Jl. of Technology and Teacher Education (2012) **20**(4), 415-440

Evaluating Corpus Literacy Training for Pre-Service Language Teachers: Six Case Studies

JULIAN HEATHER AND MARIE HELT
California State University, Sacramento, USA
jheather@csus.edu
Marie.Helt@csus.edu

Corpus literacy is the ability to use corpora—large, principled databases of spoken and written language—for language analysis and instruction. While linguists have emphasized the importance of corpus training in teacher preparation programs, few studies have investigated the process of initiating teachers into corpus literacy with the result that few guidelines exist for training teachers to make optimal use of corpus output. This paper uses a case study approach to examine six pre-service language teachers' development of multiple components of corpus literacy during a semester-long introductory grammar course through which corpus linguistics was threaded. Results showed that while corpus literacy training was generally effective, that effectiveness varied among subjects. Examining the sources of that variation suggests several practices for teacher educators planning or modifying instruction in corpus literacy.

There is no question that today's language teachers need to have an array of technology-related competencies and skills that can also be passed along to their students, thus enabling these language learners a more equitable skill set in technology, the "major cultural capital in the 21st century" (Oxford & Jung, 2007, p. 41). Standards for teacher preparation are being re-examined and revised in many states and countries to include more specific and realistic guidelines for the kinds of technological understanding that today's language teachers and their students will need (Murphy-Judy & Youngs, 2006; Oxford & Jung, 2007). Indeed, the national organization for Teachers of English to Speakers of Other Languages (TESOL) has just released new guidelines that lay out what teachers and students should know and be able to do technologically (Healey, Hanson-Smith, Hubbard, Ioannou-Georgiou, Kessler, & Ware, 2011). One of the overarching themes in the TESOL standards is that "technology should be incorporated into teaching pedagogy so that students will not only effectively acquire a second language but will also develop electronic literacy skills" (Healey, et al., 2011, p. 9).

One application of technology to language research and teaching, however, that has been neglected in the literature on Computer Assisted Language Learning (CALL) and in the TESOL Standards is corpus linguistics. This field makes use of a large, principled collection of spoken and written texts on computer (i.e., a *corpus*) to analyze the way speakers and writers actually use language for different purposes. Two major types of output can be used for and by language teachers and learners. One is information about the extent to which a linguistic structure is distributed across the registers represented in a corpus. For example, in a corpus containing academic writing and spontaneous conversation (among other registers), phrasal verbs (*break down, bring up*) were found to be roughly twice as frequent in conversation as they are in academic writing (Biber, Conrad & Leech, 2002), suggesting that speakers consider the phrasal verbs to be less formal than their single-word synonyms. The second type of useful corpus output is a concordance, or all the lines of text from the corpus that contain a particular word or phrase, with that word or phrase highlighted. These lines can be analyzed, for example, to discover patterns of co-occurence (e.g., the most common adverbs that modify the adverb *much* are *pretty, so, very* and *too*) (Biber, Conrad & Leech, 2002). Concordance lines from students' own writing can also be used to check learners' acquisition of a certain structure. To illustrate this use, Figure 1 shows the first ten concordance lines of the occurrences of *many* in a small corpus of English Learners' writing.

Teachers can use output from a corpus to help students see grammar patterns, understand semantic nuances, and discover how words interact with syntax. While many strong arguments are being made in the current literature for teachers to develop some degree of competence in this area, there are few guidelines and even fewer studies that evaluate effective teacher training in corpus literacy. Although the above-mentioned TESOL Technology Guidelines, for example, refer to "digital resources", none of the references to specific resources mention corpus linguistics or corpus literacy, nor are they included in a recent description of the scope of CALL (Hubbard & Levy, 2006).

1 who is her love. Although she has **many** ability yet a man look

2 man look down upon hers. There are **many** thing that the women

3 treated equally to man **Many** people think so women should be

4 equally to man. There are **many** organise come forward appeal to

5 can work equal. Actually there are **many** unit of work is a leader.

6 Only when meny people have a good **many** thing to do. At presen

7 Although the women work about **many** thing like the man but the

8 treated equally to men" There are **many** reasons for both sides of

9 equally to men." **Many** people have strong view and feel that

10 more than the men. **Many** the women can equivalent to the

Figure 1. Concordance lines of *many* in a Learner Corpus.

Corpus literacy, which is the ability to use the technology of corpus linguistics to investigate language and enhance the language development of students, is a complex phenomenon comprising multiple sub-skills. An attempt has been made to establish a definition of "corpus literacy" for *language students* that includes four components: understanding what a corpus is; knowing what can (and cannot) be done with a corpus; knowing how to analyze concordances; and understanding how to extrapolate general trends in language use from corpus data when appropriate (Mukherjee, 2006). However, language *teachers* have additional and differing needs. This paper presents six case studies of pre-service teachers in an introductory grammar class designed to also introduce them to corpus literacy. It illustrates the multiple ways in which corpus literacy can be developed as well as the issues which impact its development.

BACKGROUND TO THE STUDY

Corpus Linguistics is an area of technology that has had an important impact on language analysis for several decades now. Corpora (plural) have been analyzed, using published software or the researcher's own programs, to demonstrate how grammatical structures interact with lexical items as well as with larger discourse/register functions, resulting in comprehensive, descriptive corpus-based reference materials such as the *Longman gram-*

mar of spoken and written English (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Large corpora are becoming increasingly available to "average" users (i.e., *not* linguists) through online interfaces, thus allowing access to a much wider audience, including language teachers and their students.

Corpus linguists have recently begun to argue for a greater role for corpus-based findings in (second or foreign) language learning and teaching, using corpora to show that English Language Teaching (ELT) materials often emphasize characteristics of grammar that don't match how those structures are actually used by native speakers (see for example Römer (2004) on modal auxiliaries), or to demonstrate that the intuitions of native speakers about language often fail to account for important nuances of meaning (see for example the analysis of the patterns of use for *big, large, little* and *small* in Biber, Conrad, & Reppen, 1998). Corpus analysis has become instrumental in describing the co-occurrence patterns of language structures that characterize registers, such as academic writing in different disciplines, or the differences among academic lectures, study group discussions, or lab sessions (Biber, Conrad, Reppen, Byrd, & Helt, 2002), yielding information that may be critical to English Learners wishing to attain a high degree of academic literacy.

