ARTÍCULOS

**Círculo de Lingüística Aplicada a la Comunicación**
ISSN: 1576-4737

EDICIONES COMPLUTENSE

http://dx.doi.org/10.5209/clac.73731

# The Spanish Collocation Tool and Its Application in Corpus-Based Study of Spanish for Teaching and Learning

Hui-Chuan Lu 盧慧娟[1], Cai-Yu Song 宋采育[2] and An Chung Cheng 鄭安中[3]

Recibido: 8 de octubre de 2017 / Aceptado: 30 de octubre de 2017

**Abstract.** The purpose of this paper is to introduce a developed software program called the "Spanish Collocation Tool (SCT)" and its application in related corpus-based studies. The Spanish Collocation Tool (SCT) was designed to assist with the research and analysis of Spanish collocation. The SCT allows searches of collocated elements not limited to words but also parts of speech and lemmas. Furthermore, it can compare two collocation lists to detect any significant differences between them. In this study, this collocation tool, SCT, and a constructed L3 Taiwanese learners' written corpus of Spanish called CEATE were combined to create efficient access to results in a systematic approach. Furthermore, by using the SCT, the pedagogical implications of the search results for the development of on-line multimedia material for learning Spanish collocations are discussed in the end.
**Keywords:** collocation, corpus-based, Spanish, tool

**Cómo citar:** Lu, Hui-Chuan; Song, Cai-Yu; Cheng, An Chung (2021). The Spanish collocation tool and its application in corpus-based study of Spanish for teaching and learning *Círculo de lingüística aplicada a la comunicación* 85, 269-276, http://dx.doi.org/10.5209/clac.73731



## 1. Introduction

Given the importance of learning collocations and its relevance to foreign language teaching, in the last two decades, there has been a growing interest in both the area of theoretical linguistics (e.g., Aguilar-Amat Castillo, 1993; Alonso Ramos, 1995; Castillo Carballo, 1998; Corpas Pastor, 1992a, 1992b; Marcinkevicine, 1995) and in the area of applied linguistics (e.g., Kennedy, 2003; Kita & Ogata, 1997; Lewis, 2000; Navarro, 2003; Pu, 2003; Sun & Wang, 2003).

However, according to Lee (2010), there is a gap between Spanish and English studies with respect to the development of collocation tools and investigations applied in corpus-based collocation research taking into account both quantity and technique. Therefore, the purpose of this paper is to present a new software program called the "Spanish Collocation Tool (SCT)" as well as its application in related studies.

This paper is organized as follows. Section 2 reviews previous studies on collocation tools and related research. Section 3 presents the design and development of the SCT. In Section 4.1, the SCT was then used to study Taiwanese L3 learners of Spanish to provide an applied example for linguistic analysis. The pedagogical implications of the SCT in on-line teaching material are discussed in Section 4.2, and Section 5 provides some final conclusions.

## 2. Literature review

This section provides a review of previous related research in theory and practice to understand the current situation and the main functions of the developed collocation tools.

Earlier studies on collocation have explored statistical issues that occur in natural language processing. For example, Handl (2008) indicated that the structure of a collocation consists of various dimensions, including both semantic and lexical structures as well as statistics. Almela Sánchez (2006) mentioned that the extraction of collocations takes

[1] National Cheng Kung University, Taiwan. Correo electrónico: huichuanlu1@gmail.com
[2] National Cheng Kung University, Taiwan. Correo electrónico: qzop351@gmail.com
[3] The University of Toledo, U.S.A. Correo electrónico: Anchung.Cheng@utoledo.edu

into consideration the relationship between a head and its collocated element regarding co-occurrence, mutual selection, window size, frequency, and statistical significance.

Computational technology can be used for developing new search functions intended to extract systematic and generalized information from large corpora, which could provide strong evidence instead of being dependent only on linguistic intuition (Fontenelle, 1994). Fontenelle (1994) also pointed out that possible collocated elements can be found through the calculation of the collocated words that appear before or after a head. If the occurrence frequency of two words is higher than expected, they can be considered a significant combination, or a collocation. The procedure included analyzing frequency, the window size of collocated elements, POSs, and the syntax. Lin (1998) also analyzed a combination consisting of a head, its collocated element and modifiers from a text of 100-million words to find the probability of their co-occurrence.

