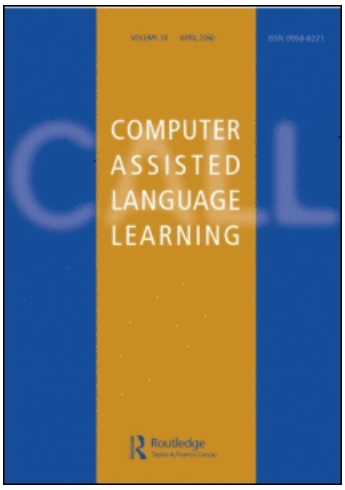

Computer Assisted Language Learning
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100697

An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology

Yu-Chia Chang [a]; Jason S. Chang [a]; Hao-Jan Chen [b]; Hsien-Chin Liou [c]
[a] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan [b] Department of English, National Taiwan Normal University, Taipei, Taiwan [c] Department of Foreign Languages and Literature, National Tsing Hua University, Hsinchu, Taiwan

Online Publication Date: 01 July 2008

To cite this Article Chang, Yu-Chia, Chang, Jason S., Chen, Hao-Jan and Liou, Hsien-Chin(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

To link to this Article: DOI: 10.1080/09588220802090337

URL: http://dx.doi.org/10.1080/09588220802090337

Computer Assisted Language Learning
Vol. 21, No. 3, July 2008, 283–299

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology

Yu-Chia Chang[a]*, Jason S. Chang[a], Hao-Jan Chen[b] and Hsien-Chin Liou[c]

[a]*Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan,* [b]*Department of English, National Taiwan Normal University, Taipei, Taiwan;* [c]*Department of Foreign Languages and Literature, National Tsing Hua University, Hsinchu, Taiwan*

Previous work in the literature reveals that EFL learners were deficient in collocations that are a hallmark of near native fluency in learner's writing. Among different types of collocations, the verb-noun (V-N) one was found to be particularly difficult to master, and learners' first language was also found to heavily influence their collocation production. In this paper, we develop an online collocation aid for EFL writers in Taiwan, aiming at detecting and correcting of learners' miscollocations attributable to L1 interference. Relevant correct collocation as feedback messages is suggested according to the translation equivalents between learner's L1 and L2. The system utilizes natural language processing (NLP) techniques to segment sentences in order to extract V-N collocations in given texts, and to derive a list of candidate English verbs that share the same Chinese translations via consulting electronic bilingual dictionaries. After combining nouns with these derived candidate verbs as V-N pairs, the system makes use of a reference corpus to exclude the inappropriate V-N pairs and single out the proper collocations. The system can effectively pinpoint the miscollocations and provide the learner with adequate collocations that the learner intends to write but misuses. It is hoped that this online assistant can facilitate EFL learner-writers' collocation use and help them transfer this essential knowledge to their future writing.

Keywords: miscollocation analysis; L1 influence; automatic error correction

Introduction

The notion of collocation has increasingly drawn researchers' attention in the field of second language learning as more scholars noted that 'language knowledge is collocational knowledge' (Nation, 2001, p. 318) by nature. Nation (2001) also argued that 'all fluent and appropriate language use required collocational knowledge' (p. 318), and native-like fluency would be enhanced through the increase in collocational competence. In addition, Conzett (2000) reported that collocational knowledge would help productive and accurate use of vocabulary. Collocational knowledge thus plays a crucial role in promoting target language proficiency.

Nattinger and DeCarrico (1992) defined collocation as 'a string of specific lexical items that co-occur with mutual expectancy' (p. 36) and the meaning of a collocation can be

*Corresponding author. Email: richtrf@gmail.com

ISSN 0958-8221 print/ISSN 1744-3210 online

DOI: 10.1080/09588220802090337
http://www.informaworld.com

inferred from the component parts. However, a collocation would be regarded as less appropriate when one of its components is replaced by another word even though collocations are less fixed than idioms. Due to its dual qualities of flexibility and restriction, collocation use becomes a prominent problem to EFL writers. Farghal and Obiedat (1995) examined EFL students' collocation performance and found that EFL learners were incompetent in collocations. In Nesselhauf's exploratory study (2003), advanced German learners of English were found to have difficulty in using proper collocations in free writing.

Examining some of Taiwanese EFL learners' miscollocations (i.e. misused collocations) in their English writing, we observed learners' misused verb collocate and the suggested correction tend to share the same translation in Chinese, such as the suggested verb 'increase' in a learner's sentence 'add more knowledge by using computers' (Liu, 2002). One possible reason to cause learners to use such collocation might be the influence of their first language (L1) (Liu, 2002). While choosing the collocate for the word 'knowledge' a learner might call the Chinese verb *zeng1 jia1* (增加, 'to become greater in quantity' in Chinese) into mind first. Learners under this circumstance might translate the phrase literally and thus have such non-native expression. Based on this assumption, our goal is to build a system capable of automatically detecting and correcting learners' misused collocations with the help from natural language processing (NLP) technology. The integration of NLP technology provides a possible way to assist learner's miscollocation correction and facilitate their self-learning.

In this paper, we describe the development of an online miscollocation aid, *Writing Assistant*, which is integrated with NLP technology and embedded in the website 'Writing Area' of the CANDLE[1] project. It is hoped that the automatic detection tool for miscollocations will help learners acquire appropriate collocations and transfer this incremental collocational knowledge into their writing. The unique feature of the *Writing Assistant* is that the system is built theoretically on first language (L1) interference concluded from previous literature on Taiwanese learners' miscollocation uses. And subsequently the more appropriate collocations are suggested according to the shared translations between Chinese (L1) and English (L2). Furthermore, it is significant that our system mainly uses data from a reference corpus (i.e. the standard corpus containing the correct entries that can serve as a criterion) rather than a learner corpus. Typical methods existing, in contrast, often utilized the errors manually tagged by teachers or specialists in learners' texts (e.g. Shei & Pain, 2000; Liu, 2002). After learners' error patterns are collected in an error library, those systems identify errors via the method of pattern matching and offer all the pre-stored suggestions. These typical methods are rather time-consuming and not easily manageable. Our approach of using the reference corpus is supposed to be far more maintainable instead.

