structures considered necessary in expository and argumentative essays, which are the input of this software. This is a valuable feature because no other AWE software has it (Burstein, 2009: 15). These structures are introductory information which forms the background for the rest of the essay ('Introductory Material'), the statement which expresses the opinion of the writer ('Thesis Statement'), the main point(s) made by the writer ('Main Point'), the statement(s) which support(s) each main point ('Support'), and the conclusion ('Conclusion'). For example, if a learner has not included a thesis statement in his or her essay, the software is likely to locate this error and inform the learner about it.

Apart from the aforementioned discourse-structure tags, the creators of *Criterion* had initially used separate tags for cases where learners had written a title for their essay, for opening and closing salutations in essays in letter format, and for content which could not be tagged with any of the other tags. These tags occurred infrequently, so such cases were lumped under the tag 'Other' (Burstein, 2009: 15; Bustein, Marcu, & Knight, 2003: 33). However, this practice obscures the number of times when the software could not categorize structures through any of the existing labels. This problem could be important because perhaps structures could not be labeled by the software because they violated the usual order of discourse structures (that is, Introductory Material, Thesis, Main Points, Conclusion), an error which should occur whenever information is ordered unusually in an essay. This possibility is likely because in *Criterion* one of the modules used to identify the discourse structures in essays is the 'global language model'. It predicts the sequence of discourse elements in an essay by seeing how well the predictions which stem from a 'local language model' - which predicts which discourse structure is likely to appear after two sections which have already been tagged as specific kinds of discourse structures - fit a final-state grammar manually created by the software creators (Burstein, Marcu, & Knight, 2003: 36).

As we have seen in section 3 of this paper, inductive, rather than deductive, content order has been claimed to characterize the writing of Chinese L1 learners of English; therefore, the

[2] *Criterion*, rather than *My Access!*, was chosen because the research reports on the latter do not give enough information about its workings for its discourse organization evaluation function to be assessable.

software's inability to locate such errors could lead to its low efficacy whenever L1 Chinese learners order their essay content inductively.

A related problem is that because most parts of the software leading to discourse evaluation in *Criterion* are probabilistic, they rely on the most frequent patterns found in essays which were commented on and graded by human graders. This means that errors which did not occur often could not be identified by the software. For example, thesis statements which are scattered in the essay instead of being expressed through one or more adjacent clauses cannot be identified. In an evaluation of the latest version of *Criterion*, the discourse structures that could not be categorized in the training data were 13% of this data (Burstein, Marcu, & Knight, 2003: 36).

One apparent problem with this software is that it presupposes that the essays written will have one paragraph as an introduction, one as a conclusion, and three paragraphs in between, each expressing and supporting one main point. This is an expected structure for a short essay, but this assumption in the software also means that it cannot locate these discourse elements in essays which have fewer or more paragraphs. For example, the writing samples from the lower-intermediate GEPT examination, which form the data for the current study, would not be able to be evaluated by *Criterion* since most of them are one-paragraph long.

Criterion also assesses how balanced the development of an essay is by calculating the proportion of the words in each discourse structure as compared to the total number of words in an essay. This measure seems useful, but it is a crude way of examining degree of development because the length of a structure in terms of its constituent words does not necessarily correlate with how rich it is in content. For example, some learners could repeat the same point in order to meet the required word limit.

Criterion can also decide whether an essay is off-topic or not and also whether content in one or more of the main-point paragraphs is off-topic.

This overview of the discourse errors which *Criterion* can identify shows that it can detect important errors of discourse organization (namely, whether the usual discourse elements occur) and content (namely, whether all or part of an essay is off-topic). This overview has also shown that this software cannot assess essays for discourse coherence, since it cannot identify cases of unusual ordering of content or content which is irrelevant to a specific segment of a text rather than to the essay topic. Recently, research has been conducted with the aim to improve *Criterion* so that it can produce more fine-grained feedback about discourse organization (Higgins, Burstein, Marcu, & Gentile, 2004), but it is still in a preliminary stage. Findings regarding the assessment of coherence inside discourse segments were not encouraging because the criterion used – whether a sentence was related to at least one other sentence in the same discourse segment – was met in the vast majority (92.81%) of sentences (Higgins, Burstein, Marcu, &

Gentile, 2004: 190).

5. A brief Introduction to RST

In their review of theoretical work on RST, Taboada and Mann (2006a: 425) give a simple definition of RST: "RST addresses text organization by means of relations that hold between parts of a text. It explains coherence by postulating a hierarchical, connected structure of texts, in which every part of a text has a role, a function to play, with respect to other parts in the text." The connections which are posited between parts of a text and which show the function of each 'part of text' in the text are called 'coherence relations'. Coherence relations show the function that the analyst thinks that the writer intended each 'part of text' to have in relation to other parts of text.

Some units are called 'nuclei' and others 'satellites'. In RST jargon, nuclei are units of analysis which are necessary parts of a text and satellites are units of analysis which modify the meaning of the nuclei. The main idea of a text needs the nuclei to be put across but if the satellites were deleted, the same main idea, more or less, would be expressed.

For example, the analyst will say that there is an elaboration coherence relation between two units of analysis, if (s)he thinks that the author wishes that the reader recognize the satellite as providing additional information for the nucleus. Figure 1 shows an extract from a paragraph from the LTTC English Learner Corpus. The second and third clauses are linked through the relationship of 'joint' because one is added to the other and jointly modify the first sentence by elaborating its meaning ('elaboration').

[Your teacher may tell you lots of ways to keep your eyes from nearsightedness.] [Such as keep thirty centimeters from your eyes to the table,] [and not to read books when it's dark.]

Figure 1. Extract from a sample paragraph from the LTTC English Learner Corpus illustrating the 'elaboration' coherence relation; each unit of analysis appears within square brackets.

As mentioned above, the coherence relations in a text are usually presented in a hierarchical structure. Figure 2 shows the structure of the extract in Figure 1. The software used to produce it is the RST Annotation Tool by Daniel Marcu[3], which is an improvement on Marc O'Donnell's RSTTool[4]. In RST diagrams, coherence relations are indicated by arrows. An arrow starts from a satellite and points to a nucleus. However, there are also some coherence relations which link units of the same kind. The relation 'joint' is such a 'multinuclear' relation.

[3] This software was downloaded from http://www.isi.edu/licensed-sw/RSTTool/index.html.

[4] This software can be downloaded from http://www.wagsoft.com/RSTTool/section2.html.

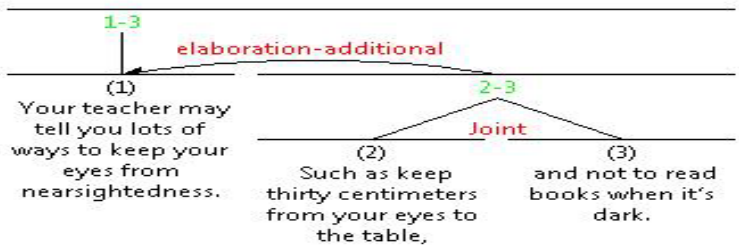


Figure 2. RST diagram indicating the coherence relations in the extract presented in Figure 1.

As mentioned earlier, the analyst chooses the coherence relation which seems to have the function that the writer intended each 'part of text' to have in relation to other parts of text. There are certain constraints on the analyst's choice of a coherence relation, but this paper will describe only one of them because, although they guided RST data analysis, the rest are not directly related to the method of this study. The constraint which helped to form the method of this project is that each text should have the structure of a coherence-relation schema. Such a schema is an abstract representation of coherence relation diagrams. The analyst tries to fit a whole text into one schema and to fit sub-schemas under this schema. Figure 3 shows the schemas which have been posited by Mann and Thompson (1987, 1988).

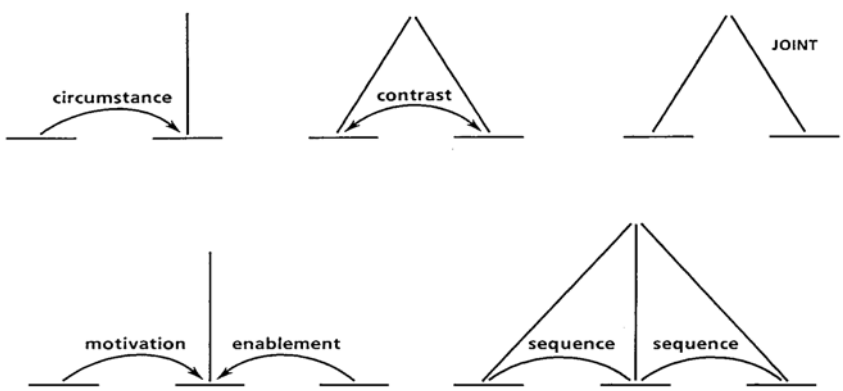


Figure 3. Schemas posited by Mann and Thompson (1987, 1988); figure taken from Mann and Thompson (1987:7).

The aforementioned schema application constraints have some consequences for the location of coherence errors. Since all these requirements must be met for a text to be considered coherent in RST, their violations indicate coherence errors. Therefore, coherence errors are expected to be indicated by diagrams which

a) do not comply with the structure of any schema,
b) include sub-diagrams which do not comply with the structure of any schema, or
c) include schemas which share units of analysis ('crossed dependencies').

This conclusion leads us to the rationale of this study: each kind of coherence error will be indicated by one of these abnormalities in the diagram. By listing the abnormalities which characterize each kind of coherence error, texts can later be tagged for coherence errors in a principled way.

6. Data

The data are 45 paragraphs written by Taiwanese lower-intermediate learners of English in Writing Task 2 of the Intermediate General English Proficiency Test (GEPT) examination, a language proficiency examination administered by the LTTC, a language testing company in Taiwan. In this task, test-takers are asked to write a 120-word paragraph. These files form part of the written section of the LTTC English Learner Corpus, which is currently under construction[5]. The corpus will consist of language samples by Taiwanese learners of English who have sat the GEPT. In the current, first phase of corpus construction, 2,000 written-production and 400 oral-production samples from the Intermediate GEPT examination have been processed.

In order to examine coherence errors in paragraphs written on more than one topic, the 45 paragraphs were equally distributed across topics. Topics were presented to test-takers in Chinese. Two of these topics are questions about personal preferences (favorite food and idol, respectively) and the third asked test-takers to explain why many elementary-school children in Taiwan are nearsighted and to propose effective ways of preventing nearsightedness.

To ensure that the data that would be analysed would vary in terms of coherence error types, samples were equally distributed across score bands in each topic. In other words, in each topic five files had low scores (ranging from 1 to 2), five files had medium scores (ranging from 2.5 to 3.5) and five had high scores (ranging from 4 to 5).

7. Method

The method involved the analysis of the aforementioned paragraphs by the author using the RST Annotation Tool software.

The units of analysis were defined in the same way as in the tagging of 385 documents of

[5] This corpus is compiled under the supervision of Professor Hintat Cheung, the director of the Graduate Institute of Linguistics at NTU. The co-directors are Professor Zhao-Ming Gao, from the Department of Foreign Languages and Literatures at NTU, and Professor Siaw-Fong Chung, from the Department of English at National Chengchi University. I am the postdoctoral research associate working on the project. The other project members are two PhD students, Ms Sally Chen and Ms Chi-Yi Wu, and the research assistant and administrator, Ms Su-Mei Chen. In the academic year 2008-9, the research assistant and administrator was Ms San-Ju Lin.

American English selected from the Penn Treebank (Carlson & Marcu, 2001). Broadly speaking, clauses were the units of analysis, except when they were complements of prepositions and verb objects. However, because the tagset that Carlson, Marcu and their collaborators used was specific to the nature of the texts which they analysed (that is, Wall Street Journal articles), I preferred to use the more neutral coherence relation categories by Bill Mann[6]. Since I combined the units of analysis from the Penn Treebank corpus and Bill Mann's categories, I had to compromise the unit-of-analysis segmentation when the units of analysis warranted a coherence relation which was not among those in Bill Mann's list. This happened when the coherence relation of 'attribution' was posited by Carlson and Marcu to link speech and thought verbs with their complements. In these cases, I considered the verb and its complement clause as one unit of analysis.

As the analysis of the texts was progressing, it became obvious that Bill Mann's list of relations could not cover all the coherence relations in the text, so they were supplemented with eight relations from the tagset by Carlson, Marcu and their collaborators (Carlson & Marcu, 2001). These additional coherence relations were: 'same-unit', 'comment', 'conclusion', 'topic-shift', 'manner', 'explanation-argumentative'.

8. Results

8.1 Qualitative Results

Table 1 summarizes the coherence breaks indicated by the main abnormalities found in the RST diagrams.

Table 1. Abnormalities found in the RST diagrams of the 45 data paragraphs and the coherence breaks indicated by them.

Diagram abnormalities	*Coherence breaks indicated by diagram abnormalities*
Dangling units of analysis	Irrelevant content Incomprehensible content 'Self-sufficiency'
Crossed dependencies	Although a sub-diagram has already been formed for one part of the text, a coherence relation arises between another text part and a unit which is a member of the first sub-diagram
Unexpected relation	Motivation
Relations occurring in unexpected parts of a diagram	Inductive content order

[6] These are the original categories posited by Mann and Thompson (1987, 1988) with some additions and can be found at this website: http://www.wagsoft.com/RSTTool/RSTDefs.htm.

Dangling units of analysis constitute the first kind of RST diagram abnormality. Clauses or larger elements which seem unrelated to the content of the rest of the text are unexpectedly linked to it through a coherence relation. Such dangling units indicated irrelevant content most of the time. There was one case where I left a unit dangling because it was impossible to understand it. Finally, there was one instance of a self-sufficient clause, which explained why the writer liked a specific foreign food in a postscriptum. This error can be categorized as one where the learner was unclear about the layout which (s)he was expected to use.

Figure 4 gives an example of a dangling unit with irrelevant content.