In speaking of statistical methods, Butler (1998) indicated three ways to measure Spanish collocations, observed or expected values, T-value and types, or tokens. In addition, McCarthy et al. (2003) suggested three statistical methods, chi-square, mutual information, and the log-likelihood ratio (LLR). A widely cited study by Manning & Schütze (1999) indicated that an $\chi 2$ (chi-square) could be used to modify the shortcoming of the t-test, in which it is assumed that the population is under a normal distribution. The $\chi 2$ can be used to examine whether the co-occurrence of two collocated words is within the confidence level. If not, the words are not considered a collocation. On the other hand, it was found that using mutual information (MI) could help in determining frequency independence. However, some errors might occur with respect to the confidence level; low-frequency words were shown to have higher scores than high-frequency words. Fontenelle (1994) also argued that the results generated through a computational program making use of complicated statistical methods to identify collocations sometimes turn out to provide non-collocations for linguists. Therefore, our study adopted the approach of generating collocations through statistical tests, checking with references and manual inspection for further selection and modification of the generated collocation list.

Finally, we evaluated two existing tools: Sketch Engine (Kilgarriff et al., 2014) and Corpus del Español (Davies, 2002-). One of the advantages of Sketch Engine is that its retrieval results display combinations with different parts of speech. Secondly, the examples of collocation can assist users with understanding of its usage. Thirdly, Sketch Engine provides a more advanced retrieval function, for instance, a comparison between two similar verbs with their collocated words. However, Sketch Engine is not free for public use, although there are four available corpora that can be selected. Furthermore, setting conditions for collocation retrieval is complicated. On the other hand, another tool with a collocation searching function, Corpus del Español, presents different ways of retrieving collocations. There are several advantages to using the Corpus del Español. It provides retrieval results according to frequency, which helps users easily select an appropriate collocation. The provided examples facilitate users' understanding of searched collocations and the retrieval results display collocation lists that are easy for users to understand. Nevertheless, users cannot import their own data for the search purpose.

## 3. Spanish Collocation Tool (SCT)

The "Spanish Collocation Tool (SCT)" was developed by the Web Mining and Multilingual Knowledge System (WMMKS) Laboratory in the Department of Computer Science and Information Engineering (CSIE) at the National Cheng Kung University (NCKU) in Taiwan.

### 3.1. Design of SCT

The SCT was written in C# programming language and was designed to offer flexibility with regard to setting window size and selecting different statistical methods for collocated elements.

To set the window size, one needs to enter a range of numbers, which refers to the distance between two collocated elements that are allowed to appear. This selection process is included in order to detect results, excluding possibilities of modifiers inserted between two words in a collocation pair. We experimented with different window sizes and found that fewer results were derived under a smaller window size setting, while larger a window size contained noisier information. However, for the different purposes of this study, a big window size range was set in order to observe more examples, followed up by manual checking to exclude the noisy information.

Furthermore, the calculation of statistical probability and correlation refers to a search for only significant pairs of two "closed" words. In the statistical design of the SCT, a chi-square estimation and mutual information were used to test whether the probability of two co-occurring elements in a combination was under the confidence level. We decided to adopt the chi-square and mutual information to calculate values of imported data, after having compared several different models of statistical methods related to collocation studies in the areas of computational linguistics and corpus linguistics through extensive research of previous studies (i.e., Manning & Schütze, 1999). In addition, the SCT software can compare two collocation lists in order to detect any significant difference between them by calculating the mutual information (MI) or chi-square ($\chi 2$) scores for each collocated pair with relative entropy (also called the Kullback-Leibler divergence). Positive and high scores in the formulas indicate a contrastive difference between the two imported corpora.

Instead of searching only for collocated pairs of raw data (i.e., words), the grouping of different parts of speech and words of the same lemmas by the SCT can provide more systematic and generalized results that will facilitate

the typological classifications used for further research analysis. Therefore, the SCT was designed to undertake the procedure of POS-tagging and lemmatization in order to develop more informative functions. We evaluated several POS taggers and found the Tree Tagger (a program developed by Helmut Schmid at the University of Stuttgart, http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php) much more efficient in comparison with other tools in terms of processing speed. Thus, it was adopted to POS-tag imported data. The Spanish tagger and lemmatizer system were preinstalled in the internal program of the SCT and could be triggered automatically for the convenience of usage. Therefore, the SCT allows searches for combinations of words, parts of speech, and lemmas. In addition, the contexts where these collocation pairs appear can be retrieved in order to obtain complete information about the usage of learners and native speakers in a corpus for further analysis.

### 3.2. Interface of results

In the interface (Figure 1), a collocation list appears on the left side of the screen. Each collocation consists of Word 1 and Word 2, of which prototypes are Lemma 1 and Lemma 2, respectively. POS1 and POS2 are their parts of speech. The Count represents the frequency at which a collocation appears in the imported data. The Score is the result of the kernel methods. W1Count and W2Count represent the frequency of occurrence of the collocated elements in the imported data.

