St. Cloud State University

theRepository at St. Cloud State

Culminating Projects in TESL

Department of English

12-2019

Graduate Assistant Corpus Literacy

Jason Badger

Follow this and additional works at: https://repository.stcloudstate.edu/tesl_etds

Recommended Citation

Graduate Assistant Corpus Literacy

by

Jason Badger

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfilment of the Requirements

for the Degree of

Master of Arts

in English: Teaching English as a Second Language

December, 2019

Thesis Committee:
John Madden, Chairperson
Michael Schwartz
Sharon Cogdill

Abstract

This study investigated awareness of, use of, and experiences with linguistic corpora by current and former master's in Teaching English as a Second Language students at a university in the Midwest. Using an online survey, participants were asked to report their knowledge of terms used in the field of corpus linguistics. An interview was also conducted to further investigate experiences with corpora resources. Results showed that most of the participants had a general, but superficial, awareness of corpus linguistics. Participants did not report use of many of the corpus resources listed in this study. Few participants reported use of corpora over time as well. This study did not gather enough data to draw conclusions from the interview. Implications for further research and for MA programs are discussed.

Table of Contents

List of Tables

Table   Page

List of Figures

Chapter I: Introduction

The term 'revolution' is often used to describe the impact corpus linguistics can have on language instruction. The term characterizes the optimism that the field has carried for around thirty years (Breyer, 2011) for the applications of corpora. These possibilities range from development of materials, empowering learners and teachers, and include new styles of teaching. Despite this, several authors have noted that there remains little teacher interaction with corpora (Reppen, 2008) (McCarthy, 2008) (Breyer, 2011), and therefore little corpora use in classrooms.

I have found the impact that corpora can have on language learning to be valuable and exciting. My first experience with corpora, as is with many people, was the Compleat Lexical Tutor. The site was mentioned in a course on technology, but little time was devoted to how to use it. Initially, I found the site confusing and difficult to use. The interface could be hard to work with, there was little offered in tutorials, and the purpose of each of the tools was not always clear. Over time and through trial and error though, I began to use the tools offered by Compleat Lexical Tutor to help me focus the classes I taught and offer clearer explanations to my students. Since then, I have enjoyed exploring other corpora besides the Compleat Lexical Tutor. The applications of sites like this seem incredibly powerful for a language teacher and perhaps more importantly, a language learner. As a teacher, corpora gave me another way to view language. Not only do I have a grammar book for structural definitions, a dictionary for word definitions, but also a corpus for descriptions of time, place, and frequency of a piece of language. As a language learner, I can use corpora to explore a chunk of language. I would like to know if other teachers have had similar experiences as I have. For this research project, I am

interested in other teachers' knowledge on this topic and their own experiences with linguistic corpora.

Problem Statement

The availability of corpus resources has grown greatly in the last decade. Corpora offer teachers another tool to use to help explain language to English language learners. Despite their availability, they don't appear to be widely used. This research will investigate attitudes and use of corpora in English language instruction by MA TESOL students at a public university in the Midwest. It will examine teachers' familiarity with corpora, their attitudes toward corpora use in the classroom, and overall their familiarity with corpora.

Chapter II: Literature Review

What is corpus linguistics? Corpus linguistics is essentially conducting computer aided analysis on large bodies of texts. McCarthy (2008, p.2) points out that computer users use corpora every day when they make searches on Google, Bing, or any other similar program. Users search for text and find examples of that text in the internet. Microsoft Word also can act like corpus software. If you wanted to know how many times the word 'the' appeared in a text, you could search a document and find the answer. This type of search is exactly what the most common corpus tool, a concordancer, can do. However, you would have a hard time searching for all verbs in the present tense. Texts in corpora have been processed to allow searches of this type. In addition, Microsoft Word isn't designed to search texts that are large enough and varied enough to allow for useful generalizations about language. Corpora are often designed to look for a specific quality such as high frequency words in academic texts. For this, a variety of academic texts from different fields would need to be gathered and analyzed. In this way, corpora allow researchers to find out specific information about language use and variation in texts. Thus, corpora allow users to find the frequency of words and text strings, as well as examples of their use in context. Corpora show language use across genres such as spoken texts or academic texts and some provide information about language use over time.

The concordancer is a basic tool of corpus linguistics. It can quickly inform a user about a search term. The output of a concordancer gives all the results of a search term as well as the context in which that term is found. The two examples below are from the Corpus of Contemporary American English (COCA) (Davies, 2008) and Lawrence Anthony's (2018)

AntConc software with COCA sample files loaded. The search terms are 'corpus' and

'linguistic'

Figure 1. COCA Key Word in Context

Figure 2. Antconc Key Word in Context

As seen in these examples, many concordancers color code search results to aid in analysis.

Concordancers also allow users to sort results to help identify patterns. The results can be sorted

in various ways such as by the words that proceed or follow the search term. These are the basic

functions of concordancers. The Brigham Young University corpora go further and color code

the text surrounding the search term by part of speech. Their software also allows for comparing

the results of searches as well as other features.

There are currently many free corpora available for teachers to use. One example is

COCA. In 2008, Mark Davies of Brigham Young University launched COCA and it is one of the

largest free American English corpora at 450 million words from sources between 1990 and

2012 (Davies, 2008). COCA is only one of over 10 corpora available from Brigham Young

University, all of which use the same search interface. The Compleat Lexical Tutor is another free online resource for teachers. As was mentioned, this site allows for vocabulary analysis. It also offers a concordancer for grammar forms and other lexical tools. The University of Nottingham provides a tool which highlights academic words from user input (Haywood, n.d.). The University of Michigan offers two corpora which could be especially helpful for teachers. One is the Michigan Corpus of Academic Spoken English (MICASE), and the other is the Michigan Corpus of Upper-Level Student Papers (MICUSP). The former is a spoken corpus, and the latter is a corpus of academic written texts which can be searched by student level, paper type, discipline, textual features, and nativeness. These resources are discussed further in Methodology.

The benefits corpora offer for teachers have been widely publicized in journals. The most obvious is having access to authentic spoken English examples whenever you need (Granath, 2009, p. 48). Römer (2006) calls this the "tireless native speaker". Instead of relying on what each teacher feels about what is common or correct, evidence can be produced for what is common and how it is used (Yoon & Hirvela, 2004, p. 260). Tsui (2004) demonstrated the impact that corpus data can have on informing non-native teachers' questions about language usage. When a teacher is unable to answer a student's question about why a word or usage is different, he or she can consult a corpus to find the answer. This might be especially attractive for non-native English-speaking teachers.

Another possible use of corpora is the creation of learner corpora. This might offer a teacher a way to analyze and display a particular student's common errors (Bernardini, 2004, p.

24). Learner corpora, once again, might provide teachers more than simply an intuitive feel for the errors their students commonly make (Mukherjee, 2006).

With corpora evidence, teachers can also examine how language and what language is presented in their textbooks. If needed, they can expand on what is presented, or even correct material that is inconsistent with what is found in the corpus data; Teachers need to be experts in their field (O'Keeffe & Farr, 2003). Teachers may also benefit from simply browsing a corpus. A teacher with better knowledge of linguistics and corpora might also find that browsing could improve their understanding of English (Breyer, 2009). In particular, teachers might expand their knowledge of language variation (Conrad, 2004), an area which textbooks often lack. If nothing else, corpora offer teachers another means of defining and explaining language to learners in addition to dictionaries and grammar textbooks (Granath, 2009, p. 47; Varley, 2009).

Despite the many benefits, there is also some risk with corpus use. Reppen (2008) and Conrad (2000, p. 556) note that teachers need to be able to understand what they find when they search a corpus. Corpora software and searches can be complex and require some training to understand. There is a risk that teachers will present misleading information to students without an adequate understanding of the data presented to them. McCarthy (2008) also points out that critiquing materials requires teachers to know if texts are corpus based, or corpus informed and what the difference is. There are drawbacks to most corpora themselves. For example, it is difficult to gather spoken language because of limitations on recording people, and the additional effort to transcribe spoken words. Therefore, no corpus can cover an entire language (Sinclair, 2004, p. x). Knowledge of what each corpus contains is an important part of corpus literacy.

Therefore, for teachers to best use corpora, they need to have a basic understanding of how to use and interpret them. For this reason, training is important. Conrad (2000, p. 556), and McCarthy (2008) argue that knowledge of this technology is a basic component to empowering teachers to be more than a receiver of research. O'Keeffe & Farr (2003) argue that teachers must acquire this knowledge during their training and go so far as to call technology education a "fundamental obligation" (p. 389). Teachers however, are not receiving training on corpus use (Reppen, 2008; McCarthy, 2008; Breyer, 2011, p. 77). Some university programs do not require students to take any courses focused directly on using corpora. In addition, technology focused courses are not in the core requirements for an MA in TESL at some universities. Authors have offered many explanations for the lack of corpora use by language teachers. These include lacking computer skills, the absence of cheap or free resources, lack of corpora knowledge, lack of classroom applications, time taken in activity design, and uncertainty about corpus results (Leńko-Szymańska's 2014; Breyer, 2009).

Teachers' attitudes towards corpora for teaching have not been investigated (Breyer, 2011, p. 5). There are few studies focusing on the teacher, most studies are learner centered. One study that has investigated teacher use of corpora is Leńko-Szymańska's 2014 study. Leńko-Szymańska followed teachers in training during a short course on using corpora. Before the course, few to none knew anything about corpus linguistics. When reporting on the course, participants' reactions were mostly positive, and some began using corpora to research their own questions. However, some issues were reported such as, participants thought the projects were time consuming and the software could be hard to use.

Breyer (2009) also studied student teacher reactions to using corpora for language teaching. Much like Leńko-Szymańska, Breyer also reported that participants in the course found corpus linguistic fascinating and relevant to their teaching, but also reported that creating activities can be time consuming, software can sometimes lack functions that would make it easier to use, and it can be difficult to find appropriate corpora. Breyer also reports that student teachers considered several issues related to teaching. They discussed the importance of teaching 'rules' compared with authentic language, at what level of student language proficiency corpora would be most helpful, and how language variation affected instruction. One issue that was encountered was that instead of reflecting on authentic language usage for activity design, participants focused on finding examples that fit how they believed language is used.

Leńko-Szymańska (2014) notes that one possible barrier to teachers using corpora is the lack of available resources. Few books exist which are aimed specifically at teachers and language instruction. Anderson and Corbett's 2009 book Exploring English with online corpora: An introduction and Randi Reppen's 2010 book Using corpora in the language classroom are two of these books. They provide a basic overview of corpora, how to interpret results, how to run searches, and possible classroom applications of corpora. There are also more comprehensive books on corpora such as Breyer's 2011 book Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges.

Research Questions

Two trends are apparent in the literature, few teachers know about and are using corpora, and of those who are, many find corpora hard to apply to language instruction. This research aims to answer the following questions:

1. Do current and former MA TESL students report awareness of the field of corpus linguistics?

2. Do current and former MA TESL students report using corpus resources?

3. What kind of experiences do current and former MA TESL students report having with the corpus resources they have used?

4. In what ways do current and former MA TESL students report using corpora?

Chapter III: Methodology

This section will present the participants, instruments, and procedures of the research. As an overview, this research used an online survey to collect initial data and then, a face to face interview to collect more detailed data.