This paper is organized into three sections:

(1) a review of some studies and computer programs of collocation in second language learning specially developed for miscollocation detection;
(2) a report of the testing of a hypothesis of L1 interference upon learners' miscollocations, and explanations for the function and the overall design of the program; and
(3) the experiment and evaluation results along with a conclusion that summarizes the system's advantages, limitations and future work.

Literature review

The term 'collocation' was first introduced by Firth (1957) and is often used to refer to word combinations such as 'make a decision' or 'endure the pain'. Lewis (1997) regarded collocation as 'combinations or words which occur naturally with greater than random frequency' (p. 8). Nation (2001) also considered collocations as 'items which frequently occur together and have some degree of semantic unpredictability' (p. 317), and he argued that knowing a word requires some expectation of the words to collocate with. Therefore, learners who lack collocational knowledge and without full awareness of the importance of collocation will fail to produce adequate word usages. Collocations thus reflect the speaker's fluency of a language, and serve as a hallmark of near native-like language capability.

Collocations are a lexical phenomenon of word combination occurring together relatively more often than other combinations. However, collocations tend to be lexicalized and have a seemingly more limited meaning than the literal interpretation (Justerson & Katz, 1995). The words in a collocation may be appearing adjacently with one another (rigid collocation) or be separated by other words (flexible/elastic collocations). Based on Benson, Benson and Ilson's definition (1986), collocations can be classified into lexical and grammatical ones. Lexical collocations are composed of content words, whereas the grammatical collocations are formed by a content word and an accompanying function word or syntactic structure.

Collocation in second language learning

From a language learning point of view, researchers proposed that collocational competence is one of the factors contributing to the differences between language competence of native speakers and that of second language learners (e.g. Shei, 2000). Improving students' collocational competence can help EFL learners cross their 'intermediate plateau' and become advanced learners (Lewis, 2000). Zhang (1993) conducted empirical studies and reported that the more proficient ESL writers used collocations more accurately and produced collocations more diversely than the less proficient writers did. Liu (2000) also found that good EFL writers of Chinese speakers in Taiwan used collocations more properly and frequently than poor EFL writers. Therefore, it is agreed that equipping students with needed collocational knowledge would enable them to achieve native-like fluency and accuracy in language production.

Nesselhauf (2003) analyzed 32 essays written by advanced German learners of English in terms of their use of verb-noun (V-N) collocation. Among all the miscollocation types, the one occurring most frequently is the wrong choice of verbs (24/65). Also Nesselhauf observed the great influence of learners' L1 on verb-noun collocations, and she noticed that non-congruent word combinations attributed to learners' L1 and L2 are far more difficult for learners to acquire. She suggested that the teaching of verb-noun collocations should focus on the verbs and that 'some semantically possible but collocationally impossible combinations with verbs should be highlighted, especially those that are possible in the learners' L1' (p. 239).

Moreover, based on the analysis of miscollocations in Taiwan EFL learners' writing, it was found that the V-N collocation is the most prevalent lexical collocation error (Liu, 1999b, 2002). Liu (2002) examined Taiwan learners' essays collected in the English Taiwan Learner Corpus (English TLC) from a web-based writing environment. She indicated that 87% of lexical miscollocations (233/265) were attributed to the misuse of V-N collocation; the result further revealed that 96% (224/233) of these V-N

miscollocations were due to misuse of verb collocates. As for the reasons behind the use of miscollocations, 56% (131/233) of these missed V-N collocations are semantically related with respect to their lexical relations in *WordNet* such as synonym, hypernym, troponym, and 38% (88/233) of them are traceable to L1 interference. It can be concluded that the V-N collocation is found to be particularly difficult for learners to acquire, and the misuse of verb collocate is worthy further investigation. In addition, learners' L1 can be a vital factor which interferes with their use of collocations.

Computer programs developed for collocation learning

Since V-N miscollocation has been indicated as the most prevalent collocation error in learner's writing, some researchers had put effort into investigating the topic by attempting to develop some effective writing platforms (e.g. Shei & Pain, 2000; Liu, 2002). These systems aim at designing a writing tool which automatically detects the errors that a learner produces; they help learners correct the misused words or provide learners with suggestions.

With a conceptual design, Shei and Pain (2000) employed a learner corpus and a reference corpus in order to facilitate error detection. The learner corpus was analyzed in order to collect error patterns used to detect identical miscollocations in input text. On the other hand, the reference corpus was used to extract correct collocations so as to exclude students' correct usages. Integrating the benefits from these two corpora, they designed a *Collocation Tutor* to systematically recognize the miscollocations. With the architecture of four deliberate procedures, Collocation Library, Error Library, Lexicon of Synonyms and Dictionary of Definitions, the system was designed to point out learners' potential collocational errors and allowed learners to observe the usages in authentic language through corpora.

Examining misused verb collocates extracted from a learner corpus, Liu (2002) proposed two possible applications to improve learners' collocation competence. One is a *Lexical Assistant*, which was implemented to help learners find the correct verb candidates for revising a given miscollocation. The other application, named as a *Grammar Checker*, was designed as a writing platform to promptly detect and correct learners' common errors (which includes miscollocation) in their essays. To implement her proposed system, she collected learners' errors – corrected manually by teachers – and categorized them as error patterns from English TLC. Using these error patterns enables the checker to detect those presorted misuses matched in learners' texts, including V-N miscollocations as well as other grammar errors.