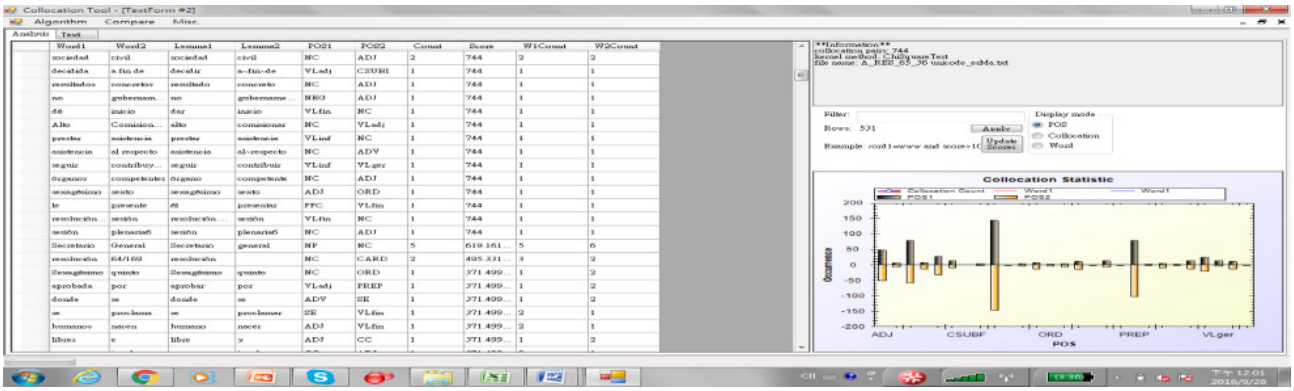



**Figure 1.** Presentation of result interface.

After viewing the first collocation result output, the user can set further required condition(s) in the "Filter" area and then click "Apply" to restrict the output and obtain a more systematic result for posterior analysis.

Examples of possible submitted commands:

lemma2=civil and score>2
word1count>2 or word2count>2
Pos1=NC
Pos1=NC and Pos2=NP and word1= sociedad

Then, the user must have two collocation lists ready to be compared. The result of comparing two collocation lists will appear in another window. According to the score sequence, the higher the absolute value of a score is, the more significant the difference between the same collocation in two imported corpora is.

Although the SCT is free for public use and can derive collocation lists of not only words but also parts of speech and lemmas, it still has certain limitations. For example, the retrieval collocations can only consist of two collocated elements. Therefore, we hope to extend the query function from a bi-gram to an N-gram collocation in order to cover a bigger range of collocated elements for studies. We have developed an error detection and revision suggestion system, and we combined the developed result with "Spanish Collocation" to enrich its assisting functions and applications. By taking advantage of combined functions, related research can be expanded to a larger scope.

## 4. The application of the Spanish Collocation Tool in teaching and learning

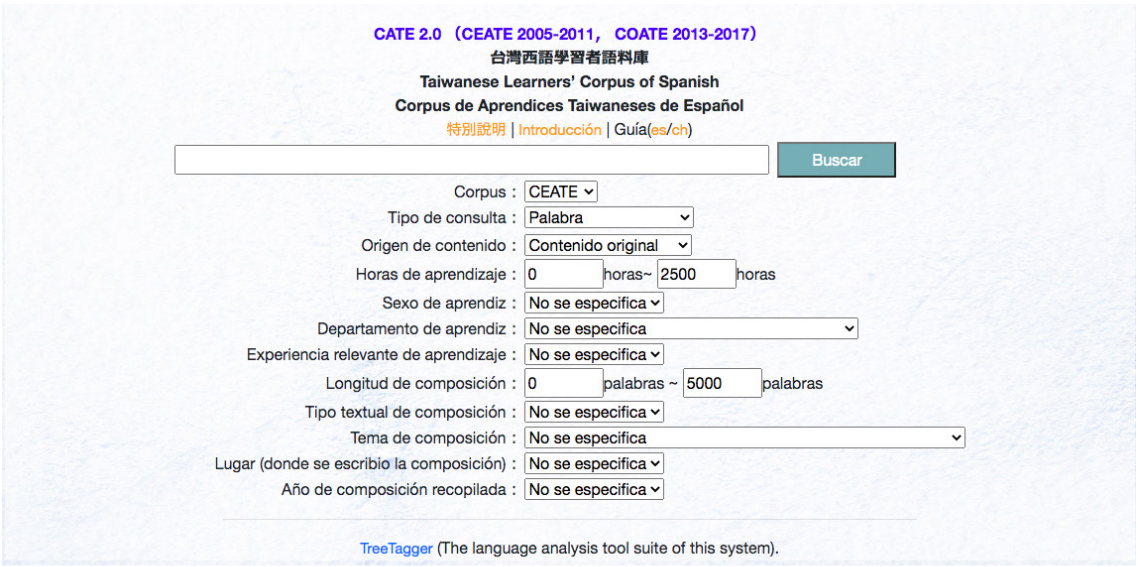
This section focuses on SCT applications in a corpus-based study and in on-line teaching material.

