Multi-Word Verbs in Prerecorded Instructor Speech: A Corpus-Informed Study

By

Tyler Theyerl

A Thesis
Submitted in Partial Fulfilment of
the Requirements for the Degree of

Master of Arts in TESOL

Dr. Douglas Margolis	Date
Dr. Annette Klemp	Date
Dr. Jodee Schaben	Date
Director, Graduate Studies	Date

University of Wisconsin – River Falls

2018

Abstract

Advances in corpus linguistics have contributed greatly to second language (L2) research and teaching. For example, evidence of high-frequency vocabulary in authentic language has been particularly useful for identifying important, general English vocabulary and words distinctive to academia. However, most academic corpora investigations use written texts and broad-based approaches, overlooking differences between written and spoken language and localized vocabulary patterns. This study examines spoken academic language in a university context by addressing a vocabulary category often associated with conversation, multi-word verbs. This examination was completed by compiling a localized, spoken academic English corpus, the Falcon Instructor Speech Corpus (FISC), containing 52,725 words from 15 different instructors at a university in the United States. Four research questions drove the investigation: (1) Which of the multi-word verbs identified in the local corpus occur most frequently? (2) How do the most frequent multi-word verbs in the local corpus compare to phrasal verb frequency lists from large English corpora? (3) What proportion of the local corpus is comprised of multi-word verbs? (4) Does multi-word verb use differ between general academic contexts and ESL contexts in the local corpus? Relevant literature on corpora and the importance of vocabulary, listening, and multi-word verbs for university English language learners (ELLs) are surveyed. Next, transcription, corpus compiling, and data gathering methods are outlined. Results suggest 68 multi-word verbs salient for ELLs at the university and provide evidence that at least 3% of words in the corpus are part of multi-word verbs. The data also shows multi-word verbs were used twice as often in general academic contexts than in ESL contexts, and that a recent, pedagogical phrasal verb list created from large corpora analyses only covered 25% of multi-word verb occurrences in FISC. Teaching implications and areas for future research are offered.

Table of Contents

List of Tables and Figures

Tables

Figures

Note: Screenshots of the online tools in these figures were used with Sketch Engine’s consent

Writing Conventions for Grammatical and Ungrammatical Examples

Example words/phrases are written in italics: *pass out*; *faint*; *talk about*; *we called after lunch*

Grammatical examples have no marking other than italics: *she came across it; the team gave up*

Ungrammatical examples are preceded by an asterisk: **she came it across*; **up the team gave*

Examples which are questionably grammatical are preceded by a question mark: *?about politics they talked*; *?she called quickly on the student*

1. Introduction

English language learners (ELLs) studying at the university level in the United States are faced with a great many challenges, not the least of which are linguistic. One focus of language study for these students has, understandably, been academic vocabulary, words used frequently in academia but not in everyday language. Studies using large language databases, also known as language corpora, have informed the TESOL field about vocabulary patterns in general English (e.g. Biber, Johansson, Leech, Conrad, & Finegan, 1999). Work from Biber et al. (1999) has informed grammars in the field about many grammatical register differences, while analyses from West (1953) up to Brezina and Goblasova (2013) have provided the field with frequent and useful general vocabulary. In the academic register, corpus work from Coxhead (2000), and more recently Dang, Coxhead, and Webb (2017), has informed instructors and learners about vocabulary which is frequent and useful across academic disciplines.

Although such work has been immensely informative for teachers and students, there has been a lack of investigations into the use of multi-word vocabulary items, such as multi-word verbs like *go on*, *come up with*, and *focus on* in spoken academic vocabulary. Furthermore, evidence from large corpus studies has often been generalized to the English language as a whole, with some emphasis on register differences (e.g. conversation, fiction, news, academic, etc.) but very little emphasis on regionally specific language variations. Investigations are needed to clarify if language patterns from large corpora match those found in specific language communities, such as a university classroom.

Investigating multi-word verbs in a local context is important for two reasons. First, multi-word verb items are problematic for ELLs on several levels. They are often comprised of words which learners know but may not be recognized as a single semantic unit (Siyanova &

Schmitt, 2007, p. 119). Multi-word verbs also vary in their transparency. For example, the meaning of *sit down* is quite clear while *come up with* (i.e. to think of or invent) is rather opaque. These verbs are also unique to English and Germanic languages, and students from first language (L1) backgrounds other than Germanic will likely have few strategies from L1 to address these forms.