Within the system proposed by Liu, she emphasized the semantic relation established in *WordNet* as her core theory which accounts for 51% of the miscollocations learners produce. However, she excluded L1 influence (38%) and other potential causes while designing the system. For precise detection, she further employed the collected errors from learner corpus to develop an error library. As long as the learner produces a similar inadequate usage, the system will promptly detect it as an inappropriate word choice. Nevertheless, the learners' corpus is not always ideally sufficient and manageable. To obtain the data and to sustain the whole system would definitely be labor-intensive and time-consuming.

In light of Liu's analysis, L1 influence which was categorized as a semantically unrelated cause merely makes up 38% of the miscollocations in English TLC. However, we observed that her semantic-related analysis (i.e. synonym, hypernym, and troponym) can also be intuitively accounted by the interference of L1. Since these semantic relations in fact also cause the verbs to share the same translation in L1, for instance, '*abide the

pain' and 'endure the pain' in Liu's English TLC corpus. Despite the fact that the misused 'abide' and correct 'endure' are semantically related as synonyms in WordNet, we can also attribute them to L1 interference for their shared Chinese translation *ren2 shou4* (忍受). We hypothesize this L1 interference might occupy an even larger portion in miscollocations, and it deserves more inspection in order to gain utmost precision of error detection. We will justify it in more detail in the later section.

So far there are few online systems coping with miscollocations ascribed to L1 interference, which in fact heavily influences learners' misuses of collocations. This study attempts to develop a web-based miscollocation-detection system as a writing aid and specially tackles learners' miscollocations attributable to the influence of L1. In addition, since most students' miscollocations are categorized into V-N collocation, it makes sense to focus on V-N collocation as the first step for the program development. Other types of miscollocations will be dealt with in the future.

System design

Hypothesis testing: L1 interference and de-lexicalized verbs

The definition of 'L1 interference' in this study is that learners may incorrectly use a certain collocation due to its corresponding translation in L1. Therefore, in English the learners' improper verb collocate and the correct counterpart may share identical translations in Chinese. Theoretically speaking, the definition coincides with two concepts related to L1 interference: one is **split category** and the other is **direct translation**.

The notion of the split category was proposed by Stockwell, Bowen and Martin (1965), which describes a situation when two words in one language are covered by only one term in another language. For instance, an inadequate collocate 'create' and its suggested one 'compose' in '*create a song' and 'compose a song' own an overlapping Chinese translation *chuang4 zuo4* (創作). Because of this shared translation between 'create' and 'compose', Taiwanese learners who incline to develop ideas according to their mother tongue would be affected by these L1 translation equivalents. As a result, learners who lack a full understanding of collocational restrictions would probably resort to L1 translation and thus produce an inappropriate V-N collocation, like '*create a song'.

On the other hand, direct translation refers to a situation where learners misuse collocations because they think in Chinese first and translate their ideas into English directly. For example, learners produce '*write homework' instead of 'do homework' because in Chinese it is verb collocate *xie3* (寫, 'write') co-occurring with *gong1 ke4* (功課, 'homework'). However, when *xie* is translated into English, its English equivalent should be 'do' rather than 'write'. Chinese learners who were trapped in Chinese translation would easily misuse collocations.

In order to verify our hypothesis that shared translations in learners' L1 can greatly affect their use of collocations, the verb pairs (i.e. inappropriate verb collocate paired with the correct one) found in Liu's study (2002) were examined to see whether miscollocations can be mostly traceable to L1 interference. After the translation matching through electronic bilingual dictionaries, the result supports that at least 84% out of the total misused verb collocates do have Chinese translations in common with those correct verb collocates (as shown in Table 1).

Both the semantically related (i.e. the relation of synonym, troponym, and hypernym) and unrelated verb pairs categorized by Liu indeed support our hypothesis. In terms of those verb pairs holding semantic relation, we observed our hypothesis occupies a great portion (92%, 121/131), which makes it plausible to generally ascribe these semantically

Table 1. The L1 interference hypothesis on Liu's miscollocation list.

Relation types categorized in Liu's analysis	Count	Verb pairs that share the same Chinese translation
Synonym relation	71	94% (67/71)
Troponym relation	33	88% (29/33)
Hypernym relation	27	93% (25/27)
No relation	88	72% (63/88)
Total	219[†]	84% (184/219)

[†]The reason the number 219 is not consistent with the one 233 mentioned earlier is because some of the miscollocations are duplicate and some of them may correspond to more than one suggestion (e.g. the manual correction for the misused collocation '*abide the pain' includes the suggested verb collocates 'endure' and 'tolerate').

related misuses to L1 interference. Moreover, while examining the rest of verb pairs sharing no identical translations, it reveals that more pairs can actually be matched if we take those with similar translations into consideration. For example, 'stir up' and 'arouse' can also be considered if we can relate them via their non-identical but similar translations in dictionaries.

Further analysis of those verb pairs failing to find matched translations indicates the misses mainly result from direct translation. Although direct translation can be categorized as the result of L1 interference, some verb pairs with such relations can hardly find the translation equivalents, for instance, the suggestion 'develop' in a learner's sentence '*establish their friendship by sharing activities'. The solution suggested might be the help of translations extracted from bilingual corpora, which will assist to cement the relation between 'develop' and 'establish'. Another small potion of misses is related to the misconceptions about de-lexicalized verb uses, such as *make, take, keep* and other light verbs. Students may regard these verbs as de-lexicalized ones and substitute for one another freely (Liu, 1999). These light verbs tend to have unfixed meanings, especially as the element of multi-word mini-idioms such as 'take your time' or 'have a ball'. It also causes a problem of proving our hypothesis.

Nevertheless, more than 84% of total verb pairs prove to share the translation equivalents, and most miscollocations can be accounted for by the concept of split category and direct translation. The result reveals that L1-accountable errors did occupy a larger proportion of miscollocations in Taiwan local learners' production; it is thus believed that L1 interference is a crucial factor causing learners' miscollocations. Our hypothesis not only provides a general description of these misuses but also facilitates the automatic computation. This hypothesis thus assists computationally correction of the misused collocations that are pervasive in learners' texts. With this concept in mind, we then design the prototype of our system based on this proposed hypothesis. The detailed function will be described in the following section.

