Using a specialized corpus and Google custom search engine (CSE) for enhancing L2 teaching and learning

Kwanghyun Park, Pennsylvania State University
kpark90@gmail.com

Abstract

This study explores how teachers can enhance L2 instruction using a specialized corpus with a popular custom search engine from Google. With a focus on the learners' lexicogrammatical performance, a corpus-based system consisting of a database of genre- and discipline-specific texts and a companion search engine are developed for an undergraduate L2 writing course. Using digital video technology, this study documents the learners' interactions with the corpus system over the course of one semester. The findings show that the learners engage in cognitive collaboration with the system and address their lexicogrammatical issues in moment-by-moment development of their linguistic competence rather than simply borrow the text from the corpus. The learners engage in intense hypothesis-testing in order to appropriate the language items from the corpus and in this endeavor, they interact with the system in a highly interactive and multi-layered way. The findings suggest the potential impact of the custom search engine on technology-enhance language learning, i.e., as a teachers’ tool to facilitate the learner's development. This study concludes by discussing implications of custom search engine as a powerful tool for corpus-enhanced language learning.

Introduction

Over the past two decades, there has been an increasing recognition of corpus linguistics and its role in language learning and teaching (Conrad 2000; Granger et al. 2002; Hunston 2002; Sinclair 2004; Flowerdew 2005). Despite the recent burgeoning of studies on corpus-assisted pedagogy, research has primarily focused on the outcomes of learning (e.g., higher test scores and a positive attitude). Learning processes (i.e., learners' interactions with corpora) remain almost entirely unexplored.

Using a corpus system, the goal of this study is to enhance and document L2 learners’ performance in academic writing with a focus on the lexicogrammar. Lexicogrammatical performance refers to the extent to which a learner makes socially appropriate choices of vocabulary and syntax in accordance with the expectations of discourse communities. To be proficient in this area, learners must advance beyond the mere production of formally accurate sentences toward an awareness of, and the ability to cope with, the socio-cultural practices of discourse communities. As such, developing lexicogrammatical performance presents a major challenge to learners.
To date, writing research has been only modestly successful in meeting this challenge. While it has been effective in delivering lexicogrammatical knowledge through explicit instructions (e.g., Swales & Feak, 1994), it has not been as effective in helping learners to use that knowledge in their performance. This study explores how teachers can enhance the instruction and facilitate the learners’ performance in academic writing using a corpus-based system. It aims to offer a means for teachers to facilitate learning in the classroom at the same time as L2 writers are engaging in independent writing work outside the classroom.

**Lexicogrammar in Academic Writing**

A socio-semiotic system that interfaces meaning and expressions (Halliday, 1978), lexicogrammar realizes meaning (semantics) through using basic building blocks (lexicon) and the rules for sequencing them (syntax). Lexicogrammar, thus, has two aspects: one representing the structural sequences, the other, the lexical realizations of those sequences. Lexicogrammatical performance, then, refers to an ability to make choices among myriads of potential intersections between the lexical and syntactic planes.

In academic writing, while lexicogrammar serves as a linguistic resource for these advanced writers, it presents a daunting challenge to novice writers, who are new to the discourse. Unlike simpler, non-academic genres (e.g., greetings and invitations), some academic genres reflect great variability in their triangular interplay of the author’s intentions, the discourse community’s expectations, and the lexicogrammar. Teaching the lexicogrammar of academic writing, thus, is not so much about simply knowing lexicogrammatical rules as it is about using them appropriately and resiliently in accordance with the particular genre and rhetorical situation.

**Corpus-based approach to writing pedagogy**

In academic writing, corpus technology holds considerable promise to provide support for teacher’s instruction. Generally, the contribution of corpus technology to academic writing pedagogy is recognized in areas such as language description, textual analysis, and teaching material development (Hyland, 2002; Upton, 2002; Upton & Connor, 2001). In classroom practice, however, the role of corpus is yet to be robustly established (Flowerdew, 2005). Currently, the role of a corpus in genre pedagogy is limited to providing exemplar texts (Lee & Swales, 2006).

Regarding the contribution of a corpus to learners’ writer’s development, there is only anecdotal evidence (Cobb, 1997; Johns, 1986). These studies have not examined the learners’ experiences while they used the corpus. Some studies attempted to examine the use of the corpus through the learners’ reflections (Chambers, 2005; Yoon, 2008; Yoon & Hirvela, 2004). These studies found that the learners built some confidence in vocabulary and writing skills through their use of concordance programs. The studies, however, did not make first-hand observations of the learners’ experiences. Instead, the studies relied on retrospective perceptions as recounted by learners in surveys, questionnaires, written reflections, and interviews.