Secondly, multi-word verbs, especially phrasal verbs, are typically associated with informal speaking registers (Biber et al., 1999). This means we might expect regional variations in multi-word verb usage, and that they are often excluded from academic vocabulary study. However, as few investigations of multi-word units in spoken academic language have been conducted, the prevalence and role of multi-word verbs in academic discourse remains unclear.

This study aims to shed light on spoken multi-word verbs in a local academic language context by compiling a corpus of instructor speech, here after called the Falcon Instructor Speech Corpus (FISC) and searching the corpus for frequent multi-word verb items. After a review of relevant literature, descriptions of the corpus compiling and searching methods and tools are provided. Next, I analyze the frequent multi-word verb items found in the corpus, compare results to other multi-word verb studies, and investigate differences between general academic teaching contexts and English as a second language (ESL) teaching contexts. Finally, discussion of findings, limitations, and directions for further study are offered.

2. Literature Review

2.1 Importance of Vocabulary Acquisition for University ELLs

Developing vocabulary knowledge is an essential aspect of second language acquisition (SLA). Although this notion is widely supported in the SLA literature (Nation, 2001; Schmitt, 2000), the lexicon had been largely ignored in favor of grammatical knowledge until recent

decades. The current importance placed on word knowledge in education in the United States is illustrated by vocabulary being mentioned nearly 200 times in *The Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects* for K-12 education (Graves, August, & Mancilla-Martinez, 2013, p. 6).

For all ELLs, both the number of words in their lexicon (i.e. vocabulary breadth) and the degree to which they know these words (i.e. vocabulary depth) will greatly affect their ability to communicate in English. For ELLs at American universities, proficiency with a wide array of vocabulary is especially important not only for daily language needs in an English-speaking culture, but also for learning content in academic subjects such as biology, economics, and computer science.

When it comes to word knowledge, teachers and students sometimes engage in a form of all or nothing thinking. Either someone knows a word, or they do not. In English language programs, conventional metrics for determining if students know a word often rely mainly on measurements of their ability to recognize visual form-meaning correspondence, using word-to-definition matching or multiple-choice tests. Although an important skill for reading, word recognition alone does not give a full picture of vocabulary proficiency. What it is to know and utilize a word is a rather complex issue, and there are several aspects of knowledge and skill involved with it (Chapelle, 1994; Nation, 2001).

For example, any English user may see the word *chaos* while reading and guess from context it has something to do with disorder and confusion; however, they may not be able to pronounce it accurately or use it effectively in spoken form. Another English user might understand the meaning conveyed from the combination of phonemes /ke·as/ and use them in speech but be unable to spell *chaos* or recognize it while reading due to its unusual use of *ch-* to

represent the /k/ phoneme and individual pronunciation of its neighboring vowels. All language users, first language (L1) and additional language (L2), have these types of variations in word knowledge when it comes to vocabulary.

Nation (2001) defines nine aspects of knowing an L2 word under the broad categories of *form*, *meaning*, and *use*. Each aspect includes (a) receptive skills, associated with listening and reading, and (b) productive skills, associated with speaking and writing. The nine aspects he identifies are: (A) *Form*: (1) spoken - word sound and production, (2) written - word form appearance and spelling, (3) word parts - parts of the word and parts needed to express meaning; (B) *Meaning*: (4) forms and meaning – word meaning and L1 translation, (5) concept and referents - word concept boundaries and items the word can reference, (6) associations - word associations and synonyms; and (C) *Use*: (7) grammatical functions - grammatical patterns with the word and places the word can be used, (8) collocations - other words that can or must appear with the word, and (9) constraints on use - where, when, and how often the word is met and used (p. 27).