Participants

This section will describe the participants recruited for this research. This research was designed to survey former and current MA TESL students at a public university in the Midwest. Although there were a total of 34 participants, 2 people did not answer any part of the survey other than the basic biographic questions. Their responses were excluded reducing the total number of participants of this study to 32 MA TESL students who were either active students or had completed their degrees. Of the 32 survey participants, one individual participated in a follow up interview.
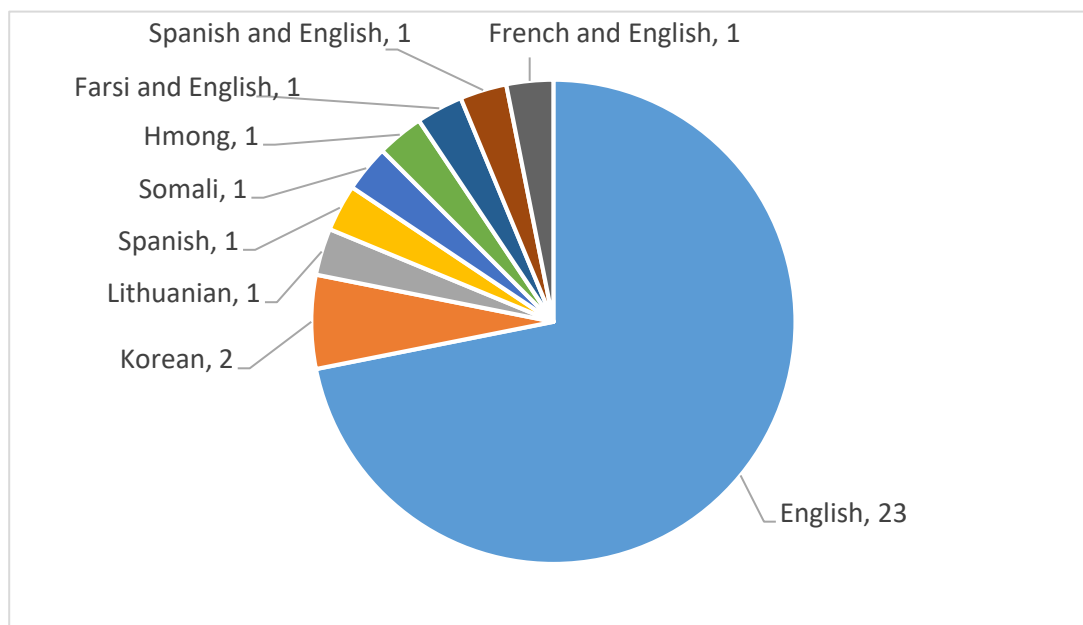
Figure 3. Participants First Language(s)

Figure 3 displays the language that participants indicated was their first language or languages. The languages reported are as follows: 23 English, 2 Korean, and 1 of the following languages: Lithuanian, Spanish, Japanese, Somali, Hmong. Three participants listed two languages: Farsi and English; Spanish and English; and French and English.

The median number of years that participants had been teaching English or another language was 2.5 years. The mean number of years teaching was 3.0. The standard deviation for the number of years taught was 2.7. The greatest number of years was 11 and the least was 0. Twenty-seven participants, or 84% of the total participants, had 5 or fewer years of teaching experience. The number of participants that indicated they were teaching a language at the time of the survey was 21 and 11 responded that they were not.

Instruments

This study used a survey and an interview to collect data. The first stage was an online anonymous survey through SurveyMonkey (Appendix A). The second stage was an interview (Appendix B). The survey was designed to quantitatively answer the first two research questions:

1. Do current and former MA TESL students report awareness of the field of corpus linguistics?

2. Do current and former MA TESL students report using corpus resources?

The interviews were designed to qualitatively answer the second two research questions:

3. What kind of experiences do current and former MA TESL students report having with the corpus resources they have used?

4. In what ways do current and former MA TESL students report using corpora?

Survey. The survey had three main sections: Biographical Information, Corpora Experience, and Corpus Resources and Frequency of Use.

Biographical information. The survey first informed participants of the research and asked for their acceptance to participate. The second question checked that they were in the target population. The target population for this study was current and former masters in TESL students at a public university in the Midwest. The third question asked participants to report their first language or languages. This question was asked to provide characteristics of the population sampled. Questions 3, 4, and 5 in the survey (Appendix A) collected data concerning participants' teaching experience. These questions were asked because the research specifically targeted teachers.

Corpora experience. Survey questions 6 – 9 were adapted from Leńko-Szymańska's (2014) research. These questions investigated participants' awareness and understanding of the field of corpus linguistics, and an important corpus research tool, the concordancer. Questions 6 through 9 were formatted in pairs. The first question in the pair was multiple choice. For example, question 6 asked, "Have you ever heard the term corpus and do you know what it is? Choose from the answers below." The second question was open-ended. Question 7 asked, "Define in your own words the term corpus. Even if you do not know or are not sure what it is, try to explain how you understand the term." Question 8 was multiple choice and asked, "Have you heard of a concordancer and do you know what it does?". Question 9 was open-ended and asked, "Define in your own words the term concordancer. Can you give any examples?"

The multiple-choice questions allowed participants to select their perceived awareness of the concept. The open-ended questions allowed the researcher to compare participants' answers with definitions of the concepts and assess what participants demonstrated that they knew. Comparing the multiple choice and open-ended questions allowed comparison of perceived and actual awareness of the concepts referred to in the question.

Figure 4. Survey Question 10

Corpus resources and frequency of use. Questions 10 and 11 were designed to give greater

detail to the kinds and frequency which resources participants used, and the frequency with

which they used each resource. Question 10 asked, "Please indicate how often you use each of

the following corpus resources". Question 11 asked, "How often do you use the resources

selected in the previous question?". However, when the paper survey was created online in

SurveyMonkey, these two questions were combined. Instead of one question for resources and

one for frequency, for each resource listed, there was a checkbox for frequency of use. Figure 4

shows the question as it appeared online. This is discussed further in the limitations section and

an example is included in Appendix A. From this point on, questions 10 and 11 are considered as

one survey question and will be referred to as question 10. Question 10 asked participants to,

"Please indicate how often you use each of the following corpus resources."

One issue identified in the literature review was a lack of available resources for teachers

to use. At the time of this research, many corpus resources had become widely available online

and free to use. Ten resources that are available online and free to use were listed. However,

there are many more than the ten listed in this research. As of 2019, Brigham Young University

has 16 English language corpora available on their website (Davies, n.d.). Other websites offer

other corpora as well. The ten resources in this study were selected by the researcher based on: their usefulness as resources for teachers or students, their free availability online, because they are well known corpus resources, or a combination of these reasons. Each resource is discussed in detail in the following paragraphs. Five of the resources listed in the survey are hosted by Brigham Young University (BYU). These resources are: The Corpus of Contemporary American English (COCA), The Wikipedia Corpus, The Time Magazine Corpus, the British National Corpus (BNC), and the Corpus of Historical American English (COHA) (Davies, 2008, 2015, 2007, 2004, 2010). Two resources are hosted by the University of Michigan. These are MICASE: Michigan Corpus of Academic Spoken English and MICUSP: Michigan Corpus of Upper-Level Student Papers (n.d.). The last three resources are The Compleat Lexical Tutor (Cobb, n.d.), The International Corpus of English (2016), and The Brown University Corpus (n.d.).

The resources from Brigham Young University were selected for several reasons. One is that BYU uses the same interface for all its corpora. This makes it easy to access different corpora because a single user interface can be learned. Different corpora can be used to answer different questions about English usage without needing to learn new tools. COCA was selected for this research because it is a large, broad corpus of American English. This is useful to both teachers and students as a source of authentic English in use. Davies (2008) also wrote that it was, "probably the most widely-used corpus of English". The British National Corpus, the Corpus of Historical American English, and the Brown University Corpus all have similar use for teachers; they are broad corpora of general English. The population surveyed for this research could have included teachers at the K-12 level, university level, and international level. For this

reason, both American English and British English corpora were included in the survey. The Wikipedia Corpus might be useful to teachers as an academic-like corpus. BYU allows users to make topic specific smaller corpora from the larger Wikipedia Corpus as well. This might be useful for a teacher in an English for Academic Purposes program with a group of students who all plan on studying the same topic. The Time Magazine Corpus could be useful for examples of written English or to look at variation in English.

The University of Michigan corpora were listed because of their possible application in university English for Academic Purposes programs. MICASE is a corpus that contains "academic speech events". (Michigan Corpus of Academic Spoken English, n.d.). Specifically, this corpus contains transcripts of individuals speaking in a university context. For example, student speech during a study group or during a lab section. This makes this corpus potentially useful to teachers of university students or students who want to study at the university level. MICUSP is a corpus of upper level academic papers. This corpus could be useful for authentic examples of university level writing.

The Compleat Lexical Tutor website is the resource with the clearest use for teachers and students. In fact, the tagline for this website is, "data-driven language learning on the Web". (Cobb, n.d.). This website provides many corpus-based tools for teachers, students of English, and researchers. The focus of the site is vocabulary analysis with a focus on high frequency words. An example of a tool the site offers is the Vocabprofile tool. This tool uses vocabulary word lists to classify the words in a text. The Compleat Lexical Tutor offers three different versions of this tool for texts aimed at different education levels. One version, VP-Classic,

highlights and sorts all the words in a sample text into categories of the first thousand words, second thousand words, academic words, and words that did not fit in the first three categories. Therefore, a teacher can focus instruction on academic vocabulary for example. The tools on the site allow users to select different corpora to use in analysis as well. Two of the corpora offered are the British National Corpus and the Corpus of Contemporary American English.

The International Corpus of English is probably most useful for teachers of international English. Many students of English will not live in an English-speaking country. Instead, they will speak English to other non-native English speakers or to English speakers from countries other than the United States. This corpus could therefore be useful to a teacher whose students will work internationally.

Interview. The interview portion of this research used a semi-structured format to allow for participants to include information they felt was important regarding their use of linguistic corpora (see Appendix B for interview questions). The semi-structured format allowed participants to elaborate on experiences that they felt were important but might not be said in a structured format.

The interview questions were from Shaw's 2011 research. The interview questions had three sections. The first section elicited participants' background with corpora. The second section elicited how corpus resources were used with students. The third set of questions elicited how teachers used corpus resources and how they felt about the resources used.

Procedures

This section will describe how this research was carried out. This research aimed to survey MA TESL former and current students at a public university in the Midwest. To recruit participants, the graduate studies office of the university, where this research was conducted, provided the researcher with a list of students who were active MA TESL students or had completed their degrees. With this list, the researcher emailed invitation to the linked SurveyMonkey survey (see appendix A) to each person. The survey was hosted through the university where this research was conducted. The survey was available for approximately one month. The email was sent two times to all participants to encourage participation. Participants accessed the survey through the link and completed the survey. At the end of the survey, questions 12 and 13 requested participation in an interview. Participants willing to be interviewed had the option of leaving an email address. These addresses were stored in a separate file to ensure the anonymity of survey results. With the email provided from this question, the researcher contacted the one participant who was willing to participate in an interview and set up a time to meet. At this meeting, the participant was given an interview consent form and the survey was conducted in a semi-structured style. Typed notes were taken of the participant's responses to interview questions.

Chapter IV: Analysis

This chapter, first, describes how the survey and interview were analyzed and then reports results of each instrument. The survey results are reported using descriptive statistics as frequency counts. The interview results are reported by grouping common themes.

Survey Analysis

Analysis of corpora experience. Questions 6 through 9 of the survey were designed to answer the first research question, "Do current and former MA TESL students report awareness of the field of corpus linguistics?" There were two pairs of questions in this section. The first pair, questions 6 and 7, was a multiple-choice question and the second was a short answer question. The first pair of questions asked participants if they had heard of corpus linguistics and then to define it. Responses to the multiple-choice question were reported as totals for each response. A description of how the short answers were analyzed follows.

To evaluate the responses to question 7, three criteria were used: 1. It is principled or representative; meaning that it represents the language that a user wants to study and has enough samples to draw meaningful conclusions from. For example, if a user wants to learn about how language is used in science, that user would not look in a corpus of newspapers. 2. It is a collection of language. 3. It can be analyzed by computer programs. These criteria were selected from Helt and Reppen's (2008) definition of a corpus, "A corpus is a representative, principled collection of language (spoken and/or written) that is stored electronically and can be analyzed both quantitatively and qualitatively using computer programs." The definition was broken down in this way because corpus researchers have pointed out how important it is to understand where the language in a corpus comes from in interpreting results of a search. A basic overview of a

corpus would tell a user that it is a collection of language stored electronically, but to be an informed user, a deeper understanding is needed.

When the responses to question 7 were downloaded from SurveyMonkey, they were recorded in an Excel spreadsheet. To this spreadsheet, three columns for each of the three parts of the definition of a corpus were added. Then, each participant's response was read. If the participant's response used language that indicated an understanding of a part of the definition, the appropriate column was given a '1'. For example, if a response stated that a corpus is a collection, the column, "Collection" would get a '1'. If the response did not include language that indicated that a corpus is a collection, the column was marked '0'. Figure 5 provides an example. Through this process, the number of responses with each part of the definition could be counted. Apart from that, the total number of parts that each response had was also counted. This was done so that definitions could be given a numerical score.

Principled	Collection	Analyzed by Computer
0	1	0
0	1	0
0	0	0
0	1	0
0	1	1

Figure 5. Question 7 Coding

Responses that indicated an understanding of criteria one, "It is principled or representative", included language that indicated that a corpus is principled. For example, statements like "subject matter", "specific scholarship", or "from a certain field" all qualify. An understanding of criteria two, "It is a collection of language", was demonstrated through use of keywords that relate to collection; for example, "body", "collection", "database", and "bank".