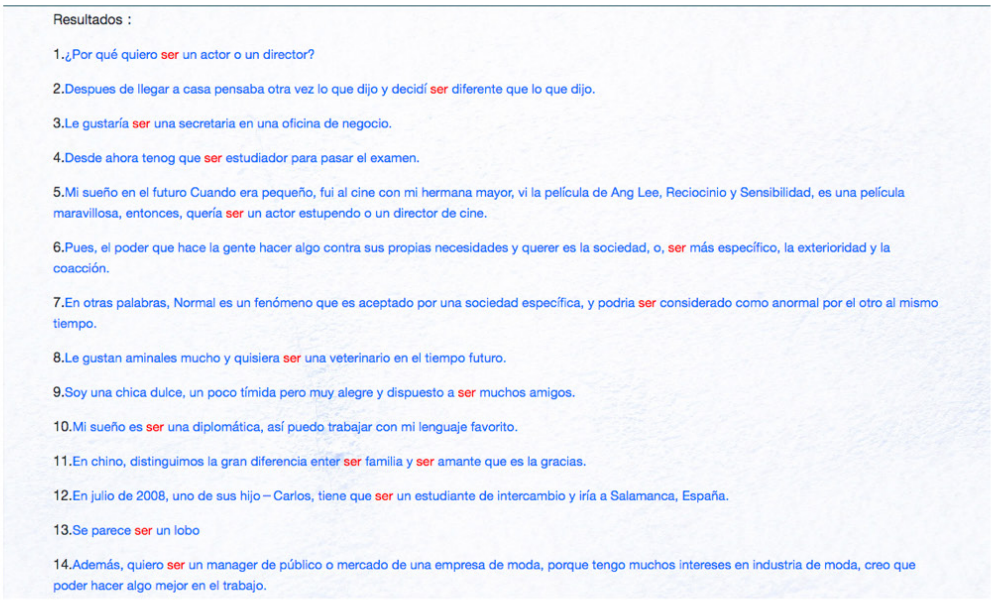
### 4.1. Corpus-based study

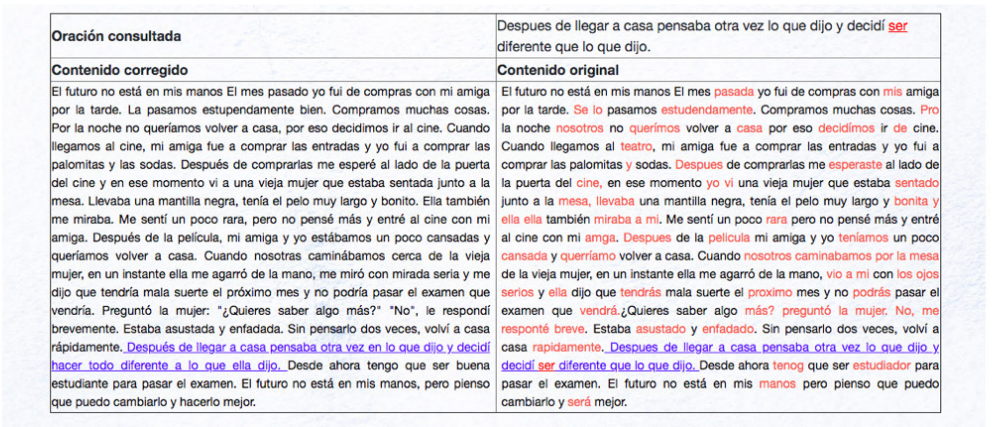
Firstly, we demonstrate the application of the SCT in a corpus-based collocation study by analyzing L3 learner data. This study was intended to examine the collocation of learned uses by analyzing the data compiled in an annotated L3

learners' corpora called Taiwanese Learners' Written Corpus of Spanish (Corpus Escrito de Aprendices Taiwaneses de Español, CEATE, Figure 2) as well as by comparing the data of learner-written texts with their revised texts corrected by native speakers of Spanish. The comparison of the two types of texts was intended to contrast the distance between the interlanguage of L3 Spanish learners and the uses in the target language.

