US-China Education Review, ISSN 1548-6613
May 2011, Vol. 8, No. 5, 659-665

Assemble Collocation and Colligation in Chinese Writing Web Tools for New Immigrants in Taiwan*

Meg Lu
National University of Tainan, Taiwan

Chien Hui Lin
Institute for Information Industry, Taiwan

Tsung Yen Chuang
National University of Tainan, Taiwan

Tsun Ku
Institute for Information Industry, Taiwan

Chia Min Tsai
National University of Tainan, Taiwan

CSL (Chinese as a second language) learning is an emergency task in Taiwan, especially when more and more new immigrants joined in Taiwan. However, there are just a few new immigrants who can finish all language courses. For this reason, this research intends to provide a training system to assist new immigrants observing and learning the phrase structure. In this paper, the researchers design a system which can help new immigrants to form Chinese sentences by applying colligation and collocation technology.

Keywords: colligation, collocation, new immigrants, CSL (Chinese as a second language) learning

Introduction

The general aims of this study are to understand the aspects of new immigrants using reference tools for Chinese writing and to examine how these tools assist them to learn Chinese. The specific purpose of this study is to examine how CSL (Chinese as a second language) learners in Taiwan use a corpus as a reference tool in conjunction with dictionaries when paraphrasing Chinese Web-related articles.

Colligation

The term "colligation" as stated in this manual is conformed to the researches of Firth (1957), Hoey (2004) and Stubbs (2002). Simply speaking, colligation is the grammatical combination of words (e.g., can) or a word category (e.g., VM (verb modifiers)). Both collocation and colligation are types of phraseologism (Gries, 2007). The distinction of colligation is concerned with the lexical company a word keeps, while it is focused with the grammatical company a word or a word category. In other words, colligation is the conceptual inter-relations of grammatical categories. Hunston, Francis and Manning (1997) concluded that there are correlations between grammatical patterns and lexical meaning. All words can be represented by specific patterns, and the meanings of words can share patterns with a lot of commons. That means a word could have a specific meaning when it

* This study is conducted under the Intelligent Web-enabled Service Research and Development Project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

Meg Lu, Ph.D., assistant professor, Department of Education, National University of Tainan.
Chien Hui Lin, IDEA (Innovative DigiTech-Enabled Applications & Services Institute), Institute for Information Industry.
Tsung Yen Chuang, Ph.D., assistant professor, Department of Information and Learning Technology, National University of Tainan.
Tsun Ku, IDEA (Innovative DigiTech-Enabled Applications & Services Institute), Institute for Information Industry.
Chia Min Tsai, Department of Information and Learning Technology, National University of Tainan.

co-occurs with a certain word. This hypothesis is followed by Hoey (2000), who mentioned that some meanings of the same word have their own grammatical patterns, which is called “colligation”. This concept was invented by Firth (1957). Furthermore, colligation is concerned with relationship between grammatical classes, whereas collocation is interested in the words which belong to these grammatical classes. Grammatical pattern “verb + to-infinitive” is an example of colligation and “dread + think” is an example to show collocation of this colligation. In short, colligation defines the grammatical company and interaction of words, as well as their preferable position in a sentence.

Table 1

The Pattern of Example for Colligations and Collocations

Colligations	Collocations
DET + N + V	This paper describes
ART + N + PREP	a pair of
ADV + ADJ + CONJ + (ART) N	as white as snow
ART + N + PRE + N	a sea of troubles
Motive (verb + particle)	ambled up, tore across
INT + ADJ	dead tired

In addition, Hoey (2000) showed how corpus analysis revealed previously unnoted types of regularity. He argued that these rules should be taught to learners. In his paper, he attempted to account word combinations for students’ reconstructions of a short text whose paragraphs had been jumbled. To explain why certain paragraphs were interpreted as possibly or not possibly text-initial, he extended the notion of colligation to cover not merely the relationship of a word to the structure of the clause and sentence, but also the paragraph and the entire text, as reflected in its tendency to occur in particular textual positions. Hoey drew the conclusion that grammar, in this case text grammar, needs to be related to choices in lexis. There is, he argued, a “hidden colligational signaling” of the text structure which is as yet unknown to us, but which is of clear pedagogical relevance for the teaching of reading and writing. In the absence of systematic research in this area, it is not yet possible to propose specific patterns to be incorporated in syllabus. However, Hoey’s study exemplifies a methodology which could be adapted by instructors and learners to explore colligational signaling for instruction. Colligation can be used to investigate the frequencies and effectiveness of different textual organizations. However, it is not just access to the organizing vocabulary that is important, but also the “control of the supporting colligation patterns (that) is also crucial” (Ravelli, 2004, p. 123). In other words, the writers need to control the level of commitment of meaning, and potentially in the flow of the discourse. A useful strategy to support awareness in this respect is to track the lexical chains from initial instances of lexis in higher level themes across subsequent phases of a model text.