However, keywords were considered within the context of the sentence. For example, a response that used the word "body" is not enough to indicate an understanding of the word as it relates to linguistics. Definitions needed to specifically relate to language, text, words, transcripts, or other references to language. The third criteria was, "It can be analyzed by computer programs". Participants demonstrated understanding through statements that indicated they could search a corpus, that it was digital, or that it could be analyzed. Use of the term "database" also indicated an understanding that it could be analyzed by computer programs because the definition of database is "usually large collection of data organized especially for rapid search and retrieval (as by a computer)" (Database, n.d). Examples of this language could be "database", "find trends", "analyze", and "searchable".

The next pair of questions, 8 and 9, asked participants if they had heard of a concordancer and then to define it. Responses to the multiple-choice question was reported as totals for each possible response. The process for analyzing short answer responses follows:

Responses to the short answer definitions of a concordancer provided by participants were evaluated as either 'yes', the participant understands what a concordancer is or 'no', the participant doesn't. This analysis differs from the analysis performed on question 7, define a linguistic corpus, because responses to question 9 either had a complete definition or nothing that indicated a participant knew what a concordancer was. There were no responses that included partial definitions. Responses that were evaluated as 'yes' demonstrated knowledge that a concordancer is a computer tool, used to search a corpus, and that it can return a search term or sequence in context.

As with question 7 in the previous section of the survey, all responses were collected and recorded in an Excel spreadsheet. However, a single column to record the response as 'yes' or 'no' was added instead of multiple columns for partial definitions. Each definition was read and marked 'yes' or 'no'. Participants who demonstrated an understanding that a concordancer is a computer tool used terms such as computerized system, tool, or software tool to describe it or what it does with language. However, these terms were considered in the context of the response. For example, a response of simply, "A tool" does not indicate an understanding of a concordancer because it does not relate the tool to language searches or analysis. Participants who demonstrated an understanding that a concordancer performs searches used terms such as, "sorts", "analyzes", "identify", "look for", and "search", to describe what it does with language. Participants who demonstrated an understanding that a concordancer returns search terms in context used terms and phrases such as "sort", "certain order", "in context", and "display the results". Again, these words were considered within the context of the response and then evaluated.

Analysis of corpus resources and frequency of use. Question 10 of the survey, which elicited resource use, was analyzed in four different ways. To answer the second research question, "Do current and former MA TESL students report using corpus resources?" frequency of use of each resource was collapsed into two categories: "Never used" and "Have used". "Never used" for responses of "Never" for a resource. "Have used" for all other responses such as every day, every week, every month, every 2-3 months, every 4-6 months, once or twice a year, and I used it in the past, but don't now. These responses indicated that a participant used a

resource at least once in the past. This clearly shows which resources MA TESL students and graduates have used.

The second way that question 10 was analyzed was by comparing this section to the previous section. Specifically, question 10 was compared with questions 6 Comparing responses provided a test on what participants might actually know. For example, if a participant reported regular use of a resource but reported that he or she had never heard of corpus linguistics, their reported use would be questionable. This section was analyzed in a third way. To further test this participant's knowledge, question 10 was also compared with question 7. Therefore, two questions were individually compared with reported frequency of use. A second reason for comparing survey responses was to investigate how experience level affects corpora use. Some participants responded to question 6 by indicating that they are already using corpora, others only responded that they have a general understanding. Comparing these participants provided insight into what resources had been used by experienced and by beginner users.

Score	Response
0	Never
1	I used it in the past, but don't now
2	Once or twice a year
3	Every 4-6 months
4	Every 2-3 months
5	Every month
6	Every week

Figure 6. Frequency of Use Coding

The frequency of use that participants reported were coded to facilitate analysis. Each response was given a numerical value. Figure 6 provides an example of the coding. The values are as follows: 0 Never, 1 I used it in the past, but don't now, 2 Once or twice a year, 3 Every 4-6 months, 4 Every 2-3 months, 5 Every month, 6 Every week.

Finally, results were presented in context of questions 6 and 7: participants' reported knowledge of a linguistic corpus and how completely participants defined a linguistic corpus.

Interview Analysis

Analysis of the interview results was limited due to only a single participant responding. For this reason, the results only reported on this individual's responses. Analysis of the interview was done by reading the participant's responses and looking for common themes. Themes were written down in a Word document as they were identified. Common themes were paraphrased for reporting. In results, these themes were compared to trends and findings in the corpus linguistics literature. It is impossible to know if this interview was representative of the population surveyed. This will be further discussed in the limitations section.

Survey Results

This section will present and examine the results that the thirty-two participants gave to the survey questions and will present the responses given by the single interview participant.

Corpora experience. This first section reports on the results of questions 6, 7, 8, and 9. This section had two pairs of questions. The pairs were questions 6 and 7, and, 8 and 9. The first question in each pair was multiple-choice and the second question was short answer.

Linguistic corpus results. Question 6 asked, "Have you ever heard the term corpus and do you know what it is? Choose from the answers below". Figure 7 shows the number of responses to each of the choices offered.
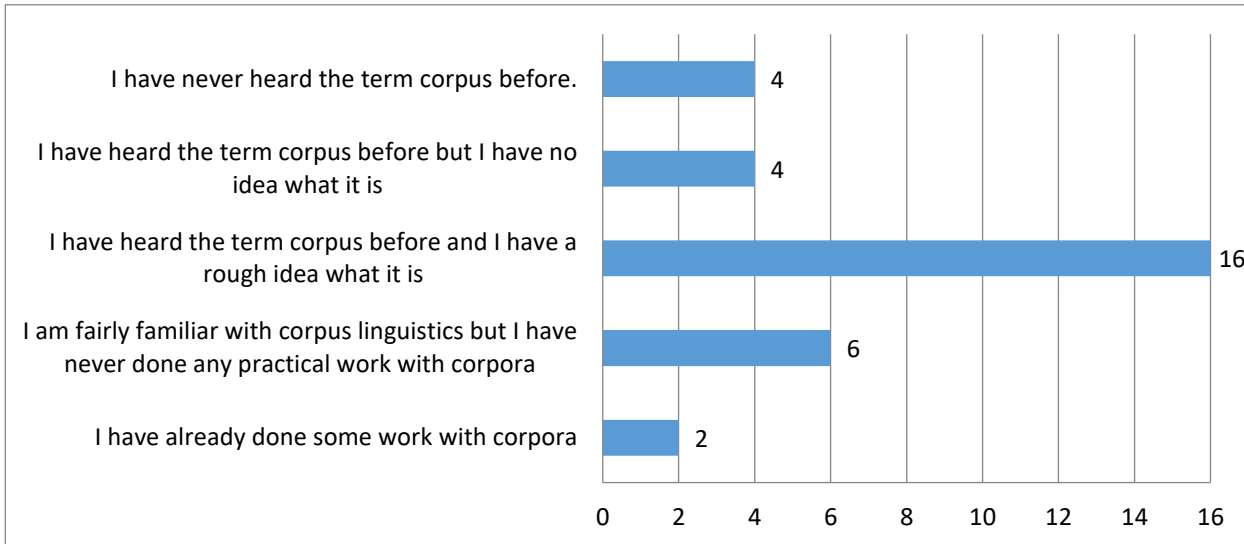
Figure 7. Multiple-Choice: Have you heard the term corpus?

Results show that the most given response was, "I have heard the term corpus before and I have a rough idea what it is"; fifty-percent of participants chose this response. Four participants reported they had never heard the term and four reported they had no idea what the term was. These responses suggest no exposure to corpus linguistics. Sixteen participants reported that they had at least a rough idea of what a corpus is. This suggests some exposure to corpora, but not in-depth knowledge. Six participants reported they were "fairly familiar" with corpora and two reported they had already done some work with corpora. In summary, 24 participants had at least a rough idea of what a corpus is and 8 had no idea what a corpus is. Figure 8 displays the results of this grouping. Seventy-five percent of participants had at least a rough idea of what a linguistic corpus is.
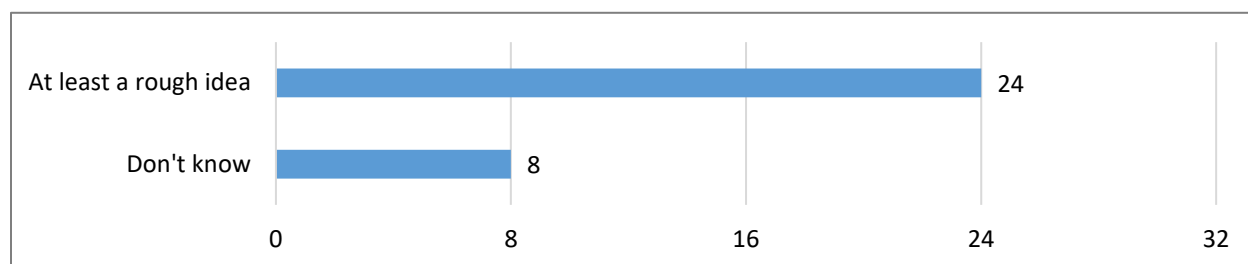
Figure 8. Grouped Multiple-Choice Responses

The short answer follow-up to question 6 was question 7. It asked participants, "Define in your own words the term corpus". As was stated in the chapter on analysis, a complete definition required the 3 parts from Helt and Reppen's (2008) definition of a corpus:

1. It is principled or representative

2. It is a collection of language

3. It can be analyzed by computer programs.

First, all participant responses were read for evidence of understanding that corpora are principled. Then, all participant responses were read for evidence of understanding that corpora are collections of language. Finally, all participant responses were read for evidence of understanding that corpora can be analyzed by a computer. Twenty-nine individuals provided responses to question 7. Three participants left this question blank. Because a response might include multiple parts of a definition, the total for Figure 9 is not twenty-nine. Figure 9 shows the total number of responses that had each individual part of the definition.
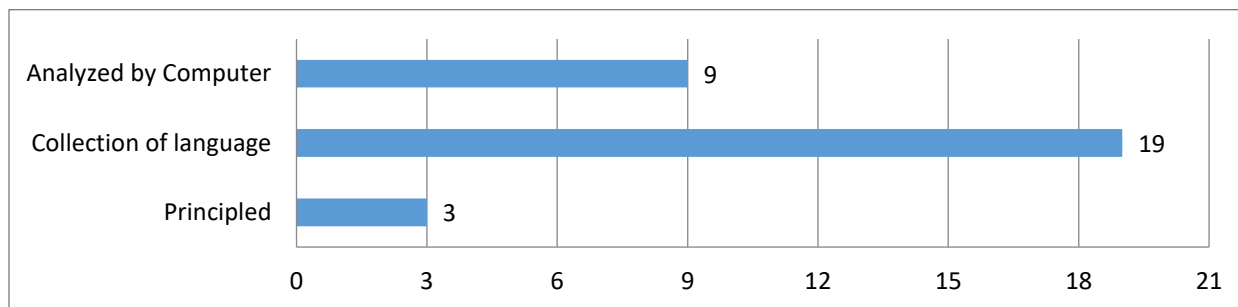
Figure 9. Corpus Definition: Individual Parts in Response

There were 19 definitions that included language that demonstrated understanding that a corpus was a collection of language. A total of 9 included language that demonstrated understanding that a corpus could be analyzed by a computer in some way. Finally, 3 stated that the language was specifically selected in some way.

Second, each response was individually analyzed to determine if all parts of the definition were present. Figure 10 shows the number of responses that included different combinations of parts of the definition.
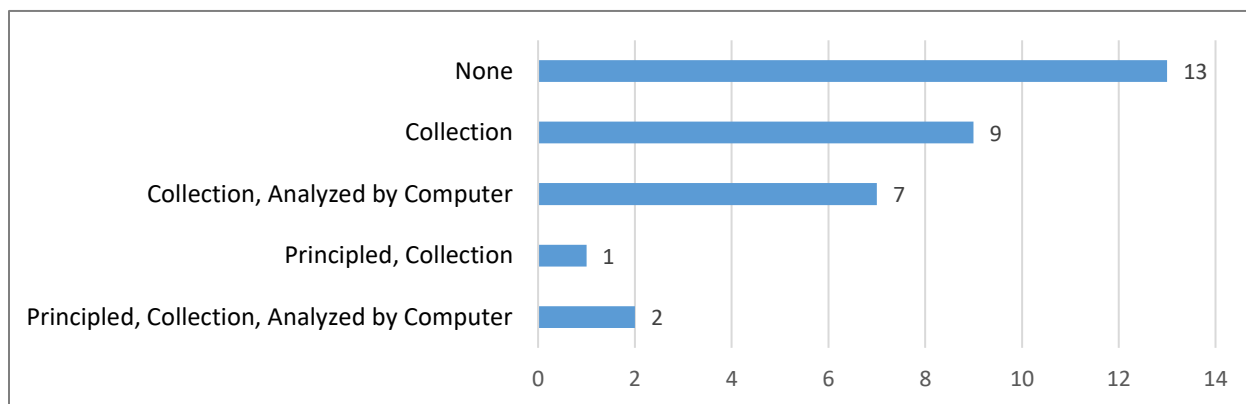
Figure 10. Corpus Definition: All Definition Parts Found in Response

The labels list the combinations of definition parts included in a response. For example, only 2 responses included that linguistic corpora are principled, a collection, and can be analyzed by a computer. One response included that corpora are principled and a collection of language but lacked that they are analyzed by a computer. Seven responses included that corpora are a collection of language analyzed by a computer but lacked that corpora are also principled. Nine responses included language that indicated a corpus is a collection of language but lacked the other two parts. Lastly, thirteen responses did not include any of the three parts of Helt and Reppen's definition.