Only a few studies provided more concrete evidences of learning in corpus-assisted activities. Gaskell and Cobb (2004) used hyperlinked, concordance-feedback to the learners’
revisions in the lieu of traditional written feedback. Then, by comparing the hyperlink records and the learners’ texts, some relations between the access to the concordance lines and the revisions were established. Although the hyperlink approach shows some benefits of using the concordances, the benefits are largely limited to correcting (rather simple) errors, rather than creating contextually appropriate texts. Hafner and Candlin (2007) is the only study that examined the learners’ corpus use based on the data collected as they used the corpus – rather than their reflective accounts - and without imposing the corpus-based assignments. The data in their study, however, consists of only the searches that the learners entered into the corpus, while missing the actual interactions. To date, no study has examined learners’ interactions with corpora through the real-time, direct observations. The role of a corpus-assisted instruction in enhancing lexicogrammatical performance, thus, remains largely unexplored.

In addressing the lack of research on the interaction between learners and a corpus, this study uses screen recordings, a computer-aided technique that records the learners’ activities on their computer screens real-time, i.e., as they are using the corpus. The screen recording technique creates digital video clips, the primary data source of this study. In addition to the issue of lacking data, the obvious problem in the corpus studies is the lack of an effective corpus-based system. In developing such a system, this study employs a custom search engine (Google CSE) rather than the traditional concordance program. In fact, the choice of the search engine is not in accordance with the research literature, which generally recommends concordance software. This study chooses a search engine over both desktop and online concordancer programs because learners need to access representative exemplars of the target academic genres. This rules out many online concordance services, as they do not allow users to modify the database or point the concordancer to a new database (but see Lu, 2009).

This study is a new attempt to use a search engine and a specialized corpus for L2-writing instruction. The proposed corpus-based system offers broad benefits to learners: the system can serve as an extensive store of representative texts that facilitate textual analysis in situ. As for lexicogrammar, the benefits of the system go beyond a simple accuracy check. The system is designed to engage students in interactive refinement of the searches, which eventually leads to the development as more competent L2 users. For teachers, this approach’s strength resides in its ability to build efficiently specialized corpora for their courses. The system allows teachers with minimum experience of a corpus to replicate the system for their own use.

**Methods**

**Context**

The context of this study is an upper-intermediate ESL writing course at a large American university. The course aimed to prepare first-year international students for university-level academic writing. This study analyzes the interactions of three focal learners with the corpus system. They are two females (Learner 1 and 3) and one male (Learner 2). All of them came from China, spoke Mandarin as their first language, and were business majors.

**The corpus system**
This study developed a corpus-based system equipped with a database of topic-specific texts as well as a companion search engine offered by Google to search the database. Chosen based on their relevance to the course readings and discussion topics, the texts in the database focus on communication and language learning. The database consists of approximately 350,000 words of academic text from online academic journals. As a companion tool for the corpus, a search interface was developed to provide access to the database. The interface in this article refers to a website linked to a custom search engine that runs in the background.

As the corpus system allowed multiple-word search, students could simply type in multiple words and phrases in order to consult the corpus. The search engine enabled users to zero in on appropriate sentences by taking multiple search words as input and allowing the users to revise the searches easily. Search results were contextualized as they were displayed in whole sentences, which were linked to the full-text documents. In order to record the searches, a program was written and connected to the system. The program saves the learner’s search terms in a database and creates a log of searches.

**Data-collection procedures**

This study collected screen recordings as well as the learners’ (oral and written) reflections. In the classroom, students recorded their computer screens while they were composing using a screen-capture program (iShowU), a program that runs in the background without interfering in writing processes. The program created video clips of the computer screen, which were saved for analysis. When the clips had been collected, the learners attended individual interview sessions, which were audio-recorded. In these sessions, the student watched the screen video clips and commented why and how they used the corpus.

During the period of six weeks, screen video clips were collected each week totaling six clips and the total length of the screen recordings was 424 minutes. In addition, six stimulated recall sessions (two sessions for each student) and nine written reflections (three reflections for each student) were collected.

**Results**

**Description of learners’ corpus searches**

The corpus search log shows that the learners used the system frequently. The three focal learners performed 329 searches in six weeks. In those searches, they searched for 772 tokens (257 token queries per student) and 342 types (114 type queries per student). The type, token, and their ratio, serve as a basic descriptor of the learner’s corpus access. Table 1 presents the descriptive statistics of the searches and the type/token ratio (TTR) of the corpus searches.