As can be seen from breaking down these many aspects of word knowledge, knowing the word *chaos* is much more than hearing it, speaking it, reading it, and writing it. Most native speakers of English intuitively know when a milder term like *disorder* is called for, can recognize word family members (*chaotic*, *chaotically*), and know *mass chaos* is appropriate while *big chaos* is not. Similarly, having an understanding that *chaos* may change meanings within certain contexts, such as Greek Mythology, is an important vocabulary skill, especially for ELLs at the university level who will meet words in fiction, non-fiction, scientific, and other contexts. Students may learn the words *chaos* and *theory*, for example, but fail to recognize that *Chaos Theory* represents specific mathematical concepts.

In both L1 and L2, these various proficiencies with vocabulary items are attained through a combination of unconscious acquisition during everyday language experiences and explicit and implicit knowledge transfer from other language users. For example, vocabulary is learned through negotiating meaning with more advanced language users (e.g. our parents and native speakers of an L2 we are using), direct classroom teaching, and engaging in explicit language study. L1 students, such as English users in the US, expand their vocabulary rapidly during adolescence and continue to add words throughout their lifetime. Graves et al. (2013) estimate the average L1 reading vocabulary size for a high school graduate is somewhere around 50,000 words, with fluctuations dependent on various factors such as socioeconomics, availability of books, and access to technology (p. 3).

For many university level ELLs, however, the lack of English exposure and intentional word learning during adolescence relative to English L1 peers, and university instructors, creates a need to quickly make up vocabulary ground if they are to keep pace with native English users in academic and general language use. Techniques such as extensive reading and engaging in leisure activities are widely recognized in the SLA field as being highly useful for improving overall L2 vocabulary and language proficiency over the long-term (Lee, 2007; Garnier & Schmitt, 2016). Such activities reinforce words that have been introduced and extend students' knowledge of word usage and collocation.

Intentional word learning, however, has shown to be more effective and efficient than activities like extensive reading for L2 vocabulary acquisition (Waring & Nation, 2004). Considered another way, university students simply do not have time to rely solely on acquiring the vocabulary items they need without intentional vocabulary learning methods. Moreover, because the primary focus of academic vocabulary in language testing and study has been the

written word, university ELLs need more resources and time to address spoken academic vocabulary (Dang, et al., 2017). To this end, language teachers need to devote time to explicit vocabulary teaching and learning in university English language programs; furthermore, instructors and ELLs at institutions of higher learning need ways to identify which words are worthy of precious classroom and study time. Several factors, including learner level, purpose, and availability will influence how teachers and students ultimately prioritize the vocabulary they choose to focus on. However, word frequency and difficulty should also play a central role.

Word frequency relates to how often a word is used or met in the language overall or in a specific genre or context, such as the university classroom. At a fundamental level, words which are used more frequently in a language, or a specific context, will be more useful to know in that language or context (Ellis, 2002; Nation, 2001). The English language contains a core vocabulary which is extremely useful for ELLs to learn because it is employed very frequently across registers and changes very slowly (West, 1953; Brezina & Goblasova, 2013). This is because the core English vocabulary contains very frequent, closed-class function words such as articles, prepositions, pronouns, and conjunctions that glue the language together, as well as frequent verbs, nouns, and adjectives central to human experiences such as *eat*, *sleep*, *tree*, and *milk*. Although languages change over time, these words are not likely to be leaving the language or changing significantly anytime soon.

West's 1953 General Service List of 2,000 widely used words in English covers between 75 and 90% of fiction, nonfiction, and academic texts in recent studies and is the basis for most current, general frequency lists in English (Coxhead, 2000, p. 213). Brezina & Goblasova (2013), however, introduced a New General Service List (NewGSL) in 2013 based on modern

corpus research which eliminated some archaic terms from West's (1953) GSL and includes more recent additions.

Focusing on this core vocabulary first can set the foundation for accessing the language in all its forms as learners develop their language skills. Instructors and university programs can gain insight into a learner's knowledge of frequent general and academic English vocabulary through admissions and placement testing, performance on classroom tasks and homework, and through the use of tools such as the Vocabulary Levels Test (VLT), created by Nation (1983, 1990) and updated by Schmitt, Schmitt, and Clapham (2001), and more recent assessments such as Webb and Sasao's (2013) New Vocabulary Levels Test (NVLT). Although these levels-tests are useful for determining a general sense of word recognition, it should be kept in mind that they are measuring neither the aural nor the oral skills needed by university level learners.