The learners' written corpus (CEATE) consisted of texts written by those who studied Spanish as their third language (L3) after learning English as their second language (L2) and Mandarin-Chinese as their mother tongue (L1). This learner corpus was POS-tagged and error-correction annotated. The corpus has been open for public search since the end of September 2009 and can be accessed at http://140.116.245.146/cate_searchpage/search.php



| **Oración consultada** | Despues de llegar a casa pensaba otra vez lo que dijo y decidí ser diferente que lo que dijo. |
|---|---|
| **Contenido corregido** | **Contenido original** |
| El futuro no está en mis manos El mes pasado yo fui de compras con mi amiga por la tarde. La pasamos estupendamente bien. Compramos muchas cosas. Por la noche no queríamos volver a casa, por eso decidimos ir al cine. Cuando llegamos al cine, mi amiga fue a comprar las entradas y yo fui a comprar las palomitas y las sodas. Después de comprarlas me esperé al lado de la puerta del cine y en ese momento vi a una vieja mujer que estaba sentada junto a la mesa. Llevaba una mantilla negra, tenía el pelo muy largo y bonito. Ella también me miraba. Me sentí un poco rara, pero no pensé más y entré al cine con mi amiga. Después de la película, mi amiga y yo estábamos un poco cansadas y queríamos volver a casa. Cuando nosotras caminábamos cerca de la vieja mujer, en un instante ella me agarró de la mano, me miró con mirada seria y me dijo que tendría mala suerte el próximo mes y no podría pasar el examen que vendría. Preguntó la mujer: "¿Quieres saber algo más?" "No", le respondí brevemente. Estaba asustada y enfadada. Sin pensarlo dos veces, volví a casa rápidamente. Después de llegar a casa pensaba otra vez en lo que dijo y decidí hacer todo diferente a lo que ella dijo. Desde ahora tengo que ser buena estudiante para pasar el examen. El futuro no está en mis manos, pero pienso que puedo cambiarlo y hacerlo mejor. | El futuro no está en mis manos El mes pasada yo fui de compras con mis amiga por la tarde. Se lo pasamos estudendamente. Compramos muchas cosas. Pro la noche nosotros no queríamos volver a casa por eso decidímos ir de cine. Cuando llegamos al teatro, mi amiga fue a comprar las entradas y yo fui a comprar las palomitas y sodas. Despues de comprarlas me esperaste al lado de la puerta del cine, en ese momento yo vi una vieja mujer que estaba sentado junto a la mesa, llevaba una mantilla negra, tenía el pelo muy largo y bonita y ella ella también miraba a mi. Me sentí un poco rara pero no pensé más y entré al cine con mi amga. Despues de la pelicula mi amiga y yo teníamos un poco cansada y querríamo volver a casa. Cuando nosotros caminabamos por la mesa de la vieja mujer, en un instante ella me agarró de la mano, vio a mi con los ojos serios y ella dijo que tendrás mala suerte el proximo mes y no podrás pasar el examen que vendrá.¿Quieres saber algo más? preguntó la mujer. No, me responté breve. Estaba asustado y enfadado. Sin pensarlo dos veces, volví a casa rapidamente. Despues de llegar a casa pensaba otra vez lo que dijo y decidí ser diferente que lo que dijo. Desde ahora tenog que ser estudiador para pasar el examen. El futuro no está en mis manos pero pienso que puedo cambiarlo y será mejor. |

**Figure 2.** CEATE interface.

### 4.1.1. Methodology

To avoid noisy information, we extracted texts using the following criteria in order to obtain consistent data characteristics: a passage of 100-200 Spanish words in length, a textual description type, and a theme related to leisure life and daily routines. To observe the collocation usage in Spanish, we focused on learners at the intermediate level with instruction of 576 to 1,088 hours, excluding learners with special backgrounds (such as immigrants, exchange, or transfer students). Ultimately, around 36,000 words in Spanish were analyzed, among which there were 17,914 words and 17,563 words from original and revised texts, respectively.

When importing data, we first fixed the maximum window size at 2 (the range between two words). Then we set the following conditions for the results of the collocation that we intended to obtain for further analysis. For example, with the purpose of observing two types of collocation, lexical and grammatical, the first group consisted of collocation types such as N-Adj/Adj-N, V-N, Adv-Adj, and V-Adv/Adv-V. The second group included the following types, N-Prep/Prep-N, Adj-Prep/Prep-Adj, V-Prep/Prep-V, and V-V. Then, texts containing collocated pairs were retrieved for further detailed analysis. Since it was unavoidable that POS-tagging of learners' texts would lower the percentage of correctness, we checked the POS-tagged results manually to maintain the basic quality of POS-tagging in the final step of extracting data.