<table>
<thead>
<tr>
<th></th>
<th>Searches</th>
<th>Token</th>
<th>Type</th>
<th>TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 1</td>
<td>48</td>
<td>92 (Mean: 1.92)</td>
<td>57 (Mean: 1.12)</td>
<td>62.0</td>
</tr>
<tr>
<td>Learner 2</td>
<td>87</td>
<td>203 (Mean: 2.34)</td>
<td>93 (Mean: 1.06)</td>
<td>45.8</td>
</tr>
</tbody>
</table>
Table 1 shows that the learners’ queries are relatively short - averaging about 2.7 words (tokens) per exchange. Learner 3 consulted the corpus most often (194 times) and her query words were longer (Mean = 3.84) than the other two learners (Mean = 2.34 and 1.92). The general impression is that the length of the query increases in proportion to the number of searches. That is, the more often learners consult the corpus, the more search words they tend to enter per query. On the other hand, the frequency of unique words (types) shows that the learners used (at least) one new word for each exchange (Mean = 1.05). It implies that the more the learners used the corpus, the fewer new words they seemed to introduce in the refined searches. This finding indicates that learners were refining searches by making a small change each time, i.e., by substituting one word with another, rather than revising the entire phrases or sentences.

The relationship between the types and tokens of the searches (TTR) offer a useful insight into the learners’ corpus use: TTR can serve as an index that reflects the learners’ searching behavior. The ratio is inversely proportional to the frequency of search term refinement. If the learners recycle more words in order to refine their searches, TTR decreases. For example, Learner 3 looked up the corpus 194 times and there were 192 new words, indicating that she refined her query by substituting approximately only one word of the previous queries. Learner 3 did this partial replacement in order to zero in on the target item. Due to the refinement, her TTR goes down to 40.3%. On the other extreme, Learner 1 did not refine her query in most cases. She consulted the corpus in 48 searches, and there were 57 new search terms. As a result, TTR goes up to 62.0%, which implies that Learner 1 did not refine her queries as often as Learner 3 did.

The search terms can be broadly divided into either content words or function words according to the learner’s search purposes. The corpus searches of the learners contain 227 function words (31.4%) out of all 722 words. The breakdown of the function words search is presented in Table 2.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Frequency</th>
<th>Examples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preposition</td>
<td>107</td>
<td>in (29), of (18), on (15), with (12), as (9)</td>
</tr>
<tr>
<td>Determiner</td>
<td>46</td>
<td>the (19), a (13), no (5), some (3), this (2)</td>
</tr>
<tr>
<td>To (infinitive)</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Pronoun</td>
<td>9</td>
<td>them (4), it (3), one (2)</td>
</tr>
<tr>
<td>Possessive pronoun</td>
<td>8</td>
<td>its (6), their (1), my (1)</td>
</tr>
<tr>
<td>Modal</td>
<td>7</td>
<td>will (7)</td>
</tr>
<tr>
<td>Subordinate conjunction</td>
<td>5</td>
<td>while (3), whereas (1), whether (1)</td>
</tr>
<tr>
<td>Coordinate conjunction</td>
<td>6</td>
<td>but (4), or (2)</td>
</tr>
<tr>
<td>Wh-Adverb</td>
<td>5</td>
<td>how (4), however (1)</td>
</tr>
<tr>
<td>Particle adverb</td>
<td>2</td>
<td>up (1), on (1)</td>
</tr>
<tr>
<td>Wh-determiner</td>
<td>2</td>
<td>which (2)</td>
</tr>
</tbody>
</table>

| 227                     |
Table 2 shows that Preposition is the most frequent item followed by Determiner and To-infinitive. When compared with the learners’ reflections, the finding indicates that learners are more aware of the uses of particular prepositions such as ‘in’ and ‘of’, than cohesive devices such as deixes and conjunctions.

The distinction is important as it reveals challenges to L2 learners in concrete details. In fact, the searches for function words seems to represent a distinct characteristic of the language learners, as the function-word searches contrast with the common understanding that the log of the general-purpose search engines largely consist of nouns and adjectives, while they rarely contain function words (Eiron & McCurley, 2003). Although the content words still outnumber the function words in the learners’ log, it still shows a remarkable difference from the general-purpose search engine queries.