Corpus Linguistics has thus added a great deal to the description of English (and other languages) as it is actually used by speakers and writers in different contexts and for different purposes, and many corpus linguists have argued that this information is vital to English learners, especially those whose goal is a native-like command of grammar, vocabulary and idiom (Gilmore, 2009). The benefits of corpus analyses, though, have been rather slow to translate into language teaching materials or techniques, resulting in very few corpus-informed ELT materials to date (but see McCarthy, McCarten, & Sandiford, 2005, for an exception). With a dearth of published materials, it becomes the responsibility of language teachers to inform themselves of the available (online) corpora and corpus-based reference materials, and to learn to generate their own corpus-based or corpus-informed teaching materials to fit their contexts (Lombardo, 2009; McCarthy, 2008; O'Keeffe & Farr, 2003).

There are several useful guides that in-service teachers can exploit on their own in order to become more adept at conducting corpus searches and using the output to inform their teaching (Anderson & Corbett, 2009; Reppen, 2010), and a number of articles have outlined ways that teachers have made use of on-line corpora in their English Language classrooms (Braun, 2007; Oksefjell Ebeling, 2009; Varley, 2009). But many in-service teachers feel overwhelmed by the task of learning to access online corpora, conduct successful searches using the interface provided, understand the output, and translate that information into useful and effective teaching materials, and teachers often cite a lack of time to invest in such endeavors as well as a lack of access to the necessary hardware and software (McCarthy, 2008).

For these reasons, researchers in the field now argue strongly that future language teachers need experience with corpus linguistics early in their Teacher Education programs (Granath, 2009; McCarthy, 2008; Mukherjee, 2006; O'Keeffe & Farr, 2003; Peters, 2006). However, there is not much guidance yet for teacher educators on how best to introduce pre-service teachers to corpus linguistics. A few articles have reported on integrating corpus linguistic techniques into teacher training courses (Helt & Reppen, 2008). In addition, several researchers have identified the skills that teachers should have in order to incorporate corpus linguistics into their language teaching, and several papers have set forth standards or guidelines for teacher education in corpus literacy (Lombardo, 2009; McCarthy, 2008; O'Keeffe & Farr, 2003; Peters, 2006; Römer, 2006). But very few studies have actually attempted to include these standards and definitions in a language teacher education context and then systematically evaluate the resulting learning outcomes in an empirical way.

Two exceptions to date are Farr (2008) and Breyer (2009). Farr (2008) integrated corpus investigation activities into a two-semester Master of Arts program in ELT in Ireland, and repeated the module with the subsequent MA group, as well, yielding a total of 25 subjects. Subjects gained access to at least five different corpora, and were given guided tasks at first, followed by more independent projects, to provide them with training in using a corpus to answer questions about grammar. By the end of the second semester, students were asked to complete a "substantial assignment in the form of a corpus-based discourse analysis paper on [a topic] of their own choice" (Farr, 2008, p. 32). Farr assessed her subjects' attitudes towards the corpus component of the MA program through a questionnaire distributed near the end of the program, and found that while the students could see clear advantages for using corpora, both as a reference for themselves as teachers, and as a tool for improving classroom ELT, they also expressed frustration at learning to conduct effective searches and at the amount of time needed to consult a corpus and analyze the output. A follow-up of five subjects as in-service teachers showed that two were employing corpus techniques in their ELT classrooms, while three expressed issues related to the availability of hardware and software and integrating corpus techniques into a pre-set methodology.

For her study of pre-service secondary school English language teachers in Germany, Breyer (2009) designed a Semester 2 elective course devoted to introducing corpus linguistics as part of a four- to five-year initial teacher training program that includes student teaching. The course met in a computer lab equipped with several corpora and five different concordancing software packages, and the 18 subjects gained experience in learning English from corpora and used those corpus-related experiences to reflect on their potential effectiveness for ELT. Breyer analyzed the subjects' reflective essays on the teaching of *some* and *any* (comparing their use in generalized corpora with their presentation in ELT textbooks used in Germany), and a questionnaire related to students' experiences creating teaching materials based on concordances. Pre-service teachers felt that some corpus-generated exceptions to accepted grammar rules should not be used with beginning learners, but they also realized that over-simplification of the rules (such as those found in the ELT textbooks) was undesirable. They appreciated the authenticity of corpus examples as compared to more artificial textbook language, and had clear ideas for using concordance output to provide authentic examples of use. Finally, they also saw their future students' independent corpus-based learning as an opportunity *and* a challenge for them as teachers.

The present study extends this prior work in several important ways. First, it follows six *individual* pre-service teachers as they develop corpus literacy over one semester, rather than aggregating the reflections of a whole class or group. This allows for greater depth of analysis while still accounting for the range of responses to the exposure to corpus linguistics. Second, because the present research design includes the collection of multiple types of data at numerous points during the semester, the researchers can investigate each subject's understanding of the application of corpus linguistics to the teaching context as that understanding develops in different ways over time. Finally, like Farr (2008), the current study integrates corpus linguistics into a pre-existing university course, but a wider range of students and future teaching goals is represented. These include undergraduates as well as graduate students who aim to teach high school English, English as a Foreign Language (EFL) abroad, community college ESL, and/or Adult Literacy. This range of student goals demonstrates the challenges of making corpus literacy relevant for pre-service teachers in a short amount of time.

With this range of students and challenges in mind, the larger study was designed to address the following questions:

- In what ways did pre-service ESL teachers develop corpus literacy?
- What were pre-service ESL teachers' attitudes towards corpus literacy?
- What factors contributed to those attitudes?

The present study focuses on the first research question, providing an in-depth analysis of six pre-service ESL teachers' development of corpus literacy over one semester.