Writing assistant

The system proposed automatically extracts the target collocations from users' texts, detects any potential misuses, and provides the adequate collocations in question as suggestions. For example, after a learner inputs a composition to the system, all verb collocates identified within V-N collocations are highlighted. The system detects potential V-N miscollocations in the learner's composition, and then supplies better V-N collocations in terms of learners' problematic V-N collocation. For instance, when a

learner produces an unacceptable V-N collocation, 'eat pills', the system would thus uncover the misused collocation and display the misused verb collocate *eat* in a blue color. As long as the learner moves the cursor to these highlighted words, the system would provide stronger V-N collocations with their Chinese equivalents in a pop-up prompt-box, such as 'take pills' (Figure 1).

In this system, after a given collocation is identified as a misuse, such suspected misuse will undergo the correction to provide the most appropriate suggestions for learners. Here, generally three cases of results are considered:

(1) For most error cases, the system can correctly provide related suggestions for learners based upon the L1 interference hypothesis. The system will also show the 'L1 interference' in the pop-up window (as shown in Figure 1).
(2) In some cases, the error is in fact not related to L1 interference; our system will show 'non-L1 interference' but still suggest high-frequency collocations in corpus as revision alternatives.
(3) In few cases, our system might fail to treat the error appropriately due to the system's limitation (details will be provided in a later section), and we will provide some hints to let users understand these limitations.

System process

An interesting property of this system is the reliance on translation equivalents between learners' first language, Chinese, and the target language, English. The suggested collocations are offered based on the shared translations between these two languages. In addition, a collocation list extracted from a large reference corpus (e.g. *British National*

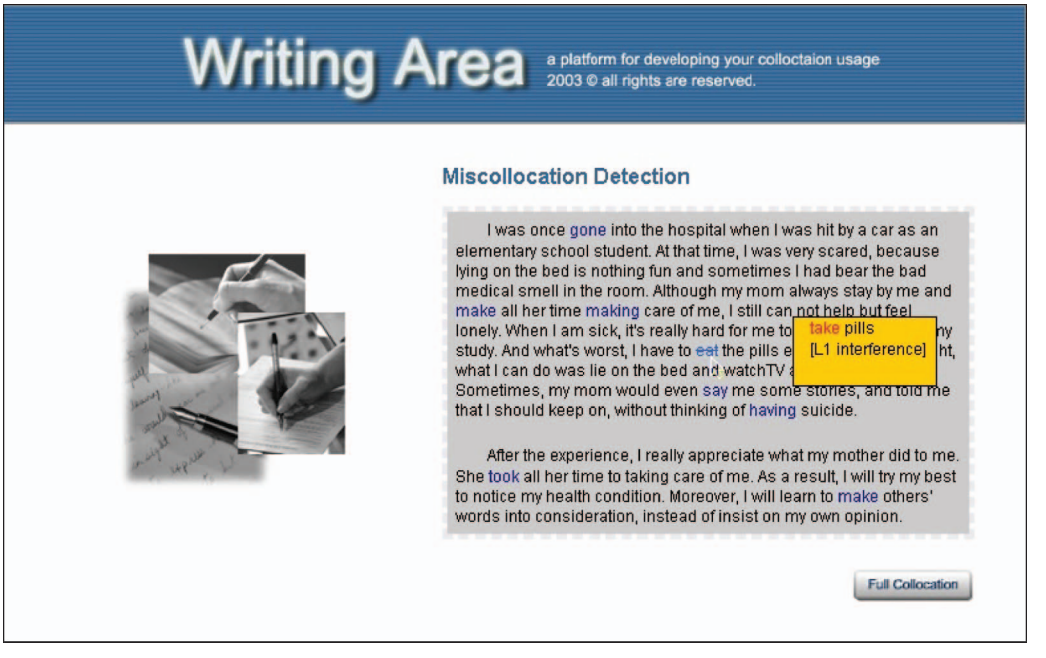

Figure 1. An example of the pop-up prompt-box with the possible answers.

Corpus) is efficiently used as a standard collocation library to provide correct collocation uses. The whole detecting process employs work in three main underlying phases described below and a supplementary stage will follow the three main stages for future modifications. We show the general architecture of the system with an example: 'The pill that the patient has to take is gone...' in Figure 2.

Collocation extraction: chunk and clause information integrated

The first stage aims at preprocessing the input texts in order to extract each verb and noun from V-N collocation. We make use of the shared task from CoNLL[2]-2000 (targeting chunk parsing) and CoNLL-2001 (targeting clause parsing) as source data to facilitate the V-N extraction from students' essays (Jian, Chang & Chang, 2004).

In CoNLL-2000 shared task, text chunking is considered as a process that divides a text into syntactically correlated parts of words. With the benefits of chunk information, we can segment the given sentence into smaller syntactic structure which facilitates precise collocation extraction. It becomes easier to identify the argument-predicate relationship between each chunk, and save more time to extract as opposed to full parsing. Take a text in CoNLL-2000 for example, the sentence 'Confidence in the pound is widely expected to take another sharp dive if trade figures for September' is annotated with chunk tags as follows:

```
Confidence/B-NP in/B-PP the/B-NP pound/I-NP is/B-VP widely/I-VP
expected/I-VP to/I-VP take/I-VP another/B-NP sharp/I-NP dive/I-NP
if/B-SBAR trade/B-NP figures/I-NP for/B-PP September/B-NP[3]
```

The words correlated with the same chunk tag can be further grouped together (as shown in Table 2). With chunk information, we can extract the target V-N collocation 'take dive' from the text by considering the last word of each adjacent VP and NP chunks.