Concordancer results. Questions 8 and 9 from the survey elicited participant knowledge about concordancers. Question 8 was multiple-choice and asked participants to indicate their familiarity of the term and then in question 9, participants were asked to define it in their own words.

Question 8 asked, "Have you heard of a concordancer and do you know what it does?" Figure 11 gives the total responses for each of the possible responses in the survey.
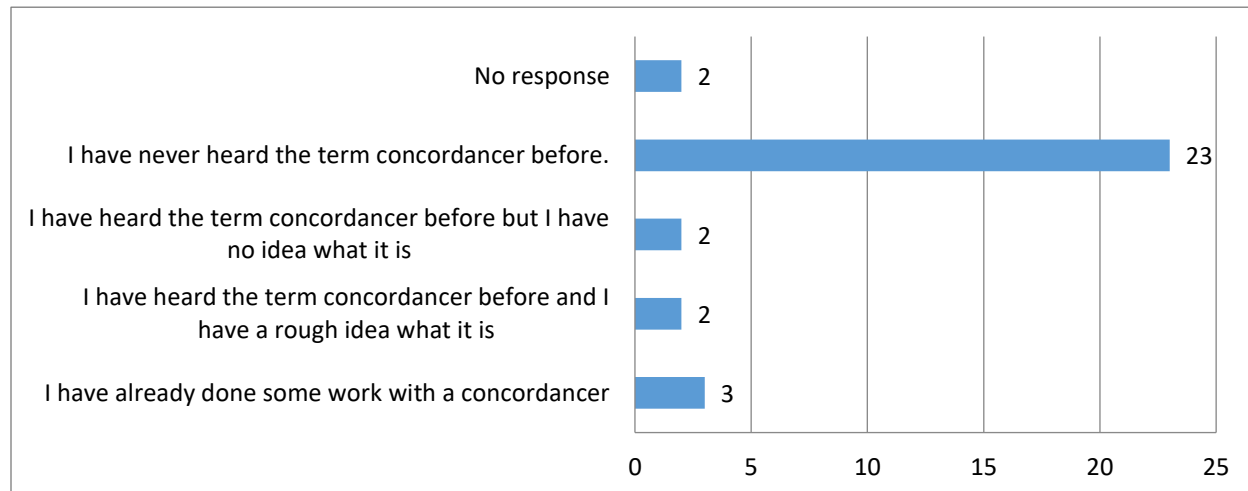
Figure 11. Multiple-Choice: Have you heard of a concordance?

Twenty-three people reported "I have never heard the term concordancer before". Two reported "I have heard the term concordancer before but I have no idea what it is". Two reported, "I have heard the term concordancer before and I have a rough idea what it is". Three reported "I have already done some work with a concordancer". Seventy-eight percent of participants had not heard the term or did not respond.

Question 9, the follow up question, elicited a short definition. In response to, "Define in your own words the term concordancer. Can you give any examples?" participants needed to include three parts: a concordancer is a computer tool, used to search a corpus, and that it can return a search term or sequence in context. 3 people were able to give complete definitions of what a concordancer is. There were no partial definitions. Only one response gave the example of AntConc. All 29 other participants didn't respond, stated that they didn't know, or made a guess of what the word means based on their understanding of the word concordancer. For example, "No idea", "I don't know", and "Someone who agrees?" Figure 12 shows "Yes" for participants that could define a concordancer, "No" for participants that could not, and "No
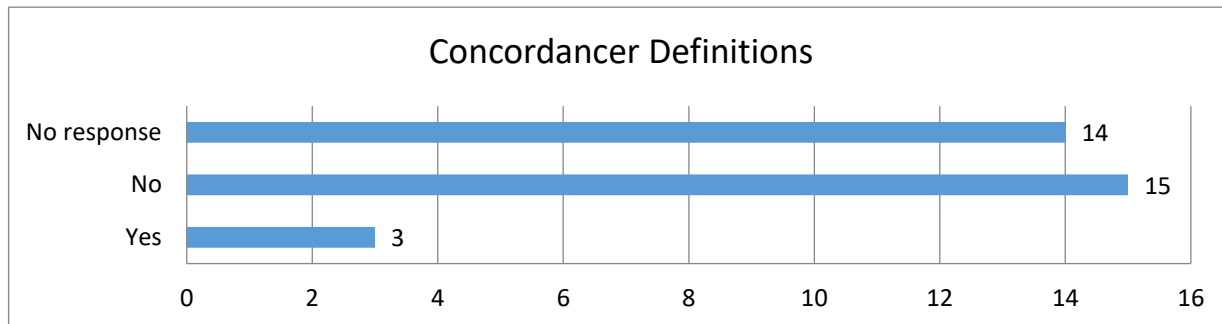
Figure 12. Concordancer Definitions

response" for participants who gave no response. Ninety percent of participants did not respond or could not define a concordancer.

***Summary*.** The results of the Corpora Experience section of the survey show that most participants have a general understanding of what a corpus is; few gave detailed descriptions however. Seventy-five percent or participants reported at least a rough idea of what a corpus is, and fifty-nine percent gave at least one part of the definition of a corpus. On the other hand, only two participants reported that they have already done some work with corpora, two gave complete definitions of a corpus, and three gave complete definitions of a concordancer.

Corpora resources and frequency of use results. Survey question 10 then asked participants about the corpus resources they used, and the frequency that they used for each resource. As described in Chapter 4: Analysis, results in this section are reported in four ways:

1. Decontextualized results.

2. Results in context of question 6

3. Results in context of question 7

4. Results in context of questions 6 and 7.

Three participants left this section of the survey partially or totally blank and did not report frequency of use for corpus resources. These responses were resolved as follows: one respondent answered question 6 of the survey by stating that they "had never heard the term corpus before" and did not report frequency of use for any of the resources. Their frequency was recorded as "never" for all resources. In addition, two others did not report frequency of use. To question 6, both participants indicated that they had, "never done any practical work with corpora". Their responses for resources left blank were recorded as "never" as well.

One participant left the Brown University Corpus check box blank for question 10, and this was recorded as no response. The total responses for this resource were 31 and not 32.

As previously stated in at the beginning of this chapter, the corpus resources offered as options were:

1. Compleat Lexical Tutor (http://www.lextutor.ca/)

2. Brigham Young Corpus of Contemporary American English (COCA)

3. MICASE: Michigan Corpus of Academic Spoken English

4. MICUSP: Michigan Corpus of Upper-Level Student Papers

5. British National Corpus (BNC)

6. TIME Magazine Corpus

7. Corpus of Historical American English

8. Wikipedia Corpus

9. Brown University Corpus of English

10. The International Corpus of English

The options for frequency of use were: Every day, Every week, Every month, Every 2-3 months, Every 4-6 months, Once or twice a year, and I used it in the past, but don't now.

1. Decontextualized results. Responses will first be reported as "Never used" for responses of "Never", and "Have used" for all other responses (Every day, Every week, Every month, Every 2-3 months, Every 4-6 months, Once or twice a year, I used it in the past, but don't now). This was done to clarify which resources had been used and which resources had not. The frequency by week, month, and year is given for the two resources that were reported as most
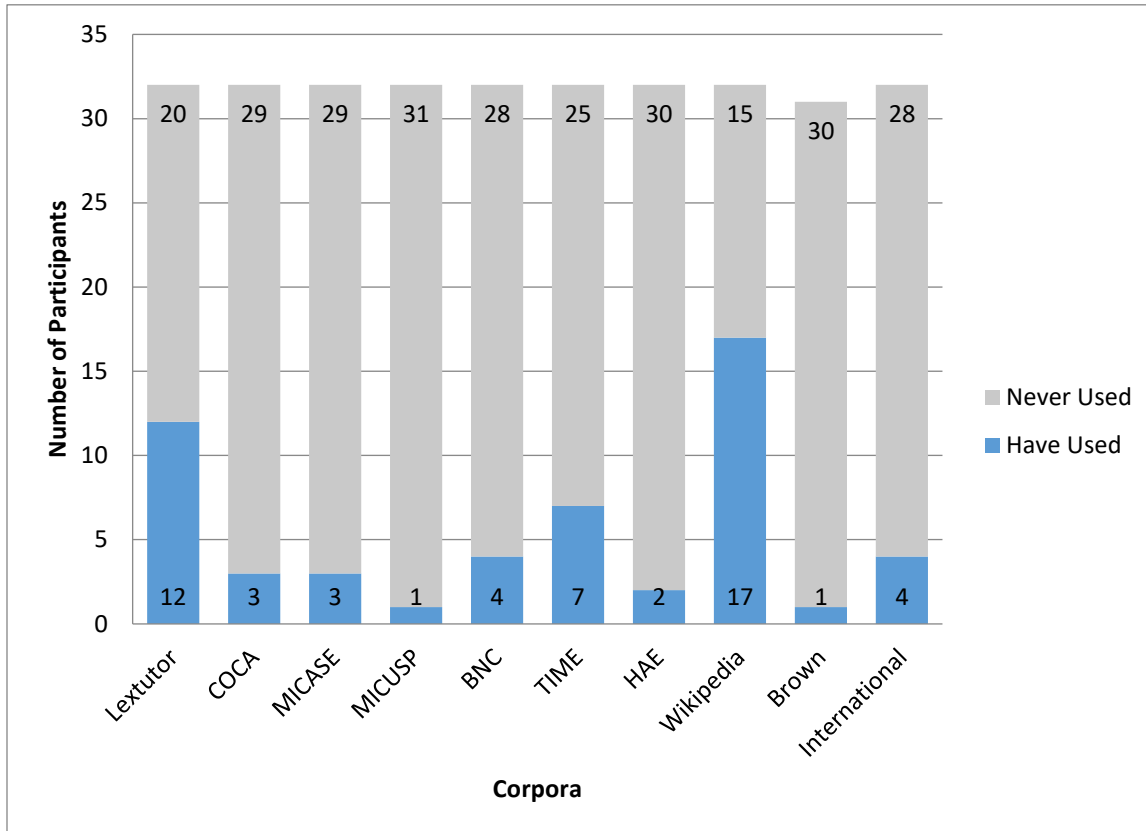
Figure 13. Corpora Resource Use

used. The reported frequency of use for all resources by week, month, and year is reported at the end of this section. Figure 13 shows the total number of participants who had used each resource.