Word frequency alone, however, is insufficient for prioritizing words for instruction. Complexity and uniqueness are also important considerations. The definite article *the* is one of the most frequent and challenging words for many ELLs. Its misuse may signal non-nativeness of the user; nevertheless, successful communication is generally possible regardless. University ELLs may want to learn how to use *a*, *an*, and *the* accurately in English, but content words, often nouns, verbs, adjectives, and adverbs, which carry central ideas in university learning tasks might be more important for learning classroom material. Similarly, *sternutation* is an interesting and some would say difficult word but falls under the category of specialized vocabulary and is exceedingly infrequent. *Sneeze* should suffice unless the learner plans to be an ear, nose, and throat doctor.

In academia, English programs, depending on the class level, tend to focus on frequent general English as well as academic and technical vocabulary from sources such as Coxhead's

(2000) Academic Word List (AWL) and textbook materials. Although the AWL has received criticism for not addressing varied word meanings dependent on academic discipline (Hyland & Tse, 2007) and being limited in its coverage of academic texts compared to more recent sources (Gardner & Davies, 2013), tools like it are useful because they focus on vocabulary items that are typically challenging to learners due to infrequent use in general English but important because they make up roughly 10% of academic texts (Coxhead, 2000). This is significant because, according to Laufer (1989) and others (Lems, Miller, & Soro, 2010; Nation, 2001; Schmitt, 2000;), having form-meaning comprehension of 95 to 98% of words in a text has shown to be needed for gaining thorough comprehension of that text. Even falling to 90-94% comprehension of words in a text has shown to significantly decrease overall comprehension (Laufer, 1989).

Addressing frequent, challenging general and academic English vocabulary items is crucial for ELLs' personal and academic success during university study; however, much of the research surrounding academic vocabulary has centered around written texts. Although reading and writing are essential skills for university ELLs, one of the greatest challenges they face is in listening to and participating in classroom instruction.

2.2 Challenge of Academic Listening

Empirical evidence and intuition suggest that listening is the language skill most often employed by university students, both in and out of the classroom (Barker, Gladney, Edwards, Holley & Gaines, 1980). The frequent use of listening skills certainly makes them important, but what is it that makes listening a challenge? Brown (2007) has compiled a short list of what makes listening difficult for ELLs from the work of Dukel (1991), Flowerdew and Miller (2005),

Richards (1983), and Ur (1984). The list includes (a) *clustering* (chunking words into meaningful thought groups that are often not sentences); (b) *redundancy* (speakers often repeat themselves or rephrase utterances); (c) *reduced forms* (phonological, morphological, syntactic, and pragmatic variations that are not as complete as they are in writing); (d) *performance variables* (hesitations, corrections, false-starts, accents, and ungrammatical forms are common in spoken language); (e) *colloquial language* (idioms, slang, and local variations used by speakers); (f) *rate of delivery* (varies widely for different speaking purposes); (g) *stress, rhythm, and intonation* (features important for meaning comprehension); and (g) *interaction* (required in many listening events) (p. 304-307). This list demonstrates the daunting task of listening even for those students who already know all the words a speaker is using, illustrating the need for ELLs to build and maintain a solid receptive vocabulary for both listening and reading.

As with most language skills, being able to cope with the issues described by Brown (2007) comes from experience and close study. Many international students in the United States have little experience with these variations in spoken language because they come from educational systems that place heavy emphasis on English grammar knowledge and the ability to read and write for academic purposes (Ferris & Hedgcock, 2014). For many reasons, both practical and cultural, listening and speaking practice with native speakers is often secondary. Learners may possess some of what Cummins (1980) first called Cognitive Academic Language Proficiency (CALP) skills when it comes to reading and writing because of the necessity to pass standardized tests in academics; however, processing and producing academic speech at native-speaker speeds can be challenging because of the emphasis the written word has had in academic language study and lack of speaking and listening practice.

