Evaluation of Corpus-Assisted Spanish Learning

Hui-Chuan Lu
FLLD, NCKU / No. 1 University Road
701 Tainan, Taiwan
huichuanlu1@gmail.com

Yu-Hsin Chu
FLLD, NCKU / No. 1 University Road
701 Tainan, Taiwan
katy0806@gmail.com

Abstract

In the development of corpus linguistics, the creation of corpora has had a critical role in corpus-based studies. The majority of created corpora have been associated with English and native languages, while other languages and types of corpora have received relatively less attention. Because an increasing number of corpora have been constructed, and each corpus is constructed for a definite purpose, this study identifies the functions of corpora and combines the values of various types of corpora for auto-learning based on the existing corpora. Specifically, the following three corpora are adopted: (a) the *Corpus of Spanish*; (b) the *Corpus of Taiwanese Learners of Spanish*; and (c) the *Parallel Corpus of Spanish, English, and Chinese*. These corpora represent a type of native, learner, and parallel language, respectively. We apply these corpora as auxiliary resources to identify the advantages of applying various types of corpora in language learning from a learner's perspective. In the environment of auto-learning, 28 participants completed frequency questions related to semantic and lexical aspects. After analyzing the questionnaire data, we obtained the following findings: (a) the native corpus requires a more advanced level of Spanish proficiency to manage ampler and deeper context; (b) the learners' corpus facilitates the distinction between error and correction during the learning process; (c) the parallel corpus assists learners in connecting form and meaning; (d) learning is more efficient if the learner can capitalizes on specific functions provided by various corpora in the application order of parallel, learner and native corpora.

1 Introduction

The trend of using corpus has expanded into all sub-areas of linguistics, including applied fields such as foreign language teaching and learning. According to Lee (2010), almost 360 corpora have been constructed for various purposes in 57 languages. Sixty-three percent of these corpora have been analyzed in previous research on language analysis and English teaching. In the past decade, the majority of corpus users have been researchers and teachers. Therefore, we are interested in extending the usage of corpus to foreign language learners, and studying how the perspective of corpus application can benefit these learners. Moreover, instead of English, we have selected Spanish as the target language of this research because the popularity of second foreign language acquisition is increasing in Taiwan, and multilingualism has become a novel research topic in applied linguistics.

Among the related literature, the application of existing corpora in teaching or learning has focused primarily on native corpus. Moreover, although there have been several studies on parallel corpus, very few have examined learners' corpus. The reason that less attention has been drawn to the evaluation of effectiveness might be attributable to the lack of access to parallel and learners' corpora. Moreover, to our knowledge, no study has compared the various types of corpora. The discussed reasons have motivated us to conduct this research. This study examines the advantages and disadvantages of the three types of corpora from the learners' perspective, and applies them complementarily to maximize the learning outcomes.

By applying extant sources, language learners can learn how to apply created corpora for the self-learning of foreign languages. As the final goal, we hope that learners can capitalize on the

27th Pacific Asia Conference on Language, Information, and Computation pages 467－473

complementary merits of various types of corpora to achieve the best results, and maximize the efficiency of their learning through the application of information technology.

2 Literature review

With the era of information technology, the corpus approach has developed rapidly over the past four decades. The first milestone of corpus research can be traced back to Kucera and Francis (1967). They constructed the Brown Corpus, which comprised one million words of modern American English. Thereafter, the interest in the study of corpus linguistics has increased over time. Kennedy (1998) stated that the corpus approach has been employed for linguistic analyses by collecting and organizing data. According to the sub-database of Proquest, *Linguistics and Language Behavior Abstracts* (*LLBA*) had exhibited an increasing publication rate from 1970 to 2010. For example, we entered "corpus" as keyword to obtain the distribution of publications during the 1970s (588 publications), 1980s (1,365 publications), 1990s (4,452 publications), and 2000s (10,886 publications). Lee (2010) indicated that the various corpus types include diachronic, contemporary, native, learner, specialized, web, monolingual, multilingual, parallel, spoken and annotated, and multimedia corpora, among others.