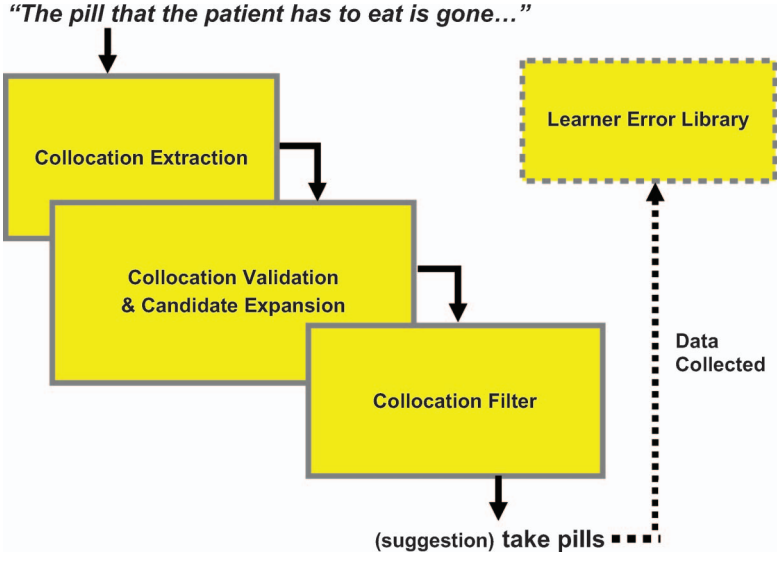

Figure 2. The processing stages of a given miscollocation '*eat medicine*' in the sentence.

In some cases, only considering the chunk information, however, is not enough. For example, the syntactic structure in the sentence '... the attitude he had towards the country is positive ...' may cause problems in extraction. With the chunk information, the system extracts the type 'have towards the country' as V + PP + NP, yet this one is erroneous due to the clause structure. To avoid this case, we further take the clause information into account.

With the training data provided by CoNLL-2001, we built an efficient clause model based on the Hidden Markov Model (Manning & Schütze, 1999) to identify the clause relation between words. This language model provides sufficient information to avoid extracting wrong V-N collocation instances. The examples show as follows (additional clause tags are annotated accordingly):

(1) .... the attitude (*S** he has **S*) toward the country
(2) (*S** I think (*S** that the people are most concerned with the question of (*S** when conditions may become ripe. **S*)*S*)*S*)

As a result, we can avoid the verb from being combined with the irrelevant objective noun (as in (1)) or erroneously extracting the adjacent noun serving as the subject of another clause (as in (2)). When the sentences in the corpus are preprocessed with the chunk and clause identification, we can subsequently assure high accuracy of collocation extraction.

By making use of the approach integrated with the linguistic information, the sentence is labeled by separate chunk and clause tags according to its syntactic structure. For example, the learner's sentence 'The pill that the patient has to take is gone' will be labeled as (Table 3):

Table 2. The examples of chunked sentence.

Sentence chunking	Chunking tag set
Confidence	NP
in	PP
the pound	NP
is expected to ***take***	**VP**
another sharp ***dive***	**NP**
if	SBAR
trade figures	NP
for	PP
September	NP

Table 3. Chunked sentence with clause information.

Sentence chunking	Chunking tag set	Clause tag set
The pill	NP	(*S**
that	SBAR	(*S**
the patient	NP	*
has to take	VP	*S*)
is gone	VP	*S*) *S*)

In this case, V-N collocations 'take medicine' will then be extracted from each last word of VP chunk 'has to take' and NP chunk 'the pill' with the constraint of clauses. Through the process of identifying chunks and clauses, each V-N collocation in input texts is found.

Collocation validation and candidate expansion

In the second stage, all extracted V-N collocations are first checked by the collocation list extracted from the reference corpus which provides the correct collocation usages. Those verb collocates validated as misused ones will undergo a series of comparisons to find their counterpart verb collocates sharing equivalent Chinese translations in electronic bilingual dictionaries. For example, the verb collocate *eat* in a collocation *eat pills* will trigger the system to search for the Chinese translation equivalents, such as *chi1* (吃), *chi1 wan2* (吃完), *shi1* (食), *chi1 guang1* (吃光), and other related translations. On the basis of these found Chinese translations, the system will continue to match the English translation equivalents; for instance, *chi1* (吃) in Chinese will match 'eat', 'finish' or 'take' in English (Table 4).

Collocation filter

In the third stage, the collocation list extracted from the reference corpus again comes into play as a filter. After matching the translation equivalents in both Chinese and English in the previous stage, the system has obtained a list of English candidate verbs associated with the English verb 'eat', such as 'finish', 'take', 'eat up', 'victual' and other candidate collocates. These candidate pairs (i.e. verb collocate listed with noun base *pills* such as *finish pills*) are then filtered by the collocation list.

Among the candidate pairs expanded via translational matching, pairs that lack referents in the collocation list are regarded as improper collocations and are left out. Only those that exactly match collocation patterns in the list will be retained. For instance, candidate pairs such as '*finish pills*', '*eat up pills*', '*victual pills*' will be excluded because they can not find matched combinations in the collocation list. The combination '*take pills*' which can be matched will be kept as a correct collocation for suggestion.

Table 4. Examples of candidate expansion based on translation equivalents.

English Input (Verb Collocate)	Chinese Translation	English Output (Candidate)
eat	吃(chi1)	eat finish take ...
	吃完(chi1 wan2)	finish eat up
	食(shi2)	victual
	吃光(chi1 guang1)	eat clean consume demolish ...
	...	...

Answer feedback and future supplement

If the extracted verb collocate is judged as correct in the validation step, the system will leave it without further processing. However, when the verb collocate written by learners is validated as misuse, the system will single out the relevant collocations found through those stages above, and show them in a pop-up box, e.g., 'take pills'. Moreover, if the suggestions provided are more than two collocations, the suggestions will be presented according to their frequencies and the measure of collocation strength calculated from the reference corpus.