CONTEXT OF THE STUDY

Participants

Participants for the study were drawn from two sections of an upper division, 15-week course on grammar for ESL teachers taught at a large state university by one of the researchers in Fall 2008. The course is required for graduate and undergraduate students in the TESOL Certificate and TESOL Minor programs, is a prerequisite for graduate students in the MA TESOL program, and is one of three options for the grammar course that English majors in a pre-credential program are required to take. A total of 52 students participated in the full study, though two subsequently withdrew from the study because they dropped the class. Thirty-seven of the participants were in TESOL programs (13 undergraduate, 24 graduate) while the remaining 15 participants were mostly English majors. Responses to a background survey administered at the start of the semester indicated that only nine students had previously heard of corpus linguistics, mostly through an introduction to linguistics or another grammar class. Four of the nine claimed to have had previous experience using corpus linguistics, but only one student could actually provide an accurate definition of the term.

Corpus Instruction

Corpus linguistics was threaded throughout the entire course through in-class activities, course projects, and readings assigned from two recent corpus-based grammar reference books: *Longman student grammar of spoken and written English* (henceforth: LSG) (Biber, Conrad, & Leech, 2002) and *The teacher's grammar of English* (Cowan, 2008). At the fourth class meeting, students were provided with a systematic introduction to classroom use of corpus linguistics. This instruction defined key terminology such as *corpus linguistics, corpus, concordance*, and *left/right collocate*; identified the types of questions a corpus can answer and how they are answered; and introduced corpus-based grammatical analysis through discussion of five examples. A handout provided information about corpus resources for ESL/EFL teachers. Subsequently, corpus output (concordances and information about frequency and distribution) was used throughout the semester to address class questions about grammar. Concordance examples from a small learner corpus (a collection of English Learners' writing, for example) were also used to demonstrate how a teacher might assess learners' progress in controlling certain grammatical structures.

The present study focuses especially on two major corpus-based projects that students completed. For the first project (due in week 7), they worked in groups to use information in their corpus-based course textbooks (Biber et al, 2002; Cowan, 2008) to (a) critique the treatment of negation in a grammar textbook and (b) design supplemental teaching materials for use with that text in an intermediate-mid-level ESL grammar class. Concordances from the Corpus of Contemporary American English (COCA) (Davies, 2008) were provided for *no*, *not,* and *n't* (90-100 lines of data for each form), but use of this data for the project was optional. Each group submitted two group products: a lesson plan and a handout with supplemental teaching materials. Each individual student also submitted a two-page narrative containing their critique of the published materials and a description of their use of corpus materials.

The second project (due in week 14) asked students to critique the treatment of adjectives and adverbs in two grammar textbooks and design supplemental teaching materials. For this project, however, students worked individually and were required to use COCA to investigate one of four aspects of adjectives and adverbs: two-syllable adjectives which can be superlative/comparative using either *-er/-est* or *more/most* (e.g. *clever*); comparative and superlative forms of single syllable words (e.g. *fun*); *farther* vs. *further*; and adverbs and adjectives with the same form (e.g. *hard*). Prior to starting this project, students received one day (75 minutes) of training on COCA in a computer lab so they would be able to generate the frequency, distribution, and concordance data necessary for its completion. In addition to the critique of the published materials, students' narratives for this project discussed their search strategies for finding corpus information and justified their use of that information.

METHOD

Data Collection

All products from the two course projects (lesson plans, handouts, and narratives) were collected from all of the participants, with their permission. After the first (group) project, 51 of the participants completed a brief survey which focused on their confidence in their ability to perform corpus-based activities. At the end of the semester (i.e. after Project 2), 48 of the participants completed a longer survey which again asked them their confidence levels with a broader range of corpus-based activities which reflected additional training received for the project; it also sought to identify their general attitudes towards corpus linguistics and its future use in their teaching. Finally, 12 participants volunteered for follow-up interviews, which were recorded and subsequently transcribed.

To minimize the possibility that one researcher's role as course instructor would influence participants' responses, all surveys and interviews were conducted by the other researcher (who was *not* the instructor), and participants were informed that the data would not be shared with the instructor/researcher until the following semester.

Case Selection and Data Analysis

Case selection followed a three-step process. In the first step, the researchers identified 23 students who represented a typical cross-section in terms of program/level of study, teaching background, and performance in the class. This sub-group was divided between the researchers, and a chronological summary of each student's data was written; each researcher checked the other's summaries for accuracy and completeness. In the second step, three students with incomplete data for Project 2 were omitted; the remaining 20 students were divided into three groups representing strong, average, and weak performance in the course; this categorization was initially based on grades on the final project and subsequently confirmed by a close reading of the chronological summaries. In the third step, one researcher reread all the chronological summaries to identify two subjects from each group, making a total of six subjects who represented typical and/or interesting cases.

The Subjects

Holly and Tracey were selected from the “strong” group, Claire and Owen from the “average” group, and Emily and George from the “weak” group. While Emily and Tracey were in the same group for Project 1, none of the other subjects worked together. None of the six subjects had any knowledge or experience of corpus linguistics prior to the course.

Holly was a graduate student who was taking the class as a prerequisite for courses in an MA TESOL program, which she was pursuing because she wanted to teach English overseas. While she had limited experience tutoring English and French, she had no experience as a classroom language teacher. For Project 1, her group “looked up *not* and *n’t* using the conversational register of the corpus” [Holly: Narrative 1]. For Project 2, Holly focused on two-syllable adjectives that can take both inflectional *(-er/-est*) and phrasal (*more/most*) forms of comparative and superlative in a lesson plan that incorporated corpus data in several ways.

Tracey and Emily were both undergraduate students who were pursuing a certificate in TESOL and had no previous teaching experience. They had different career goals: teaching EFL for Tracey and teaching high school or college ESL for Emily. Their group for Project 1 used the provided concordances, whose use was optional, of *not*, *n’t*, and *no* in an inductive activity in their lesson on negation. For the individual project on adjectives and adverbs, Emily focused on the uses of *farther* and *further*, which she had identified as important in academic reading and writing, while Tracey’s lesson covered the use of *pretty* as both an adjective and an adverb.