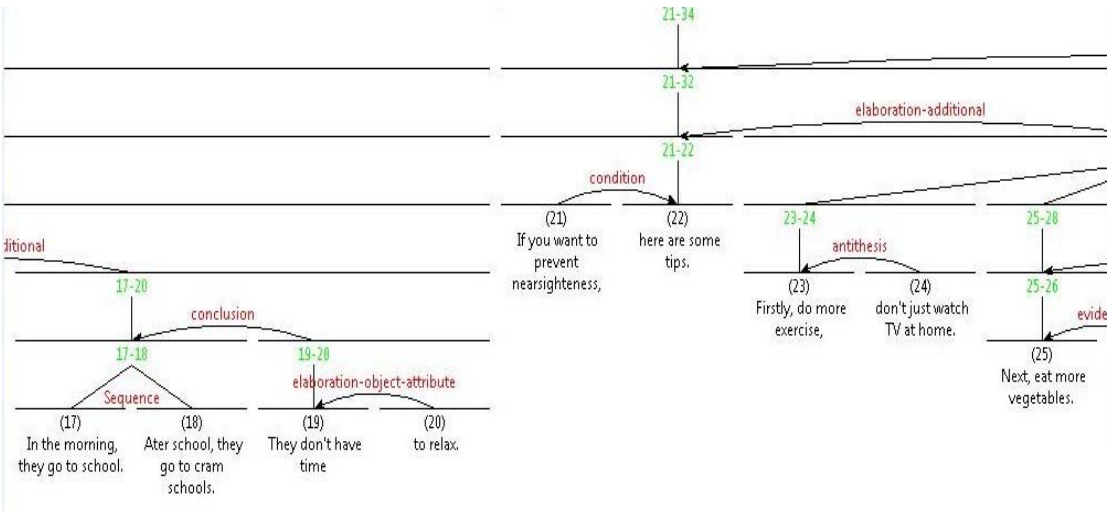


Figure 4. Extract from a diagram where the structure consisting from units 21-34 is dangling because it is irrelevant to preceding text.

This extract comes from a paragraph written on the topic about the nearsightedness of elementary school children in Taiwan. Since the topic asked test-takers to propose effective methods of preventing nearsightedness in general, the advice which the writer gives to the reader in the sub-diagram consisting of units 21 to 34 is irrelevant.

An example of crossed dependencies cannot be illustrated diagrammatically because the RST Annotation Tool automatically corrects such abnormalities in a diagram. However, one can consider the coherence relations among the units in the extract in Figure 5. This figure shows the first lines written in a paragraph on the favorite exotic food topic. In this figure, the units are numbered for ease of reference to them in the discussion that follows.

1. [Taiwan is a special country.] 2. [We can eat a lot of foods from other countries.] 3. [They are gathered in this small island.] 4. [Like Japan, America, Tailand and more.]

Figure 5. Extract from a paragraph on the favorite exotic food topic; each unit of analysis appears within square brackets and the number of each unit of analysis precedes it.

Unit 3 restates information given in unit 2, so 3 is the satellite and 2 the nucleus of a 'restatement' coherence relation. Together, they express a result which stems from the fact that Taiwan is a special country, expressed in unit 1. Therefore, units 2 and 3 together form the satellite of a 'result' coherence relation, where 1 is the nucleus. Unit 4 exemplifies the countries whose food the Taiwanese can eat in Taiwan, so it is the satellite of an 'elaboration' coherence relation and 2 is the nucleus. This coherence relation is problematic because unit 4 intrudes in the sub-diagram which has already been formed by units 2 and 3.

Unwarranted coherence relations constitute the next RST diagram abnormality. The only such coherence relation which was found in the pilot was 'motivation'. It is a coherence relation between a nucleus and a satellite, the latter of which offers a reason why the reader should do something which is expressed in the former. This relation is found in argumentative discourse (Azar, 1999) and not in expository and narrative discourse, which the GEPT test-takers were expected to produce. Figure 6 gives two examples of this error in an extract from a paragraph on the nearsightedness of elementary students in Taiwan.

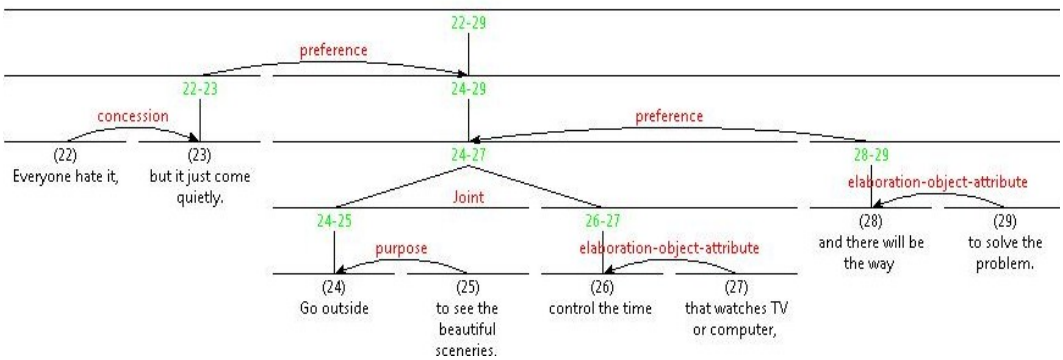


Figure 6. Extract from a diagram where a) the structure consisting of units 22-23 is the satellite in a 'motivation' coherence relation and the structure consisting of units 24-29 is its nucleus and b) the structure consisting of units 28-29 is the satellite in a 'motivation' coherence relation and the structure consisting of units 24-27 is its nucleus.

In units 22 and 23, 'it' refers to nearsightedness. These units jointly form a sub-diagram which serves as the satellite in a 'motivation' relation because they give a reason why someone should do the actions described in the units 24-29.[7] Units 28 and 29 have the same function for units 24-27, so they are the satellite in a 'motivation' relation as well.

Finally, coherence relations in inappropriate parts of a text are the last RST diagram

[7] The relations connecting units 22-23 to units 24-29 and units 28-29 to 24-27 are called 'preference' in this diagram only because the relation 'motivation' is not in the list of coherence relations in the RST Annotation Tool. Throughout the RST analysis of the data, the tag 'preference' was too stand for 'motivation'.

abnormality. These coherence relations are acceptable if they occur in the right parts of a text but there were cases where their location was inappropriate and indicated inductive content order. The 'conclusion' coherence relation indicates a relation where the satellite is a reasoned judgment, inference, necessary consequence or final decision. For example, a student explained why Taiwanese elementary pupils are nearsighted by giving the example of what happened to her younger brother and concluded that "playing video games and watching television too much may be closely related to the cause of elementary students' nearsightedness problem." The 'background' coherence relation usually appears in introductions or briefly in later parts of a text but when students use it extensively in the main body of a text, it may lead to inductive content.

It should be noted that although the RST analysis yielded a wealth of diagram problems which indicate coherence errors in the data, some coherence errors did not show up as problems in the RST diagrams. In other words, the aforementioned diagram abnormalities are not enough to pinpoint all the coherence errors in the data. There are cases where the writer inappropriately addresses the reader but this does not lead to a structural error in the diagram and cases where a topic sentence is missing or scattered in different parts of the text without affecting the diagram. Therefore, the intuition of the error tagger is always necessary for the location of coherence errors.

8.2 Quantitative Results

The variety of coherence errors which were indicated in the preceding qualitative analysis would not be meaningful in this study if it were not supplemented with an analysis aiming to see which errors are the most frequent. The rationale is that those errors which seem to occur often in the data may warrant further investigation in later, large-scale studies. However, these results should be interpreted with caution because they are based on an RST analysis which has been conducted by only one person and only once. In other words, they are based on data which have not been checked for this validity and reliability. Moreover, the number of writing samples analysed is small, so the descriptive statistics which will be presented here are far from statistically reliable. For this reason, inferential statistics have not been conducted on the data.

As it has been mentioned in the overview of the qualitative results, dangling structures usually indicated irrelevant content. However, there was also one case where I could not link a structure to a preceding sub-diagram because this structure was incomprehensible and another because it appeared in a post-scriptum. Because these two errors occurred only once each, I have excluded them from the calculations which resulted in the figures in Table 2. In this table, because written samples varied in terms of their length, the number of occurrences of dangling structures was divided by the total number of units of RST analysis in each sample. Thus, the frequency of this diagram abnormality was normalized in a way appropriate to the way texts

were analysed. The last column is a coarser estimation of the frequency of dangling structures per topic because it is the count of the texts which included at least one dangling structure.

Table 2. Irrelevant content instances across topics according to the RST analysis of 45 paragraphs.

Topic	Cumulative 'dangling' structures normalized per RST units of analysis	Mean 'dangling structures' normalized per RST units of analysis	Writing samples with at least one dangling structure; percentage of texts per topic is given within parentheses
Nearsightedness	0.377	0.021	5 (33.33%)
Idol	0.059	0.004	1 (6.66%)
Exotic food	0.111	0.012	2 (13.33%)

All three measures of frequency agree with each other in that in the 'nearsightedness' topic there are more dangling structures than in the other topics and that the 'exotic food' topic contains more dangling structures than the 'idol' topic. This finding can be seen as indicating that topic affects the occurrence of irrelevant content. Especially in terms of the last frequency measure, it is impressive that in one topic one third of the samples contained irrelevant content. All the frequencies are small, but it should be kept in mind that the maximum number of words in this task was only 120 words. In other words, the short word length created few 'opportunities' for irrelevant content to occur.

It was interesting to examine whether these differences in the frequency of dangling structures also seem related to the score band (low, medium, or high) under which the samples fall. In Table 3 below, the cumulative percentages of the dangling structures are presented in terms of essay topic and score band.

Table 3. Cumulative percentage of 'dangling' structures per topic and score band.

Score band	Essay topic		
	Nearsightedness	Idol	Exotic food
Low score	37.92%	0%	50%
Mid score	38.65%	100%	0%
High score	23.42%	0%	50%
Total percentage	100%*	100%	100%

The total number from the percentages in this column is 99.99% because these numbers are rounded. The exact total number is 100%.

The breakdown of samples which contain dangling structures in the nearsightedness topics is as expected, since one would expect that learners with low and mid scores would be more likely to include irrelevant content in their writing than the high-performing learners. The data from the other two topics is more complicated, since all cases of irrelevant content in the idol topic occurred in the middle-score paragraphs and half of them in the low- and the other half in the high-score paragraphs in the food topic. However, this finding can be easily explained by the very few occurrences of dangling structures for the idol and food topics. There was only one occurrence of a dangling structure in the idol topic and it was in a middle-score paragraph and there were only two occurrences in the food topic, one in a low- and the other in a middle-score paragraph.

In sum, results on dangling structures show that this type of RST diagram abnormality indicates irrelevant content and that the frequency of such errors depends on the essay topic, at least in the writing of these low-intermediate Taiwanese learners of English.

Coherence errors stemming from crossed dependencies are likely to be very rare since this data contains only one such error.

As explained in the previous section, the coherence relation of 'motivation' was unexpected because it normally occurs in argumentative text types whereas the essay topics were expository. This coherence relation occurred only in the paragraphs written on the 'nearsightedness' topic. This finding is congruent with the previous finding that irrelevant content made manifest by dangling structures was much more frequent in the nearsightedness than in the other texts. Indeed, it seems that there is some interrelation between dangling structures and the existence of a 'motivation' relation in nearsightedness texts, as shown in Table 4.

Table 4. Total and mean number of Motivation relation instances in the paragraphs written on the Nearsightedness topic according to the RST analysis and percentage of paragraphs which included both at least one 'motivation' relation and at least one 'dangling' structure.

Cumulative instances of 'motivation' coherence relation normalised per RST unit of analysis	Mean instances of 'motivation' relation normalized per RST units of analysis	Percentage of paragraphs with 'motivation' relation which also have 'dangling' structures
0.112	0.007	66.67%

In terms of to the coherence errors due to the occurrence of a coherence relation in an inappropriate part of a paragraph, Table 5 presents the same kinds of normalized data as Table 2 but for the inappropriate occurrences of the 'background' coherence relation.

Table 5. Inappropriate uses of the 'background' coherence relation across topics according to the RST analysis of 45 paragraphs.

Topic	Cumulative inappropriate uses of the 'background' coherence relation normalized per RST units of analysis	Mean inappropriate uses of the 'background' coherence relation normalized per RST units of analysis	Number of writing samples with at least one instance of an inappropriate use of the 'background' coherence relation; percentage of texts per topic is given within parentheses
Nearsightedness	0.059	0.004	1 (6.67%)
Idol	0.184	0.012	3 (20%)
Exotic food	0	0	0 (0%)

As it can be seen, the majority of cases occur in the idol topic, so it seems that the occurrence of such errors also depends on topic. To see whether there was a score-band effect as well, in Table 6 below, the cumulative percentages of the dangling structures are presented in terms of essay topic and score band.

Table 6. Cumulative percentage of cases of 'background' coherence relation per topic and score band (there were no cases in the 'exotic food' topic).

Score band	Essay topic	
	Nearsightedness	Idol
Low score	0%	0%
Mid score	0%	42.83%
High score	100%	57.17%
Total percentage	100%	100%

This table indicates that the 'background' coherence relation occurred in the wrong part of the text for paragraphs which achieved medium and high scores. This finding may not be significant in the 'nearsightedness' topic since this error was only found in one paragraph, but it seems to be more important in the 'idol' topic since this error occurred in one fifth of these paragraphs.

The last coherence error indicated by the RST analysis is the use of the 'conclusion' coherence relation in an inappropriate part of the text. This error occurred only twice and only in two middle-score paragraphs, so it seems that this error occurs rarely. Moreover, it occurred only in the 'nearsightedness' topic, so this error is also possibly due to a topic effect.

9. Discussion

9.1 RST Analysis as a Means to Coherence Error Detection

The results of this pilot study indicate that different kinds of abnormalities in RST diagrams built on writing samples of low-intermediate GEPT test-takers indicate various coherence errors. In particular, dangling units and unexpected coherence relations in the diagrams are indications of irrelevant content. Coherence relations in inappropriate parts of the text indicate inductive content order. Finally, the crossed dependencies indicate local coherence errors because they apply to coherence relations within rather than across sub-diagrams. Consequently, this method of textual analysis seems promising. However, as it has been mentioned in section 8.2 above, this method cannot detect all coherence errors that a human analyst can. Moreover, it is labor-intensive, so it may be impractical to use. Therefore, if this method proves effective – that is, if in a large-scale study inter- and intra-judge reliability are high and the agreement between the coherence errors located by the RST analysis and the judgments of language teachers and/or native speakers is high – it should be used by skilled analysts and only when fine-grained analyses of coherence errors are desirable.