**Enhancing Lexicogrammatical Performance**

In addressing their lexicogrammatical issues, the learners focused on two aspects of academic writing: lexicogrammatical accuracy (e.g., verb conjugation) and rhetorical appropriateness (e.g., formality). Accuracy was the first and foremost concern in the learners’ use of the corpus. Syntactic, lexical, and morphological issues appeared to be the major areas of concern - with syntactic concerns more frequent than lexical and morphological concerns. These concerns motivate and shape their interaction with the corpus. Students’ syntactic motivations frequently revolved around choice of part-of-speech items (e.g., preposition) and grammatical form (e.g., adverbial). Figure 2 shows an example of syntactic motivation that revolves around preposition:

<table>
<thead>
<tr>
<th>Original text</th>
<th>Oppositely, the author suggested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus search 1</td>
<td>‘opposite’</td>
</tr>
<tr>
<td>Corpus search 2</td>
<td>‘oppositely’</td>
</tr>
</tbody>
</table>

| Revised text           | In opposition, the author suggested |

**Figure 2. Corpus-assisted revision for accuracy**

Regarding the corpus searches and the following revision, Learner 1 commented that she was looking for an adverbial connective in the sentence-initial position (e.g., however). She queried for ‘opposite,’ and then its adverbial form, ‘oppositely.’ As the corpus results did not show ‘oppositely’ as an adverbial connective, she discarded it and chose to revise her original sentence using another expression, ‘In opposition.’ The screen recordings show that her revision is based on the results that the system displayed during her corpus consultation.

Appropriateness was a less frequent but more difficult issue for the learners. As it is not an issue of correctness, addressing the appropriateness issue requires more proactive evaluation of the corpus-search results. For example, Learner 3 entered three related searches and all of them revolve around the key search word, “task” (see Figure 3):

| Original text | Writing an essay in English is a tough task |
The goal of corpus search in Figure 3 is not to produce an error-free text but a more appropriate text. Three adjectives, ‘tough’, ‘hard’, and ‘difficult’, are interchangeable without resulting in any lexicogrammatical error. The challenge in this task is that it is not always possible to decide which one is more appropriate than the others, as the evaluation of choice depends on the learners’ intention as well as the target audience. The evaluation, thus, relies heavily on the learner’s rhetorical awareness. Learner 3 stated in her reflection that she was trying to find a better adjective for ‘task,’ and that she revised the searches because ‘the revised one shows that the sentences become more academic-like.’ Yet, what is more ‘academic-like’ than the other is still difficult to decide and the corpus system alone is not sufficient to address the issue.

Despite the effectiveness of the system, some challenges still remain. According to the focal learners, for example, the most difficult challenge is to ‘build a sentence,’ i.e., to produce a sentence with limited linguistic repertoire. Regarding the production of sentences, the screen recordings show a clear benefit of a search-based corpus system, beyond less interactive concordance software (see Figure 4).

<table>
<thead>
<tr>
<th>Original text</th>
<th>Lately, as technology is developing in a fast speed, there are more tools for L2 users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus search 1</td>
<td>technology is developing in a fast speed</td>
</tr>
<tr>
<td>Corpus search 2</td>
<td>society develop fast</td>
</tr>
<tr>
<td>Corpus search 3</td>
<td>in a fast developing society</td>
</tr>
<tr>
<td>Corpus search 4</td>
<td>technology develops fast</td>
</tr>
</tbody>
</table>

| Revised text | Lately, as technology is developing in a fast speed, more tools are invented for L2 users |

In Figure 4, the challenge is not limited to lexical concern, i.e., word choice, but it involves broader aspects of the lexicogrammar touching upon syntax, semantics, and rhetorical structure. Syntactically, Corpus search 1 (“technology is developing in a fast speed”) and Corpus search 2 (“society develop fast”) involve at least two kinds of choices: a choice between the progressive and simple aspect of the verbs (“developing,” and “develop”) and another choice between the adverbial and adverb (“in a fast speed” and “fast”). Semantically, Corpus search 2 (“society develop fast “) and Corpus search 4 (“technology develops fast”) involve a choice between two agents, technology and society, for the verb, “develop.” On the other hand, “society” serves as a circumstance and becomes a background in Corpus search 3 (“in a fast developing society”). We also notice that these choices are connected with the learners’ rhetorical decision with regard to structuring the sentence. She has an option to begin her sentence with either “society” or “technology,” depending on which she decides to highlight. Based on her searches, she made a decision to keep her original sentence.
The series of the searches and the query refinement suggest that the enhanced performance is a result of the negotiated performance. The learner does not copy the texts from the corpus verbatim but actively refine the queries and make choices according to her communicative purposes.