The results of the Frequency of Use section of the survey showed that the three most used resources were the Wikipedia Corpus, Compleat Lexical Tutor, and the TIME Magazine Corpus. Fifty-three percent of participants reported use of the Wikipedia Corpus and 37 percent of participants reported using the Compleat Lexical Tutor. However, the mean reported use of all resources was 17 percent. In addition, excluding the two most used resources, the Compleat Lexical Tutor and the Wikipedia Corpus, a mean of 10 percent of all participants reported having used a resource. Figure 13shows that few of these resources had been used by participants of this survey.
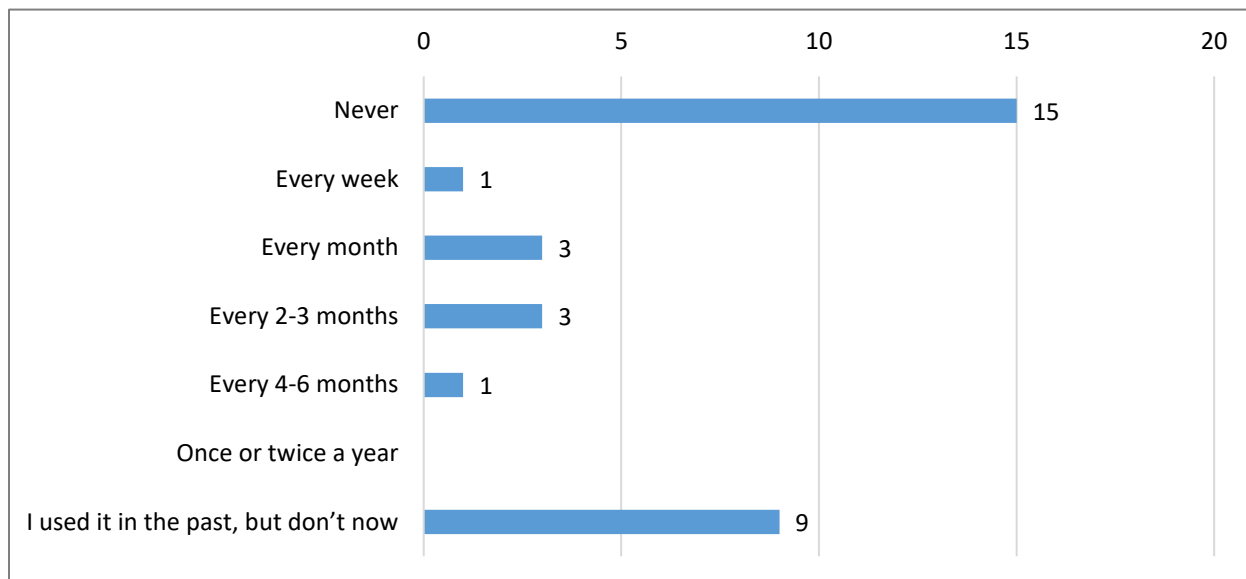
Figure 14. Wikipedia Corpus Use

Out of the three, the most used resource was the Wikipedia Corpus. Figure 14 shows the frequency of use of the Wikipedia Corpus by week, month, and year.

Seventeen out of thirty-two people reported using this resource. The Wikipedia Corpus also recorded the greatest number of people, nine, who reported using it in the past, but not anymore. Eight people also reported some level of ongoing use of this resource either weekly or over months.

The second most commonly used resource was Compleat Lexical Tutor. Figure 15 shows the frequency of use of Compleat Lexical Tutor.
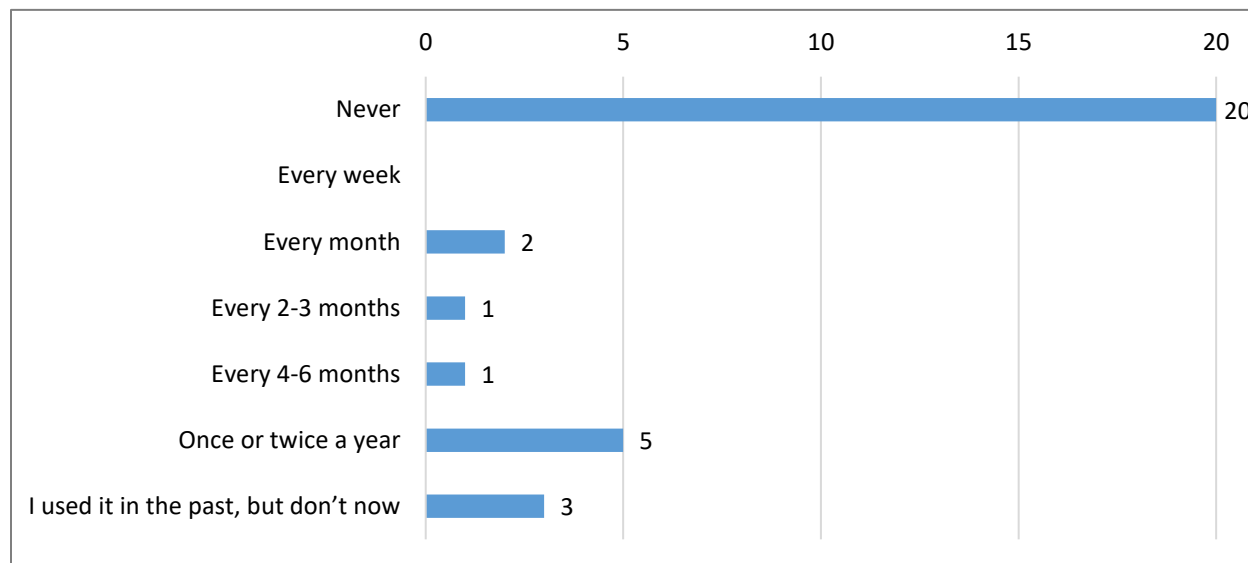
Figure 15. Complete Lexical Tutor Use

Twelve participants had used this resource. Nine reported that they continue to use it at least periodically each year.

The TIME Magazine Corpus also stands out with seven people reporting having used it and six reporting continued use at least once or twice a year. The mean number or participants who reported using each resource for the eight resources excluding the Wikipedia Corpus and Compleat Lexical Tutor was three people. Table 1 shows the frequency of use reported for each resource.

Table 1

Frequency of use for corpus resources

	Never	Every week	Every month	Every 2-3 months	Every 4-6 months	Once or twice a year	I used it in the past, but don't now	No Response	**Total**
Compleat Lexical Tutor	20		2	1	1	5	3		**32**
COCA	29					2	1		**32**
MICASE	29			1		2			**32**
MICUSP	31					1			**32**
BNC	28			1		1	2		**32**
TIME	25		2	1	1	2	1		**32**
COHA	30		1			1			**32**
Wikipedia Corpus	15	1	3	3	1		9		**32**
Brown University Corpus	30						1	1	**32**
International Corpus	28		1			2	1		**32**
Total	**265**	**1**	**9**	**7**	**3**	**16**	**18**	**1**	

Table one shows that the resources that were used were not used frequently. One participant reported using the Wikipedia Corpus every week, whereas 17 participants in total reported using resources in the past but not any longer and 16 in total reported using resources once or twice a year.

   2. Results in context of question 6. This section reports results of question 10 of the survey in context of participant responses to question 6 of the survey, "Have you ever heard the term corpus and do you know what it is? Choose from the answers below". This analysis was

performed by filtering by each possible response to question 6 and counting the number of

participants who reported using each resource. The response options for question 6 were:

a. I have never heard the term corpus before.

b. I have heard the term corpus before but I have no idea what it is

c. I have heard the term corpus before and I have a rough idea what it is

d. I am fairly familiar with corpus linguistics but I have never done any practical work

with corpora

e. I have already done some work with corpora

Results for this analysis are reported as total participants for each resource.

The response that indicates the greatest familiarity with linguistic corpora was, "I have

already done some work with corpora". Figure 16 shows the resources these participants reported
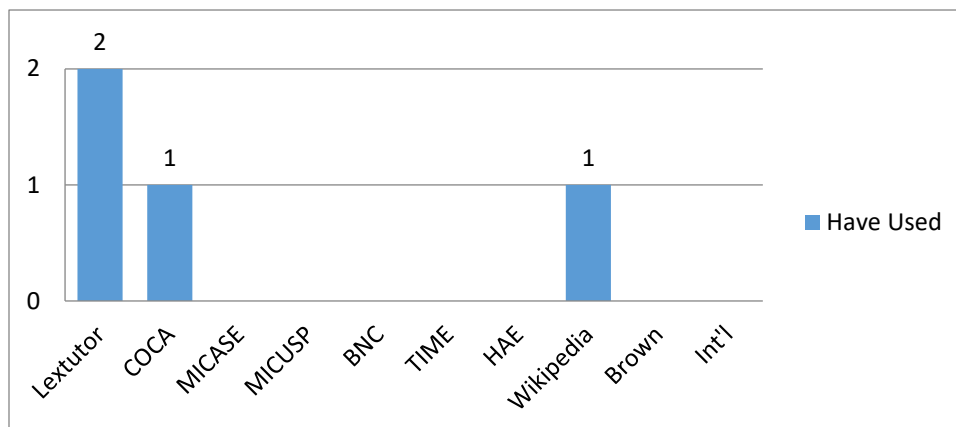
using.

Figure 16. "Already done some work"

Two participants reported already having done some work. These individuals indicated that the resources they were using were Compleat Lexical Tutor, COCA, and the Wikipedia Corpus. Both respondents reported having used Compleat Lexical Tutor.

Six participants reported, "I am fairly familiar with corpus linguistics but I have never done any practical work with corpora". These participants reported having used more resources than the participants who gave the above response. Figure 17 shows the resources these participants reported using. The greatest number of people reported using Compleat Lexical Tutor.
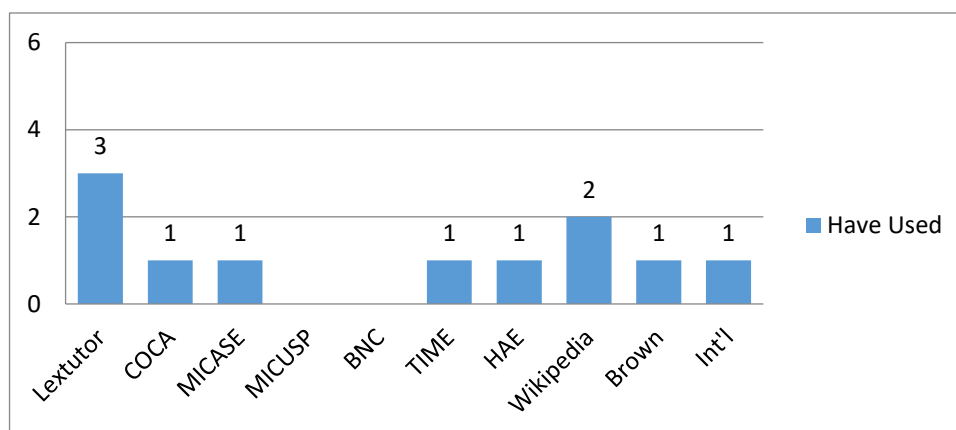
Figure 17. "Familiar, but never worked with"

Sixteen participants reported, "I have heard the term corpus before and I have a rough idea what it is". Figure 18 shows the resources these participants reported using.

Figure 18. "Hear the term…rough idea"

For this response, the greatest number of people reported being familiar with the Wikipedia Corpus, followed by Compleat Lexical Tutor.

Four participants indicated, "I have heard the term corpus before but I have no idea what it is". Figure 19 shows the resources these participants reported using.

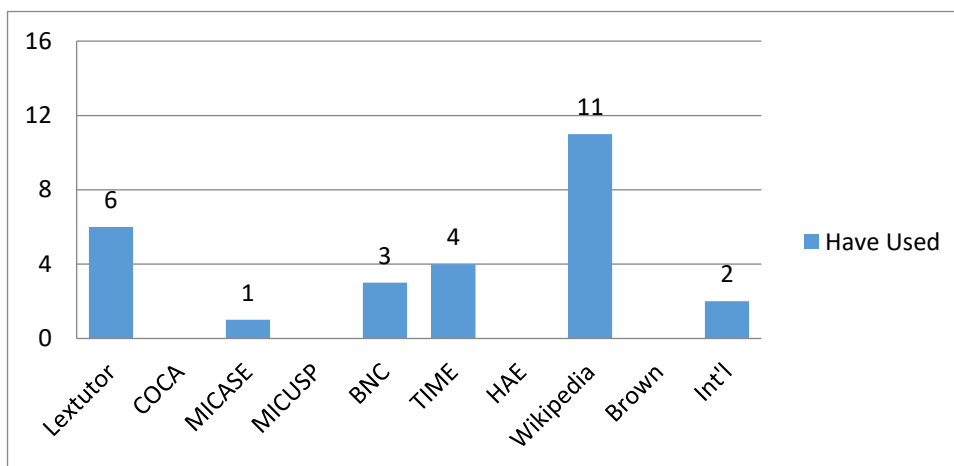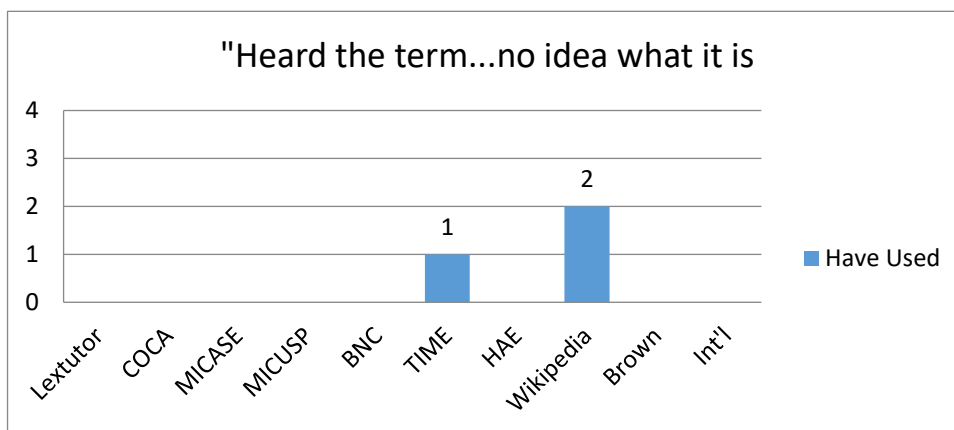*Figure 19.* "Heard the term…no idea what it is"

Although these participants indicated they have no idea what a linguistic corpus is, two individuals reported that they had used the Wikipedia Corpus, and one the TIME Magazine Corpus. This result suggests that the frequency of use for the TIME Corpus and the Wikipedia Corpus is overreported. This will be further analyzed in the discussion chapter.