By examining the compared results of the original texts written by learners and the texts corrected by native speakers of Spanish through the SCT, we were able to derive lists related to learners' overuse and underuse of collocation. The results of collocation overuse showed that combinations appeared more in the original text than in the revised version. In classroom teaching, the results suggested that the items of overuses should be discouraged as learners learn to use them. On the other hand, the underused items on the collocation list should be stressed in teaching. We considered not only the scores (high chi-square score) but also the counts (≥ four times) of collocated pairs to exclude any collocation pairs that appeared only once but had high value. In addition, the result lists with high positive values of relative entropy (KL values) were considered of importance and they should be stressed in teaching and learning.

### 4.1.2. Results

In order to find out the correct uses of different combinations, we derived a lexical collocation scale consisting of N+Adj>V+Adv>Adj+N>V+N among the texts written by learners at the intermediate level. That is, the combination of N+Adj shows the highest accuracy rate, whereas the V+N combination has the lowest rate among all combinations. We also developed the following hierarchical order of grammatical collocation uses by the learners at the intermediate level: V+P>Adv+P>V+V. If we consider 50% of similarity as the threshold for stable development, we could draw the conclusion that N+Adj might have been developed at the beginning level, whereas the combination of Adj+N enters a more stable stage at the intermediate level; the combination of V+N, is still on the way to be developed at a later stage. These observations may imply an order of priority for the design of instructional materials.

With respect to the structure of V+N, combinations of "solucionar problema, tomar sol" were found at the intermediate level. For the combination of Adj+N, we could find that learners at the intermediate level might either know how to distinguish the pre-noun adjective from the post-noun—for example, "buen/mal humor, nuevos amigos," or they were already familiar with these fixed expressions. In the N+Adj combination, we observed many examples related to abstract concepts at the intermediate level, for example, "experiencia especial, comercio exterior".

**Table 1.** Examples of collocations.

| | **Intermediate level** |
|---|---|
| V-N | escuchar música, solucionar problema, tomar sol, leer libros, leer revistas, estudiar inglés, aprender lenguaje |
| N-Adj | moto acuática, mercado nocturno, orejas iguales, ojos pequeños, universidades diferentes, comercio exterior, ciudad acogedora, experiencia especial, hospital católico, cosa orgullosa, chica normal, cara pequeña, hermana mayor, asignatura principal, oficina exterior, persona trabajadora, estudio español, escuela secundaria, estudios primarios, lenguaje nuevo |
| Adj-N | mala nota, buenas notas, misma forma, gran interés, pequeño pueblo, mejor amigo, grande boca, mismo carácter, buen humor, pequeña comunidad, mal humor, mucha gente, mismo año, nuevos amigos |

By applying our new software to complete the above referenced corpus-based contrastive research between learners' written texts and native speakers' revised texts, the derived collocation lists can provide a clearer direction for designing pedagogical materials for the learning of collocation.

**Table 2.** Results of the comparison between learners' original and revised texts.

| Collocation | | Texts | Example | KL value | Count | χ2 value | Word 1 | Word 2 |
|---|---|---|---|---|---|---|---|---|
| Lexical | V-N | Original | tengo pelo | 7.439138E-05 | 5 | 66.3197 | 120 | 39 |
| | | Revised | tengo tiempo | -6.576873E-05 | 6 | 57.0643 | 140 | 54 |
| | N-Adj | Original | tiempo libre | -0.002388381 | 19 | 4693.08 | 45 | 24 |
| | | | pelo largo | -0.0007633751 | 12 | 3302.57 | 36 | 17 |
| | | | hermana mayor | 0.0003415697 | 11 | 4499.16 | 18 | 21 |
| | | Revised | tiempo libre | 0.003937931 | 23 | 7116.01 | 44 | 24 |
| Grammatical | V-V | Original | gusta hacer | -0.0006338776 | 8 | 232.019 | 102 | 36 |
| | | Revised | gusta ir | 0.003447111 | 10 | 400.337 | 114 | 30 |

With the SCT and extracted L3 data, we were able to obtain a more conclusive generalization of collocation usage for a specific level of Taiwanese learners of Spanish.