Focusing on the target language of Spanish, *Reference Corpus of Current Spanish* and *Corpus of Spanish* are two well-known Spanish corpora of Hispanic native speakers. Howe and Ranson (2010) and Lavid, Arús, and Zamorano-Mansilla (2010) applied native corpus by extracting and analyzing the data from both of these corpora for different linguistic purposes. Howe and Ranson (2010) analyzed temporal modifiers in Spanish, whereas Zamorano–Mansilla (2010) contrasted Spanish grammar usage with English. Although previous studies have utilized existing corpora for research; investigations on the application of corpus to facilitate language learning are scarce. Therefore, we selected *Corpus of Spanish* because it has rich data and offers powerful search functions, as one of the linguistic resource to evaluate the effectiveness of using this corpus for assisting learning.

Different from native corpus as *Corpus of Spanish*, the learners' corpus, which is the collection of production of foreign language learners has its distinguished characteristics. Granger, Kraif, Ponton, Antoniadis, and Zampa (2007) indicated the help of error-tagged learners' corpora in both teaching and learning languages. Gilquin, Granger, and Paquot (2007) emphasized the importance of learners' corpora in English for academic writing purposes. A variety of data can be drawn from learners' corpus to discover leaner-specific patterns such as lexical, grammatical, wording and reliance, etc. Teachers and researchers can identify the tendency of the language usage of learners through corpus. Mukherjee (2008) showed that learners should take advantage of the resources of the learners' corpora. Dalziel and Helm (2008) indicated that the learners' corpora can guide learners through self-inquiry. These studies confirmed the positive value of utilizing the learners' corpora. However, few empirical studies have provided concrete evidence to prove its effectiveness in assisting learning. *L2 Spanish Written Corpus* and *Spanish Learner Language Oral Corpora* are representative of two learners' corpora of Spanish. Both collected data from learners whose native language is English. However, the *L2 Spanish Written Corpus* is not available to the public, whereas *Spanish Learner Language Oral Corpora* only contains spoken data. Therefore, we applied our constructed learners' corpus to research taking the learners' background and resource availability into consideration.

Moreover, using parallel corpus as a reference database is beneficial for contrastive analysis, translation study and language learning (Baker, 1993; Malmkjaer, 2005). Zhang, Wu, Gao and Vines (2006) suggested that parallel corpora can be used for various purposes, such as cross-language information retrieval, and data-driven natural language processing systems. Because Spanish is the target language and Chinese as the first language of our learners, we required a parallel corpus containing Spanish and Chinese. Although Spanish–English or English–Chinese parallel corpora could be found, we could not locate a Spanish–Chinese parallel corpus for us to employ before we dedicated to its construction.

Consequently, in this paper, besides (a) the *Corpus of Spanish*, we introduce (b) the *Corpus of Taiwanese Learners of Spanish*, and (c) the *Parallel Corpus of Spanish, English and Chinese*. Furthermore, we compare the effectiveness of their utilization as assistant resource for language learning. By investigating various types of corpora, this study answers the following

research questions: (a) by comparing three types of corpora, what are the advantages and disadvantages of each corpus from the users' point of view? (b) by combining three types of corpora, how do they complement each other to obtain the optimum learning result?

3 Methodology

3.1 Participants

Twenty-eight Taiwanese learners of Spanish who studied in the Department of Foreign Languages and Literature participated in the survey. Their mother tongue is Chinese, and English and Spanish were learned as their first and second foreign languages, respectively. They learned Spanish in a classroom for 300 to 400 hours, and *Dos Mundos* was used as the textbook for learning Spanish in a classroom environment. The Wisconsin Placement Test was administered to identify the Spanish proficiency level of all participants. Table 1 shows the characteristics of the participants.

Type				
Year	Third year		Forth year	
	20 (71%)		8 (29%)	
Sex	Female		Male	
	23 (82%)		5 (18%)	
Profic .level	1	2	3	4
	0 (0%)	19 (68%)	8 (29%)	1 (3%)

Table 1. Characteristics of Participants.

3.2 Instruments

The following three corpora were adopted as assisting resources; (a) the *Corpus of Spanish*, (b) the *Corpus of Taiwanese Learners of Spanish* and (c) the *Parallel Corpus of Spanish, English, and Chinese*; that represent a type of native, learners and parallel language, respectively. The first one was created by Mark Davies of BYU, and the other two were constructed by the National Cheng Kung University (NCKU) team in Taiwan.