**Learners’ evaluation of the corpus system**

The analysis of the data in the previous section showed a clear pragmatic value of the corpus system. Then, how do the learners describe their experiences and what are their evaluations of the system? The learners have idiosyncratic difficulties and used the corpus system in different ways according to their divergent needs. The idiosyncrasy has a visible influence on their evaluation of the corpus system, which varies according to the extent to which the system addresses their issues.

Learner 1 commented that she was using the system simply as an ‘error checker.’ Comparing the corpus search engine with electronic dictionary, she reflected: ‘But in some cases, electronic dictionary does not help. . . Because of this kind of problem [of the dictionary], *I always made wording mistakes, and I could not receive a high mark for my paper* . . . After I knew this tool [the corpus system], I used it every time when I write. As I know, all of my classmates use this tool as well. *They all feel this tool helped them a lot when they write their paper.*’ In her response, the learner clearly indicated that her use of the corpus was motivated by a realistic evaluation of the corpus in terms of its effectiveness in helping her ‘receive a high mark.’

Learner 2, unlike Learner 1, described the system as ‘as advanced linguistic reference.’ He wrote that he chose to use the system in addition to a bilingual translation website because ‘*It [the system] really helps me a lot when I am not sure about the usages.* . . . But if I don't know the meaning of the words at all, the searching engine [of the corpus system] won’t help me very much.* The use of it is based on abundant knowledge of English.’ Learner 2, thus, presented his critical evaluation of the corpus by pointing out that the corpus is effective only when he had some previous knowledge of the target vocabulary. Thus, his evaluation is that the corpus system works for advanced users or those with ‘abundant knowledge,’ and for him, linguistic awareness is the key to an effective corpus use. Based on this evaluation, he established his own tool-use policy: the corpus system for ‘usage’ and dictionary for definition.

Learner 3 was the most proactive use of the system. She found that the corpus system addressed her needs in academic writing and used it as her primary reference: ‘Before my teacher introduced [the corpus system] to our class, I did not know how to revise my essay by using a language tool except looking up vocabulary in certain website which was not quite helpful in terms of writing. *As long as I know Corpus [system], I started searching word in the Corpus [system]. It gave me amounts of references that I can check that whether the scholar[s] use the word in the same way as I did.*’

She used the corpus system to look for scholarly use of language, i.e., the way advanced writers use language in academic writing. She paid a lot of attention to the appropriateness of her writing and used the corpus to achieve the appropriateness. She was also concerned with stylistic issues
such as formality, clarity, as well as word and grammar choice. Addressing these issues requires sensitivity to the lexicogrammar of the academic writing and she found that the corpus system was the only tool that gave her access to the lexicogrammar of the target discourse.

**Discussion and Conclusion**

This study sought answers to the following questions: How can we enhance learners’ lexicogrammatical performance in L2 academic writing? What is the role of the corpus-assisted instruction in enhancing the performance? The previous corpus-based research focused on either the final outcome or on the perceptive responses based on the learners’ reflections. The alternative, as posited by the present study, is to improve the performance through a corpus-based system equipped with a topic-specific corpus and a companion search engine. The results show that the system was effective and efficient in helping the learners meet their lexicogrammatical challenges and thereby improve their performance in this regard.

A pedagogical implication of the results is that teachers can incorporate a corpus with a search engine in their writing instruction for the practical benefit of students. Considering the time constraints on teachers, it is a demanding task to prepare example texts manually and provide individualized feedback on lexical/grammatical issues. For students, too, even concerted attempts with much time invested can still result in only modest improvements in lexicogrammatical performance. In trying to mitigate this difficulty, a corpus system can be introduced to learners as a supplementary resources as well as self-directed corpus-based assignments.

Future research on corpus-enhanced language learning should focus on providing a more efficient system than traditional concordance software. Despite its usefulness in form-focusing activities and error correction (Gaskel & Cobb, 2004), concordance software is less effective in addressing global issues in academic writing such as formality, tense and aspect choices, and sentence production.

The data showed that interactivity of a system is important, as learners enhance their performance through a negotiation with the corpus system (see Figure 4): they interactively review the search results, compare the results against their authorial intentions, and make revisions. Therefore, it is necessary to examine the potential contributions of the search-engine interface to corpus-enhanced academic writing instruction. Future implementation of a corpus system should address the limitations of the search engine as well by exploring the teachers’ role in offering the learners further help with a corpus system to address the lexicogrammatical issues as linguistic and social choices.

**References**