## 4.2. Teaching material

Furthermore, the implications of the search model presented in the previous sections and the generalization of usage tendency of Taiwanese learners of Spanish obtained through the search results by using the SCT led to the development of on-line multimedia material specifically for learning Spanish collocations (Figure 3).

Los recuerdos con mi familia

Para mi familia[1], es importantísimo[2] viajar juntos cada año. En las vacaciones del año nuevo, mis padres, mi hermano y yo, hicimos un pequ viaje[3] por Taiwán. Estábamos muy emocionados, ya que era la primera vez que visitábamos el este de Asia. Como nos quedamos poco tiempo, visitamos Taipei, que es la capital[4] de Taiwán. Conocimos muchos sitios, por ejemplo: El Taipei 101, el Museo del Palacio Nacional y el Parque Nacional Yangmingshan. Este viaje a Taiwán es un lindo recuerdo para nosotros.

Nuestro primer destino[1] fue el Taipei 101. Taipei 101 tiene 508 metros de altura y es uno de los lugares[2] en los cuales, se da uno de los mejo espectáculos de fuegos artificiales para la celebración de Año Nuevo. A muchas personas les gusta quedarse cerca del Taipei 101 para ver los artificiales y hacer juntos la cuenta regresiva para recibir el nuevo año. Al bajarnos del avión, tomamos el metro para ir al Taipei 101, ya que n queríamos perdernos el espectáculo. Sacamos muchas fotos[3] de los fuegos artificiales[4]. Como todavía estamos desorientados, por la diferenci horario, una vez que terminó el espectáculo, regresamos al hotel para dormir antes de continuar con el resto de nuestro viaje.

Palabras clave

| viaje | |
|---|---|
| 1. | El paisaje más famoso es la Montaña Ali. Especialmente en marzo, está todo muy bonito, porque las flores brotan en ese momento. Es una muy buena elección que hagáis un viaje a Chayi. ¡Buen viaje! |
| | 阿里山是最有名的風景區。尤其在3月，因爲所有的花都在那個時間盛開，一切都美不勝收。去嘉義旅行是個很棒的選擇。祝你旅途愉快！ |
| 2. | Hace cinco años hice un largo viaje a Irak con mi mascota, una pulga, se llama Colón. |
| 3. | Así que quería hacer un viaje maravilloso para relajarme. |

| fotos | |
|---|---|
| 1. | En las vacaciones de invierno pasadas, mi hermano mayor y yo fuimos a una playa que no supimos cómo se llamaba. Como nuestros padres tenían que trabajar, no fueron con nosotros. Las vistas allí eran muy bonitas. Al otro lado de la playa había una montaña muy alta y grande. Tomamos muchas fotos. Allí había un grupo de jóvenes que estaban asando carne y pescado. Nosotros comimos y bebimos algunos refrescos con ellos. Ellos empezaron a charlar con nosotros y así nos conocimos, ahora nos llevamos muy bien y nos hicimos buenos amigos. |
| | 上個寒假，我和我哥哥去一個不知名的海灘。因爲我們的父母要工作，所以沒有去。那裡的風景很漂亮。海灘的另一邊有座又高又大的山。我們拍了很多照片。那裡有一群年輕人在烤肉。我們和他們一起吃烤肉和、喝飲料。他們開始與我們聊天，我們因此而認識，也相處得很融洽，就這樣成爲了很好的朋友。 |
| 2. | La capital de Irak, Bagdad, era muy hermosa. La plaza tenía una estatua de Sadam Housein y me tomé unas cuantas fotos con ella, porque me parece que él fue muy atrevido al oponerse a los EE. UU.. |
| 3. | Durante el viaje en tren, charlamos, sacamos fotos y nos comimos los bocadillos. |

**Figure 3.** Spanish on-line collocation learning course.

It begins with a text entitled "Los recuerdos con mi familia" by integrating ten keywords that were intended to be included in the instruction. In addition to the connected videos and the vocabulary list, further information related to their collocated pairs was included in hyperlinks. Take two keywords, "viaje" and "foto", as examples. In the learning material, different types of collocations such as N-Adj/Adj-N ("buen viaje, largo viaje, viaje maravilloso") and V-N ("hacer viaje, tomar fotos, sacar fotos") can be demonstrated based on the text extracted from learner data and modified by native speakers of Spanish. Furthermore, the context (i.e., sentences) where these collocated elements can appear and their correspondent translations in Chinese, the mother tongue of our Taiwanese learners, can be served as a learning option.

The sample of collocation teaching materials based on the development of the SCT and its application in a corpus-based contrastive study served as a prototype for designing more related pedagogical material. In the future, other references to the native corpus, Corpus del Español (Davies, 2002) can be added to obtain a more generalized view of the natural language by examining the frequency and tendency toward native usages.