9.2 Coherence Errors by Low-intermediate Taiwanese Learners of English

As mentioned in section 3, this pilot study also aimed to examine which coherence errors are made by a specific population of English language learners, namely, low-intermediate Taiwanese learners of English. As mentioned above, errors of irrelevant content, inductive content order, local coherence errors due to crossed dependencies, use of an inappropriate coherence relation (i.e., 'motivation') and the occurrence of coherence relations 'background' and 'conclusion' in inappropriate parts of a text have been detected via the RST analysis. However, as mentioned in section 3, inappropriate addresses to the reader were not detected through the RST analysis. These addresses are inappropriate because the topics were expository and addresses to the reader are common in argumentative writing. Cases where a topic sentence is missing or scattered in different parts of the text could also not be detected through the RST analysis.

The frequency of the coherence error types which could be detected through RST in the data should be considered with caution given the small number of paragraphs, their short word length, and the fact that the analysis was not checked for inter- and intra-judge reliability. Keeping this caveat in mind, one can note that the 'dangling structure' RST diagram abnormality is the most frequent one. The second most frequent RST abnormality was the inappropriate use of the 'background' coherence relation. The third most frequent RST

abnormality was the unwarranted occurrence of the 'motivation' coherence relation.[8] All the other RST diagram abnormalities occurred so infrequently that it seems that they are unlikely to occur frequently in a large-scale study. There was only one case of crossed dependencies and only two cases of inappropriate use of the 'conclusion' coherence relation.

As indicated in the quantitative analysis of the data in section 8.2, all coherence errors located in the data seem to vary depending on topic. Most dangling structures occur in paragraphs on the 'nearsightedness' topic and most cases of inappropriate use of the 'background' coherence relation occur in paragraphs on the 'idol' topic; the 'motivation' coherence relation and the inappropriate use of the 'conclusion' coherence relation occur only in paragraphs on the nearsightedness topic. These indications of topic effects – which could be due to topic content, phrasing or other topic characteristics – point to the need to investigate the occurrence of coherence errors for this population further. Analysing larger numbers of data and writing samples from a larger variety of topics will be able to indicate which coherence errors are frequent irrespective of writing topic and, therefore, warrant more attention from English language teachers and AWE software.

The quantitative analysis by both score band and topic for the two most frequent errors, namely, 'dangling structures' and the inappropriate use of the 'background' coherence relation, showed a different trend for each of these errors. For the 'dangling structure' error, this analysis was done only for the answers to the 'nearsightedness' topic because, contrary to the other two topics, it received a number of such errors big enough for this data analysis to be meaningful. The breakdown of error numbers according to score bands was as expected since most 'dangling structures' occurred in the low- and mid-score paragraphs. In terms of the inappropriate use of the 'background' coherence relation, only the answers to the 'idol' topic received enough answers for the analysis per score band to be meaningful. Here, the results were different from those for the 'dangling stucture' errors because most such errors occurred in the paragraphs which had received high scores. Moreover, the second most error-populated score band was the mid one and no such error occurred in the low-score band. As mentioned in section 8.1, 'dangling structure' errors indicate irrelevant content and the inappropriate use of the 'background' coherence relation indicates inductive content order. The contrasting aforementioned results between the two error types can be explained through a consideration of the literature on the criteria used in essay marking. Research indicates that in L1 essay grading, the focus is on discourse organization whereas in L2 essay grading the focus is on syntactic and lexical errors (e.g., Breland & Jones, 1982; Gonzáles, Chen, & Sanchez, 2001). This fact may

[8] These frequency comparisons were made according to all measures used in this study (that is, cumulative occurrences normalized by RST units of analysis, mean occurrences normalized per RST units of analysis, and number of writing samples with at least one occurrence).

explain why some students received mid- and high-grades although their paragraphs were partly organized inductively. Another possible explanation is that these paragraphs formed part of students' answers to the level of the GEPT examination aimed to low-intermediate learners of English; markers considered that more local errors should weigh more in the marking than discourse errors. Both these possible explanations indicate the need for an examination of errors other than discourse errors in the data, so that the gravity of these errors in each score-band can be compared against that of the discourse errors. Therefore, to reach a conclusion about whether the discourse errors located often in this pilot study occur frequently in the writing of low-intermediate Taiwanese learners of English in general, we do not only need to repeat this analysis with more writing samples and two analysts, but also the writing samples should be tagged for other errors as well. The development of an error tagging system for the LTTC English Learner Corpus is currently under way.

Another central aim of the study was to examine whether inductive order errors are frequent in the writing of low-intermediate Taiwanese learners of English. The relatively frequent occurrences of the 'background' coherence relation in inappropriate parts of a paragraph indicate that these learners make such errors. However, the concentration of such errors in the paragraphs written on the 'idol' topic indicates the possibility of topic effects. As mentioned above, a large-scale study with samples written on a larger variety of topics is necessary to measure the frequency of inductive order errors and their relation to topic effects.

9.3 Coherence Errors Detected via RST Analysis and *Criterion*

As mentioned in section 4, a secondary aim of this pilot study is to examine the extent to which the errors located through the RST analysis can also be located via *Criterion*. This section will examine this issue by considering each coherence error separately.

The most frequent kind of RST diagram abnormality in the data of the present study is 'dangling structures'. In all except two cases, dangling structures indicate irrelevant content, so if this finding proves valid through a large-scale study, AWE software should be able to locate irrelevant-content errors. It is unclear to me whether *Criterion* would be able to categorize such cases as off-topic. The first method used in *Criterion* to locate off-topic essays and segments is through comparisons of the vocabulary used in the essays used for training the software to score and give feedback on a specific topic. The second method is a comparison between the proportion of times in which a word is used in a variety of topics and that where a word is used in a specific topic. The third method does not require training data and relies on a comparison of the vocabulary in the essay prompt and in the essays (Burstein, 2009: 8-12). In segments like the dangling one in Figure 4, topic-related vocabulary is used, so such segments would probably not be categorized as irrelevant to the topic by any of these methods. However, one should keep in mind that the high concentration of irrelevant content errors in the paragraphs written on the

nearsightedness topic indicates the possibility of topic effects on the occurrence of this error. Large-scale studies are necessary to clarify whether this error occurs often in the writing of low-intermediate L1 Chinese learners of English irrespective of topic-related factors.

The inappropriate use of the 'background' coherence relation is the second most frequent coherence error located through the RST analysis of the data. As explained earlier, it indicates inductive content order. Inductive content order cannot be detected by *Criterion*. However, the high concentration of this error in paragraphs written on the 'idol' topic may mean that this finding is just due to a topic effect and may not occur across topics. If large-scale studies indicate this, *Criterion* and other AWE software would not need to detect this error. Moreover, Chen's (2008) finding that paragraphs with inductive content order were acceptable by half the English native speakers participants in his study may mean that inductive order content should not be considered an error and calls for further examination of what makes inductive content order more or less acceptable for a native English speaker.

The next most frequent coherence error detected through the RST analysis is the unwarranted occurrence of the 'motivation' coherence relation. It is unclear whether the methods which detect off-topic content would be able to locate unwarranted instances of the 'motivation' relation in an essay. However, in any case, the fact that this relation occurred only in paragraphs written on the 'nearsightedness' topic may indicate a strong topic effect and large-scale studies which manipulate topic characteristics should be conducted to examine whether such coherence errors occur often enough and across topics to warrant the creation of AWE software which can detect them.

The inappropriate use of the 'conclusion' coherence relation occurred only twice. Its very low frequency probably means that AWE software would not need to locate this coherence error. As for the inappropriate use of the 'background' coherence relation, it cannot be detected by *Criterion*. This error was the second most frequent one, but again, most of its instances occurred in paragraphs written on only one topic, the 'idol' one. This finding warrants large scale studies which will examine whether such errors occur frequently across topics for this learner population.

The local coherence error caused by crossed dependencies occurred only once in the data. There is some controversy over whether such diagrammatic structures should be considered erroneous, because it has been claimed that crossed dependencies occur in the productions of native speakers as well (Wolf & Gibson, 2004, 2005). Therefore, such errors probably do not warrant further investigation or location through AWE software.

10. Conclusion

The main finding of this study is that an RST analysis of short texts written by low-intermediate L1 Chinese learners of English can provide detailed information about coherence errors. Nevertheless, the study presented here is only a pilot, so it was conducted on a small number of texts and only by one researcher. Therefore, this paper mainly serves as an index of research questions that need to be addressed through further research. As mentioned above, the inter- and intra-judge reliability of the analysis remains to be tested and the frequency of each kind of coherence error located needs to be measured through the analysis of larger numbers of texts.

The results of this pilot study strongly indicate the possibility of topic effects on coherence error occurrence. Therefore, further examination is also necessary to examine whether the topic effects on coherence errors occur when more samples are analysed. Moreover, texts written on topics which vary in terms of various dimensions (e.g., text type associated with a topic, how clearly the topic question explains what the essay structure should be) should be examined to examine whether certain kinds of topics lead to certain kinds of coherence errors. If these topic effects are confirmed through further research, attempts should be made to explain them. This research would be beneficial for AWE design, since if the topic-related factors shown to influence these errors could be detected by AWE software, essay scoring and feedback would be refined.

References

Azar, M. (1999). Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. *Argumentation*, 13(1), 97-114.

Burstein, J. (2009). Opportunities for Natural Language Processing Research in Education. In Gebulkh, A. (Ed.), *Springer lecture notes in computer science* (Vol. 5449, pp. 6-27). Springer: New York, NY.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online Writing Evaluation service. *AI Magazine*, 25(3), 27-36.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. In Harabagiu, S. & Ciravegna, F. (Eds.), *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 18(1), 32-39.

Carlson, L. & Marcu, D. (2001). *Discourse Tagging Manual.* ISI Tech Report ISI-TR-545.

Chen, J.P. (2001). *Markedness in intercultural discourse: a study of Chinese EFL students' discourse patterns.* PhD thesis, Guangdong University of Foreign Studies, in ERIC, RIE June 2001.

Chen, J.P. (2008). An investigation into the preference for discourse patterns in the Chinese EFL learning context. *International Journal of Applied Linguistics*, 18(2), 188-211.

Díaz-Negrillo, A. & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, 19, 83-102.

González, V., Chen, C-Y., & Sanchez, C. (2001). Cultural Thinking and Discourse Organizational Patterns Influencing Writing Skills in a Chinese English-as-a-Foreign-Language (EFL) Learner. *Bilingual Research Journal*, 25(4), 417-442.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, 63(4), 480-499.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In Dumais, S., Marcu, D. & Roukos, S. (Eds.), *Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL) 2004: Main Proceedings* (pp. 185-192). Boston, MA: Association for Computational Linguistics.

Kaplan, R. (1966). Cultural thought patterns in intercultural education. *Language Learning*, 16(1), 1-20.

Mann, W.C. & Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87-190). Marina del Rey, CA: Information Sciences Institute.

Matalene, C. (1985). Contrastive rhetoric: an American writing teacher in China. *College English*, 47(8), 789-808.

Mohan, B.A. & Lo, W.A.-Y. (1985). Academic writing and Chinese students: transfer and developmental factors. *TESOL Quarterly*, 19(3), 515-34.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stamping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2), 103-134.

Scollon, R. & Wong-Scollon, S. (1995). *Intercultural communication: A discourse approach.* Blackwell: Oxford, U.K. and Cambridge, MA, U.S.A.

Taboada, M. & Mann, W.C. (2006a). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3), 423-459.

Taboada, M. and Mann, W.C. (2006b). Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4), 567-588.

Vantage Learning. (2007). *MY Access! Efficacy Report***.** Newtown, PA: Vantage Learning. Retrieved on August 6 2009 from http://www.vantagelearning.com/docs/myaccess/myaccess.research.efficacy.report.200709.pdf.

Ware, P. (2005). Automated writing evaluation as a pedagogical tool for writing assessment. In A. Pandian, G. Chakravarthy, P. Kell, & S. Kaur (Eds.), *Strategies and practices for improving learning literacy* (pp. 174-184). Selangor, Malaysia: Universiti Putra Malaysia Press.

Watson Todd, R., Khongput, S., & Drasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing*, 12(1), 10-25.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York: Palgrave MacMillan.

Wolf, F. & Gibson, E. (2004). Representing Discourse Coherence: A Corpus-based Analysis. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (article no. 134), Geneva, Switzerland.

Wolf, F. & Gibson, E. (2005). Representing Discourse Coherence: A Corpus-based Analysis. *Computational Linguistics*, 31(2), 249-287.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 205-220

Effects of Collocation Information on Learning Lexical Semantics for Near Synonym Distinction

Ching-Ying Lee[*,+] and Jyi-Shane Liu[#]

Abstract

One of the most common lexical misuse problems in the second language context concerns near synonyms. Dictionaries and thesauri often overlook the nuances of near synonyms and make reference to near synonyms in providing definitions. The semantic differences and implications of near synonyms are not easily recognized and often fail to be acquired by L2 learners. This study addressed the distinctions of synonymous semantics in the context of second language learning and use. The purpose is to examine the effects of lexical collocation behaviors on identifying salient semantic features and revealing subtle difference between near synonyms. We conducted both analytical evaluation and empirical evaluation to verify that proper use of collocation information leads to learners' successful comprehension of lexical semantics. Both results suggest that the process of organizing and identifying salient semantic features is favorable for and is accessible to a good portion of L2 learners, and thereby, improving near-synonym distinction.

Keywords: Lexical Semantics, Near-synonym Distinction, Lexical Collocation Behavior.