Collocation

Halliday and Hasan (1976) have defined that collocation is an associative meaning relationship between regularly co-occurring lexical items. Even there exists a lot of collocation extraction paper in the field of natural language processing (unfinished sentence). But this is two kinds of application. Collocation is one of the most difficult aspects in second language learning. However, it has been largely neglected by researchers and practitioners. Although, the role that collocation plays in language acquisition is an important topic, very few systematic studies can be found to address this issue. One recent study written by ZHANG (1993) indicated that a series of experiments were conducted to explore the relationship between the knowledge of collocation and proficiency in writing. It was found that more proficient second language writers use significantly more collocations more accurately and in more varieties than less proficient learners. Gitsaki (1996) further identified some factors which could affect the development of collocational ability during language acquisition, for example, frequency in the input, complexity of the collocations, degree of L1 (first language)-L2 (second language) difference and the order of collocational parts (e.g., Prep. Noun was found to be more difficult than Noun Prep. collocations). Despite the lack of empirical studies, researchers generally agreed that collocational knowledge is one of the things which contribute to the difference between native speakers and second language learners.

The Web as a Corpus for Language Learning

The abundant and varied texts of the WWW (World Wide Web) tantalize linguists and language instructors alike: The Web's ever-expanding and self-renewing machine-readable body of Web pages in scores of languages is easy to retrieve, also, challenging to sift through and exploit efficiently. Yet, there are compelling reasons to supplement existing corpora with online materials. Once compiled, a corpus represents a snapshot of language usage and issues at the time the content is produced. The great expense of setting up a large corpus precludes frequent replacement or updating, and content can age surprisingly quickly. In contrast, countless new documents appear on the Web daily, which are examples of current language usage and contemporary issues abound. In addition, even a large corpus might include few examples, if any of a relatively infrequent expression or construction that would not be difficult to locate online. Furthermore, certain domains or text genres may be underrepresented or missing entirely in an existing corpus. Using the Web as a source can easily compile an ad-hoc corpus to meet the specific needs of groups of learners or translators. Finally, while off-the-shelf corpora and corpus tools may entail significant fees and often require expensive hardware, the Web is virtually free, and desktop computers to perform the necessary processing are now within the reach of researchers and students.

Chinese Language Grammar

The grammatical knowledge-base of contemporary Chinese serves as a basic linguistic knowledge-base for Chinese information processing. It passed the technical appraisement in November, 1995. Through the continuous development in the past over four years, it is extended to 73,000 entries from 50,000, and the classification of these 70,000 words is accomplished. In addition, a new morpheme database has been developed for the undefined word recognition. Up to now, the distinct grammatical descriptions in every class have been carefully checked and corrected, while more than 20 new attributes as well as a great quantity of examples are added. Therefore, the scale and quality of the whole knowledge-base are improved remarkably.

Design Methodology

In this section, the function of the tutoring system is described, and the main architecture is presented (see Figure 1). In our opinion, a supporting system for writers should aim to do more than just point out the errors. For this reason, the motivation for our software program is trying to help learners to observe the usage of the verb which can help them to make sentence. In order to achieve this goal, the authors design a system architecture which adapts the concept of Web as corpus and filters the information by grammar knowledge. The software is a part of course of "Writing Right in Chinese" for CSL learners.