The *Corpus of Spanish* ("Corpus del Español" in Spanish, CdE) comprises 100 million words. The powerful search functions of the corpus such as lemma and collocation surpass other available native corpora of Spanish. We set data of the year 1900 as our source for users' searches to obtain more contemporary data.

.

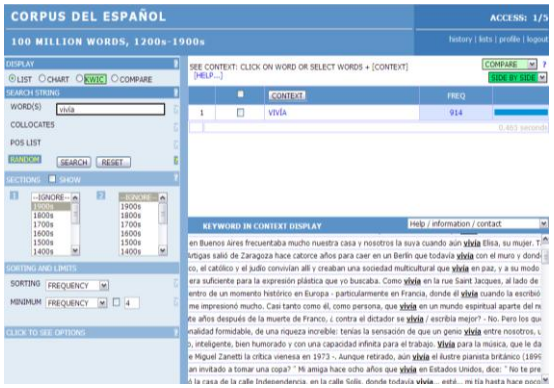

Figure 1. The Interface of CdE.

The second corpus is the *Written Corpus of Taiwanese Learners of Spanish* ("Corpus Escrito de Aprendices Taiwaneses de Español", CEATE). It was created by the NCKU corpus team in 2005, and contains 2,425 texts, and approximately 446,694 words. It was POS-tagged and corrections were added for every error made in the learners' version. For the questionnaire, "revised compositions" were chosen as a condition set for users' searches.

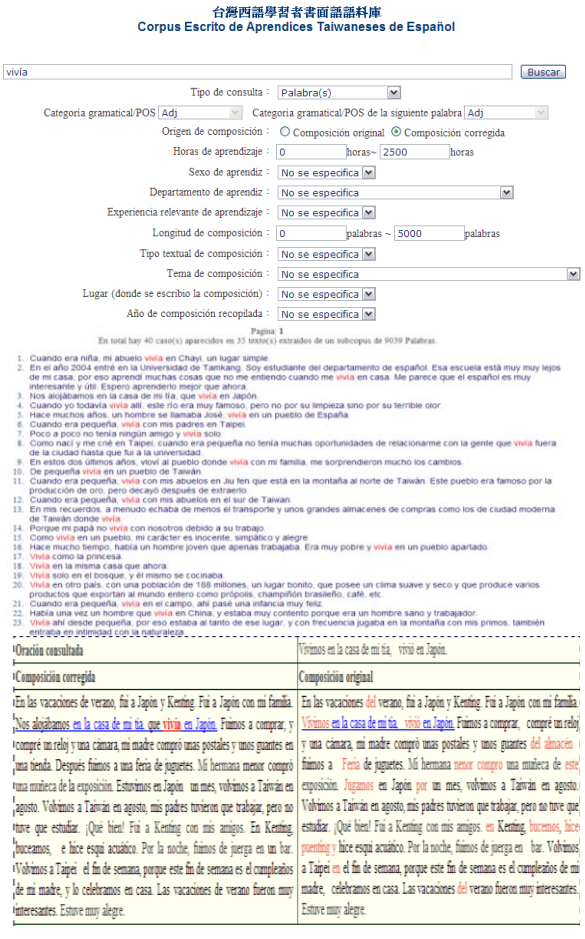

Figure 2. The Interface of CEATE.

The third corpus is the *Parallel Corpus of Spanish, English, and Chinese* ("Corpus Paralelo de Español, Inglés y Chino", CPEIC). It was constructed by the NCKU corpus team in 2012. A tri-lingual parallel corpus contains written data from the Bible and various fairy tales, with 755,461 words in Spanish, 794,571 words in Engish, and 923,509 words in Chinese. Data of Spanish, English, and Chinese were individually POS-tagged and word-aligned among these three languages. Searches can be conducted by setting single or multiple keywords of various languages, and their part of speech. From the search result, it can be observed that the syntactic and lexical contrasts of parallel meanings among them.

Figure 3. The Interface of CPEIC.