1. Introduction

One of the most common lexical misuse problems in the second language context concerns near synonyms. Near synonyms are lexical pairs or sets that have very similar cognitive or denotational meanings. Dictionaries and thesauri often overlook the evaluative distinctions among near synonyms and 'end up showing certain circularity' in providing semantic meaning (Tognini-Bonelli, 2001). L2 learners are left with individual judgment and preference in lexical choices of almost synonymous words. Near synonyms, however, may vary in

[*] Department of English, National Taiwan Normal University, Taipei, Taiwan
[+] Department of Applied Foreign Languages, Kang Ning Junior College, Taipei, Taiwan
E-mail: chingying.lee1212@gmail.com, cylee@knjc.edu.tw
[#] Department of Computer Science, National Chengchi University, Taipei, Taiwan
E-mail : jsliu@cs.nccu.edu.tw

[Received July 1, 2009; Revised January 22, 2010; Accepted January 28, 2010]

collocational or implicative behavior (Partington, 2004). Among a group of nearly synonymous words, some may indicate favorable conditions while others refer to unfavorable situations, and some may show approval while others imply disapproval. These subtle distinctions between near synonyms are not easily identified and may never be acquired by L2 learners.

Lexical use is an area where L2 learners frequently demonstrate a number of errors. Many L2 learners rely on dictionaries and thesauri to provide denotational meaning of a lexical item without being aware of the subtle implications embedded in contexts. Implicit knowledge of lexical items is not easily taught. Semantic infelicities due to inappropriate lexical use leads to miscommunication and unfavorable social consequences. Therefore, misuse of lexical items, particularly among near synonyms, calls for more attention and treatment in L2 lexical learning.

The purpose of this research is to explore the potential of applying computerized linguistic resources and observing collocation behaviors in semantic learning for near synonym distinction. We propose a categorized collocation profile with graded association strength to filter and organize salient semantic features. It serves as a guided process to help develop concrete conceptual links so semantic meaning and unique features of lexical items become more easily accessible to L2 learners. Both analytical evaluation and empirical evaluation are performed to examine the effects of collocation information on near synonym distinction. Observations and implications in regards to L2 semantics learning are described.

2. Literature Review

Knowledge of the appropriate contextual use of the particular languages' resources is a crucial component of linguistic competence (Barron, 2003). L2 learners often face difficulties in understanding subtle and elusive nuances of appropriateness (Dewaele, 2008). The task of making proper lexical decisions between near synonyms is particularly challenging for L2 learners and requires adequate semantic competence. It is inadequate to only know a word meaning or definition. A core lexical competence is characterized by appropriateness of word choices, particularly between near synonyms.

The idea of using collocation information to observe the word sense has been developed in post-Firthian corpus linguistics. The relevant studies investigate how a lexical item functions to convey semantic meanings, or how it carries out its discursive or evaluative properties (Sinclair, 2003; Channell, 2000; Stubbs, 2001; Partington, 2004). L2 learners should be aware that lexical meanings cannot be determined only by semantics. Therefore, it is helpful to examine the effects of collocation information on lexical meaning and functions.

According to Stubbs, 'there are always semantic relations between node and collocates

and among collocates themselves' (2001). The collocational information is interpreted through the proximity of a consistent series of collocates (Louw, 2000). Its main function is to convey the speaker or writer's attitude or evaluation. According to priming theory, Partington (2004) indicates that a person has a set of mental rules in the priming process, combined with the mental lexicon, of how items should collocate. In addition, the process by which lexical items are primed in one's mind is highly contextually dependent. The corpus linguistic techniques for lexical collocation provide a distinctive way to study semantic profiles.

The problem of near synonym distinction and appropriate lexical choice is especially daunting for second language learners (Mackay, 1980). The majority of vocabulary errors made by advanced language learners reflect learners' confusion among similar lexical items in the second language. The language of explanations in dictionaries is somewhat arcane such that it becomes limited in accessibility and usefulness in practical L2 contexts. Martin (1984) discussed instructional approaches to synonym teaching and suggested the importance of providing common collocates to students. With the availability of computerized corpora, recent research has exploited concordances and collocation data for advising L2 learners in lexical choice (Yeh, *et. al.*, 2007; Chang, *et. al.*, 2008). Through enquiry into the interplay between lexical semantics of near synonyms and their collocation information, this study provides analytic and empirical observations and contributes to reducing L2 learners' confusion of sophisticated lexical connotations and applications.

3. Methodology

Corpus-based approaches to applied linguistics assert that lexical semantics can be revealed by study of a large corpus. The analysis of the corpus uses computational techniques to identify words that typically co-occur with a lexical item under investigation. Our study attempts to understand the potential of adopting corpus linguistics for the purpose of improving learners' performance in lexical semantics. In particular, we focus on investigating the effects of lexical collocation information on near-synonym distinction in either the self-learning or the classroom context.

Recent developments in concordancing tools include web-based systems that provide online access to query and retrieval. Both Sketch Engine (Kilgarriff, *et. al.*, 2004) and VIEW (Davies, 2008a) are powerful tools for corpus-based language research. Research issues concerning lexical behavior, collocational pattern, syntax, and semantics can all be facilitated by the language data access capability and the statistical summarization functions of these state-of-the-art concordancing tools. For the purpose of exploring the potential of lexical collocation information for semantic grounding and synonym distinction, we adopted VIEW as the concordancing tool in our study and used it to retrieve collocation information based on its access to two large corpora, BNC (Burnard, 1995) and COCA (Davies, 2008b).

The notion of collocational profile is proposed to provide an organized description of collocation behavior. Collocates are grouped by POS categories and graded by association strength with a keyword. The statistical measure chosen to gauge association strength in the study was the mutual information (MI) measure (Church & Hanks, 1990). The MI measure compares the probability of two words occurring together through intention with the probability of the two words occurring together by chance. Higher MI scores indicate strong association between two words. An MI score greater than 2 can be considered high enough to show a substantial association between two words. The MI measure, however, has been known to unduly overvalue infrequent words. The list of words considered in the collocational profile is restricted to the top 20 with the highest frequency of occurrence and has a minimum number of 5. These adjustments have allowed us to partly offset the drawbacks of MI measure.

For transitive verbs such as *affect/influence*, we focus on the basic syntactic pattern of S (subject noun) +V (transitive verb) + O (object noun) and a few extended patterns, such as Adv (adverb) + V + O, and V + Adv + O. Words that meet the constraints of POS tags and occurrence positions with respect to the keyword (transitive verb) are retrieved by VIEW and classified into three categories: subject collocates, object collocates, and adverb collocates. The positional constraint for subject collocates is the left horizon of the keyword within a span of five words. Object collocates are restricted to the right horizon of the keyword within a span of five words. Adverb collocates must be immediately before or after the keyword.

When the list of most frequent collocates is retrieved, the collocates are further graded by their MI scores. Collocates with MI scores higher than 5.5 are graded as dominant collocates. Collocates with MI scores lower than 3.5 are graded as moderate collocates. Those in between are graded as strong collocates. The grade order of dominant, strong, and moderate indicates the decreasing strength of association between the collocates and the keyword. The POS categorization and the graded association strength of collocates provide a profile that highlights the significant semantic links and illustrates the interactive network of semantic meaning. This will help enhance a concept map of the keyword where semantic features become more recognizable and synonym distinction is clarified.

Figure 1 is a screenshot of VIEW with BNC, where collocation information for the keyword *affect* was retrieved. The search string portion specifies the targeted collocation constraint as the adverb (POS) occurring in the span of one word in both directions (left and right) of *affect* as verb. The upper right portion of the window shows the search result, which is a list of collocated adverbs sorted by MI value. This constitutes the lexis list and MI-BNC value in the collocational profile of *affect*, as shown in Table 3. The complete collocation profile of a keyword is constructed by multiple uses of VIEW with various collocation constraints and corpora.

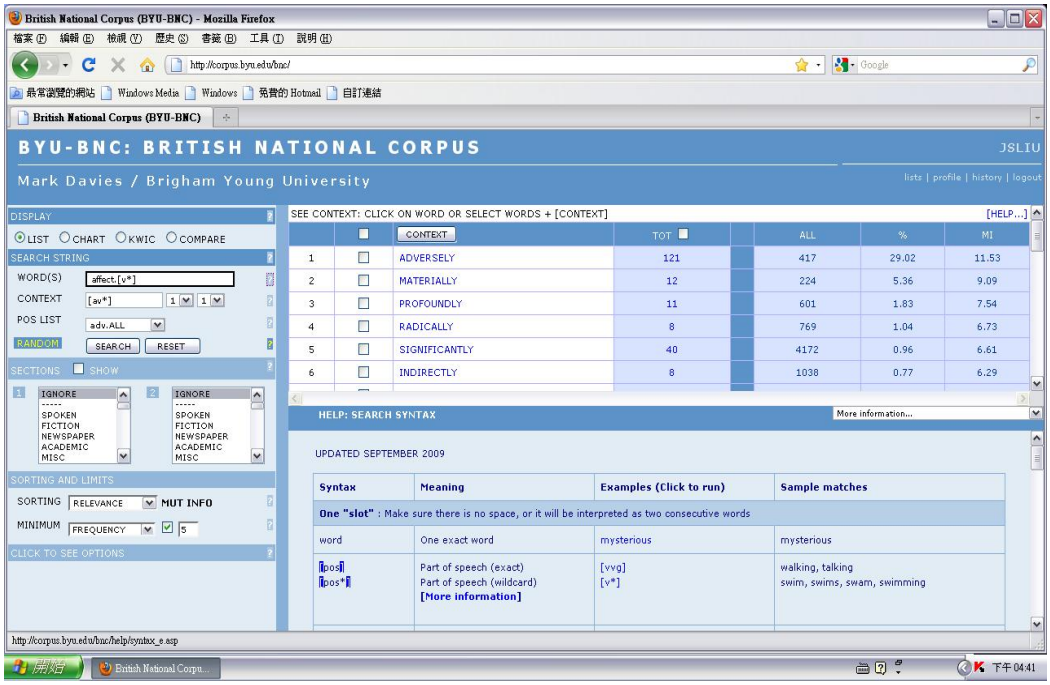


Figure 1. Screenshot of VIEW providing collocation information.

4. Evaluation

Two sets of tests are conducted to explore and verify the effects of collocation information on lexical semantics acquisition and near synonym distinction. In the first test, we walked through the process of producing a collocational profile, acquiring semantic features, and illuminating semantic distinction between near synonyms. The purposes were performing an objective analysis on the effects of collocational profiles in leading to a clear description of semantic features and allowing comparative induction that reveals subtle semantic differences between near synonyms. The second test involved a written test and survey given to a group of recruited test subjects. The purpose was to solicit language learners' actual experience and observe the effects of collocational profiles on language learners' performance in near synonym distinction tasks. By conducting both analytical and empirical verification, we hoped to achieve a sound investigation to better understand the extent to which collocational profiles can help reveal semantic distinctions of near synonyms to L2 learners.

4.1 Analytical Verification

The near-synonyms, *affect* and *influence*, were chosen for the study based on the degree of difficulty for L2 learners and their fitness in serving as a representative lexical semantics learning task. Dictionary definitions given by Merriam-Webster are: -*affect*, 1. to act upon; to

produce an effect or change upon; 2. to influence or move, as the feelings or passions; -*influence*, 1. to control or move by power, physical or moral; 2. to affect by gentle action, to exert an influence upon. Webster's New Dictionary of Synonyms gives the following discrimination: -*affect*: 1. always presupposes a stimulus powerful enough to evoke a response or elicit a reaction; 2. implies a definite alteration or modification; -*influence*: always presupposes an agent that moves a person or thing in some way or to some degree from a course, or effects changes in nature, character, or behavior. Unfortunately, these abstract explanations of discrimination are confusing to most L2 learners and do not provide definite clarification.

Table 1. Comparison of subject collocates of near-synonyms (affect, influence).

affect				*influence*			
type	lexis	MI -BNC	MI -COCA	type	lexis	MI -BNC	MI -COCA
dominant	--	--	--	dominant	factor	6.29	6.21
strong	factor	4.92	5.02	strong	variable	--	5.12
	variable	--	4.22		government	5.00	--
	disease	(3.07)	3.68		ability	3.57	3.78
	decision	(3.34)	3.61		--	--	--
moderate	condition	2.36	3.32	moderate	attitude	--	3.44
	issue	2.73	3.02		behavior	--	3.37
	policy	2.17	2.73		decision	2.81	2.22
	matter	2.59	--		culture	--	2.61
	behavior	--	2.46		policy	2.08	1.89
	change	2.42	2.12		teacher	--	1.92
	action	--	2.29		process	1.99	1.81
	problem	1.66	1.91		experience	--	1.90

Tables 1, 2, and 3 show comparisons of subject, object, and adverb collocates of *affect* and *influence*. Collocational profiles seem to provide contextual evidence that can be used by L2 learners to derive grounding features for concrete discrimination. The following observations were made based on comparison of collocations: 1. The subject of *affect* seems to be a stimulus that would evoke changes, while the subject of *influence* tends to be physical or abstract entity that has power to cause changes. 2. Objects' status changes accomplished by *affect* seem to be more obvious for recognition, while changes caused by *influence* are more related to some inner status. 3. The listed adverbs are all dominant collocates, indicating the manner of making changes is an important parameter of semantic features of the two near-synonyms. 4. The manner of making changes in *affect* seem to be related to the magnitude of effects, while the extent of control is the focus in describing changes done by *influence*. 5. The association of *adversely* with *affect* is outstanding with MI scores higher than 11. *Affect* also has a unique collocate of *severely* and the stronger association of *negatively* than *positively*. These are compelling evidence to the unfavorable (negative) prosody of *affect*.

Overall, we derived the following distinction based on collocational profile evidence. *Affect* implies mostly negative impact or disturbance caused by a strong stimulus. *Influence* assumes some entity that has a power to exert subtle control over the object.