Like Tracey and Emily, Claire, an undergraduate student who was pursuing the TESOL Certificate in order to teach in adult literacy programs or adult education, was in a group which chose to use the provided concordances for Project 1. For Project 2, Claire focused on single syllable adjective forms such as *big*, *warm*, and *fun*.

Owen was a graduate student in literature whose goal was to teach literature at the college level. He did not indicate any previous teaching experience. His group didn’t use the provided concordances in Project 1; instead, they conducted an original search to generate examples of double negatives, the focus of their lesson. Owen’s lesson plan for Project 2 covered comparative and superlative forms of single syllable adjectives.

The final subject, George, was a non-degree student who had very little prior teaching or tutoring experience and was completing a TESOL Certificate because he wanted to teach in Korea. His group for Project 1 covered a lot of material in their lesson on three different aspects of negation--*do*-insertion, double negatives, and negative prefixes—but did not use corpus data in any way. For Project 2, George focused on teaching comparative and superlative forms of single syllable adjectives ending in –y such as *angry*, but again he did not include any corpus data in his lesson plan.

DATA ANALYSIS

This paper focuses on the following data from the six case study subjects: Project 1 (lesson plan, handout, individual narrative); the survey completed after Project 1; Project 2 (lesson plan, handout, individual narrative); the end-of-semester survey; and interviews with three of the subjects (Claire, George, and Holly). These data were analyzed for five components of corpus literacy. The first two components came from Mukherjee's (2006) definition of corpus literacy for *students*:

1. knowing what can (and cannot) be done with a corpus
2. knowing how to analyze concordances

For *language teacher* preparation, the following additional components were also included:

3. selecting an appropriate corpus
4. understanding how to make sense of basic frequency information; and
5. using corpus output to generate appropriate teaching materials

While the data was initially coded for all five of these aspects of corpus literacy development, subsequent reanalysis indicated that the second and fourth components interacted in ways which allowed them to be collapsed into a larger single category: "understand and analyze corpus data."

Select An Appropriate Corpus

Although the instructor did introduce students to multiple corpora, including a learner corpus, the vast majority of their interactions with a corpus involved the Corpus of Contemporary American English (COCA). Not surprisingly, therefore, none of the subjects demonstrated the ability to choose an appropriate corpus. The one exception was Holly, whose discussion of the use of different types of corpora, which will be discussed in the next section, suggests an awareness of the need to select corpora based on pedagogical goals.

Understand What a Corpus Can and Cannot Do

Data for this section comes exclusively from the comments in surveys, interviews, and narratives that reveal—implicitly or explicitly—subjects' theoretical understanding of the uses and limitations of language corpora as well as how that theoretical understanding was affected by their experiences with corpus searches. These comments covered three general topics: the information provided by corpus linguistics; the interaction of corpus linguistics with language instruction; and the limitations of corpus linguistics.

Information Provided by Corpus Linguistics

The two strong students (Tracey and Holly) and the stronger of the average students (Owen) showed the best understanding of the types of information provided by corpora. As early as Project 1, Tracey noted that corpus-based materials "would give examples of real speech,… would show patterns of usage, how word order is structured around negation, and would help lead to student discussion of how *no* is used differently than *not/n't*" [Tracey: Narrative 1]. By the end of the semester, Tracey was also commenting on her use of corpus data to test assumptions about language:

> I chose the last 40 sentences from each category [she means 'register'] and was surprised at the statistics… pretty was used as an adjective only 25% of the time, mostly in the fiction category (70%). I had thought that the adjective form would be the most prevalent, but pretty was most frequently used as a qualifier, an adverb that indicates the degree of another modifier. [Tracey: Narrative 2]

Holly's Project 1 offered little insight into her views on corpus-based data, but by the end of the semester, she clearly saw its potential to allow her to explore both her intuitively-held beliefs about language [Holly: Survey 3 Q1; Interview] and differences in usage [Holly: Survey 3 Q1; Interview] as well as to obtain authentic examples [Holly: Survey 3 Q10; Interview]. Similarly, Owen primarily saw corpus linguistics as providing "an (almost) endless source of examples of living language" [Owen: Survey 3 Q11] which also can reveal register differences. He may also have seen corpora as providing windows on societal/cultural trends by tracking usage changes over time: for Project 2, Owen tried to explain a growth in the use of *greener* by referring to "an increasing environmental awareness [and]… the transition to more eco-friendly practices in society" [Owen: Narrative 2].

The remaining three subjects all referred to the value of corpus-based data, but their enthusiasm was tempered by a lack of skills or was contradicted by how they used that data.

For example, Claire saw several potential ways in which corpora could allow exploration of words: "it was useful to see how people used different words and the frequency of their use…[and] to check on meaning differences of different words and sometimes the same word" [Claire: Survey 3 Q1]. In her narrative for Project 2, she referred specifically to the information on register distribution in the *Longman student grammar* (LSG) for adverbial uses of *good*, which suggests that this was another aspect of corpus linguistics which she saw as being useful (she noted that the grammar textbook does not include this information).

At first, Claire was very uncertain about using the corpus, even after receiving instruction in the computer lab, which she characterized as helpful but insufficient:

> I have more difficulty understanding what how how to negotiate the COC- the site. Um, my daughter actually helped me, she went through and, when we went through the tutorial she read the directions, and explained them re-interpreted them to me, what the directions were actually saying, that helped. [Claire: Interview]

By the end of the semester Claire referred to her ability as that of a "novice" [Claire: Survey 3 Q5] who is only able to perform "the simplest searches" [Claire: Survey 3 Q1]. Nevertheless, she demonstrated the ability to search COCA in several different ways:

> I looked up words by part of speech and was successful once I found the POS [part-of-speech] symbols. Then I tried the 'lemma" and discovered I could bring up all forms of comparative adjectives by bracketing the base word. This opened up a whole new way of researching for me. [Claire: Narrative 2]

Claire's evaluation of her own abilities appeared to arise from a realization that corpus linguistics offers more than she was capable of researching. At one point, she stated that she "know[s] there is much more available because the charts in Longman show that" [Claire: Survey 3 Q8], but she also acknowledged that "some of the graphs that they had in the book, I wouldn't even know where to begin to look to find a graph like that" [Claire: Interview].