3.3 Exercise and evaluation

To ensure that participants were familiar with the search functions of various corpora, they practiced with an exercise prior to the formal evaluation. In the exercise, participants were required to do at home a similar practice (Appendix A) in which eight pairs of words were listed to be differentiated and selected according to their frequency of usage. These questions can be classified into the following two groups: (a) past tense, preterit or imperfect: *vivió/vivía* `lived´, *comió/comía* `ate´, *preguntó/preguntaba* `asked´, *murió/moría* `died´; and (b) copular verbs SER or ESTAR ´to be´ with the adjectives: *ser/estar possible* ´to be possible´, *ser/estar feliz* `to be happy´, *ser/estar limpio* `to be clean´, *ser/estar enamorado* `to fall in love´. Finally, the participants needed to evaluate different corpora in a questionnaire with open questions after experiencing the practice process for each question.

One week later, in the classroom, participants were limited to 45 minutes to finish evaluating these corpora through searching seven pairs of words that appeared in the formal evaluation. As those questions listed in the exercise, these questions were grouped into two categories: (a) past tense: *hubo/había* `there was´ + *N*, *fui/iba a* + *destino* `went to + destination´, *dijo/decía* `said´, *llegó/llegaba* `arrived´; and (b) copular verbs SER or ESTAR ´to be´ with adjectives: *ser/estar conveniente* `to be convenient´, *ser/estar seguro* `to be sure´, *ser/estar contento* `to be glad´. Upon completion, the survey participants were asked to evaluate three corpora by contrasting their advantages and disadvantages.

The pairs of words used in this exercise and the formal evaluation were selected based on the frequency of search result from *Corpus of Spanish*. Two specific categories, past tense and copular verb, were included in the exercise and evaluation because both are difficult for learners to distinguish the two similar elements of each pair according to our teaching experience. Moreover, a contrast exists among the three languages; that is, there are two copular verbs (SER/ESTAR) in Spanish, one (BE) in English, and none in Chinese. The same occurs for past tense. Two (preterit and imperfect) in Spanish, one in English, and zero in Chinese. And these three languages are target language (L3), first foreign language (L2) and mother language (L1) of our participants respectively.

Compared with English learners, the number of Spanish learners is relatively less in Taiwan. Moreover, a complete exercise and training program for using the corpus tools should be addressed to participants before the formal evaluation. Furthermore, although only seven or eight questions were listed in the exercise and the formal evaluation, each question took a participant at least five minutes to complete the search activity, fill the result, and write down the user experience. Hence, considering these limitations, we only had two Spanish classes with a total number of 28 students from the same university for this preliminary study of evaluation work covering only two Spanish grammatical categories.

4 Results and discussion

4.1 Exercise and evaluation

Tables 2 and 3 show the search results and user satisfaction, respectively.

Q	CdE	CEATE	CPEIC
	indefinido/imperfecto		
1	*había* (100%)	*había* (96%)	*había* (53%)
2	*iba* (77%)	*fui* (100%)	*iba* (67%)
3	*dijo* (100%)	*dijo* (100%)	*dijo* (86%)
4	*llegó* (100%)	*llegó* (100%)	*llegó* (79%)
	SER/ESTAR		
5	*Ser* (100%)	*Ser* (100%)	*Ser* (100%)
6	*estar* (96%)	*ser* (100%)	*estar* (100%)
7	*estar* (100%)	*estar* (100%)	*estar* (100%)

Table 2. Search Results of Frequency.

The search result for the frequency shown in Table 2 indicates the inclination of high frequency usage in two related elements of one pair. From this table, we observe the similarities (Questions 1, 3, 4, 5, and 7) and differences (Questions 2 and 6) for usage inclination among the three corpora through the participants' search results. Learners' corpus seems to have different result from the other two types of corpora, the native and parallel corpora. The participants had the chance to understand that different results could be searched with distinct corpora used. Generally speaking three types of corpora would help, in different degrees, the distinction between two elements of each pair. All three corpora could provide information of sentence and paragraph levels for learners to obtain more details and lexical meanings to distinguish two elements of the same pair.

Q	CdE	CEATE	CPEIC
1	80%	92%	78%
2	75%	89%	76%
3	86%	93%	69%
4	87%	88%	83%
5	96%	100%	93%
6	90%	60%	91%
7	91%	100%	100%

Table 3. User Satisfaction.

Then, based on the search experience, the majority of participants (> 60%) consented that these three corpora, CdE, CEATE, and CPEIC, were useful in helping learners to gain linguistic knowledge, as shown in Table 3.