Table 2. Comparison of object collocates of near-synonyms (affect, influence).

affect				*influence*			
type	lexis	MI -BNC	MI -COCA	type	lexis	MI -BNC	MI -COCA
dominant	--	--	--	dominant	government	5.50	--
strong	life	5.42	(2.75)	strong	outcome	5.39	5.39
	outcome	--	4.76		perception	--	5.37
	ability	(3.38)	4.12		behavior	5.11	4.97
	performance	(3.27)	4.11		decision	4.48	4.90
	behavior	3.70	3.61		attitude	4.82	4.64
	quality	(2.37)	3.69		life	4.29	--
	--				policy	3.56	4.14
					opinion	4.07	--
					choice	3.90	3.71
					direction	3.61	--
moderate	decision	2.83	3.30	moderate	development	3.03	3.48
	health	2.58	3.02		--		
	rights	2.93	--				
	relationship	--	2.73				
	policy	1.58	2.35				
	development	1.60	2.32				

Table 3. Comparison of adverb collocates of near-synonyms (affect, influence).

affect				*influence*			
type	lexis	MI -BNC	MI -COCA	type	lexis	MI -BNC	MI -COCA
dominant	adversely	11.53	11.87	dominant	unduly	7.74	9.25
	negatively	--	9.36		profoundly	8.84	8.63
	materially	9.09	--		greatly	7.54	8.26
	profoundly	7.54	7.97		positively	(5.00)	8.17
	positively	--	7.77		strongly	7.98	8.05
	radically	6.73			heavily	7.19	7.66
	significantly	6.61	6.63		negatively	--	7.60
	indirectly	6.29	7.23		indirectly	--	7.03
	seriously	5.78	4.92		significantly	5.75	6.35
	dramatically	5.33	5.72		deeply	6.34	5.85
	directly	5.29	6.26		directly	(4.95)	5.68

4.2 Empirical Verification

We constructed a set of ten test questions concerning contextual lexical choice of *affect* and *influence*. Each test question was composed of an independent sentence in which one of the near synonyms is the intended component as a verb and the test part is highlighted as a lexical choice between the near synonym pair. For example, how did your past experiences **affect** or **influence** the way you coped with changes? Test subjects were asked to decide which of the two near-synonyms was the correct lexical use in the sentential context.

The same test questions were administered to the subjects in three phases with different contexts. In phase one, the subjects answered the test questions with L1 translation of the near-synonyms and their own lexical recognition. In phase two, the same set of test questions were given to the subjects with L2 denotation of the near-synonyms from two dictionaries, one being an English-English dictionary (the Merriam-Webster's), denoted as D1, and the other being a dictionary of synonyms (Webster's new dictionary of synonyms), denoted as D2. After answering the test questions, subjects were asked their opinion of whether each type of dictionary was useful in distinguishing the near-synonyms and making the correct lexical choice. In phase three, collocation information of the near-synonyms was provided and the same set of test questions were used again. At the end of the test questions, subjects were asked to indicate whether collocation information was useful in near-synonym distinction. The full questionnaire is shown in the Appendix. The test subjects recruited were 40 English-major freshmen at a top-tier university in Taiwan. The test was taken in a self-learning context.

Table 4. Overall test results of "affect/influence" distinction.

	Test Score	Confidence		Usefulness
subjects' recognition	6.25/10 (1.63)	4.3/10 (3.09)	dictionary	52.5% (21/40)
with dictionaries	5.53/10 (1.63)	6.48/10 (3.04)	synonym dictionary	42.5% (17/40)
with collocation	6.15/10 (1.73)	5.9/10 (3.22)	collocation profile	67.5% (27/40)

Table 4 shows the summarization of the group performance and overall effects of additional semantic information with respect to the task of near-synonym distinction. We make the following observations.

1. The subjects scored 6.25 points (out of 10) on average in making the correct lexical choices between *affect* and *influence* with a standard deviation of 1.63. The lexical decisions were deem confident only 4.3 times (out of 10) on average with a standard deviation of 3.09. The performance in making correct lexical choices is not particularly satisfactory. The low confidence level also indicates noticeable difficulty perceived by the subjects. More than half of the test questions were answered without confidence.

2. When dictionary definitions were provided for consultation, the subjects scored lower (from 6.25 to 5.53) than the phase one test with self lexical recognition. Nevertheless, confidence levels in performing the task show a considerable increase (from 4.3 to 6.48). The performance degradation in phase two seems to indicate that dictionary consultation for *affect* and *influence* does not result in better understanding and seems to bring difficulty to subjects' near-synonym recognition. Yet, the subjects made lexical choices with higher confidence and seemed to not be aware of the newly-created misuse. This reveals a significant problem in the near-synonym self-learning context. A considerable portion of language learners may not be capable of using L2 dictionaries for successful near-synonym distinction and may perceive inaccurate lexical knowledge.
3. The effects of providing collocation information seem to be positive in the near-synonym distinction task. In the phase three test, the subjects became more cautious (5.9 vs. 6.48) on potential lexical misunderstanding but scored better (6.15 vs. 5.53) than the phase two test with dictionaries. Although the test score does not show improvement over the phase one test with subjects' self lexical recognition, the lexical decision was made with higher confidence (5.9 vs. 4.3), indicating the subjects did gain useful information from the collocational profile for distinguishing the near-synonyms.
4. The subjects' perception of the usefulness of additional semantic information seems to be consistent with the test scores. The subjects perceived the lowest usefulness (42.5%) in the distinction task with dictionaries, which were fittingly accompanied by the lowest test score (5.53). More than two-thirds (67.5%) of the subjects perceived the collocational profile as useful information in near-synonym distinction.

The empirical study reveals that language learners do experience difficulty in near-synonym semantic recognition. The problem should be brought to the attention of language teachers and needs to be addressed adequately. The overall test results support the positive effects of collocation information on near-synonym semantic distinction. Both the test scores and the subjects' perception show meaningful enhancement in better understanding of lexical semantics. The positive effects are not as evident in difficult near-synonyms, and this can be logically expected. Some collocation information for distinguishing similar semantics may not be obvious to language learners. This indicates that the positive effects of collocation information on near-synonym semantic recognition may be greatly improved by pedagogical instruction over self-learning for most language learners.

5. Discussion and Conclusion

With both analytical and empirical verification, we show that collocation observation is useful in recognizing semantic features of a word of interest. Syntactic patterns and POS categories provide a structure for anchoring and characterizing the semantic links between collocates and

the target word. The scale of collocate association strength helps distinguish salient semantic features that are conducive to L2 learners' comprehension of the target word. When the target word is a transitive verb, the collocational profile of subject, object, adverb, and adjective collocates with graded association strength serves as an effective instrument in revealing the semantics and improving learners' recognition of the target word. The collocational profile also provides analytical evidence for L2 learners in comparing and discriminating near-synonyms. In self-learning with dictionary consultation, L2 learners are often briefed by abstract definition and left with vague and shallow lexical recognition. Collocational profiles, together with denotational meaning in dictionaries, give a solid conceptual grounding of target word for L2 learners in getting full grasp of the lexical semantics.

In this study, we used VIEW as a concordancing tool, to retrieve collocation information related to the targeted words for investigation. Our position is not to design and develop a new system that outperforms current concordancing tools, such as VIEW and SKETCH ENGINE. Instead, we attempt to point out that there is a gap between L2 learners' proficiency and the powerful investigative functions provided by these concordancing tools. We addressed the problem of how the linguistic resources and the computational functions, as provided by current concordancing tools, can be further built upon to benefit L2 learners.

We proposed a categorized collocational profile with graded association strength to filter and organize salient semantic features. It serves as a guided process to help develop concrete conceptual links such that semantic meaning and unique features of lexical items becomes more easily accessible to L2 learners. The process of constructing collocational profiles that we manually simulated on top of VIEW can be automated by a computer program and can be potentially developed as an online lexical query instrument for L2 learners in pedagogical and self-learning contexts. The development of such a software system, however, is not within the scope of the paper.

Lexical misuse has been a tenacious problem for generations of L2 learners. Most L2 learners are unaware of the subtle semantic distinctions among near-synonyms. The approach we propose can potentially fill in the gap for improving L2 learners' lexical recognition and reducing semantic infelicities. We conducted analytical evaluation to simulate L2 learners' cognitive standpoint and performed the process of deriving insightful semantic information from target words' collocational profiles. We also carried out an empirical evaluation to observe the response from actual language learners and verify that proper use of collocation information leads to learners' successful comprehension of lexical semantics. Both results suggest that the process of organizing and identifying salient semantic features is favorable for and is accessible to a good portion of L2 learners. In addition, pedagogical instruction, as an enhancement to the use of a collocational profile, may benefit an even larger portion of L2 learners.

Acknowledgments

The authors would like to thank Prof. Miao-Hsia Chang of NTNU for her research advice and Prof. Huei-Ling Lai of NCCU for her assistance in questionnaire administration. Anonymous reviewers' comments also helped improve the content of the paper.

Reference

Barron, A. (2003). *Acquisition in interlanguage pragmatics.* Amsterdam: Benjamins.

Burnard, L. (1995). *British National Corpus: User's reference guide for the British National Corpus.* Oxford: Oxford University Computing Service.

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3), 283-299.

Channell, J. (2000). Corpus-based analysis of evaluative lexis. In S. Hunston & G. Thompson (Eds.), *Evaluation in Text* (pp. 38-55). Oxford University Press.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22-29.

Davies, M. (2008a). VIEW. Available online at http://view.byu.edu/.

Davies, M. (2008b). The Corpus of Contemporary American English: 385 million words, 1990-present. Available online at http://www.americancorpus.org.

Dewaele, J. (2008). Appropriateness in foreign language acquisition and use: Some theoretical, methodological and ethical considerations. *International Review of Applied Linguistics in Language Teaching*, 46(3), 245-265.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D., (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, 105-116. Lrient.

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Louw, B. (2000). Contextual prosodic theory: Bring semantic prosodies to life. In C. Heffer, H. Sauntson & G. Fox (Eds.), *Words in Context: A Tribute to John Sinclair on his Retirement.* Birmingham: University of Birmingham.

Mackay, S. (1980). Teaching the syntactic, semantic, and pragmatic dimensions of Verbs. *TESOL Quarterly*, 14, 17-26.

Marin, M. (1984). Advanced vocabulary teaching: The problem of synonyms. *The Modern Language Journal*, 68, 130-137.

Partington, A. (2004). Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9(1), 131-156.

Sinclair, J. (2003). *Corpora for lexicography.* In Sterkenberg, P. Van (Ed.), A Pactical Guide to Lexicography. Amsterdam: John Benjamins.

Stubbs, M. (2001). Text, corpora, and problems of interpretation: a response to Widdowson. *Applied Linguistics*, 22(2), 149-172.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Yeh, Y., Liou, H. C., & Li, Y. H. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, 20(2), 131-152.

Appendix. Questionnaire

Part I. Circle the word which is appropriate for the context of the sentence.

Both “affect” and “influence” are translated as 影響 in Chinese and both are used as **verb** in the following sentences.

1.	How did your past experiences **affect** or **influence** the way you coped with changes?	☐ sure	☐ not sure
2.	Environmental issues continue to **affect** or **influence** us all.	☐ sure	☐ not sure
3.	It's going to **affect** or **influence** the quality of the lives of people in Taipei.	☐ sure	☐ not sure
4.	We believe that the culture and language of individualism **affect** or **influence** these trends.	☐ sure	☐ not sure
5.	Price and easy availability heavily **affect** or **influence** consumers' choices.	☐ sure	☐ not sure
6.	Local networks have more power to **affect** or **influence** public opinion than any other media.	☐ sure	☐ not sure
7.	The amount and type of fat that you eat can **affect** or **influence** the health of your heart.	☐ sure	☐ not sure
8.	A market leader’s actions may greatly **affect** or **influence** the industry structure.	☐ sure	☐ not sure
9.	They could severely **affect** or **influence** the success or failure of the program.	☐ sure	☐ not sure
10.	The party can heavily **affect** or **influence** the political agenda.	☐ sure	☐ not sure

Part II. Given the dictionary definitions of the two words, circle the word which is appropriate for the context of the sentence.

Dictionary 1
affect
1. to act upon; to produce an effect or change upon
2. to influence or move, as the feelings or passions

influence
1. to control or move by power, physical or moral
2. to affect by gentle action, to exert an influence upon

Dictionary 2
affect
- always presupposes a stimulus powerful enough to evoke a response or elicit a reaction
- implies a definite alteration or modification

influence
- always presupposes an agent that moves a person or thing in some way or to some degree from a course, or effects changes in nature, character, or behavior

1.	How did your past experiences **affect** or **influence** the way you coped with changes?	☐ sure	☐ not sure
2.	Environmental issues continue to **affect** or **influence** us all.	☐ sure	☐ not sure
3.	It's going to **affect** or **influence** the quality of the lives of people in Taipei.	☐ sure	☐ not sure
4.	We believe that the culture and language of individualism **affect** or **influence** these trends.	☐ sure	☐ not sure
5.	Price and easy availability heavily **affect** or **influence** consumers' choices.	☐ sure	☐ not sure
6.	Local networks have more power to **affect** or **influence** public opinion than any other media.	☐ sure	☐ not sure
7.	The amount and type of fat that you eat can **affect** or **influence** the health of your heart.	☐ sure	☐ not sure
8.	A market leader's actions may greatly **affect** or **influence** the industry structure.	☐ sure	☐ not sure
9.	They could severely **affect** or **influence** the success or failure of the program.	☐ sure	☐ not sure
10.	The party can heavily **affect** or **influence** the political agenda.	☐ sure	☐ not sure

你是否能解讀 Dictionary 1 進而區別 affect 與 influence? ☐ yes ☐ no

你是否能解讀 Dictionary 2 進而區別 affect 與 influence? ☐ yes ☐ no

Part III.

搭配**主詞**		搭配**副詞**		**Verb**	搭配**受詞**	
相同	不同	相同	不同		相同	不同
factor variable decision policy behavior	disease condition matter change problem	profoundly greatly directly indirectly	**adversely** materially differentially disproportionately	**affect**	life outcome behavior decision policy development	ability performance relationship rights
factor variable decision policy behavior	government ability attitude teacher process	profoundly greatly directly indirectly	unduly deeply strongly significantly	**influence**	life outcome behavior decision policy development	perception attitude direction

Given the collocations of the two words, circle the word which is appropriate for the context of the sentence.