Lastly, four respondents indicated, "I have never heard the term corpus before". Figure 20 shows the resources these individuals reported using.
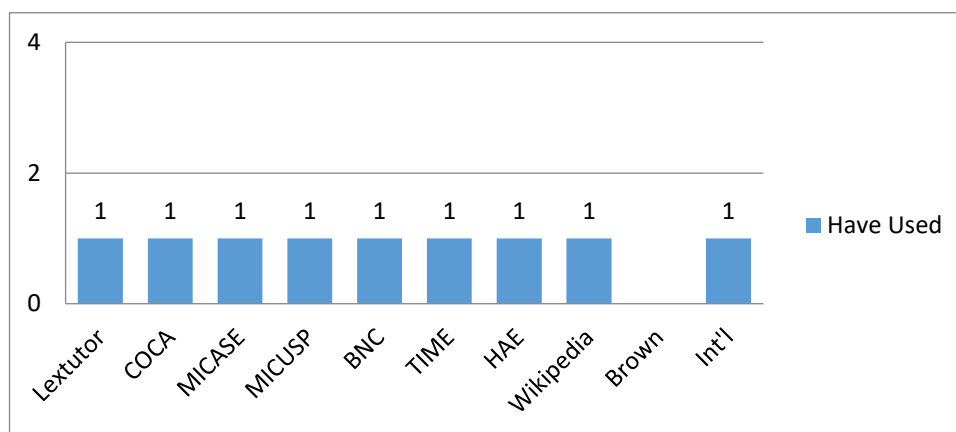
Figure 20. "Never heard the term before"

Despite this response, one individual reported having used all but the Brown corpus. This individual reported using all resources, "Once or twice a year" except the TIME Magazine Corpus and Wikipedia Corpus. These were reported as being used, "Every 4-6 months" and "Every month" respectively. This individual's response is also interesting considering their response to, "Define in your own words the term corpus". The participant wrote, "Collection of written text on a particular subject". This indicated that this person knew a linguistic corpus is a collection of texts and suggested that they know it is principled as well. It is unlikely that a

participant uses corpus resources without having heard the term. Considering the definition this participant gave for a linguistic corpus and their reported use of resources, the most likely reason for this result is that the participant accidentally hit the wrong radio button when answering the multiple-choice question, "Have you ever heard the term corpus and do you know what it is? Choose from the answers below".

3. Results in context of question 7. This section reports the Frequency of Use results in context of participants' definitions of a linguistic corpus (survey question 7). Responses were grouped in different ways to compare results. The completeness of the definition given determined the group. These groups are: no parts of the definition, one part of the definition, and two or more parts of the definition.

The frequency of use that participants reported was coded to facilitate analysis. Each response was given a numerical value. The values are as follows: 0 Never, 1 I used it in the past, but don't now, 2 Once or twice a year, 3 Every 4-6 months, 4 Every 2-3 months, 5 Every month, 6 Every week.

The three resources that will be analyzed in context of participant definitions of a corpus are the Wikipedia Corpus, Compleat Lexical Tutor, and the TIME Magazine Corpus. This is due to their high reported use.

Wikipedia Corpus. Many participants who could not define a corpus or gave imprecise responses reported use of the Wikipedia Corpus. Participants who gave imprecise definitions also reported more frequent use of this resource. This is further analyzed in Discussion. Figure 21 shows a comparison of the number of responses for those who gave zero parts of the

definition, one part, and two or three parts. The frequency that each response reported is also included in the bar graph.
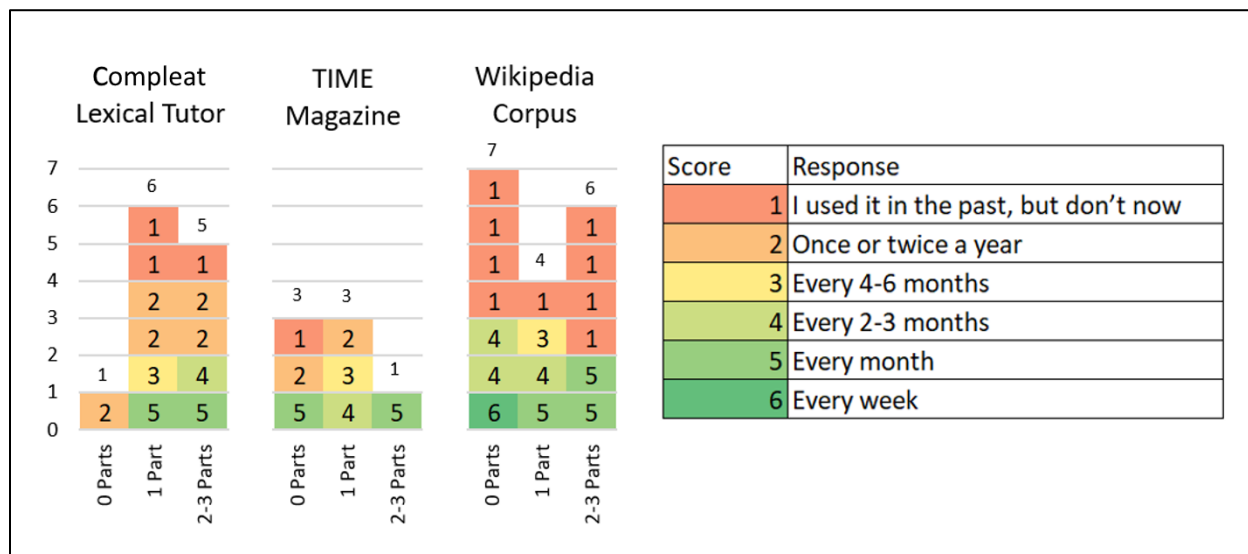
Figure 21. Comparison of Compleat Lexical Tutor, Time, and Wikipedia

As seen in Figure 21 on the right, the results for the Wikipedia Corpus are compared. The column totals show 7, 4, and 6 people reported use of the resource; seventeen people in total. Seven of these individuals gave none of the parts of the definition of a corpus. Four individuals gave one part of the definition, and six gave two or three parts. The greatest number of participants who reported use of the Wikipedia Corpus were those who gave 0 parts of the definition. This group also reported frequent use. Four people who gave no parts and four people who gave 2-3 parts responded, "I used it in the past, but don't now". However, of the participants who gave zero parts, three people reported using this resource monthly or weekly (scores 4, 4, 6). The most frequent use (score 6) was given by an individual who did not define a corpus. Only two participants who gave one or more parts of the definition reported frequent use of this resource (scores 5, 5).

Compleat Lexical Tutor. The second most used resource was the Compleat Lexical Tutor website. The cluster on the left of Figure 21 compares this group. The column totals in Figure 21 show 1, 6, and 5 individuals reported using this resource; a total of 12. The difference in use for this resource was noticeable between participants who gave zero parts of the definition, one part, and two parts. Only one participant who did not define a corpus reported having used this resource and this individual reported having used it, "Once or twice a year", whereas six and five individuals with more precise definitions reported using this resource. One individual reported using the Compleat Lexical Tutor every 2-3 months and one every month (scores 4, 5).

TIME Magazine Corpus. Time magazine was the resource with the third highest reported use. The cluster in the middle of Figure 21 compares these responses. The column totals in Figure 21 show 3, 3, and 1 individual reported using this resource; a total of 7. However, like the Wikipedia Corpus, people who did not demonstrate an understanding of what a corpus is reported using this resource. Three participants who provided none of the definition parts in their response reported using this resource. In contrast, of the participants with the detailed definitions, only one reported use of this resource. This suggests that use of this resource may also have been overreported.

4. Results in context of questions 6 and 7. Finally, it is important to consider resource use in the context of both questions 6 and 7, "Have you ever heard the term corpus and do you know what it is?" and the open-ended question, "Define in your own words the term corpus" to estimate accurate resource use. Figure 22 shows resource use only for people who reported at least a rough idea of what a linguistic corpus is and were able give at least one piece of the definition.

Frequency of Use: Corpora Resources

Score	Response
1	I used it in the past, but don't now
2	Once or twice a year
3	Every 4-6 months
4	Every 2-3 months
5	Every month
6	Every week
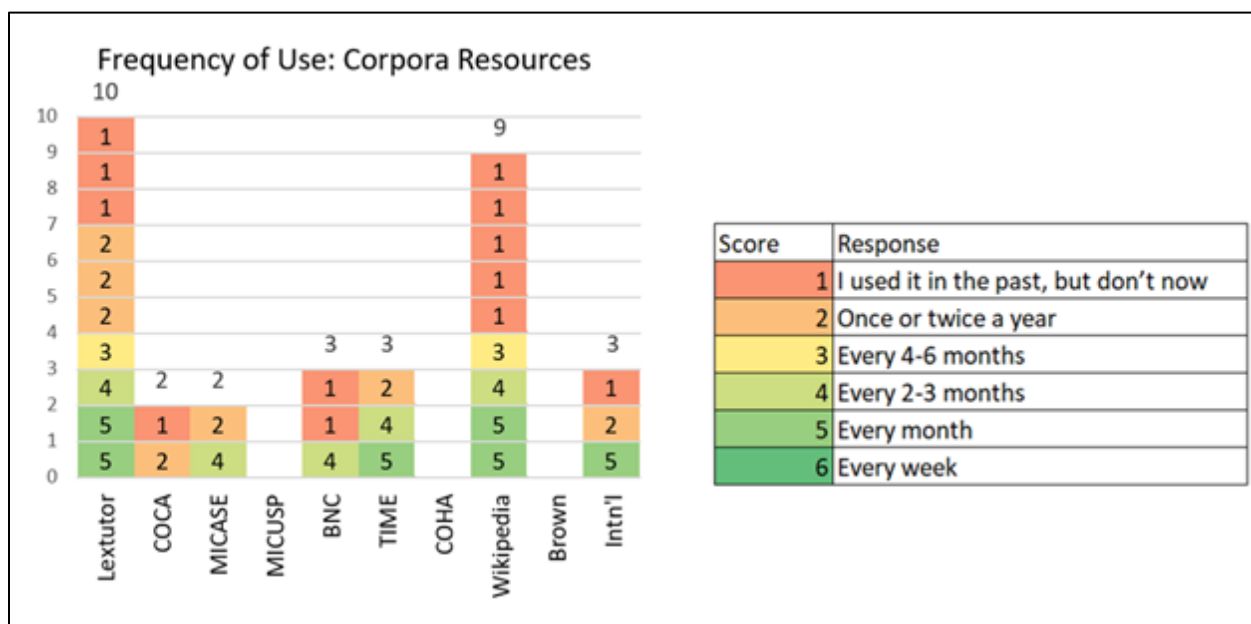
Figure 22. Frequency of Use in Context of questions 6 and 7

Seven resources in total were reported to have been used. Compleat Lexical Tutor is the resource with the highest reported use; ten participants reported using this resource. The Wikipedia Corpus is the second most used. Three resources had never been used. These were MICUSP, COHA, and the Brown University Corpus.

Figure 22 also shows the frequency participants reported for each of these resources. For resources used at least once in the last year (scores 2-6), Compleat Lexical Tutor shows the

greatest use with seven participants reporting use. Four participants reported using the Wikipedia Corpus at least once in the last year.

Summary. Apart from Compleat Lexical Tutor and the Wikipedia corpus, zero to three participants reported use of the other resources. This finding is consistent with the results of the Corpora Experience section of the survey which found two and three participants had already worked with corpora and gave complete definitions of corpus linguistics and a concordancer.