George also showed this contradiction between what is achievable in theory versus in practice. He indicated that the "five-minute tour" of COCA (an online introduction) at the start of Project 2 raised his awareness of the usefulness of corpus searches:

> *George*: there's many ways you can do searches. You can do, uh, combined word searches, adjective searches, adverb searches, so I thought that would be very useful.
> *Interviewer*: In what way?
> *George*: Well. If you needed uh specific examples of particular verb fo-or word forms classifications and so forth, you could isolate them in that way. [George: Interview]

However, this awareness of the possibilities of corpora did not translate into an ability to use the corpus: "I couldn't figure out how to do a lot of different searches even though I knew they were there, they were available, I couldn't figure them out and I was pressed for time, so I didn't, spend a lot of time, trying to resolve the uh confusion" [George: Interview].

The final subject, Emily, showed a different type of contradiction. On the one hand, she valued "the opportunity to see where 'further' and 'farther' are used in authentic language" [Emily: Narrative 2]. On the other hand, this position appears to be contradicted elsewhere in her narrative where she writes:

> While searching for adequate examples in which "further" and "farther" were used correctly, I found that in many cases these two words were used interchangeably and thus incorrectly according to the definition I would teach the class. According to the Merriam-Webster's dictionary these two words can be used interchangeably, however, the corpus indicated that "farther" is most commonly used when describing physical distances, while "further" is used for figurative distance. I chose to include only the sentences that corresponded with the lesson, but the teacher would address how some use them interchangeably during the lecture (as noted on step 4 of the lesson plan). [Emily: Narrative 2]

There seems to be a tension here between her initial comment that the two terms are interchangeable and the subsequent comment about the corpus indicating a difference in usage based on physical vs. figurative distance. Emily appears to have resolved this tension by carefully selecting concordance lines to fit her pre-established ideas of the rule governing usage of these words. In doing so, she contradicted one of the benefits of this type of "data-driven" approach.

The Interaction of Corpus Linguistics with Published Teaching Materials

Three of the students discussed the theoretical ways in which corpus linguistics intersects with language teaching materials. For Emily, the most interesting/useful aspect of Project 1 was that it "allowed us to see how even the most popular books used in teaching ESL can still have a lot of necessary materials missing" [Emily: Survey 2 Q1]. Similarly, Owen saw a role for corpus linguistics in supplementing the information found in language textbooks. At mid-semester, he pointed out that "the [published] materials only discourage double negatives whereas the corpus-informed texts discuss the appearance of accepted double negatives in spoken English" and "by utilizing the Corpus of Contemporary American English, we may present examples which shore up this shortcoming of the [published] materials" [Owen: Narrative 1].

Holly appreciated that the corpus allowed her to look critically at curriculum, which was a new experience for her [Holly: Survey 2 Q1]. In fact, this new-found critical stance led Holly to suggest:

> it would be really interesting to have a curriculum based on a corpus instead of just a straight grammar book that's standard English because I don't think that, most people who speak… native English, speak the way grammar book says to. [Holly: Interview]

She also explained that textbooks and corpus linguistics worked well together to help guide her in creating searches where the textbooks could give her "some words to begin with" [Holly: Survey 3 Q7] when she "had trouble thinking of words to search for" [Holly: Survey 3 Q3].

Beyond curricular decisions, Holly identified two other possible roles for corpora in instructional materials. She suggested that a corpus of spoken language might help in identifying correct forms to teach when tutoring pronunciation, but a comment in the interview— "I don't think you can do that with a corpus"—showed her awareness that corpora cannot answer every question. She was more certain about the option of using learner corpora "to look at what is really going on in the classroom and help focus in on what needs to be focused on and not spend so much time on things that they're really doing well" [Holly: Interview]. Thus, she showed an awareness of the different corpus types and why one might choose each for different purposes.

Limitations of Corpus Linguistics

Two students specifically addressed the limitations of corpus linguistics. The first was Holly's comment in the preceding paragraph about the inability of corpora to provide guidance about pronunciation. The second came from Tracey, who showed an awareness of how the content of a corpus can affect the quantitative findings as well as qualitative interpretations of those findings:

> As an aside, I was surprised to see that the spoken category contained data that was primarily from television – news and talk shows – and very little "everyday" person to person, on the street language use….In addition, the LSG described how adverbs and adjectives differ across registers according to the Longman SWE Corpus [Longman Corpus of Spoken and Written English], an interesting fact not evident in the COCA. [Tracey: Narrative 2]

Understand And Analyze Corpus Data

For Project 1 (the group project), it was anticipated that students might use two types of corpus data: the concordances provided for *not*, *n't*, and *no*, the use of which was optional; and any frequency data from LSG that was relevant to the focus of their lesson plans. The second (individual) project required students to generate and use corpus data such as frequencies, register distribution, and concordances. The six subjects generally showed the best development of this aspect of corpus literacy, though it was by no means uniform across students or time.

One of the surprises from Project 1 was that Owen's group performed their own COCA search on double negatives in spoken language. Owen's Project 1 narrative does not discuss search strategies or data analysis, so it is impossible to discern his role in performing this search, but it is striking that his group chose to do their own search since the project did not require students to do this or even to use the provided concordances for *not, no*, and *n't*, and students had not yet had any training in using COCA. Owen's Project 2 provided good support for his claim that he became "quite adept at searching the corpus and analyzing the results" [Owen: Survey 3 Q7]. His narrative noted his own surprise at the infrequent occurrence of *funner* and *funnest*, which he supported by including statistics on their occurrence. His evident comfort with corpus data was confirmed by the inclusion in his teaching materials of concordance lines and a frequency table, as well as by his comment on frequency data showing that “the use of the adjective ‘greener’ has doubled since 1999” [Owen: Narrative 2].