1. How did your past experiences **affect** or **influence** the way you coped with changes? ☐ sure ☐ not sure
2. Environmental issues continue to **affect** or **influence** us all. ☐ sure ☐ not sure
3. It's going to **affect** or **influence** the quality of the lives of people in Taipei. ☐ sure ☐ not sure
4. We believe that the culture and language of individualism **affect** or **influence** these trends. ☐ sure ☐ not sure
5. Price and easy availability heavily **affect** or **influence** consumers' choices. ☐ sure ☐ not sure
6. Local networks have more power to **affect** or **influence** public opinion than any other media. ☐ sure ☐ not sure
7. The amount and type of fat that you eat can **affect** or **influence** the health of your heart. ☐ sure ☐ not sure
8. A market leader's actions may greatly **affect** or **influence** the industry structure. ☐ sure ☐ not sure
9. They could severely **affect** or **influence** the success or failure of the program. ☐ sure ☐ not sure
10. The party can heavily **affect** or **influence** the political agenda. ☐ sure ☐ not sure

搭配詞資訊是否有助於區別 affect 與 influence? ☐ yes ☐ no

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 221-236

A Corpus-based Study on Figurative Language through the Chinese Five Elements and Body Part Terms

Siaw-Fong Chung*

Abstract

Using a corpus-based approach, this paper analyzes figurative language through observing the Chinese five elements (五行) of 金 'metal,' 木 'wood,' 水 'water,' 火 'fire' and 土 'earth.' This work found that there are at least two types of figurative language in Mandarin Chinese – one of which occurs at the morphosyntactic level and the other occurs during the mappings between two domains (between the body part terms and these five elements). When the figurative uses of the co-occurring five elements with body part terms were tested in a psycholinguistic experiment composed of two groups of subjects (non-native and native speakers of Mandarin), a majority of the non-native speakers were unable to comprehend these figurative uses. This study attempts to prove that a linguistically-driven understanding of the five elements will be of great help to teaching or learning figurative language in a Mandarin L2 context.

Keywords: Corpus, Five Elements, Figurative Language, Body Part, Learners of Chinese, Psycholinguistic Experiment.

1. Introduction

The relationship between body part terms and emotion metaphors was discovered by early psychologists, such as William James (1884) and Carl Lange (1884), who suggested that the origin of emotions is inside one's body. Linguists of present days, such as Kovecses (2003) and Wierzbicka (1999), have also examined emotions in English and compared them to those in different languages. In Yu's (1995: 85) inspection of Mandarin metaphorical expressions related to anger and happiness in Chinese, he noted that the "underlying cognitive model based on the fundamental theories of Chinese medicine has led to a cultural emphasis in China of sensitivity to the physiological effects of emotions on the internal organs." Therefore, it holds that Chinese people are aware of the relatedness between the five elements and emotions

* Department of English, National Chengchi University, Taipei, Taiwan
E-mail: sfchung@nccu.edu.tw

[Received August 4, 2009; Revised January 29, 2010; Accepted February 1, 2010]

in Chinese. Our current study is different from Yu's by addressing the following questions.

(1) (a) What are the distributional patterns of the Chinese five elements in corpora data?
(b) To what extent will a corpus-based method help to extract figurative language containing the Chinese five elements?
(c) How will a linguistic analysis contribute to the understanding of figurative language by learners of Mandarin as a second language?

In addition to extracting figurative language, our work aims to explain how a corpus-based method can be used to assist teaching and learning. We intend to see the extent to which corpora and collocational understanding help in extracting these figurative patterns and how these patterns can be applied to teaching and learning of Mandarin to foreigners.

2. The Chinese Five Elements (五行)

Traditional Chinese medicine believes that the five elements also control one's internal body – they "are said to vanquish one another and to produce one another" (Veith, 2002: 19). These elements are also reckoned by philosophers to be phenomena that rule nature. Table 1 provides these resonances (of mapping), according to traditional Chinese beliefs.

From Table 1, it can be seen that the five elements are related to emotions (last column of Table 1) and to body parts (shaded).

Table 1. Five element resonances

自然界						五行	人體				
方位	氣候	發展過程	五色	五味	時令		臟	腑	五官	形體	情志
東	風	生	青	酸	春	木	肝	膽	目	筋	怒
南	暑	長	赤	苦	夏	火	心	小腸	舌	血脈	喜
中	濕	化	黃	甘	長夏	土	脾	胃	口	肌肉	思
西	燥	收	白	辛	秋	金	肺	大腸	鼻	皮	悲
北	寒	藏	黑	鹹	冬	水	腎	膀胱	耳	骨	恐

Hicks, Hicks, and Mole (2004: 28) said that, in Chinese, "emotions create movement and disturbance in a person's *qi*." Yu (1995: 81) has also commented that "[w]herever *qi* is locally impeded, it will affect the circulation of blood and local pain may occur as a result of increased internal pressure in that area" and "[t]his may point to the reason why *qi* is one of the basic words for the emotion of anger." From here, one can see how the Chinese relate emotions to the five elements (Table 1) and to body parts. Yet, despite the traditional beliefs about the five elements and body part terms, we found that the denotation of body parts and emotions may sometimes not be in accordance with our linguistic knowledge, except for some that we can immediately relate based on physiological knowledge. While the connectivity of some pairings (such as that between 火 'fire' (heat) and 心 'heart;' as well as 水 'water'

and 膀胱 'bladder') can be easily explained, many others, such as the combinations of 腎 'kidney' and 水 'water,' as well as 肺 'lung' and 金 'metal,' are not entirely linguistically-driven. Chinese speakers, however, do not seem to find this a problem – that is, they can use 肚子 'stomach' and 水 'water' on the one hand and believe that 腎 'kidney' and 水 'water' are closely related on the other. This discrepancy between world versus linguistic knowledge may be confusing to a learner of Mandarin. Therefore, we hope to provide some insights to explain these apparent 'discrepancies' from a linguistic perspective, further supported by empirical data from corpora and a psycholinguistic experiment[1]. It is also through a metaphor framework (Lakoff & Johnson, 1980; Lakoff, 1999) that we hope to explain the mapped meanings of these five elements when they appear as physical entities (of metal, wood, water, fire, and earth) and as abstract elements.

This study claims that collocational data from corpora can be utilized to raise awareness amongst foreign learners of Mandarin so that patterns in the target language can be recognized. These patterns may cause difficulty for learners both at word formation and at sentential levels. For example, some non-existent associations in English (*e.g.*, 肝 'liver' with 火 'fire' to mean 'irascibility') can be better explained with corpora data[2]. By providing quantitative data, our research can shed light on the differing conceptualizations a foreign learner of Mandarin may need to overcome. The following expresses the methodology used in this work.

3. Methodology and Results for Corpora Analyses

All single- (*e.g.*, 火 'fire') and multiple- (*e.g.*, 肝火 liver-fire') character expressions containing the five elements were extracted from the Academia Sinica Balanced Corpus of Modern Chinese (hereafter Sinica Corpus), shown in Table 2. From Table 2, a total of 25,079 instances were found containing these five elements either as single-character expressions (Column 4) or in multiple-character morphemes (Column 6). Among these, 水 'water' constitutes the biggest proportion, with about 40% of the total number of instances. This may be due to the fact that water has a wide applications of functions – to drink, to wash, to flow, to move, to flood, *etc*., not mentioning its possibilities of combination with different morphemes ranging from aquatic-related attributes (*e.g.*, 水田 'paddy field' and 水產

[1] This observation was made based on the resonances in Table 1 versus the linguistic data observed. This did not include other relations amongst the elements such as the 克 'control' cycle, which may explain some conflict between elements.

[2] Nevertheless, there may also be another level of metaphoricity because 肝火 'liver-fire' can mean both 'bodily heat' and 'irascibility' (in addition to the mapping between a body part term (肝 'liver') and the fire element (火 'fire')). These different levels of mappings, however, are not the focus of the current work. Figurative language was identified and accumulated once a metaphorical meaning was detected (regardless of the level of mappings involved).

'aquatic products') to watery (水水 'juicy') and to other instrumental meanings (*e.g.*, 水貨 'smuggled goods'). As one can see, 土 'earth' constitutes the lowest percentage, with only about 8% of the total instances, suggesting that it perhaps has less frequent applications or may appear in limited, usually soil-related, contexts. The second highest, 金 'metal' (28%), often denotes finance-related terms (金融/基金/資金) and all types of metals (金屬).

Table 2. Number of instances from Sinica Corpus

Elements	Total Instances	%	Single Character	%	Morphemes	%
金 Metal	6,997	27.90	230	3.29	6,767	96.71
木 Wood	3,463	13.81	80	2.31	3,383	97.69
水 Water	9,999	39.87	1,436	14.36	8,563	85.64
火 Fire	2,709	10.80	246	9.08	2,463	90.92
土 Earth	1,911	7.62	149	7.80	1,762	92.20
Total	**25,079**	**100.00**	**2,141**	**8.54**	**22,938**	**91.46**

On the right of Table 2, we can see that, for 水 'water,' about 14% of its instances appear as a single character and this constitutes the highest percentage among all five elements. The other four elements appear as single-character expressions in no more than 9% of their respective total hits. In addition, we found that 木 'wood' rarely appears on its own (2.3%). In order to see the word combinations formed by the five elements, analysis of their positions in an expression was carried out (Table 3).

Table 3. The five elements as morphemes in the Sinica Corpus

Five Elements (E)	Two-characted Expressions				Three-characted Expressions						Total	%
	Initial		Final		Initial		Medial		Final			
	E? (*e.g.*, 金錢 'money')		?E (*e.g.*, 黃金 'gold')		E?? (*e.g.*, 金字塔 'pyramid')		?E? (*e.g.*, 基金會 'foundation')		??E (*e.g.*, 獎學金 'scholarship')			
	Tk.	Ty.	Tk.	Ty.	Tk.	Ty.	Tk.	Ty.	Tk.	Ty.		
金 Metal	**2,225**	146	1,819	73	748	200	1,110	175	506	75	6,408	30.69
木 Wood	**760**	99	502	61	119	43	210	67	113	49	1,704	8.16
水 Water	**3,209**	178	2,594	154	480	137	788	225	227	66	7,298	34.96
火 Fire	**1,029**	70	818	83	227	45	185	58	9	7	2,268	10.86
土 Earth	**1,710**	65	1,105	63	155	44	142	30	88	20	3,200	15.33
Total	8,933	N/A[3]	6,838	N/A	1729	N/A	2,435	N/A	943	N/A	20,878	100.00

In Table 3, the use of the five elements in expressions with two- to three- characters is shown. The number of tokens (Tk.) refers to the instances found, including repeated ones. The number of types (Ty.) refers to the number of varied forms found. From Table 3, we can see

[3] The symbol '?' refers to any Chinese character appearing before and/or after the five elements. 'N/A' because it is uncommon to add up the different types from different elements. The total in Table 3 does not add up to the total hits in Table 2 because we only considered up to three characters.

that, in terms of tokens, all of the five elements appear consistently at the initial position of two-character expressions (in bold). (Note that comparisons can be made only within each element since different elements are shown to have different overall frequency in the corpus.)

In terms of types, the highest numbers of types within each element are shaded. We found that 金 'metal,' 木 'wood,' and 土 'earth' appear with more varied forms at the initial position of the expressions (as in 金牛座 'Taurus,' 木板 'plank' and 土地 'ground')[4]. On the other hand, 火 'fire' appears most often in the final position of two-character expressions such as in 香火 'burning joss stick' and 烈火 'raging fire.' 水 'water' appears most often in the medial position of three-character expressions (*e.g.*, 淡水魚 'freshwater fish' and 排水管 'a drain'). In addition, 火 'fire' seldom appears in the final position in three-character expressions, except in names (*e.g.*, 陳樹火 'Chen Shu-Huo'). In fact, all of the nine instances for '??E' are proper nouns of human names. The analysis in Table 3 will help predict the behavior of the five elements in word formation. A corpus-based study like this can display linguistic phenomena that we seldom notice in daily use. In addition, we also found that, while some of the words retain the physical meanings of the five elements (*e.g.*, 木箱 'wooden chest' and 木材 'lumber'), some show meaning extensions to denote more figurative use such that in 水準 'standard.' As for 土 'earth,' it seems to have different meanings, including soil (紅土 reddish earth), territory (國土 territory), local (土狗 Formosan/local dog), and not fashionable (老土 old-fashioned).

In addition, in order to observe whether or not these five elements also co-occur with body part terms and how they pattern in the corpus, we first selected a list of body part terms as our reference list (given in Table 4).

Table 4. List of body parts (translated from the English Swadesh list)

Body Parts	Gloss	Body Parts	Gloss	Body Parts	Gloss	Body Parts	Gloss	Body Parts	Gloss
皮	skin	嘴/口	mouth	手	hand	背	back	毛/頭(髮)	hair
指甲	fingernail	耳(朵)	ear	舌(頭)	tongue	心(臟)	heart		
眼(睛)	eye	牙 (齒)	teeth	頭/臉/面	head	胸(膛)	breast	腹部/肚子/胃	belly
腸	intestines	頸/脖子	neck	鼻(子)	nose	骨骼	skeleton		
血	blood	膝	knee	身	body	肝	liver		
骨	bone	肉	flesh	腿/腳	leg	足	foot		

We took the English body part terms from the Swadesh list (Swadesh, 1971) because this list constitutes the basic concepts which are claimed to exist in various languages. The Chinese translations were borrowed from the annotations by a research group at Academia

[4] Note that, at this stage, we did not distinguish the literal from the figurative use since distinguishing the figurative from the literal at the morphological level may sometimes introduce extraneous problems. Furthermore, existence of proper names (*e.g.*, 鄭木金, 黃木添, 彭木城, 鍾木郎, *etc*.) may affect the overall results.

Sinica (with expansion by the author). From Table 4, there are forty-two Mandarin body part terms used in this part of the research. The number of co-occurrences of the body part terms with the five elements in ±5 window span (for words, not characters) was recorded (see Table 5). When more than one body part term was found appearing within the designated window size (as in 伸手捧了些清水洗去臉上沙塵 'to stretch and hold some clean water to wash away the dirt on the face'), the body part terms ('hand' and 'face,' in this case) were counted in each category, respectively.