Interview Results

Questions 12 and 13 on the survey were designed to recruit participants for a follow up interview to learn in greater detail about experience with linguistic corpora. To the question, "Would you be willing to participate in a follow up interview?" Ten individuals indicated, "Yes". To the question, "Are you able to meet for an interview?" Seven Individuals indicated, "Yes". However, only one person left an email to be contacted at.

Throughout the interview the subject discussed many corpus resources including: Compleat Lexical Tutor, COCA, AntConc, and MICASE. They stated that they had used corpora tools for tagging texts for their thesis research as well as for teaching.

In response to questions 1 and 2 of the interview, how did you learn about corpora and how did you learn to use corpora software, the subject stated that they initially learned about linguistic corpora during university coursework. One of the classes that introduced corpora focused on vocabulary instruction, the other class title was not mentioned. Beyond these courses, they learned of other resources from classmates and sought out other resources independently.

The instruction in the vocabulary class focused on Compleat Lexical Tutor; learning to use other resources was done independently.

Questions 3, 4, and 5 asked about using corpora as a teacher. Question 3 asked, "Do you ask students to consult a corpus as part of homework?", question 4 asked, "Do you use corpora to correct student errors?", and question 5 asked, "Do you see corpora as primarily a tool for teachers, a tool for students, or a tool that both can take advantage of equally?". As a teaching tool, the interview subject used Compleat Lexical Tutor's Vocab Profile tool to check texts for K1, K2, and Academic Word List words to ensure that texts were comprehensible. They also attempted to use COCA as a tool for students to use at home to correct errors. However, they found it time consuming to plan, the software difficult to use, and it confused students. Because of this, the subject concluded that corpora tools were mostly for teachers, rather than students. Regarding MICASE, the subject found the difficulty level of the texts too high for their students.

Questions 6 and 7 asked, "What kinds of searches do you most often perform when using a corpus?", and "How often and in what ways do you use corpora". The subject reported that the most frequent kinds of searches they performed were for frequency, collocations, and key word in context (KWIC).

Interview questions 8, 9, and 10 asked, "What do you find difficult about using corpora?", "How comfortable do you feel using corpora?", and "How comfortable do you feel using corpus data to create materials for class?" The interview subject questioned the quality and reliability of corpora. They noted that texts do not have accurate grammar, there are misspellings and typos, and that articles are taken from arbitrary sources. Also, they questioned if texts had

been tagged correctly in the first place. Due to these criticisms, the subject concluded that corpora were better for vocabulary and not for grammar.

Despite the difficulties the interview subject encountered, the interview subject reported that they would be willing to try to use corpora tools again in teaching.

Chapter V: Discussion

This chapter will interpret the results of the data collection as they relate to the research questions, discuss areas for future research, and discuss limitations of the study.

Research Questions

1. Do current and former MA TESL students report awareness of the field of corpus linguistics? The multiple-choice question of the survey shows that MA TESL current and former students who were surveyed think they have at least a rough idea of what a corpus is; twenty-four of thirty-two participants gave this response. Participants' definitions of corpus linguistics support this conclusion too. Nineteen responses included some part of a complete definition of a linguistic corpus. The greatest number of responses stated that linguistic corpora are collections of language. The second greatest number of responses stated that they are collections of language that can be analyzed by a computer. However, definitions missed that they are principled collections of language. There are many resources available today to use, but a user needs to be aware of what the corpus contains.

The survey results suggest that MA TESL students surveyed are getting exposure to corpora mainly through a university course which introduces the Compleat Lexical Tutor website. This result was also found in the single interview response which stated that initial introduction to linguistic corpora was through a vocabulary class and that exploration of other resources was done independently.

2. Do current and former MA TESL students report using corpus resources? The results of the survey indicate that few teachers have used any of the ten resources listed.

Although Figure 13 on page 41 shows all resource use, Figure 22 on page 52 is likely most representative of actual corpus resource use by participants. This Figure filters participants who did not indicate they know what corpus linguistic is through the multiple-choice response, and the short answer definition. While it is possible that some participants are using websites without knowing what they are, it is unlikely given the specific nature and knowledge required to use them. When responses that did not demonstrate an understanding of what corpus linguistics is were filtered out, seven out of ten resources had been used by participants. However, five of these resources had been used by no more than two or three participants out of thirty-two who responded to the survey. The remaining two resources were used by ten participants for Compleat Lexical Tutor and nine for the Wikipedia Corpus.

For these two resources, there was little sustained use over time. Seven of ten people reported having used Compleat Lexical Tutor at least once in the last year. Four out of nine participants reported using the Wikipedia Corpus at least once in the last year. Four participants for each of these two resources reported using them at least once in the last six months.

The results of this survey agree with what researchers have identified. Research has shown that language teachers are not using corpus resources (Leńko-Szymańska, 2014; Breyer, 2009), despite the availability of online resources today. One reason that more students are not using resources might be because there are no specific courses available. One course at the university where this research was conducted introduces students to Compleat Lexical Tutor but is not specifically aimed at introducing corpus linguistics or available resources.

The survey results also suggest overreported use for two resources. It is likely that participants overreported use of the Wikipedia Corpus and the TIME Magazine Corpus. The

Wikipedia Encyclopedia is a highly recognized website but is different from the BYU Wikipedia

Corpus. Participants may have thought that the survey referred to the encyclopedia site and not

the corpus site. TIME Magazine is also well recognized, although not as well recognized as the

Wikipedia Encyclopedia. Participants may have thought that TIME Magazine Corpus referred to

TIME Magazine itself.

Conversely, the reported use for the Compleat Lexical Tutor is likely accurate. When

compared with previous responses, participants who had at least some knowledge of corpus

linguistics reported using the Compleat Lexical Tutor whereas those without knowledge did not

report using this resource. The reason for this might be because participants would recognize the

names Wikipedia and TIME but would be unlikely to recognize Compleat Lexical Tutor unless

they already knew what it was. Those who were familiar with Compleat Lexical Tutor probably

had it introduced to them through their university. Survey responses suggest this is the case. One

individual who gave a partial definition of a corpus and reported use of Compleat Lexical Tutor

wrote, "I understand corpus through my first semester in EAP program". Part of the definition

given by this participant included the phrase, "relates the most common academic words". Three

other responses described a corpus as, "a vocabulary "bank" of words", "group of words ... the

academic word list", and "A body of words that could be classified in categories such as most

frequent X". Because Compleat Lexical Tutor has tools that focus on analysis of vocabulary

words, these responses suggest that this is the resource participants had in mind when taking the

survey. These answers also suggest that this resource was introduced to them through a

university course.

This does not entirely account for the results of the survey however. Even when responses that did not demonstrate understanding of corpus linguistics were removed, the number of participants who reported use of the Wikipedia Corpus (9 participants) were almost equal to those who reported use of the Compleat Lexical Tutor (10 participants). The survey and interview responses did not give any hints as to why this resource was so highly reported. One reason participants might be using the Wikipedia Corpus is because among the resources listed, it has a few unique features. The description of the Wikipedia corpus states, "Most importantly, you can create and use virtual corpora from any of the 4,400,000 articles in the corpus. For example, in less than a minute you could create a corpus with 500-1,000 pages… related to microbiology, economics, basketball, Buddhism, or thousands of other topics…You can then search within that virtual corpus, compare the frequency of a word, phrase or grammatical construction in your different virtual corpora, and also create "keyword lists" based on the texts in your virtual corpus" (Davies, 2015). This feature might be useful for teachers teaching academic English to a university student population. Because this survey sampled MA TESL students at a public university in the Midwest, and because the university offered an Academic English Program at the time of the survey, it is possible that this is the reason participants used this resource. A second reason for this result might be that the Wikipedia Corpus is introduced in coursework at the university surveyed. Alternatively, even for participants who demonstrated knowledge of corpus linguistics, the Wikipedia Corpus might be subject to overreporting due to confusion. This might be the most likely reason because of the overwhelming popularity of the Wikipedia Encyclopedia.

3. What kind of experiences do current and former MA TESL students report having with the corpus resources they have used? This research offers a very limited response to this question due to the lack of participants. The participant who was interviewed appeared to have read the same materials as this study and even tried to employ corpora in the classroom. However, the participant echoed the concerns of researchers like Leńko-Szymańska (2014) and Breyer (2009). These researchers have noted possible reasons for why teachers don't use corpora such as it being time consuming to design activities, the software can be hard to use, the results can be uncertain, and there are few classroom applications. These concerns are echoed in the interview conducted in this study.

4. In what ways do current and former MA TESL students report using corpora? Compleat Lexical Tutor is also clearly the resource that the greatest number of participants in this survey are familiar with. It is apparent in the definitions that cite the academic word list and word banks as well as in the frequency of use of corpus resources. The interview subject also mentioned Compleat Lexical Tutor as a starting place for linguistic corpora use. Given that a class at the university surveyed introduces this resource, this finding is not surprising. However, because Compleat Lexical Tutor uses the British National Corpus and the Corpus of Contemporary American English, and that they are both easily accessed, it is surprising that more students didn't report having used these sites. This suggests that students may not understand how the Compleat Lexical Tutor tools produce results, even if students use these tools.

Future Research

One area that future research should focus on is the use of Compleat Lexical Tutor. The results of this survey show that this is a resource that is being used by MA TESL students but tells little of their experiences with it. Compleat Lexical Tutor is the most accessible and practical resource in this study because it provides simple tools that answer specific questions about language. It is also a resource that many participants already reported using. Future research should focus on Compleat Lexical Tutor to learn how it is being used, what experiences users have had with the site, and what users understand about how the website provides the tools it has available. Compleat Lexical Tutor is a great introductory tool to corpora. The university also provides a course that includes some instruction on this resource which would provide a population to study.

Future research might also try to explain the results of this survey that found that the Wikipedia Corpus is the second most used resource among the ten listed. If this resource was not overreported by participants, future research might investigate how and why participants are using this resource.

In addition, this research could be repeated to recruit more interview participants. This would add more qualitative data to the findings in the interview portion of this research.

Recommendations

This research and prior research have concluded that MA TESL students are not using corpus resources and are not aware of the field of corpus linguistics. In part this is because some universities do not offer specific courses on corpus linguistics. It is recommended that short

introductions to corpus linguistics and available resources be provided. This would give participants knowledge and experience with linguistic corpora and would indicate if participants found the field worth learning about. This research might inform a university if MA TESL students believe that courses should be included in their MA program. It might also show if learning about resources increases the frequency that participants use corpora resources. Leńko-Szymańska's 2014 study offers an example of this type of research and a starting point for future research. The 2014 study was "designed to introduce students to the concept of a corpus and its analysis, and to outline various applications of corpora in language education, with special emphasis placed on the in-house preparation of courses, teaching materials and class activities." (Leńko-Szymańska's, 2014, p. 263). Leńko-Szymańska's study used many of the same resources in this research and gave wide ranging instruction for possible uses.

Limitations

The survey instruments in this study were a limitation. The corpus resources section of the survey appears to have been a source of confusion for participants. Because participants who indicated they did not know what a linguistic corpus was were required to fill out the corpus resources section, they may have seen familiar names like Wikipedia and TIME Magazine and indicated they were familiar with them, rather than the Corpora sites that use Wikipedia and TIME Magazine as resources. If they had been exited from the survey based on their responses to if they knew what a corpus and concordancer were, the survey results may have been clearer.

It is possible that even participants who were somewhat familiar with corpora were confused by the names like Wikipedia and TIME. This could be improved by providing participants screenshots of the actual corpus sites when asking about their use of the resources

and clearly stating in the question that the corpus resource is different from the encyclopedia website and magazine.

A limitation of this study is that in answering the short answer questions 7 and 9, it is not possible to know what participants don't know, only what they report. It cannot be assumed that because a partial definition was given, that a participant doesn't know what a linguistic corpus is. Interviews would have clarified what participants know and don't know about linguistic corpora, but only one participant was interviewed.

Another limitation of the survey was question thirteen, "How often do you use the resources selected in the previous question?" Because of the way that question twelve was designed in the actual survey, question thirteen did not provide useful data because it did not ask about any specific resource. When the survey was created in SurveyMonkey, questions twelve and thirteen were combined. Question thirteen was not removed however. See Appendix A for an example of how questions 12 and 13 appeared on SurveyMonkey. The screenshots show different numbering because the numbering was slightly different in the online format. To draw conclusions from this data, responses to the questions in the Corpora Experience section of the survey were compared to responses on Corpus resources and Frequency of Use. Those who did not report awareness of corpus linguistics or were unable to define what a corpus was were considered unlikely to be familiar with and use corpus resources.