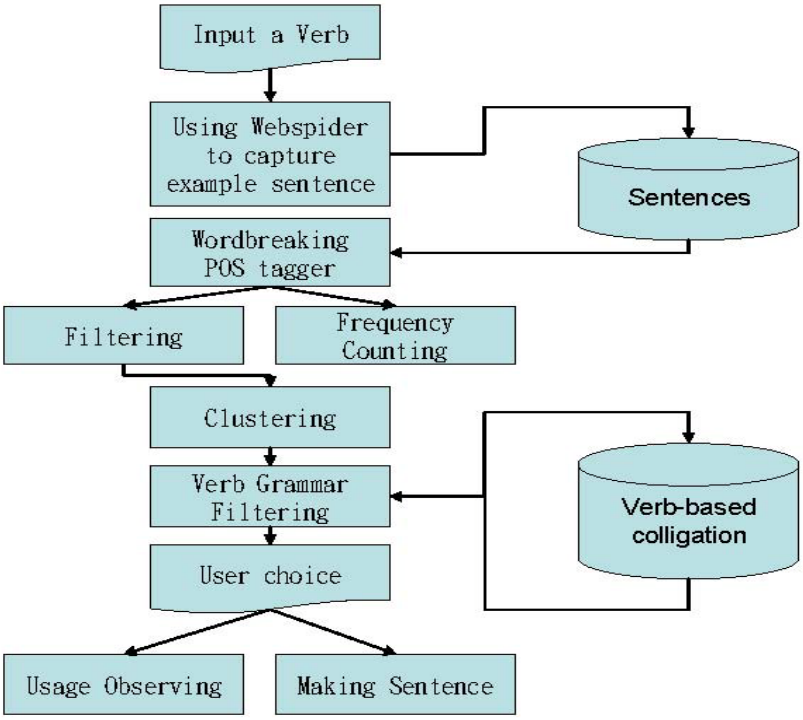


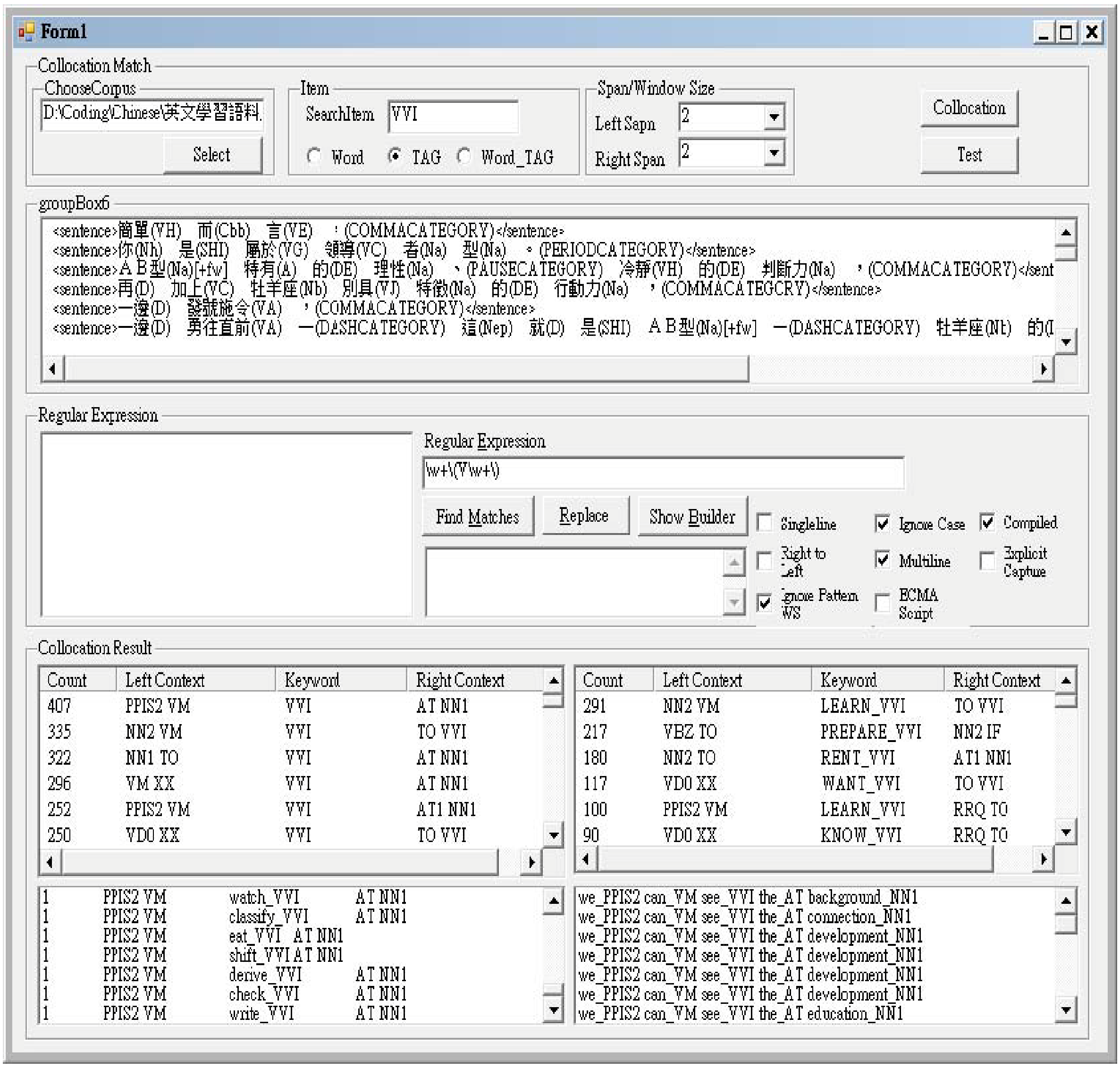
Figure 1. System architecture.