For the supplementary stage, the system automatically builds and sustains a learner error library in the process of detecting miscollocations. As the system detects collocation errors in learners' texts, those error patterns will be collected and stored in the learner error library in the system, which increasingly enlarges the library itself. The aid of a learner error library can reduce the time of the detecting process in a web-based environment and facilitate users to get prompt feedback. Afterward, the first action the system should take is to match learners' written collocations with this error library. If error patterns exactly match those in the error library, the problematic verb collocates will be recognized and the corresponding suggestions will be provided as well. Only the rest will continue with the original detecting process above. In this way, students' compositions can also be directly colleted in the digitalized form for the convenience of doing error analyses in the future.

Experiment and discussion

Experiment setup

According to our system design, the whole process will enable users to input the essays and obtain some more appropriate collocations as suggestions. The evaluation of our system consists of:

(1) collocation extraction;
(2) collocation validation;
(3) performance of corrections and quality of the suggestions.

A previous study has shown a promising performance on the collocation extraction with chunk and clause information (Jian et al., 2004). Therefore, the key factors that can determine the performance of our module will be the latter two evaluation approaches. Moreover, the focus in our study is to confirm the hypothesis about L1 interference; we assume the reliability of suspected collocation extraction can be ensured in the system (though in some cases this collocation extraction module might not function perfectly with ungrammatical sentences), and we can put more focus on the later two evaluations to manifest how our correction scheme works.

As the approach we presented in the last section, we require the machine-readable bilingual dictionary and the reference corpus for our experiment setup. The bilingual dictionary is employed for the generation of candidate English verbs. In our experiment, the electronic bilingual dictionary mainly consists of some effective dictionaries (such as *Linguistic Data Consortium* (LDC) English-Chinese wordlists available online) as well as bilingual word lists extracted based on word-alignment technique (such as Wu & Chang, 2004). We also noted the word-aligned bilingual word list is found to be more effective as it is extracted from an authentic bilingual corpus.

More corresponding translations of a given word can accordingly be matched in the processing.

On the other hand, the collocation list extracted from the reference corpus is utilized for collocation validation and collocation filter. From the *British National Corpus* (BNC), the BNC Collocation List is abundantly extracted via the natural language processing techniques (Jian et al., 2004). 661,040 collocation types are obtained with 87% precision. Most of the erroneous extractions often appear with low counts. Thus, though these errors do slightly interfere with the system's correction, the interference can be prevented by the presentation according to frequencies.

Evaluation

The precision of collocation validation

To evaluate the performance of the collocation validation, we randomly selected 200 manually checked V-N collocations extracted from the *Sinorma* corpus[4] and employed the 226 V-N miscollocations in the students' writing (Liu, 2002).[5] In the validation of correct collocations, the BNC Collocation List can help achieve the performance of 97.5% precision; in checking wrong collocations, we can have a satisfactory result of 90.7% precision (Table 5). All in all, among 426 adequate and misused collocations, 93.9% precision of validation can be achieved based on the BNC Collocation List.

The performance of miscollocation correction

The data of students' misused verb collocates and teachers' suggested alternatives from English TLC collected by Liu (2002) can also serve as the testing data to evaluate the performance of correcting miscollocaitons. This original source data consist of 233 verb-noun miscollocations (219 misuses caused by the anomalous verb uses).

With closer examination of Liu's data, we found that some nouns serving as the objects of the same verb can be treated separately as different miscollocations (e.g. two entries of misuses along with suggestions, '*talk/tell news' and '*talk/tell sadness', can be singled out in the same sentence 'talk their good news or sadness to each other'). Moreover, while the suggestions provided are sometimes more than one, we will also divide them as different entries (e.g. two suggestions for the sentence 'gain their own aims' are 'reach' and 'attain', and these two suggestions will thus be divided into '*gain/reach aim' and '*gain/attain aim'). We also excluded the duplicate miscollocations (e.g. only one entry '*make/cause misunderstanding' is selected from both the sentences, 'from making such misunderstanding' and 'you make a great misunderstanding). As a result, among 233 miscollocations paired with suggested corrections, 226 erroneous collocations caused by wrong verb choice are left as testing data.

Table 5. The performance of validation.

Validation	Count	BNC Collocation List Judgment		Precision	Overall
Correct collocation	200	(True)	195	97.5%	
		(False)	5		400/426
Wrong collocation	226	(True)	205	90.7%	(93.9%)
		(False)	21		

We evaluated the performance of our system according to the teachers' suggestion annotated in English TLC. Based on our hypothesis, all the 226 verb pairs were processed with an electronic bilingual dictionary to obtain verb candidates, and these verbs paired with nouns were further filtered by the reference corpus, BNC Collocation List. While presenting the more adequate collocations as suggestions, we ranked our suggestions by collocation co-occurrence (i.e. the frequency of the co-occurring verb and noun) and Log Likelihood Ratio score (i.e. the measure of the strength of association between the verb and the noun in a given collocation) (Jian et al., 2004). Only the top 10 ranking results will be taken into consideration.

We further used two indicators to measure the correction performance: the precision of correction and the mean reciprocal rank (MRR) (Voorhees & Tice, 1999).