This study also has limited generalizability due to the target population and small sample size. This study targeted current and former students at a public university in the Midwest. Although 150 participants were contacted to participate in this study, only 32 provided usable

responses resulting in a response rate of 21%. A typical response rate for this type of survey is approximately 30% (Dörnyei, 2003, p. 76).

The lack of participants in the interview stage of this research is a limitation of this study. Given that few responses indicated experience with linguistic corpora beyond Compleat Lexical Tutor, it is unlikely that much could have been learned about participants' experience with a variety of corpora. However, more could have been learned about experiences with Compleat Lexical Tutor.

Chapter VI: Conclusion

This research investigated what awareness thirty-two current and former MA TESL students at a public university in the Midwest reported of the field of corpus linguistics and of corpus resources. It also aimed to investigate their experiences with corpus resources. This was necessary because MA TESL graduates are expected to be experts in their field, linguistic corpora are more easily accessed and used than ever before, and because they are powerful tools for researching language use.

The results indicate that while most of the participants had a general awareness of corpus linguistics, their understanding is superficial. Few participants gave detailed responses to the questions in this survey. Participants knew what an individual would notice from visiting a corpus site, that they are collections of language that can be searched by a computer. Almost no participants defined corpora as principled collections and almost no participants knew what a concordancer is.

Data from this study show that most of the corpus resources listed in this study are not being used. Only two of ten corpus resources had significant numbers of participants that reported usage. Even for these corpus resources, continued use was low. Many participants reported that they had used one of the two resources in the past but no longer do.

Regarding participants' experiences with corpus resources, this study did not collect enough data to draw conclusions.

Despite decades of optimism over linguistic corpora use in the classroom, there still appears to be little sustained use by English teachers. Many teachers leave the Master's in

Teaching English as a Second Language program with a general understanding of what a

linguistic corpus is and brief exposure to a few tools, but few have the depth of understanding to

regularly use and interpret results of a corpus.

References

Anderson, W., & Corbett, J. (2009). Exploring English with online corpora: An introduction. Basingstoke [England: Palgrave Macmillan.

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Bernardini, S. (2004). Corpora in the classroom: An overview and some reactions on future developments. In How to Use Corpora in Language Teaching (pp. 15-36). Philadelphia: John Benjamins Publishing Company. Retrieved from http://www.ebrary.com

Breyer, Y. (2009). Learning and Teaching with Corpora: Reflections by Student Teachers. Computer Assisted Language Learning, 22(2), 153-172.

Breyer, Y. A. (2011). Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges. English Corpus Linguistics. Volume 13. Peter Lang Frankfurt.

British National Corpus. (2009, January 26). Retrieved January 26, 2019, from http://www.natcorp.ox.ac.uk/

Cobb, T. (n.d.). Compleat Lexical Tutor. Retrieved January 26, 2019, from https://www.lextutor.ca/

Conrad, S. (2000). Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century?. TESOL Quarterly, (3). 548.

Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. In How to Use Corpora in Language Teaching (pp. 15-36). Philadelphia: John Benjamins Publishing Company. Retrieved from http://www.ebrary.com

Database. (n.d.). In Merriam-Webster. Retrieved January 1, 2019, from https://www.merriam-webster.com/dictionary/database

Davies, Mark. (2004-) British National Corpus (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/.

Davies, M. (2008). Corpus of Contemporary American English. Overview. (n.d.). Retrieved January 26, 2019, from https://corpus.byu.edu/coca/

Davies, M. (2010). Corpus of Historical American English (COHA). Retrieved January 26, 2019, from https://corpus.byu.edu/coha/

Davies, M. (n.d.). English-Corpora.org *. Retrieved May 26, 2019, from https://www.english-corpora.org/

Davies, M. (2007). TIME Magazine Corpus. (n.d.). Retrieved January 26, 2019, from https://corpus.byu.edu/time/

Davies, M. (2015). Wikipedia Corpus. (n.d.). Retrieved January 26, 2019, from https://corpus.byu.edu/wiki/

Dörnyei, Z. (2003). Questionnaires in Second Language Research: Construction, Administration, and Processing. Mahwah, New Jersey: Lawrence Erlbaum Associantes.

Granath, S. (2009). Who benefits from learning how to use corpora? In Corpora and Language Teaching. International Journal Of Corpus Linguistics, 14(4), 47-66

Haywood, S. (n.d.). ACADEMIC VOCABULARY. Retrieved April 29, 2015, from http://www.nottingham.ac.uk/alzsh3/acvocab/awlhighlighter.htm

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a

    teacher-training course on the use of corpora in language education. Recall, 26(2), 260-

    278. doi:10.1017/S095834401400010X

Mccarthy, M. (2008). Accessing and interpreting corpus information in the teacher education

    context. Language Teaching, 41, 563-574. Retrieved February 23, 2015, from

    http://journals.cambridge.org/abstract_S0261444808005247

Michigan Corpus of Upper-Level Papers (MICUSP). (n.d.). Retrieved April 29, 2015, from

    http://elicorpora.info/

Michigan Corpus of Academic Spoken English (MICASE). (n.d.). Retrieved January 26, 2019,

    from https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art - and

    beyond. In Corpus technology and language pedagogy: New resources, new tools, new

    methods (pp. 5-24). Frankfurt am Main: Lang.

O'Keeffe, A., & Farr, F. (2003). Using Language Corpora in Initial Teacher Education:

    Pedagogic Issues and Practical Applications. TESOL Quarterly, (3). 389.

Reppen, R., & Helt, M. (2008). Training teachers to use corpus resources. TEIS Newsletter,

    23(2). Retrieved February 23, 2015, from http://www.tesol.org/read-and-

    publish/newsletters-other-publications/interest-section-newsletters/teis-newsletter/

Reppen, R. (2010). Using corpora in the language classroom. New York: Cambridge University

    Press.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope

    and a wish list for future developments. Zeitschrift Fur Anglistik Und Amerikanistik,

    54(2), 121-134.

Shaw, E., "Teaching Vocabulary Through Data-Driven Learning" (2011). All Theses

    and Dissertations. Paper 3024. Retrieved from http://scholarsarchive.byu.edu/etd/3024

Sinclair, J. M. (2004). How to Use Corpora in Language Teaching. Philadelphia, PA, USA: John

    Benjamins Publishing Company. Retrieved from http://www.ebrary.com

Tsui, A.B.M. (2004). What teachers have always wanted to know – and how corpora can help.

    In How to Use Corpora in Language Teaching (pp. 15-36). Philadelphia: John Benjamins

    Publishing Company. Retrieved from http://www.ebrary.com

The International Corpus of English (ICE). (2016, February 9). Retrieved January 26, 2019,

    from http://ice-corpora.net/ice/

The Michigan Corpora: MICASE MICUSP. (n.d.). Retrieved January 26, 2019, from

    https://lsa.umich.edu/eli/language-resources/micase-micusp.html

The Standard Corpus of Present-Day Edited American English (the Brown Corpus). (n.d.).

    Retrieved January 26, 2019, from

    http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html

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(2), 133-

    152. doi:10.1080/09588220902778294

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. Journal

    Of Second Language Writing, 13257-283. doi:10.1016/j.jslw.2004.06.002

Appendix A: Survey

Sample Survey

-Questions 6, 7, 8, and 9 from Leńko-Szymańska, 2014

1. Biographical Information:

1. By selecting "Yes", you indicate that you are at least 18 years of age, you have read the information provided above, and you have consented to participate. You may withdraw from the study at any time without penalty after accepting to participate by exiting the survey.
2. Are you currently an MA TESL student or have you completed you MA TESL work?
3. What do you consider to be your first language or languages?
4. How many years have you been teaching English or another language?
5. Are you currently teaching a language?

2. Corpora Experience:

6. Have you ever heard the term corpus and do you know what it is? Choose from the

   answers below.

   a. I have never heard the term corpus before.

   b. I have heard the term corpus before but I have no idea what it is

   c. I have heard the term corpus before and I have a rough idea what it is

   d. I am fairly familiar with corpus linguistics but I have never done any practical

      work with corpora

   e. I have already done some work with corpora

7. Define in your own words the term corpus. Even if you do not know or are not sure

   what it is, try to explain how you understand the term.

   _____
   _____
   _____
   _____
   _____

8. Have you heard of a concordancer and do you know what it does?

   a. I have never heard the term concordancer before.

   b. I have heard the term concordancer before but I have no idea what it is

   c. I have heard the term concordancer before and I have a rough idea what it is

   d. I am fairly familiar with concordancer but I have never done used one.

   e. I have already done some work with a concordancer

9. Define in your own words the term concordancer. Can you give any examples?

_____

_____

_____

_____

_____

3. Corpus resources:

10. Please indicate how often you use each of the following corpus resources:

1. Compleat Lexical Tutor (http://www.lextutor.ca/)	☐
2. Brigham Young Corpus of Contemporary American English (COCA)	☐
3. MICASE: Michigan Corpus of Academic Spoken English	☐
4. MICUSP: Michigan Corpus of Upper-Level Student Papers	☐
5. British National Corpus (BNC)	☐
6. TIME Magazine Corpus	☐
7. Corpus of Historical American English	☐
8. Wikipedia Corpus	☐
9. Brown University Corpus of English	☐
10. The International Corpus of English	☐
Other (please list):	
None	☐

4. Frequency of Use:

11. How often do you use the resources selected in the previous question?

	Every day	Every week	Every month	Every 2-3 months	Every 4-6 months	Once or twice a year	I used it in the past, but don't now
Answer from previous question	☐	☐	☐	☐	☐	☐	☐

12. Would you be willing to participate in a follow up interview? Y/N

13. Are you able to meet in the for an interview? Y/N

Questions 10 and 11 as they appeared on SurveyMonkey (The online number was questions 10, 11, and 12)

10. Please indicate how often you use each of the following corpus resources:

	Every day	Every week	Every month	Every 2-3 months	Every 4-6 months	Once or twice a year	I used it in the past, but don't now	Never
Complete Lexical Tutor (http://www.lextutor.ca/)	○	○	○	○	○	○	○	○
Brigham Young Corpus of Contemporary American English (COCA)	○	○	○	○	○	○	○	○
MICASE: Michigan Corpus of Academic Spoken English	○	○	○	○	○	○	○	○
MICUSP: Michigan Corpus of Upper-Level Student Papers	○	○	○	○	○	○	○	○
British National Corpus (BNC)	○	○	○	○	○	○	○	○

Other (please specify)

Prev   Next

11. Please indicate how often you use each of the following corpus resources:

	Every day	Every week	Every month	Every 2-3 months	Every 4-6 months	Once or twice a year	I used it in the past, but don't now	Never
TIME Magazine corpus	○	○	○	○	○	○	○	○
Corpus of Historical American English	○	○	○	○	○	○	○	○
Wikipedia Corpus	○	○	○	○	○	○	○	○
Brown University Corpus of English	○	○	○	○	○	○	○	○
The International Corpus of English	○	○	○	○	○	○	○	○

Other (please specify)

Prev   Next

12. How often do you use the resources selected in the previous question?

☐ Every day
☐ Every week
☐ Every month
☐ Every 2-3 months
☐ Every 4-6 months
☐ Once or twice a year
☐ I used it in the past, but don't now

Prev   Next

Appendix B: Interview Questions

Sample Interview Questions:

-Questions adapted from Shaw (2011)

Background:

1. How did you learn about corpora?

2. How did you learn to use corpora software?

Corpora and students:

3. Do you ask students to consult a corpus as part of homework?

    a. How often?

    b. How are students asked to use corpora?

4. Do you use corpora to correct student errors?

    a. What tools and websites do you use for error correction?

    b. How do you use these for error correction?

5. Do you see corpora as primarily a tool for teachers, a tool for students, or a tool that

    both can take advantage of equally?

Corpora and teachers:

6. What kinds of searches do you most often perform when using a corpus?

    a. Frequency

    b. Collocations

    c. KWIC

    d. Vocabulary profile

    e. Keyword list

    f. Other (explain):

7. How often and in what ways do you use corpora for:

    a. Vocabulary questions?

    b. Grammar questions?

    c. Register questions?

    d. Language variation questions?

8. What do you find difficult about using corpora?

9. How comfortable do you feel using corpora?

10. How comfortable do you feel using corpus data to create materials for class?

    a. What kinds of materials do you create for class using corpora?

    b. Do you use corpora-based materials produced by other teachers/books/textbooks?

        i. Where do you get them?

11. Overall, what do you think about using corpora for English instruction?