Web as Corpus

The concept of "Web as corpus" is adapted based on two reasons. The first reason is many post-intermediate CSL learners write correct Chinese grammatically, but their writing often lacks a native-like flavor or is not versatile in diction. The content retrieved from real world can solve this problem. The second reason is that we do not want that the example sentence is provided only by the teachers, because this may limit students' creativity and imagination.

Colligation and Collocation

This study adapts both collocation and colligation in the system (see Figure 2). This is depending on the word class of their constituent lexical collocations consisting of a windows size of open class words. The reason to adapt both collocation and colligation is that even formal frequency of co-occurrence is helpful, but there still has ultimately limited paradigm by using collocation. Therefore, further progress may require something closer to a construction grammar approach (as colligation done), in which form and function are seen as equally important aspects of linguistic items. But the retrieval of colligational patterns in texts is much more difficult than that of collocational patterns. Also, the regular expression engine embedded in the software enables the tool to be customized to meet users' needs to calculate the frequency of lexicon, syntax and discourse relations in texts.



ChooseCorpus: D:\Coding\Chinese\英文學習語料.

SearchItem: VVI

○ Word ● TAG ○ Word_TAG

Left Sapn: 2

Right Span: 2

```
<sentence>簡單(VH)   而(Cbb)   言(VE)   ：(COMMACATEGORY)</sentence>
<sentence>你(Nh)   是(SHI)   屬於(VG)   領導(VC)   者(Na)   型(Na)   。(PERIODCATEGORY)</sentence>
<sentence>ＡＢ型(Na)[+fw]   特有(A)   的(DE)   理性(Na)   、(PAUSECATEGORY)   冷靜(VH)   的(DE)   判斷力(Na)   ，(COMMACATEGORY)</sent
<sentence>再(D)   加上(VC)   牡羊座(Nb)   別具(VJ)   特徵(Na)   的(DE)   行動力(Na)   ，(COMMACATEGCRY)</sentence>
<sentence>一邊(D)   發號施令(VA)   ，(COMMACATEGORY)</sentence>
<sentence>一邊(D)   勇往直前(VA)   —(DASHCATEGORY)   這(Nep)   就(D)   是(SHI)   ＡＢ型(Na)[+fw]   —(DASHCATEGORY)   牡羊座(Nt)   的(I
```

Regular Expression: \w+\(\/\w+\)

☐ Singleline ☑ Ignore Case ☑ Compiled
☐ Right to Left ☑ Multiline ☐ Explicit Capture
☑ Ignore Pattern WS ☐ ECMA Script

Count	Left Context	Keyword	Right Context
407	PPIS2 VM	VVI	AT NN1
335	NN2 VM	VVI	TO VVI
322	NN1 TO	VVI	AT NN1
296	VM XX	VVI	AT NN1
252	PPIS2 VM	VVI	AT1 NN1
250	VD0 XX	VVI	TO VVI

Count	Left Context	Keyword	Right Context
291	NN2 VM	LEARN_VVI	TO VVI
217	VBZ TO	PREPARE_VVI	NN2 IF
180	NN2 TO	RENT_VVI	AT1 NN1
117	VD0 XX	WANT_VVI	TO VVI
100	PPIS2 VM	LEARN_VVI	RRQ TO
90	VD0 XX	KNOW_VVI	RRQ TO

```
1        PPIS2 VM        watch_VVI        AT NN1
1        PPIS2 VM        classify_VVI     AT NN1
1        PPIS2 VM        eat_VVI   AT NN1
1        PPIS2 VM        shift_VVI AT NN1
1        PPIS2 VM        derive_VVI       AT NN1
1        PPIS2 VM        check_VVI        AT NN1
1        PPIS2 VM        write_VVI        AT NN1
```

```
we_PPIS2 can_VM see_VVI the_AT background_NN1
we_PPIS2 can_VM see_VVI the_AT connection_NN1
we_PPIS2 can_VM see_VVI the_AT development_NN1
we_PPIS2 can_VM see_VVI the_AT development_NN1
we_PPIS2 can_VM see_VVI the_AT development_NN1
we_PPIS2 can_VM see_VVI the_AT development_NN1
we_PPIS2 can_VM see_VVI the_AT education_NN1
```

Figure 2. Screenshot of system interface.