Holly was one of the four subjects whose group examined the provided concordances for *not*, *no*, and *n’t* for Project 1. However, it was not clear from her project how these concordances were used, so no inferences can be made about Holly’s ability at that time. Holly’s interaction with corpus data in Project 2, however, did suggest a strong ability to understand and analyze corpus data. She not only created a chart to track frequencies of forms across registers, but also formed and tested hypotheses about the data she gathered to complete that chart:

> At first I only looked at the numbers, unless there was something curious about the results. For instance, “more narrow” was common, but “most narrow” only had eight occurrences. I wondered if the “more” was being used as a determiner rather than as part of a comparative. The interesting thing was that it was a comparative, but in the “most narrow” concordances some of the “mosts” did show up as determiners. I also thought it was interesting that “narrow” in the comparative form was more common as a phrasal adjective, but in the superlative form it became inflectional. [Holly: Narrative 2]

Holly’s comment demonstrated a deeper grasp of the grammar than many of the other students appeared to hold, which probably contributed to her more nuanced understanding of the corpus and its output.

Tracey and Emily were in a group for Project 1 that also used the provided concordances. While their group enthusiastically integrated several of the concordances into their lesson, it was not clear which individuals were responsible for analyzing and selecting the corpus lines used in the project, so, again, no inferences can be made about either subject’s abilities. Data from Project 2 provided more insight into Tracey and Emily’s development.

Tracey’s comment about using COCA to test assumptions about language (see quote in previous section) indicated her apparent facility with corpus data. She clearly comprehended register distribution information in LSG and was able to analyze concordance output well enough to test her assumptions. However, the focus on looking for “words” in the following two comments makes one wonder whether her ability to perform searches may have extended more to using the corpus as a tool for lexical investigation than for grammatical investigation:

> I enjoyed using the site and have gone back to see how other words are used, just to satisfy my curiosity. [Tracey: Narrative 2]
>
> Performing a search in the COCA was an easy process. I played

> around with lots of words to see their frequency, distribution, and to read the concordances, seeing how words are used. [Tracey: Survey 3 Q7]

While neither project provided data concerning Emily's ability to use frequency information, Project 2 clearly showed that Emily had little difficulty searching COCA to generate concordance lines for *further* and *farther*, which she was able to analyze well enough to identify that the terms were used interchangeably in the corpus (see quote in the previous section). However, her handout included adverbial uses of *further* that did not match her lesson's pedagogical focus on *further/farther* as distance terms; unfortunately, the data is insufficient to determine whether this issue reflects her inexperience as a teacher, a lack of grammatical knowledge, difficulty analyzing concordances to identify appropriate examples, or some combination of these factors.

Claire also used the provided concordances for Project 1 as well as frequency information in LSG. At first glance, the latter seemed to cause Claire little difficulty: she attached frequency charts from LSG to her narrative and explained that in examining them, "my group partners… and I discovered the negation form more frequently used in all forms of the English language is not/n't." [Claire: Narrative 1]. However, Claire acknowledged that she was not always as comfortable with frequency information as the above quote would suggest:

> you know when I went to COCA it was ok, I could understand them somewhat, but in our book, the way um in the Longman Student Grammar, they have some charts in there, some graphs, that I could not understand. [Claire: Interview]

Claire's Project 2 narrative contained a very detailed description of her search strategy which supported her assertion that she had less difficulty understanding corpus output:

> I checked the word fun. I thought I would frequently find the nonstandard comparative words, funner and funnest. This was not the case. There were only four corpus lines for funner and eleven for funnest. The correct comparatives more and most fun had hundreds of corpus lines. [Claire: Narrative 2]

The final subject, George, was in a group that did not use any corpus data for Project 1. For Project 2, George researched comparative and superlative forms of *quiet, gentle* and *angry* in the Spoken and Academic registers with some success. He produced a handout containing the frequencies he obtained from COCA (including information on changes across time) as well as an evaluation of patterns he saw. Even though he did not provide any references to concordances, it is clear that he had some facility with analyzing them since the differences he cited could only be determined from examining usage in context: for example, for the form *gentlest*, he states "Spoken: most often for people. Academic: as often for ideas, etc, as for people." [George: Lesson Plan 2].

In his Project 2 narrative, George provided quite a bit of detail about his searches and the range of frequency and distribution data. For example, in discussing the three target adjectives, he stated, "the obvious tendency for greater usage of the inflectional variety of comparative and superlative over the phrasal variety in Spoken and Academic situations is conclusive." He further explained, "I attribute the disparity of usage between the inflection versions and the phrasal versions of comparatives and superlatives as the difference between *casual* and *formal*. They represent stylistic opposites or extremes…" [George: Narrative 2]. Unfortunately, George failed to realize that his theory does not take into account the difference in formality represented by the Academic and Spoken registers and which should, according to his theory, have led to frequency differences between them (i.e., one would then expect more inflectional forms in Spoken, but more phrasal forms in Academic).

Use of Corpus Data To Develop Teaching Materials

Owen clearly saw the value of incorporating corpus-based materials into language teaching, as evidenced by his group's lesson plan for Project 1, which included a two-minute teacher-led explanation of the value of corpus linguistics prior to examination of five examples of double-negatives taken from COCA. Owen's lesson plan for Project 2 included a good example of data driven learning, where students created hypotheses about the most likely comparative/superlative forms of *fun* which were then tested through examination of a table containing usage frequency for the various possibilities in COCA, broken down by register. Owen appeared to have given some thought to the best way of presenting this information: "I decided a table would best show how the evolution of the language, through common usage, seems to have selected "funner' and 'funnest' (words both found in the dictionary as gradations of 'fun') for extinction" [Owen: Narrative 2].

Although Holly's Project 1 narrative explained how her group "looked up *not* and *n't* using the conversational register of the corpus" [Holly: Narrative 1], it's not clear from either her narrative or the group's lesson plan how this information was used in creating instructional materials. For Project 2, Holly utilized quite a lot of corpus information for inductive, language analysis activities in her lesson, including information on register variation from LSG, register distribution from COCA, frequency information from COCA, and concordances for *likely* and *narrow*. One notable feature of her lesson was her decision to include all of the concordances for *most narrow*, rather than omitting the ones that were determiners (e.g., Most narrow alleys have driveways). Though this would give her students more of a challenge in terms of inductively discovering patterns, it would also provide a much richer language analysis problem for them and is an appropriate use of concordances (Reppen, 2010).