Furthermore, these students may possess weak Basic Interpersonal Communication Skills (BICS), associated with listening, speaking, and pragmatic knowledge, again because of the emphasis on CALP for testing purposes and a lack of experience with authentic communication in English (Ferris & Hedgcock, 2014). In terms of vocabulary, it is colloquial language such as phrasal verbs, slang, and idioms specific to geographical regions and contexts that can cause major confusion for ELLs because learners are often unaware of these forms before moving to the community which uses them.

Additionally, listening comprehension in the classroom involves much more than simply decoding the sounds produced by the speaker. Linguistic and non-linguistic information comes at students from multiple sources simultaneously and includes written text, ambient sound, body language, visual aids on PowerPoints, and much more (Celce-Murcia, Brinton, Goodwin & Griner, 2010, p. 366). The full scope of this linguistic and other stimuli should be considered when creating language curriculum. In this section, however, the focus is on the importance and challenges with the process of listening to spoken language in academic settings. Flowerdew (1994) suggests that listeners engage five kinds of linguistic knowledge that interact and enable each other during listening events: phonological (i.e. sounds), lexical (i.e. words), syntactic (word relationships), semantic (i.e. meaning), and pragmatic (i.e. contextual). Utilizing all five of these knowledge areas during listening gives listeners the greatest chance of achieving their goal, comprehension of another speaker's meaning.

For example, students bring to the classroom a pragmatic, or schematic, background knowledge of how a Western university lecture should be structured. If an instructor begins, "Last time…", the listener might predict that a review of the previous lesson is starting. A logical utterance to follow might be, "…we discussed," followed by the topic of the previous lesson.

However, phonological, lexical, and syntactic processes are required to construct semantic meaning from the first part of the utterance and to confirm if the second part indeed follows (Flowerdew, 1994, p. 8-9). Without the pragmatic knowledge of lecture discourse, students might lag behind as they piece phonological, lexical, and syntactic input into semantic meaning to understand that a lesson review is occurring. On the other hand, if students lack phonological or lexical knowledge to construct semantic meaning, they might be able to guess that the first part of the lesson is a review, but the meaning of what is being said is not likely to be understood.

Field (2011) describes academic listening comprehension in a similar way. He suggests three phases of listening: decoding (phonetic signals are grouped into words), parsing (syntactic pattern is imposed on the group of words that was just decoded), and meaning construction (raw meaning is enriched by background knowledge and integrated into what has already been heard by the listener) (p. 109). Skills such as decoding and parsing are often referred to as bottom-up processing skills, the ability to decipher spoken language by combining phonological data into meaningful words, sentences, and texts (Celce-Murcia et al., 2010, p. 366).

Conversely, top-down skills include the pragmatic and other background knowledge learners use to make meaning of what they hear, such as knowing the structure of a typical college lecture or guessing meanings of words from the surrounding context. Essentially, both sets of listening skills are crucial for university ELLs. Moreover, without vocabulary knowledge, a learner with excellent bottom-up decoding and top-down skills is likely unable to construct semantic meaning.

Several studies have demonstrated bottom-up decoding issues faced by ELLs during academic listening. Hansen (1994) found that notes taken by graduate-level ELLs training to be

English teachers included fewer major topics and minor points than L1 peers during lectures. Furthermore, ELLs had instances of recording wrong information from the lecture, both phonologically (mishearing a word) and at the meaning level (correct words paraphrased in a way that suggests misunderstanding of the lecturer's meaning). L1 students had no such errors in their notes (p. 137-143).

Similar to note-taking, transcription studies can offer insight into students' bottom-up decoding skills. Field (2011) and Harada (1998) have conducted transcription studies with university students in the UK and the US, respectively. Field (2011) compared L1 student transcriptions and intermediate-level L2 student transcriptions of a non-specialist topic (cinema attendance) university lecture and found the L2 students to be far less capable of completing the task accurately, with many mistakes at the phonological decoding level. Perhaps it is not very surprising that native speakers (NS) of English can transcribe English more effectively than non-native speakers (NNS); however, the participants in this study were all students at the same university studying in similar programs. These results demonstrate how ELLs may be getting far less information than peers during lectures due to ineffective listening skills. Of course, this type of study is limited by the fact that participants are also necessarily writing, not only listening.