The precision of correction. In evaluating the precision of correction, we only take into account those miscollocations validated by our system. Therefore, in addition to six validation errors caused by our system (i.e. failure to detect genuine students' errors), after examining the data in detail, we excluded some conditions that are not considered in our system to obtain the more exact result of the performance: (1) We observe that some students' errors can be acceptable; for instance, 'broaden his knowledge' in the source data. Under this condition, our system will treat this collocation as a correct one, and no more correction will be carried out. In addition, (2) while the noun collocate is a pronoun or reflexive (e.g. 'talks me' and 'just do myself'), these entries will not be considered as well. Moreover, our reference corpus, the BNC Collocation List, extracted based on the Verb-Noun $_{\text{Obj}}$. structure can only cope with the correction with the same structure. We thus only tackle the problems with Verb-Noun $_{\text{Obj}}$. structure in our system design. Therefore, (3) if the structure of a miscollocation is Noun $_{\text{Subj}}$. -Verb Structure, our system will make no correction on these miscollocations (e.g. in the miscollocation instance 'I felt my heart couldn't move anymore', the entry 'heart *move/beat' is thus excluded.) Overall, 21 exceptions are discarded (the distribution is concluded in Table 6).

While comparing the suggestion with teachers' manual correction, we regard a result as a correct one only if our system's suggestion exactly matches the manual judgment. This basic but critical evaluation shows 84.4% precision (173 correct answers out of 205 testing entries) (Table 7). Some of the good examples such as '*publish/release album' or '*create/compose song' are shown in the Table 8.

Table 6. The conditions excluded in the evaluation of correction precision.

Exception Type	Count
(1) Validation error	6
(2) Acceptable collocation	6
(3) Noun not in consideration	5
(4) Noun $_{\text{Subj}}$. -Verb structure	4
Total	21

Table 7. Evaluation on the performance of correcting miscollocations.

Evaluation Type	Precision
Basic evaluation	84.4% (173/205)
Manual examination	94.1% (193/205)

Moreover, further analyzing the results, we notice that some suggestions provided by our system are acceptable. One of these examples can be seen in the entry '*show/state opinion'; the system's suggested correction 'express opinion' is also adequate in the context. A human judge, who is an English teacher in senior high school, assessed the system's correction without the original answers; all 205 miscollocations as well as their correction are manually re-examined to obtain a more precise result. The examiner's result shows that 193 corrections prove to be appropriate and a higher overall precision (94.1%) can be achieved.

The mean reciprocal rank. To evaluate the quality of the suggestions provided by our system, we further take another indicator into consideration. In light of the precision obtained from the last section, even though 94.1% of suggestions contain the appropriate corrections, no evidence shows whether our system can provide the most relevant answers with better ranking or not. Correction presentation with good ranking or adequate arrangement certainly could save a user's time of screening the suggestion lists.

The second indicator is the mean reciprocal rank (MRR) (Voorhees & Tice, 1999) of the first correct suggestions returned, so as to assess how precisely the first batch of returned results meet users' expectations. For each test miscollocation, we calculate its reciprocal of the rank at which its correct suggestion was first found. If the r-th answer returned is adequate, the reciprocal rank of the answer will be $1/r$. The MRR is the average reciprocal ranking score for all the test miscollocations. The MRR of our system performance is 0.66 (Table 9) which shows that our system is effective in the answer ordering. After a query is submitted, the user could easily find the answers in the first or second ranking in average without consuming much time.[6]

Table 8. Examples of the collocations suggested by the system.

Misused Verb + Noun	Teacher's Suggestion	System Suggestion
*publish album	release	release album do album issue album
*create song	compose	write song compose song contribute song
*conquer difficulty	overcome	overcome difficulty surmount difficulty
*abide pain	suffer	take pain suffer pain bear pain
*release pressure	relieve	put pressure relieve pressure ease pressure

Table 9. Evaluation on the quality of the suggestion ordering.

Evaluation Type	MRR
Basic evaluation	0.50
Manual examination	0.66

Table 10. Examples of the inappropriate suggestions occurring in the system.

Types	Misused Verb + Noun	Teacher's Suggestion	Incorrect System's Suggestion
No translation equivalents in the dictionary	*open computer	turn on	get computer build computer operate computer
No referents in the collocation list[†]	*prevail plot	dominate	sell plot sell off plot work out plot
De-lexicalized verb	*do life	live	make life have life spend life

[†]In this example, though the verb candidate 'dominate' was successfully generated, it was filtered out in the end because of the lack of a referent in the collocation list.

System limitations

The system is built on an assumption that most of students' miscollocations can be attributed to L1 translation interference. Although the system with 94.1% of coverage can solve the problems of most misused collocations, there are still some misses that need further error analysis and discussion. Most prevalent problems can be categorized into three primary types (Table 10):

(1) Some miscollocations which might result from L1 translation interference failed to find adequate collocate candidates through translation matching due to insufficient bilingual translation data available. Even though some English verbs are perceived to be similarly translated in Chinese, these verbs without identical translation in the source data cannot adequately generate expected candidates and then fail to support our hypothesis. Therefore, a better bilingual dictionary generated from more effective bilingual corpora with robust alignment techniques might help to solve this problem.
(2) Even if the English candidate verbs can be generated based on translation equivalents, the system still fails to offer correct suggestions since these correct candidates paired with nouns can not match the referents in the collocation list. Although the collocation list extracted from BNC corpus contains more than 0.6 million collocation types, the list still cannot completely cover all the collocations used in English. An updated collocation list with high-precision extraction or an increase of collocation resources such as electronic collocational dictionaries might also improve the performance of collocation suggestion.
(3) De-lexicalized verbs which denote a variety of meanings are not easily matched with specific Chinese translation equivalents; thus, these errors might not be detected and corrected by the *Writing Assistant*.

Conclusion

In this paper, we address the problem of detecting and correcting V-N miscollocations, which are a gap in current automatic writing systems. The proposed strategy has a number of innovative ideas. First, we utilize the translation equivalents shared by the misused and adequate verb collocates so as to revise the collocation that learners intend to write but

misuse. Second, the approach will be a work of tailoring to errors caused by a specific L1 (such Mandarin Chinese). With little modification, the strategy is likely to generalize to other EFL learners with a different L1. Third, we prove the feasibility of collocation checking based on a very large corpus, such as BNC. Finally, we make use of this reference corpus as the criterion to confirm the correctness of collocations and to offer more proper alternatives as suggestions. Compared with the maintenance of a learner corpus, the reference corpus can spare the time of the system implementation without any manual error tagging in advance.