4.2 General evaluation

General evaluation regarding to three different types of corpora is shown in Table 4.

	Advantages	Disadvantages
CdE	* rich examples * POS and lemma tagging * frequency order	* difficult vocabulary and sentence structure
CEATE	* errors vs. correction * easy comprehension * context	* lack of diversification * insufficient examples
CPEIC	* three languages	* lack of diversification * insufficient examples * not applicable to daily usage (Bible)

Table 4. Advantages and Disadvantages of 3 Corpora.

To answer research Question 1, we discuss the advantages and disadvantages of each corpus. Through various powerful search functions, CdE provided numerous systematic examples for learners. However, overly complex functions and an excessive number of examples sometimes causes more difficulties and obstacles for learners.

CEATE facilitated the distinction of contrasting usages between two aspects of the past tense (preterit and imperfect) or two copular verbs (SER/ESTAR) through the errors made by students, and the correction revised by Spanish native speakers. However, limited examples could not cover the infinite possibilities of learning situations because of the arduous work of corpus creation.

CPEIC was especially helpful in distinguishing contrastive types of adjectives such as "listo" ("intelligent and ready" in English) because different meanings were clearly revealed in English and Chinese in word level with no need of going further to the sentence or paragraph level. The main problem of this corpus was related to the technical problem of correctly

matching the parallel meanings among the three languages.

With respect to research Question 2, learning could be more efficient if the complementary advantages of specific functions were provided by the various corpora. The parallel corpus assisted in forming connections between form and meaning. The learners' corpus facilitated the distinction of error and correction in the learning process, and the native corpus required a more advanced level of Spanish to manage an ampler context. Therefore, the recommended order of using these three types of corpora is (1) parallel corpus, (2) learners' corpus, and (3) native corpus. Without the first two types of corpora, only more advanced learners can be benefited because the native corpus required a higher level of language knowledge.

5 Limitation and Future Works

The first limitation was related to the participants. We only had 28 Spanish language learners who participated in evaluating the three corpora. The results are not representative enough; in future studies, we plan to conduct an evaluation task with more participants to make the conclusion more valid and reliable. Moreover, our participants were from the Department of Foreign Languages, and they were enrolled at the same university. We need to expand the evaluation work to learners of multiple universities and from different levels of language proficiency, including learners in Spanish departments and other universities in Taiwan.

Second, in the environment of a computer room, when more than 20 participants worked simultaneously using the three corpora, the corpora might collapse and so intervened the search process. This situation did not occur when the exercises were conducted individually at home, or when less than 10 users were working simultaneously. A technical team is currently taking the responsibility to determine and solve the problem.

Finally, the questions listed in the exercise and formal evaluation were limited to only two types: past tense and copular verbs. Future studies should include more linguistic varieties such as various syntactic and semantic aspects for users to evaluate the general effectiveness of the three corpora.

6 Conclusion

Existing constructed corpora have contributed to corpus-based studies. Their applied value should not be restricted to only researchers or teachers. Foreign language learners should also be considered as beneficial users if they are pre-trained and familiar with instructions and functions of distinct corpora.

Various types of corpora can benefit users in learning foreign languages if they are applied in a complementary way to capitalize on the best results of various functions and purposes of existing corpora. Parallel corpus can supply the translation of parallel meanings through similarities or differences of structures and lexical expressions. Learners' corpus can offer a base to contrast the errors made by learners and corrections revised by natives to impress the learners, and the numerous examples of native corpus provide a helpful source to enrich learners' linguistic knowledge and performance.

In future studies, a greater number of participants with various language proficiencies and from different campuses should be included in such studies to make the findings more generalizable.

References

Baker, M. 1993. Corpora in translation studies: An overview and some suggestions for future research. Target, 7(2): 223-243.

Dalziel, F. and Helm, F. 2008. Exploring modality in a learner corpus of online writing. Linguistic Insights - Studies in Language and Communication, 74: 287-302.

Gilquin, G., Granger, S., and Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. Journal of English for Academic Purposes, 6(4): 319-335.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., and Zampa, V. 2007. Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. ReCALL, 19(3): 252-268.

Howe, Chad and Ranson, L. D. 2010. The evolution of clausal temporal modifiers in Spanish and French. Romance Philology, 64(2): 197-207.