Harada's (1998) study examined the mishearings of content words by eighteen advanced-level ELLs at the University of California Los Angeles. First, participants listened to a video lecture and produced a summary. Next, based on the summaries, a challenging 1.5-minute segment of the lecture was chosen, played three times, and transcribed by participants. The clip was played first at a normal rate, then with pauses between intonation patterns, and finally at a normal rate again so that participants could check their transcriptions.

The transcripts were analyzed and mishearings were categorized into simple (i.e. one

factor) and multiple (i.e. more than one factor) mishearings. The errors were further categorized into twelve types of errors (e.g. syllable substitution, segment deletion, missegmentation, etc.). The most common type of error was segment substitution (i.e. one or more sound is substituted for another within the word). For example, the word *march* misheard as *match*, where the phoneme [t] is substituted for [r]. Other examples from the study include hearing *reality* in place of *rally*, confusing *defending* with *depending*, and replacing *quit* with *quick* (Harada, 1998).

Overall, even with listening to the lecture a total of four times (three for transcription), more than 150 content words, an average of 8 words per student, were misheard in just a 1.5-minute stretch of lecture. Given that function words like prepositions are more frequently misheard by ESL students due to reduction in natural speech, Harada (1998) argues that the total number of misheard words would likely triple had they been counted (p. 57). In a real-world classroom situation, students typically only listen to a lecture once at native speed with a wide variety of delivery styles. Consequently, the potential to miss many more content words is possible. As can be seen by the examples given above, the mishearing of just one sound in a content word can drastically change the meaning of the word and possibly the entire stream of speech.

With the segment substitution issue described in Harada's (1998) study, it is difficult to say if subjects misheard phonemes as described in the study, misspelled words when transcribing, or simply knew the word *reality* and did not know, or had limited knowledge of, the word *rally*. In the last scenario, rather than recognizing that a new, unfamiliar word had been used, participants might have consciously or unconsciously guessed or selected a word which was more familiar to them. Although few ELLs will get to the level where they will know 100% of the vocabulary in a lecture, arming students with many words that they are likely to hear and

are likely to not know due to difficulty can help get them move closer to full understanding.

Another way to inform ourselves about the listening skills university ELLs need is to recognize discourse patterns in spoken academic language. Analyses of academic lectures have found that lecture formats can vary greatly across disciplines (Dudley-Evans, 1994). After comparing four lectures of biology and engineering, Dudley-Evans (1994) suggests that the biology lectures observed followed an "information-driven" structure of categorization and listing terms and main ideas, while the engineering lectures followed an argument or "problem-solution" structure where a problem is proposed, and several ways to address the problem are discussed (p. 150-157). These lecture frameworks require students to think in different ways and deal with different, often vocabulary-related discourse markers. Having an understanding of the framework a lecture is likely to take and the vocabulary common in those frameworks can ease the listening burden by providing students with valuable, pragmatic knowledge before they enter the classroom.

Further evidence of what is happening with listening in the university classroom comes from Ferris and Tagg (1996), who surveyed over 900 professors at four different higher-education institutions (i.e. a community college, a public teaching-oriented university, a public research-based university, and a private university). The researchers wanted to examine the classroom expectations and requirements and surveyed instructors from the disciplines that had the highest ESL student enrollment: business, engineering, math, computer science, music, natural sciences, and a miscellaneous grouping of others. Responses to the surveys showed several trends.

According to Ferris and Tagg (1996), the first trend was that lecture note-taking was viewed as important in all fields. This came as no surprise and many university preparatory ESL

classes spend time teaching note-taking skills. In terms of interaction within the classroom, business seemed to be the most interactive while science courses were the least so. Professors indicated that business students were required to participate by asking questions, engaging in small-group discussions, and collaborating with peers. On the other hand, science professors indicated students spend class time listening, taking notes, and working individually, with minimal active participation (Ferris & Tagg, 1996, p. 48). It should be recognized, however, that this study focused solely on lectures. More active collaboration and interaction may take place in the lab sections of science courses.