The *Writing Assistant* as described in this paper is suitable for learners to use as an online collocation helper while writing their composition. After they complete their writing pieces, learners can use the *Writing Assistant* to see whether they misuse V-N collocations; they can then apply given suggestions to revise their composition. Since the *Writing Assistant* is developed primarily on the ground of translation interference, the system will be a great help for learners of Chinese speakers. Besides, the system *Writing Assistant* merely relied on the reference corpus, and, therefore, it can save much effort of recruiting human correcting. In this way, the possibility of some errors that are not collected in a learner corpus will be reduced; thus detective accuracy will be enhanced. In sum, it is hoped the system *Writing Assistant* will guide learners to learn appropriate uses of collocation and will also equip learners with better collocational knowledge. In the future, the problem of other types of miscollocations can also be taken into consideration, such as Adjective-Noun or Adverb-Adjective collocations. With the capability of handling more miscollocations, a comprehensive writing system can be developed.

Acknowledgement

This work is carried out under the project 'CANDLE' funded by National Science Council in Taiwan (NSC93–2524-S007–001). Further information about CANDLE is available at http://candle.cs.nthu.edu.tw/.

Notes

1. The CANDLE project is a national e-Learning project, which uses corpora and various natural language processing (NLP) tools to construct advanced English learning tasks for intermediate learners in Taiwan such as college freshman students or those with a similar English proficiency level. The detailed content is shown in the website (http://candle.cs.nthu.edu.tw/).
2. CoNLL is the acronym of the *Conference on Computational Natural Language Learning*. The shared task data for CoNLL–2000 and CoNLL-2001 is available at (http://cnts.uia.ac.be/conll2000/) and (http://cnts.uia.ac.be/conll2001/) respectively.
3. Most chunk types might consist of two different chunk tags: B-CHUNK for the initial word of a chunk and I-CHUNK for the other words in the same chunk. Considering the example in question, the chunk type of NP, *the pound*, can be derived by grouping *the*/*B-NP* and *pound*/*I-NP* as a whole
4. *Sinorama* is a balanced bilingual corpus containing the English text (about 50,000 sentences) and translation in Chinese.
5. The original number of miscollocations provided by Liu (2002) should be 233 but only 226 are left for testing with exceptions filtered (details provided in the following section).
6. We chose MRR as our scoring metric because it reflects the quality of returned answers. MRR is mostly used by the evaluation of current information retrieval systems, such as Google, to evaluate the ranking of returned pages. The mean reciprocal is bounded between 0 and 1. Apparently, if a system's MRR score is near 1.0, it is supposed to always return answers in the first place. If the system performs with MRR score of 0.5 approximately, most of its answer will be returned in the second ranking.

Notes on contributors

Yu-Chia Chang is currently a PhD candidate in the Institute of Information Systems and Applications in National Tsing Hua University, Taiwan. His research mostly focuses on the cross-disciplinary study with respect to language technology.

Prof. Jason S. Chang specializes in Natural Language Processing and teaches at Department of Computer Science of National Tsing Hua University.

Prof. Hao-Jen Chen's research interests mostly concern CALL and he teaches at the Department of English in the National Taiwan Normal University in Taiwan.

Prof. Hsien-Chin Liou teaches Foreign Languages and Literature at National Tsing Hua University with a research focus on CALL.

References

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Philadelphia: John Benjamins.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 70–86). London: Language Teaching Publications.

Farghal, M., & Obiedat, H. (1995). Collocations: A neglected variables in EFL. *International Review of Applied Linguistics*, *33*, 313–331.

Firth, J.R. (1957). Modes of meaning. *Papers in linguistics* (pp. 1934–1951). Oxford: Oxford University Press.

Jian, J.Y., Chang, Y.C., & Chang, J.S. (2004, September). *Collocational translation memory extraction based on statistical linguistic information*. Paper presented in ROCLING 2004, Conference on Computational Linguistics and Speech Processing, Taipei.

Justeson, J.S., & Katz, S.M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*(1), 9–27.

Lewis, M. (2000). *Teaching collocation: Further development in lexical approach*. London: Language Teaching Publications.

Linguistic Data Consortium. (2002). English-to-Chinese wordlist (version 2.0) and Chinese-to-English wordlist (version 2.0). Retrieved May 26, 2007 from http://www.ldc.upenn.edu/Projects/Chinese/.

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

Liu, C.P. (2000). A study of strategy use in producing lexical collocations. Selected papers from the Ninth International Symposium on English Teaching (pp. 481–492). Taipei: Crane.

Liu, L.E. (2002). A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English. Master's thesis, Tamkang University, Taipei.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MA: MIT Press.

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

Nattinger, J.R., & DeCarrico, J.D. (1992). *Lexical phrase and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*, 223–242.

Shei, C.C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning*, *13*(2), 167–182.

Stockwell, R., Bowen, J., & Martin, J. (1965). *The grammatical structures of English and Spanish*. Chicago: Chicago University Press.

Voorhees, E.M., & Tice, D.M. (1999)*The TREC-8 question answering track evaluation*. In E.M. Voorchees & D.K. Harman (Eds.), Proceedings of the Eighth Text REtrieval Conference (TREC-8) (pp. 83–105). NIST Special Publication 500–246.

Wu, C.C., & Chang, J.S. (2004). Bilingual collocation extraction based on syntactic and statistical analyses. *Computational Linguistics and Chinese Language Processing*, *9*(1), 1–20.

Zhang, X. (1993). English collocations and their effect on the writing of native and non-native college freshmen. (Ph.D. dissertation, Indiana University of Pennsylvania, 1993). *Dissertation Abstracts International*, *54-03A*, 0910 (UMI No. AAI9319454).