Kennedy, G. 1998. Introduction to Corpus Linguistics. London: Longman.

Kucera, H., and Francis, W. N. 1967. Computational analysis of present-day American English. Providence, RI: Brown University Press.

Lavid, J., Arús, J., and Zamorano-Mansilla, J. R. 2010. Systemic functional grammar of Spanish: A contrastive study with English. London: Continuum International Publishing Group.

Lee, D. 2010. Bookmarks for Corpus-based Linguists. Retrieved from http://www.uow.edu.au/~dlee/CBLLinks.htm

Malmkjaer, K. 2005. Linguistics and the language of translation. UK: Edinburgh University Press.

Mukherjee, J. 2008. English corpus linguistics and foreign language research: Line of development and perspectives. ZFF, Zeitschrift fur Fremdsprachenforschung, 19(1): 31-60.

Zhang, Y., Wu, K., Gao, J., Vines, P. 2006. Automatic acquisition of Chinese–English parallel corpus from the web. In Advances in Information Retrieval (pp. 420-431). London: Springer. doi: 10.1007/11735106_37.

CEATE,
http://corpora.flld.ncku.edu.tw

CEDEL2,
http://www.uam.es/proyectosinv/woslac/cedel2.htm

Corpus del Español,
http://www.corpusdelespanol.org/

CPEIC,
http://140.116.245.228/FW/tri_lingual_index.html

CREA,
http://corpus.rae.es/creanet.html

Linguistics and Language Behavior Abstracts (LLBA)
http://search.proquest.com/

SPLLOC,
http://www.splloc.soton.ac.uk/

Appendices

Appendix A: Pre-training

Aplicación de 3 córpora (Frec. de uso)-preprueba
Número: Nombre:
Pregunta 1: *vivió/vivía…*

Explicación: ¿Por qué se selecciona? ¿Qué diferencia hay en el sentido o uso? Evaluación de ayuda: Sí o No ¿Cómo? ¿En qué aspecto?		
A. CdE：vivió & vivía	*vivió/vivía*	□ No □ Sí
B. CEATE：vivió & vivía	*vivió/vivía*	□ No □ Sí
C. CPEIC：vivió & vivía	*vivió/vivía*	□ No □ Sí

Pregunta 2: *comió/comía…*
Pregunta 3: *preguntó/preguntaba…*
Pregunta 4: *murió/moría…*
Pregunta 5: *ser/estar posible…*

Explicación: ¿Por qué se selecciona? ¿Qué diferencia hay en el sentido o uso? Evaluación de ayuda: Sí o No ¿Cómo? ¿En qué aspecto?		
A. CdE：[ser] posible & [estar] posible	*ser/estar*	□ No □ Sí
B. CEATE：posible	*ser/estar*	□ No □ Sí
C. CPEIC：posible	*ser/estar*	□ No □ Sí

Pregunta 6: *ser/estar feliz...*
Pregunta 7: *ser/estar limpio…*
Pregunta 8: *ser/estar enamorado…*

Appendix B: Questionnaire

Aplicación de 3 córpora (Frec. de uso)
Pregunta 1: *hubo/había + N*

Selección: Circule el que se usa con más frecuencia y tache el que no se usa. Evaluación de ayuda: ¿ayuda o no? (X vs. Y)		
A. CdE：hubo [NN*] 和 había [NN*]	*hubo/había*	□ No □ Sí
B. CEATE：hubo 和 había	*hubo/había*	□ No □ Sí
C. CPEIC：hubo 和 había	*hubo/había*	□ No □ Sí

Pregunta 2: *fui/iba a + destino*
Pregunta 3: *dijo/decía...*
Pregunta 4: *llegó/llegaba...*

A. CdE：[ser] conveniente & [estar] conveniente	*ser/estar*	□ No □ Sí
B. CEATE：conveniente	*ser/estar*	□ No □ Sí
C. CPEIC：conveniente	*ser/estar*	□ No □ Sí

Pregunta 5: *ser/estar conveniente...*
Pregunta 6: *ser/estar seguro...*
Pregunta 7: *ser/estar contento...*
Evaluación en general:

Córpora	Ventajas	Desventajas
A. CdE		
B. CEATE		
C. CPEIC		