For Project 1, Emily and Tracey's group used the provided corpus data in an inductive activity for which they selected groups of three concordance lines per target item and asked guiding questions to help learners see patterns of usage. However, although both subjects commented in their narratives on the value of this type of activity and the use of concordance lines in this way, their group's attempt to do so was not totally successful. In particular, some of the groupings of concordance lines did not fit together well. For example, one grouping includes three examples of *be* + *not*, but in one case, *be* is used as an auxiliary verb (*he was not speaking*), while in the other two cases, it is the main verb but is followed by different forms—an infinitive (*is not to have*) and an adjective (*is not possible*).

By Project 2, Emily had become much more effective at grouping concordances for inductive activities. However, as has already been discussed, she included concordance lines for an irrelevant meaning of *further* and may have selected other concordance lines based on a preexisting and inaccurate conception of what the corpus data should show.

Regarding Project 1, Tracey stated, "I had a hard time understanding how to use corpus material in my lesson plan" [Tracey: Narrative 1]. For Project 2, Tracey included a corpus-based activity that focused on the use of *pretty* as both an adjective and an adverb. Her handout included groups of five concordance lines in each of three sections; students identified the word that *pretty* modified, the word class of the modified word, and patterns in each section of concordance lines. Students then chose one sentence from each section and replaced *pretty* with a synonym (i.e. they focused on both meaning and function). Tracey also understood that instructors need to carefully consider how they will use corpus output based on the needs and level of their students:

> I did have a difficult time choosing sentences to use in my corpus activity since much of the concordance examples were hard to understand out of context or they were from sources at a level more

> advanced than I thought these intermediate students could follow. [Tracey: Narrative 2]

In spite of Tracey's perception of the difficulty of selecting from concordance output, the activity which she included in her lesson plan illustrated her success at choosing appropriate lines and grouping them well.

For Project 1, Claire's group used corpus information in several ways: frequency information was used to determine which forms to focus on; 10 lines "taken from Corpus of Contemporary American English" [Claire: Lesson Plan 1] were used for a mechanical practice activity; and oral activities focused on negative words used in "sample sentences inspired by" COCA [Claire: Lesson Plan 1].

In Project 2, Claire again used results of her COCA searches to guide her decisions about instructional focus: "After checking these various words, I concluded that most spoken comparative words are used correctly, and my lesson plan should be general practice of typical single syllable adjective forms" [Claire: Narrative 2]. Her instructional materials for Project 2 used 21 lines of corpus output, grouped by grammar point, for an inductive activity on comparative/superlative adjectives which she mistakenly refers to as "deductive":

> In my lesson plan, instead of reviewing the rules for the comparative adjectives, I am going to use some corpus lines in a deductive [sic] activity and see if they can discover the rule for 'adjective, comparative, superlative.' I will incorporate the words *more fun* and see if they can discover that this is an irregular form of one-syllable adjectives. [Claire: Narrative 2]

Claire also demonstrated an awareness of the need to be selective with using corpus output, noting that she chose "simple sentences from COCA for a corpus based exercise because the low-to-mid-intermediate ESL student has difficulty in reading comprehension. Sentences that are too complex could be a source of confusion" [Claire: Lesson Plan 2].

George was the least successful of the six subjects in terms of using corpus data to develop teaching materials. His group did not use any corpus-based information in their teaching materials for Project 1. George's COCA research for Project 2 did not fit very well with the lesson plan itself, in which he chose to completely ignore two of his research foci and concentrate solely on adjectives that end in –y such as *angry*. His lesson plan did not mention any use at all of his COCA findings, not even of the handout in which he summarized his COCA findings. Thus, while George appeared to have acquired the ability to conduct some types of corpus searches and read their results, he showed little evidence of being able to apply those findings to classroom instruction.

DISCUSSION AND IMPLICATIONS FOR LANGUAGE TEACHER EDUCATORS

Considering all the case studies together, the corpus literacy training received by these pre-service teachers appears to have been effective while also highlighting a need for improvement in several areas. Additionally, though not surprisingly, the effectiveness of the training was not uniform across all six subjects, reflecting the differing strengths each subject brought to both the course content (English grammar) and the technological issues related to corpus linguistics. Several positive findings provide encouragement for continuing this type of training for pre-service teachers.

During the semester, these subjects interacted with concordance lines perhaps more than any other type of corpus output. This extensive exposure appears to have had a positive effect, in that several subjects attempted (with varying degrees of success) to utilize concordance lines to plan a data-driven learning activity in their lessons (see Holly, Owen, & Tracey, for example). It is evident, however, that more focused instruction is needed to help teachers develop two skill areas: (a) utilizing concordance output to identify lexico-grammatical patterns rather than focusing only on lexical items and their meanings; and (b) organizing and presenting concordance data in ways that lead more clearly to autonomous learning for their students. To address the first skill area, pre-service teachers will need to be exposed to more lexico-grammatical patterns that have been identified and discussed by corpus linguists (Biber, Conrad & Reppen, 1998; O'Keeffe, McCarthy & Carter, 2007). Once teachers begin to notice such patterns, they will be more apt to look beyond the focused lexical word in concordance output. One way to address the second problem would be to provide more practice in designing questions that guide learners to notice the patterns represented in different groupings of concordance lines (Reppen, 2010).

One unexpected positive outcome of this training was that some case study subjects began to see themselves as empowered to critique published teaching resources and to create their own materials, based on frequency or distribution information from a reliable corpus, as a supplement when that missing information seems important or critical. While this evidence comes mostly from the stronger subjects (see for example Emily, Owen, and Holly), more could be done to encourage this type of thinking in all pre-service teachers.

While all six subjects discussed the topic of corpus-generated information, only the stronger subjects seemed to address the interaction of corpus linguistics and language instruction, or were able to infer some of the limitations of corpus linguistics (especially Holly and Tracey). Since the latter two topics are also essential to well-rounded corpus literacy (Mukherjee, 2006), it is evident that more specific training is needed in those areas.