Table 5. Co-occurrences of body part terms with the five elements

Five Elements	Total Instances	Instances with Body part Terms up to ±5	Per 1,000 Instances	Types of Body part Terms
金 Metal	6,997	73	10.43	22
木 Wood	3,463	38	10.97	15
水 Water	9,999	144	14.40	35
火 Fire	2,709	34	12.55	14
土 Earth	1,911	25	13.08	12
Total	**25,079**	**314**	**12.52**	**N/A**

As displayed in Table 5, there are only 314 instances from the total hits in which these body parts were found in the designated contexts of the five elements. This frequency is rather low as there are, on average, only 13 instances of body part terms appearing in every 1,000 instances of the (combined) five elements.

From Table 5, we can see that 水 'water' is the most frequently used element with body part terms compared to the other four elements (14 instances per 1,000 instances). This is followed by 火 'fire' and 土 'earth,' each with 13 instances in every 1,000 instances. Sample sentences for 水 'water' and 火 'fire' are, respectively, 增加肚子裡的『墨水』'increase the ink (knowledge) in one's stomach' and 眼睛幾乎要冒出火 'fire seems to be bursting out from his/her eyes.' These examples show the co-occurrences of the body part terms with the five elements (regardless whether the five elements appear in single- or multiple- character expressions, or whether they are literal or figurative). A non-figurative use of 土 'earth' can be seen in 他滿口滿鼻都是沙土 'his mouth and nose are full of sand.' "Types of Body part Terms" (last column of Table 5) refers to the number of types of body parts found with a particular element. For instance, 水 'water' co-appears with 35 (83.33%) out of the 42 body part terms selected for the analysis of this work, indicating that 水 'water' appears most frequently with body part terms. 金 'metal,' the second most frequent, is found with twenty-two (52.38%) body part terms.

In order to see whether a certain body part term is used particularly frequently with an element, Table 6 lists the types of body part terms co-appearing more than 5% of the time with the five elements. The data are body part, frequency, and percentage. The most frequently occurring body part terms are shaded. For 金 'metal,' we found that 面 'face' is itself a

classifier often used with 金牌 'gold medal,' as in 各摘下一面金牌 'each has won a gold medal.' In this example, 面 'face' in 一面金牌 'a gold medal' has also had metaphorical extension from 'face' to 'surface.' Earlier examples of 土 'earth' have also shown its metaphorical extension to the meanings of local and not fashionable. Therefore, future studies on the metaphorical extension of the five elements would be interesting.

Table 6. Types of body part terms found with the five elements in ±5 window size

金 Metal (Total=73)			木 Wood (Total=38)			水 Water (Total=144)			火 Fire (Total=34)			土 Earth (Total=25)		
面 face	23	31.51	手 hand	9	23.68	口 mouth	29	20.14	心 heart	6	17.65	身 body	4	16.00
身 body	11	15.07	身 body	5	13.16	身 body	20	13.89	身 body	6	17.65	腳 leg	3	12.00
手 hand	6	8.22	頭 head	4	10.53	頭 head	12	8.33	手 hand	4	11.76	頭 head	3	12.00
心 heart	5	6.85	眼 eye	4	10.53	手 hand	9	6.25	眼睛 eye	4	11.76	口 mouth	3	12.00
眼 eye	4	5.48	腳 leg	2	5.26				臉 face	3	8.82	手 hand	3	12.00
			手指 fingernail	2	5.26				肉 meat	3	8.82	鼻 nose	2	8.00
			腿 leg	2	5.26							心 heart	2	8.00
			口 mouth	2	5.26									
			臉 face	2	5.26									

As for 木 'wood,' its most often occurring body part term is 手 'hand' (as in 一隻手揮動著木杖 'with one of his hands waving the wooden stick'), suggesting that 木 'wood' often is used to refer to something that can be held by the hands (thus having a functional use). 水 'water,' on the other hand, often collocates with 口 'mouth,' indicating that these two are often used together. One classic example can be seen in 我不禁吞了一口口水 'I couldn't help but swallow one mouth of saliva (I couldn't help but swallow hard)' in which the first 口 'mouth' is a classifier. As for 火 'fire,' its most frequently appearing body parts are 心 'heart' and 身 'body,' such as 就是秉持這一把心中之火 'it is to adhere to the fire in one's heart' and 抱住身上有火的小孩子 '(someone) is hugging the kid that is on fire,' with the first example used figuratively and the second used literally. 土 'earth' is also frequently used with 身 'body' (*e.g.*, 撣一撣身上的塵土 'to brush away the dust on (one's) body').

From the analysis in Table 6, one can see that certain body parts are more commonly used with a certain element. Their co-occurrences here are mainly driven by cognitive motivations, *i.e.*, one knows that 木 'wood' is handy, 水 'water' is drinkable, 塵土 'dust' can cover one's body, 火 'fire' can burn one's body, *etc*. It is possible that these elements pre-select a certain body part to co-occur with due to the nature of the physical elements. Analysis as such will also provide a good example for presenting cognitive mechanisms through linguistic realizations. From these collocations, linguistic predictions can also be

made. For instance, we predicted a higher percentage of figurative language could possibly be found with 火 'fire' when it co-appears with body part terms such as 心 'heart' and 眼睛 'eye.' We also predicted that 金 'metal,' 木 'wood,' and 土 'earth' would be used less figuratively, based on their most often appearing (body part) collocates being a classifier, hand, and body, respectively, each possessing a relation that is likely to be literal. These predictions were made based on the collocational patterns found in a corpus. Nevertheless, we could not make a solid prediction regarding 水 'water' since its collocates of 口 'mouth,' 身 'body,' 頭 'head,' and 手 'hand' can be used both literally and figuratively. These (linguistic) collocates in Table 6 are obviously different from the resonances of the five elements presented in Table 1 earlier, further confirming that language use and traditional beliefs might be two separate knowledge systems for the Chinese.

4. The Five Elements, Body Part Terms and Figurative Language

This section carries out an analysis of figurative language, calculating the number of co-occurrences of body part terms and the five elements which are non-literal. We used the term 'figurative language' to refer to the above phenomenon of figurative use, focusing particularly on instances where the Chinese five elements co-appear with body part terms, especially when they carry a figurative meaning. Our definitions of figurative language are also in accordance with the following two important features listed by Liu (2008: 23) for idioms (a term he uses to refer generally to figurative language)[5].

> (1) Idioms are often non-literal or semi-literal in meaning – that is, an idiom's meaning is often not completely derivable from the interpretation of its components. (2) They are generally rigid in structure – that is, some of them are completely invariant but others allow some restricted variance in composition... (Liu, 2008: 23)

The linguistic data of our concern are also non-literal (opaque) or semi-literal (semi-opaque). Their meaning cannot be derived completely from their components. Opaque instances including four-character idioms in Chinese such as 冷水澆頭 'to pour cold water on one's head (to discourage).' These four-character idioms were checked against the Ministry of Education's Dictionary of Chinese Idioms (because not all four-character expressions in Mandarin are idioms)[6]. Figurative language concerned in this work is generally rigid in nature but does allow for some restricted variance in composition. For instance, both 眼睛冒金星

[5] Liu listed three features with the last one being "[i]dioms are multiword expressions consisting minimally of two words, including compound words" (pg. 23) which refers mainly to English and is not applicable here.

[6] Available at http://dict.idioms.moe.edu.tw/sort_pho.htm.

'Venus is at view (to be dazed)' and 撞了個滿頭金星亂冒 'Venus appears above one's head due to a collision' are two variant forms of 冒金星 'Venus appears'[7]. Our analyses of figurative language also included similes, which usually appear in the construction 'body part X is like Y,' as in 祂的心像死水 'His heart is like dead (still) water.' In this example, even though 死水 'still water' itself is a personification of the water by giving it a feature of death, we concentrated on the figurative language found between the mappings of body part terms and the five elements. We also included examples in which the relationship between the body part terms and the five element terms is implicit. For instance, in 清肝退火 'to clean up liver and to reduce internal bodily heat,' 火 'fire' has no explicit reference to 肝 'liver' but the implied meaning is 退肝火 'to recede the fire of the liver' (also, 'to cool down').

Based on the above criteria, our final results concerning the figurative uses of the five elements with the body part terms are given in Table 7.

Table 7. Literal and figurative uses of body part terms with the five elements

Five Elements	Literal	Figurative	Total
金 Metal	69 (95%)	4 (5%)	73 (100%)
木 Wood	32 (84%)	6 (16%)	38 (100%)
水 Water	134 (93%)	10 (7%)	144 (100%)
火 Fire	18 (53%)	16 (47%)	34 (100%)
土 Earth	25 (100%)	0 (0%)	25 (100%)
Total	**278 (89%)**	**36 (11%)**	**314** (100%)

Based on the total 314 instances for all five elements (from Table 5), we can see the distributions of literal versus figurative usage in Table 7. From this total, about 89% are literal and only 11% are figurative. Previous work (Chung, 2009: 77) found that about 30% of metaphorical expressions are used in newspapers, and the percentages in Table 7 are obviously lower, except for 火 'fire,' which distributes differently with half (53%) of its instances being figurative and the other half (47%) literal. This displays the possibility that 火 'fire' not only is used more often with body part terms but also that half of its instances in a corpus are likely to be figurative. All of the other four elements pattern the same – with more than 84% of the uses carrying literal meanings. This demonstrates that most of their co-occurrences with body part terms refer to their concrete entities, rather than the abstract elements. Co-occurrences of 火 'fire' with body parts are as along the lines of 他胸中的熱火何等地狂燒 'the hot fire inside his chest is burning crazily' and 坦利一時心嫉如火

[7] The selection of the body part terms is, however, non-arbitrary, a feature shown in most studies on preference selection of collocation. Nevertheless, it is uncertain whether this is due to extralinguistic knowledge caused by Chinese traditional medicine or it is based on a purely linguistically-driven model, as we found counterexamples for a pure extralinguistically-driven model. Therefore, we intend to look into this issue in terms of rigid versus less rigid figurative use.

'Terry's heart was momentarily jealous like fire burning.' We also predicted that 金 'metal,' 木 'wood,' and 土 'earth' would be used less figuratively, based on their most commonly appearing (body part) collocates. The results in Table 7 make clear that, from all the co-occurrences of body part terms with the five elements, only 5% of the instances of 金 'metal' are used figuratively, *i.e.* 心被金錢佔據 'the heart is invaded by money (gold and money,' in which a mapping between 心 'heart' and 金錢 'money' is found through the action of 'invading.' As for 土 'earth,' surprisingly, all of its instances have literal meanings in our corpus (*e.g.*, 哥哥洗去父親滿身的泥土 '(my) brother washed away the soil all over father's body'). An intuitive observation did find instances such as 面如土灰 'a face like grey soil (earth),' but uses such as this were not present in our data. One reason could be that 土 'earth' is not used with body parts but with other aspects of humans, such as 'aspiration' (*e.g.*, 土氣 'to be unrefined in appearance' and 'language' 土話 or 土語 'the local language'). As for 木 'wood,' 16% of its instances are used figuratively and the most commonly seen figurative use is 麻木 'become numb/numbness.' (Even though both 麻 'hemp' and 木 'wood' can refer to a type of crop or plant, respectively, when they are combined, a new meaning of 'being numb' is derived.) Intuitive investigation found examples such as 石木心腸 'a heart as hard as stone and wood' and 心如木石 'a heart like wood and stone,' but these examples were again not found in the data set of the corpus. In the following, four out of the five elements (excluding 土 'earth,' which consists of zero instances of figurative use) that were used figuratively are laid out in Table 8. The words in which the five elements were found are displayed in the first row of each element. The second row of each element shows the body part terms used with these four elements to form figurative language.[8]

From Table 8, one can see the most commonly found figurative language for all four elements. The results differ slightly from those in Table 6. In 金 'metal,' no particular pattern is displayed, as all instances were sparsely found. For 木 'wood,' 麻木 'being numb' and 手指 'fingernail' are highlighted to be the most frequent in their respective cells. For 水 'water,' its appearance as a single word is used most commonly in the figurative sense, while the corresponding body part terms are 肚子 'belly' and 心 'heart' (*e.g.*, 他們有一肚子的苦水 'they have one full stomach of bitter water (complaints) and 江水像跳動的心臟般 'the river water is pumping like the heart'). As for 火 'fire,' it is most frequently used in a figurative sense as a single word, followed by 退火 'recede fire.' The corresponding body part terms for 火 'fire' are 心 'heart' and 眼睛 'eyes.' If one contrasts this table with Table 6, one can produce several observations which are important for the learning of Mandarin, for example, when 金 'metal' co-appears with 面 'face' (see Table 6), it is likely to be used

[8] By listing them this way, the table by no means shows that any items from the first row can be freely combined with the items in the second row. The table merely provides a calculation of the expressions found.

literally because it is not found in Table 8, which consists of figurative use and when 木 'wood' co-appears with 手 'hand,' it usually refers to the physical property of wood, which is handy (literal). Conversely, when 木 'wood' appears in a figurative use, it is more like to denote 麻木 'numbness,' and this corresponds also to the limbs.