## 5. Conclusions

In light of the current trend in corpus linguistics research, we have developed a Spanish Collocation Tool (SCT) to facilitate the analysis of collocations for Spanish. In comparison with other existing tools, the SCT is a downloadable free service for selecting statistical methods, setting window sizes, searching queries including POS-tagged and lemma information, and viewing source texts.

The present study also combined the corpus tool, SCT, and a constructed L3 Taiwanese student-written corpus of Spanish, CEATE, to create efficient access to results in a systematic approach. By using the SCT in the study, we derived a learning sequence and found patterns in learners' collocation uses in contrast to those used by native speakers.

Furthermore, by using the SCT, the pedagogical implications of the search results and generalization of usage inclination for Taiwanese learners of Spanish led to the development of on-line multimedia material for learning Spanish collocations. The provided sample served as a prototype for designing effective pedagogical materials to facilitate Taiwanese learners' L3 learning.

Finally, it is hoped that researchers and teachers with the same interests can benefit by applying our software in data analysis and in teaching material design.

## Acknowledgments

We are grateful for the grant for the research project MOST103-2410-H-006-059-MY2. Additionally, we appreciate the assistance provided by professor Wen-Hsiang Lu for his expertise in computational linguistics. Also, we would like to express our sincere gratitude to our research assistant, Ting-Xuan Wang.

## References

Aguilar-Amat Castillo, A. (1993). En torno a la combinatoria del léxico: Los conceptos de colocación e idiomatismo. In *Lenguajes naturales y lenguajes formales: Actas del IX congreso de lenguajes naturales y lenguajes formales* (pp. 267-272). Promociones y Publicaciones Universitarias, PPU.

Almela Sánchez, M. (2006). *From words to lexical units. A corpus-driven account of collocation and idiomatic patterning in English and English-Spanish* (Vol. 35). Frankfurt: Peter Lang.

Alonso Ramos, M. (1995). Hacia una definición del concepto de colocación: de J. R. Firth I. A. Melcuki. *Revista de Lexicografía*, *1*, 9-28.

Butler, C. S. (1998). Collocational frameworks in Spanish. *International Journal of Corpus Linguistics*, *3*(1), 1-32.

Castillo Carballo, M. A. (1998). El término colocación en la lingüística actual. *LEA*, *20.1*, 41-54.

Corpas Pastor, G. (1992a). Las colocaciones como problema en la traducción actual (inglés/español). *Revista del Departamento de Filología Moderna*, *2/3*, 179-186.

Corpas Pastor, G. (1992b). Tratamiento de las colocaciones del tipo a+s/s+a en diccionarios bilingües y monolingües (español-inglés). In *Proceedings of the EURALEX'90 Actas del IV Congreso Internacional*, pp. 331-340.

Davies, M. (2002-). *Corpus del Español: 100 million words, 1200s-1900s*. Available online at http://www.corpusdelespanol.org.

Fontenelle, T. (1994). What on earth are collocations? An assessment of the ways in which certain words co-occur and others do not. *English Today*, *10*(04), 42-48.

Handl, S. (2008). Essential collocation for learners of English-The role of collocational direction and weight. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 43-66). John Benjamins Publishing.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly*, *37.3*, 467-870.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. *(2014).* The Sketch Engine: ten years on. In *Lexicography*: 1–30. Available online at https://www.sketchengine.co.uk/

Kita, K. & Ogata, H. (1997). Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning*, *10.3*, 229-38.

Lee, D. (2010). *Bookmarks for corpus-based linguistics.* http://www.uow.edu.au/~dlee/CBLLinks.htm

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

Lin, D. (1998). Extracting collocations from text corpora. Paper presented at the *First workshop on computational terminology.*

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

Marcinkevicine, R. (1995). Collocation: The object of analysis, aspects, and methods. *Lituanistica*, *2.22*, 40-54.

McCarthy, D., Keller, B., & Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In Proceedings of the *ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment-Volume 18*.

Navarro, C. (2003). Didáctica de las unidades fraseológicas. In M. V. Calvi, F. San Vicente & M. Baroni (Eds), *Didáctica del léxico y nuevas tecnologías* (pp. 99-115).

Pu, Jianzhong. (2003). Colligation, collocation, and chunk in ESL vocabulary teaching and learning. *Foreign Language Teaching and Research*, *35.6*, 438-445.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of *International Conference on New Methods in Language Processing*. Manchester, UK.

Sun, Y.-C. & Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16(1)*, 83-94.