Several over-arching issues have been brought into focus through the analysis of these case studies. Both Claire and George noted their difficulties with the online interface with COCA and general searching requirements, and Claire also needed help from a family member to navigate the COCA site, while other subjects became quite adept at conducting more sophisticated searches. Additional training (workshops in a computer lab or very detailed written instructions) should be offered for those students who feel less comfortable with computing in general. Another possibility would be to break the projects into smaller steps and offer feedback and revision possibilities at each step, so that the projects seem less overwhelming. This is especially important for the second project where students are required to work alone *and* to incorporate corpus-based information into their lesson plans and handouts.

A final, perhaps more serious problem relates to integrating corpus literacy training into an entry-level grammar class. While that still seems like the most logical place for introducing pre-service teachers to corpus linguistics, it is also true that those subjects with a weaker grasp of grammar, like George and Emily, have more difficulty conducting corpus searches (especially by part-of-speech or inflection) and are more likely to misrepresent their findings to their students. The challenge lies in finding a way to infuse more of the instruction of the course content itself (i.e., basic pedagogical English grammar) with corpus-based information, using more corpus-based materials that the instructor creates to specifically address the course content, all the while using that instruction and those materials to introduce students to how to use a corpus to enhance instruction. It requires an iterative process, and the design of the curriculum likely will continue to evolve as the researchers continue to teach this course. At the same time, the ideal situation would be to continue to integrate corpus linguistics (*and* grammar) in several other courses so that pre-service teachers have more opportunities to develop corpus literacy in addition to other technological skills (Desjardins & Peters, 2007).

On a positive note: In her interview, Claire discussed ways in which corpus-based, data-driven learning could be a useful resource to her as a teacher for checking usage to identify problem areas to address in instruction, and also useful to students who could investigate words:

> This is the way I would sell it to a student, and I'm using the word *sell* seriously, it's a mystery, it's like finding a mystery, like clues to a mystery. That's how actually I would market it to my students, if my students were old enough, I would introduce them to the corpus. [Claire: Interview]

Claire represents the average student in this study, and it is extremely encouraging to note that she has made very strong connections between corpus linguistics and *language instruction*. Her comments demonstrate that she is engaged in a revision of her whole language teaching *philosophy*, amending it to include corpus linguistics as a central focus. She acknowledges the possibilities that corpus linguistics presents for promoting autonomous learning in her future students and, implicitly, for herself as their teacher.

References

Anderson, W. & Corbett, J. (2009). *Exploring English with online corpora: An introduction.* New York: Palgrave Macmillan.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English.* London: Longman.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* New York: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36,* 9-48.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL, 19,* 307-328.

Breyer, Y. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning, 22*(2), 153-172.

Cowan, R. (2008). *The teacher's grammar of English: A course book and reference guide.* New York: Cambridge University Press.

Davies, M. (2008). Corpus of Contemporary American English. www.american-corpus.org

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness, 17*(1), 25-43.

Desjardins, F. & Peters, M. (2007). Single-course approach versus a program approach to develop technological competencies in preservice language teachers. In M.A. Kassen, R. Lavine, K. Murphy-Judy, & M. Peters (Eds.), *Preparing and developing technology-proficient L2 teachers* (pp. 3-21). San Marcos, Texas: Texas State University Press.

Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal, 63*(4), 363-372.

Granath, S. (2009). Who benefits from learning how to use corpora? In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 47-65). Amsterdam: John Benjamins.

Healey, D., Hanson-Smith, E., Hubbard, P., Ioannou-Georgiou, S., Kessler, G., & Ware, P. (2011). *TESOL technology standards: Description, implementation, integration*. Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Helt, M. & Reppen, R. (2008). Training teachers to use corpus resources. *Teacher Education Interest Section (TEIS) Newsletter*, *23*(2). Retrieved June 28, 2011, from http://www.tesol.org/NewsletterSite/view.asp?nid=3091

Hubbard, P. & Levy, M. (2006). The scope of CALL education. In P. Hubbard & M. Levy (Eds.), *Teacher education in CALL* (pp. 3-20). Amsterdam: John Benjamins.

Lombardo, L. (2009). Introduction: Establishing guidelines for the use of corpora as resources for learners (and their teachers). In L. Lombardo (Ed.), *Using corpora to learn about language and discourse* (pp. 7-37). Frankfurt am Main: Peter Lang.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching, 41,* 563-574.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone*. Cambridge: Cambridge University Press.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art – and beyond. In S. Braun, K. Kohn & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 5-23). Frankfurt am Main: Peter Lang.

Murphy-Judy, K. & Youngs, B. L. (2006). Technology standards for teacher education. In P. Hubbard & M. Levy (Eds.), *Teacher education in CALL* (pp. 45-60). Amsterdam: John Benjamins.

O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly, 37*(3), 389-418.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press.

Oksefjell Ebeling, S. (2009). 'Oslo Interactive English': Corpus-driven exercises on the Web. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 67-82). Amsterdam: John Benjamins.

Oxford, R. & Jung, S. (2007). National guidelines for technology integration in TESOL programs: Factors affecting (non)implementation. In M.A. Kassen, R. Lavine, K. Murphy-Judy, & M. Peters (Eds.), *Preparing and developing technology-proficient L2 teachers* (pp. 23-48). San Marcos, Texas: Texas State University Press.

Peters, M. (2006). Developing computer competencies for pre-service language teachers: Is one course enough? In P. Hubbard & M. Levy (Eds.), *Teacher education in CALL* (pp. 53-65). Amsterdam: John Benjamins.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge, UK: Cambridge University Press.

Römer, U. (2004). A corpus-driven approach to modal auxiliaries and their didactics. In J. Mc.H. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 185-199). Amsterdam: John Benjamins.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik: A quarterly of language, literature and culture, 54*, 121-134.

Varley, S. (2009). I'll just look that up in the concordancer: Integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning, 22,* 133-152.

Author Note

Julian Heather, Department of English, California State University, Sacramento; Marie Helt, Department of English, California State University, Sacramento.

Correspondence about this article should be addressed to Julian Heather, Department of English, CSU Sacramento, 6000 J Street, Sacramento, CA 95819-6075. Email: jheather@csus.edu