Table 8. Figurative use of the four elements and their body part terms

金 Metal (Total=4)		**木 Wood (Total=6)**	
金錢 'money' (1)	貼金 'paste-gold' (1)	麻木 'numb' (3)	木 'wood' (1)
金星 'Venus' (1)	金星亂冒 'to see stars' (1)	麻木感 'numbness' (1)	樹木 'trees' (1)
心 'heart' (1)	頭 'head' (1)	手指 'fingernail' (2)	心 'heart' (1)
臉 'face' (1)	眼睛 'eye' (1)	手 'hand' (1)	身 'body' (1)
		腳 'leg' (1)	
水 Water (Total=10)		**火 Fire (Total=16)**	
水 'water' (3)	墨水 'ink' (1)	火 'fire' (5)	熱火 'hot fire' (1)
水流 'water-flowing'(1)	冷水 'cold water' (1)	退火 'recede-fire'(3)	火焰 'flames' (1)
江水 'river water' (1)	淚水 'tear' (1)	火苗 'flames' (1)	火燒 'fire-burn' (1)
苦水 'bitter water' (1)		肝火 'liver-fire' (1)	火光 'fire-light' (1)
止水 'still water' (1)		怒火 'anger-fire' (1)	火氣 'internal bodily heat' (1)
肚(子) 'belly' (4)	頭 'head' (1)	心 'heart' (6)	肝 'liver' (1)
心 'heart' (3)	肚 'belly' (1)	眼睛 'eye' (4)	胸 'breast' (1)
心臟 'heart' (1)		臉 'face' (2)	嘴唇 'lip' (1)
		毛 'hair' (1)	

When 水 'water' co-appears with 口 'mouth,' a literal meaning is usually derived. When it co-appears with 肚子 'belly' and 心 'heart,' it is likely to be figurative. When 火 'fire' co-appears with 身 'body,' it is likely to refer to the physical 'fire' (Table 6). When it co-appears with 心 'heart' and 眼睛 'eye,' it usually refers to the figurative anger. Finally, when 土 'earth' co-appears with 身 'body,' it is likely to be literal. It is never used in a figurative sense.

If we examine Tables 6 and 8 against Table 1 in terms of the resonances of the five elements, only 火 'fire' and 心 'heart' seem to show consistent co-appearance both as the resonant and in linguistic terms. There are also some occurrences of 土 'earth' and 口 'mouth,' as well as 木 'wood' and 眼 'eye'[9]. Hence, overall, some 'conflicting' use of body part terms seems to be found co-occurring with the five elements in real language and in the resonances of the five elements. Without a proper explanation differentiating the extralinguistic and linguistic knowledge to second learners of Chinese, they are likely to be confused if they happen to read something about the five elements in their learning process. A corpus-based study like the current one will help distinguish the cultural phenomena from the

[9] Nevertheless, the form 目 'eye' was not collected in our body part list. When we searched for this term manually in the same window size of 木 'wood,' zero results were found.

linguistic ones. Furthermore, a corpus-based study will also help discover characteristics that are often implicit in the language. For a second language learner of Mandarin, these implicit uses can be made clearer if their linguistic patterns are displayed, as shown in this work. In addition to being able to predict language usage, this study has also found that there are at least two types of figurative language in Mandarin Chinese – namely, those occurring at the morphosyntactic level and those occurring during the mappings between two domains (the body part terms and the five elements). At both levels, we found mappings from the concrete meaning of the five elements to their less concrete meaning, although there might be one or more levels of abstractness involved. Our analyses also show that figurative language in Chinese involves complex domain mappings, which can prompt discussion regarding the theoretical issues related to metaphor mappings.

5. Figurative Language and Foreign Learners of Mandarin

In order to examine the understanding of figurative language by native and non-native speakers of Mandarin, we conducted a psycholinguistic experiment based on a translation task. In this task, we asked both (foreign) learners and native speakers of Mandarin to translate from Mandarin to English some figurative sentences containing the five elements and the body parts. Only subjects who truly understood the figurative meanings would be able to translate these sentences. A questionnaire was created for this purpose, with examples taken or modified from the Sinica Corpus. Subjects were asked to translate the Mandarin sentences in (2) into fluent English. All of the keywords are highlighted in (2) but were not highlighted in the questionnaire. All subjects were told not to refer to dictionaries while answering[10].

(2)
(a) 老李最擅長的就是往自己臉上貼金了。
(b) 整個嘴唇因爲休息太少而火氣上升腫了起來。
(c) 肝火旺盛會導致口乾舌燥。
(d) 這種中藥吃了之後退火顧眼睛。
(e) 廣泛閱讀可以增加肚子裡的墨水。
(f) 一直坐在電腦桌前，容易造成四肢麻木。
(g) 他們在這場比賽中輸得灰頭土臉。
(h) 上一場失敗的戀愛後，小華心如止水。
(i) 她找朋友吐了一肚子的苦水。
(j) 興奮的他頭上被澆了一盆冷水。
(k) 心裡愛的火苗一下子滅了。
(l) 水深火熱
(m) 一頭霧水
(n) 大動肝火
(o) 眼冒金星

Only six non-native (NN) speakers of Mandarin were recruited, and all of them were advanced Mandarin learners at National Chengchi University (average age=29.5). Their

[10] All sentences except for (2(l)) contain at least one body part term and one of the five elements.

answers were contrasted with the answers provided by six other native (N) speakers of Mandarin (average age=30). The two groups of subjects differed in their language proficiencies based on a scale of 1 to 7, with 1 being least fluent and 7 being most fluent, in Mandarin (NN=5.0; N=6.8) and in English (NN=6.5; N=5.5)[11]. We hypothesized that the native speakers of Mandarin would fully understand the figurative stimuli in (2) and would express their meanings in English adequately. The non-native speakers, however, would only partially understand these figurative uses and as a result, their answers might differ slightly from the original meanings of the stimuli. These hypotheses were tested in terms of how many out of the total six subjects in the respective group answered adequately according to the figurative meaning of each stimulus. By 'adequate answer,' we referred to cases when an English translation fully expressed the figurative meaning of the idioms, even if the target words might not be directly translated, as in (3a) and (4a). Inadequate answers were such as (3b), (4b) and (4c).

(3) 老李最擅長的就是往自己臉上貼金了。
(a) "Old Lee's expertise is flattering himself." (S5, N)
(b) "Lao Li's habit is to pretend to be rich." (S4, NN)

(4) 上一場失敗的戀愛後，小華心如止水。
(a) "After the last love disappointment, Xiaohua's heart is like a still water." (S2, NN)
(b) "After the failure of the last relationship, her heart feels like running water." (S4, NN)
(c) "After the crazy love affair ends, the heart bleeds water." (S1, NN)

Missing translation in other parts of the sentences which did not affect the understanding of the target words was acceptable (as in (5a)). Scores were either one or zero. Unanswered or missing information in any of the target words (*e.g.*, 火氣上升 in (5b) (translated as 'swollen lips')) was considered inadequate; thus, an answer falling into this category would be accorded a zero score. In some cases, over interpretation (5c) occurred, and these cases were also considered inadequate.

(5) (a) 廣泛閱讀可以增加肚子裡的墨水。
"Broad range literature can increase one's knowledge." (S4, NN)
(b) 整個嘴唇因爲休息太少而火氣上升腫了起來。
"My lips are entirely swollen due to lack of rest." (S5, NN)

[11] Even though we tried to recruit more non-native speakers with differing countries of origin, most of the subjects were unable to answer the questions, as they found the task difficult. This further indicates that figurative expressions in a foreign or second language deserve further research.

(c) 一直坐在電腦桌前，容易造成四肢麻木。
"If you sit in front of the computer all day, you will become unfit and begin to progress your body shape in a horizontal manner." (S5, NN)

All of the answers were then marked as adequate or inadequate (one for adequate and zero for inadequate). The results showed that the non-native speaker group only obtained a 41% (SD=20.77) score for adequate answers. The native speaker group, in contrast, obtained high performance with 92% (SD=8.61) of adequate answers, indicating almost all the answers were correct. Nonetheless, a higher standard deviation value (SD) for the non-native speaker group means that the subjects' answers in this group varied greatly compared to those given by the native speaker group. When tested using a Mann-Whitney test, significance was found, $U(28)=0.00$, $p<.05$, suggesting that the two groups differed significantly from each other. From the experiment, we found that most non-native speakers had problems with the following stimuli in (6), as each had only one adequate answer (16.67%) from the six subjects who participated. Some of these stimuli were left empty.

(6) (a) 整個嘴唇因爲休息太少而火氣上升腫了起來。(b)
"She always talks and that's why she's getting angry so easily." (S3, NN)
(b) 這種中藥吃了之後退火顧眼睛。(d)
"Eating this traditional Chinese medicine will help healing conjunctivitis." (S5, NN)
"Having eaten herbal medicine, your eyes will feel as if it suddenly can see clearly again." (S6, NN)
(c) 大動肝火 (n)
"Being quick to reacting to emotions leads to distress." (S5, NN)
"He has the guts to take a risk." (S1, NN)
(d) 眼冒金星 (l)
"reach for the stars." (S3, NN)
"eye twinkling." (S4, NN)

From this experiment, we found that the figurative language studied in this paper is indeed difficult for non-native speakers of Mandarin. For instance, the translated meaning for 水深火熱 'predicament' cannot be formulated based on any conceptual metaphors. Sometimes, even though both Mandarin and English sentences may possess a similar literal meaning, the translated English sentences may become a different or sometimes novel use with meanings 'forced or borrowed' from the translated source language such that in 'head stuck in the clouds' (S4, NN) for 一頭霧水, which, albeit being analyzable (to mean 'daydreaming' or 'not thinking realistically' in English), does not have equivalent meanings in the source and target languages. That is, when a body part is not understood in the same way

in a different language, it is very hard for a foreign or second language learner to master the meanings in the target language. Therefore, analyzing the similarities and differences between any two languages is important as the learning of metaphors not only involves learning new vocabulary but also learning a different culture. Since learners do not know many of the opaque or semi-opaque meanings of these figurative expressions, by understanding the relationships between the five elements and the body part terms, learners are likely to improve in their ability at guessing the figurative meanings of these uses.

6. Conclusion and Future Work

Our paper proposes a criteria-based method to identify figurative language through observing co-occurrences of body part terms with the Chinese five elements. The research questions of this work were answered based on a detailed analysis of the five elements and their appearances in a corpus either as a single-character expression or as a morpheme. Our study also finds results regarding figurativeness in word formation and that metaphors may occur at units as small as morphemes. The findings of this work also show the different uses of the five elements – these five elements are not treated equally when formulating figurative language. For instance, we found that, in the Sinica Corpus, 水 'water' is the most frequently occurring element compared to the other four elements. When examined with body part terms, however, the element of 火 'fire' stands out and also comprises the highest percentage of figurative usage. Additionally, the research herein also shows that a corpus can be of great help to language learners, as it presents linguistic data in the form of statistics to them. A corpus-based study is also able to present distributions of collocated data, through which we predict the possible occurrences of literal versus figurative usage. Through a psycholinguistic experiment, we found that linguistic analysis is needed in teaching and learning of Mandarin since figurative language constantly causes great difficulty to learners of Mandarin.

Since our study also finds results regarding occurrences of figurativeness in word formation, for future research, we intend to analyze figurativeness at the morphosyntactic level, as we found that there are many uses of 木 'wood' in the sense of 'stupidity' (*e.g.*, 木頭木腦 'one without expression' and 木頭人 'a blockhead'). In addition, 木舌 'a tongue that is made of wood' is also used to mean 'someone who is silent.' These examples may be low in frequency and, therefore, not collected in the corpus we used. Another explanation for this may be attributed to their denotation of negative meanings (usually used to mock people). These uses are considered improper or impolite, resulting in lower production both in speech and in writing.

For future work, the hypotheses regarding the ease and difficulty of learning certain body part metaphors will be further tested. Further studies can also focus on extending the corpus to the World Wide Web in order to find the linguistic phenomena outside the precompiled corpus.

An extension of this work can examine the relation of body parts and the five elements in English. The English phrase ‘my heart is on fire’ seems to differ in meaning from its Chinese equivalent (’to be angry’). Therefore, a cross-cultural investigation is also feasible. In addition, this paper finds ambiguity with regard to translating the Chinese 金 to ‘gold’ or ‘metal’ and 土 to ‘soil’ or ‘earth’ in some phrases. It would, therefore, be interesting to see how English translation deals with this ambiguity and how this can become useful to studies in machine translation. The paper is also able to pinpoint the existence of traditional Chinese concepts in Mandarin and how they can be contrasted with linguistic data for the purpose of computer-assisted language learning.

Acknowledgments

This research was supported in part by National Science Council under Grant NSC 97-2410-H-004-001-. Acknowledgements also go to Professor Chu-Ren Huang and the reviewers for their comments on the previous versions of this work.

References

Chung, S.-F. (2009). *A Corpus-driven Approach to Source Domain Determination. Language and Linguistics Monograph Series*. Nankang, Academia Sinica.

Hicks, A, Hicks, J. & Mole, P. (2004). *Five Element Constitutional Acupuncture*. Edinburgh & New York, Churchill Livingstone.

James, W. (1884). What is Emotion?. *Mind*, ix, 188-205.

Kovecses, Z. (2003). *Metaphor and Emotion. Language, Culture and Body in Human Feeling*. New York, Cambridge University Press.

Lakoff, G. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York, Basic Book.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*., Chicago & London, The University of Chicago Press.

Lange, C. (1887). *Ueber Gemuthsbewgungen*, 3(8).

Liu, D. (2008). *Idioms: Description, Comprehension, Acquisition, and Pedagogy*. New York & London, Routledge.

Swadesh, M. (1971). *The Origin and Diversification of Language*. Edited post mortem by Joel Sherzer. Chicago, Aldine.

Veith, Z. (Translator). (2002). *The Yellow Emperor’s Classic of Internal Medicine*. Berkeley, Los Angeles & London, University of California Press.

Wierzbicka, A. (1999). *Emotions across Languages and Cultures: Diversity and Universals*. New York, Cambridge University Press.

Yu, N. (1995). Metaphorical Expressions of Anger and Happiness in English and Chinese. *Metaphor and Symbol Activity*. 10(2), 59-92.

The Association for Computational Linguistics and Chinese Language Processing

(new members welcome)

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 invites submission of original research papers in the area of computational linguistics and Chinese speech and language processing. All papers must be written in English. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, Chinese speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The accepted papers are divided into the categories of regular papers, short paper, and survey papers. There is no strict length limitation on the regular papers but it is suggested that manuscripts not exceed 40 double-spaced A4 pages. 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/public.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.14 no.2 June 2009

Contents

Special Issue Articles:
Computer Assisted Language Learning

Papers

Short Papers

Regular Issue Articles:

Short Papers

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