Using Grammatical Knowledge to Filter

Finally, this study mainly modifies the factoid detection rules and adds the GKB (The Grammatical Knowledge-Base of Contemporary Chinese) dictionary to filter the example sentences.

Instructional Scenario

"Writing Right in Chinese" is a course designed for new immigrants in Taiwan. Those new immigrants have some similar characteristics: They not only do not have a lot of time to participate in school coursework, but also are adults with their own language and culture experiences. Hence those reasons, instructional designs should help new immigrants apply a self-learning tool rather than a step-by-step teaching strategy. A scenario is shown in Figure 3. Whenever a new immigrant learner logins the system, he/she can observe the pattern of the information provided in system. After he/she has completed the lesson, he/she can create a sentence using these words or grammar and share with other learners.

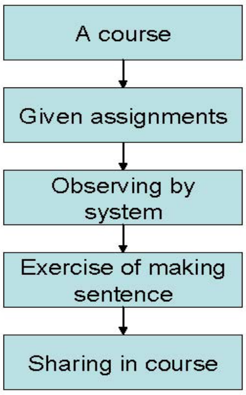


Figure 3. Instructional scenario.

Findings and Challenges

In our pilot study, this research finds several challenges which need to be solved. The first challenge is whether high frequency of occurrence in a general language corpus indicates that a word combination is likely to be part of most native speakers' mental systems. Our system and previous study showed that frequency data do reliably predict psychological associations between words, as evidenced by word association norms. But not all frequency data are suitable for learning. The frequency data need selecting and filtering by professional mechanism and teachers. This is the main reason that this research integrates collocation, colligation and GKB in this system. The second challenge addressed by this research is that whether CSL learners tend to acquire the relation of collocations/colligation they have to type in the system. The results of this study shows that contrary to the claims of other researchers, learners may indeed acquire many of the collocations they meet on a regular basis. However, it seems that the relatively low levels of input to which learners are typically exposed tend to leave them with a distinctively "non-native-like" profile of collocational or colligational knowledge. Taken together, these results suggest that, though learners can be expected to pick up some collocations and colligations implicitly, teachers should also provide an explicit focus on students' learning, especially, on the learning of those salient pairs and relations which learners appear to have difficulty in acquiring. This result of collocation and colligation has showed some powerful advantages. One advantage is that it allows the outcome to be identified rapidly, and with a high degree of reliability from samples of text far too large for human analysts to reliably handle. The other advantage is that it helps us to identify those collocations and colligations which are semantically and syntactically regular, but which many psychologically-oriented views of language maintain are likely to have a special "holistic" status in the language system because of their high frequencies of occurrence.

Conclusions and Future Work

According to the purpose of this research, one writing observing environment has been created to assist CSL learners by way of collocational and colligational interface. The architecture of this system has been described. The design of this system intends to guide new immigrants how to improve their Chinese writing by pointing out their potential errors in collocations and colligations, and allowing CSL learners to observe the use of collocations and colligations in authentic language. The results of this pilot study reveal a minuscule part of the new immigrants' language learning that is necessary in order to become fluent Chinese writers.

References

Firth, J. R. (1957). Modes of meaning. In J. R. Firth, *Papers in linguistics 1934-1951* (pp. 190-215). Oxford: Oxford University Press.

Gitsaki, C. (1996). The development of ESL collocational knowledge (Unpublished doctoral dissertation, The University of Queensland).

Gries, S. (2007). New perspectives on old alternations. In E. C. Jonathan, L. F. Amy, & W. K. David (Eds.), Papers from *the 39th Regional Meeting of the Chicago Linguistics Society (Vol. 2), The Panels* (pp. 274-292). Chicago IL: Chicago Linguistics Society.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hoey, M. (2000). The hidden lexical clues of textual organization: A preliminary investigation into an unusual text from a corpus perspective. In L. Burnard, & T. McEnry (Eds.), *Rethinking language pedagogy from corpus perspective* (pp. 31-42). New York: Peter Lang.

Hoey, M. (2004). The textual priming of lexis. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 21-41). Amsterdam/Philadelphia: John Benjamins.

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal, 51*(3), 208-216.

Ravelli, L. J. (2004). Signaling the organization of written texts: Hyper-themes in management and history. In L. Ravelli, & R. Ellis (Eds.), *Analyzing academic writing: Contextualized frameworks* (pp. 105-130). London: Continuum.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

ZHANG, X. (1993). English collocations and their effect on the writing of native and non-native college freshmen (Unpublished doctoral dissertation, Indiana University of Pennsylvania).