Next, tasks common in English for Academic Purposes (EAP) classes, such as in-class debates, student-led discussions, and out-of-class assignments that require interaction with native speakers, were reported as uncommon in all disciplines surveyed according to Ferris and Tagg (1996). The surveys also suggested that class size may correlate with interactive activities (e.g. smaller classes are more interactive than larger classes). Finally, professors reported assigning formal speaking assignments less frequently than hypothesized by the researchers, and oral presentations seem to be moving away from the individual and toward pair or group work (p. 49).

This survey model has some limitations, such as using a Likert scale to ask if professors *always*, *often*, *sometimes*, *never* include an activity: one professor's *often* might be another's *sometimes*. Nevertheless, it offers some interesting insights into what is happening in content classes. Because things like the discourse structures of lectures, professors' expectations, and class activities vary widely, it should not be assumed that students across disciplines engage in the same types of listening tasks. Authentic examples of classroom language are needed to investigate lecture structure, activity types, and commonly used vocabulary further. Tasks that

stretch ELLs' language skills, such as interacting with native-speakers on campus, are good for language overall; however, time spent on tasks they will need to do for major courses and earn a grade might be more motivating and worthwhile.

Evolution of pedagogical practices is another reason to continually investigate the tasks university ELLs are asked to do. In a survey of academic listening research from 2000 to 2010, Lynch (2011) suggests that the traditional one-way lecture format is increasingly moving from a one-way spoken text to a social gathering where all in attendance are encouraged and often expected to contribute (p. 84). This could make the classroom listening task a different kind of challenge as ELLs are expected to not only listen and take notes but also share their ideas and respond to others in class. Furthermore, the traditional importance placed on written academic vocabulary leaves students lacking terms and phrases that are more prevalent in spoken English, such as phrasal verbs (Biber et al., 1999). On the other hand, the movement to online and flipped courses because of economic, geographic, and pedagogical considerations might find students engaging in more independent work and viewing of prerecorded materials.

As has been demonstrated thus far, vocabulary knowledge and effective listening skills are essential for university ELLs. Therefore, identifying the vocabulary used in spoken academic contexts is crucial for preparing students to enter the English-medium university classroom. As previously discussed, much of the research focus for academic vocabulary has been on written registers; however, that has begun to change. Lexicogrammatical analysis of academic language conducted by Biber (2006) has demonstrated a distinctness between written and spoken academic English, an issue discussed in further detail later. Furthermore, Dang and Webb (2014) have demonstrated that Coxhead's (2000) AWL, which covers about 10% of academic texts, only covers about 4% of the words in academic speech. As is the case with reading, these

numbers are significant because research has shown that learners should know 95% of words in spoken language for a high degree of stable listening comprehension (Schmitt, Cobb, Horst, & Schmitt, 2017; van Zeeland & Schmitt, 2013).

Fortunately, analysis of spoken academic vocabulary is becoming easier with advances in technology and the creation of large databases of spoken academic language such as the British Academic Spoken English Corpus (BASE), the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL), and the Michigan Corpus of Academic Spoken English (MICASE) (Nesi & Thompson, 2006; Biber et al, 2004; Simpson, Briggs, Ovens, & Swales, 2002). Recent work by Dang et al. (2017) has provided ELL instructors and learners with the Academic Spoken Word List (ASWL) which includes 1,741 word families grouped by learner level. The ASWL also contains frequent function words and covers around 90% of large academic corpora created for the study (p. 982). Analyzing authentic language with these corpora and creating lists to prioritize which vocabulary to spend time on can be very valuable for university ELLs.

Although recognition and progress in identifying spoken academic vocabulary is positive, some categories of vocabulary have received more attention than others. The ASWL (Dang et al., 2017), for example, addresses general academic vocabulary consisting of single words. Several studies have addressed multi-word units in academic speech, including collocations and idioms (Ackermann & Chen, 2013; Liu, 2003; Martinez & Schmitt, 2012; Simpson & Mendis, 2003; Simpson-Vlach & Ellis, 2010). However, the use of multi-word verbs, one of the most challenging and frequent vocabulary items in spoken English, have not been addressed directly in the speech of university instructors, likely the speakers that university students are hearing the most in academic contexts.

2.3 Multi-Word Verbs and Idiomatic Language

Many researchers group multi-word verbs and idioms under the broader category of formulaic language or the study of phraseology (Cowie, 1998; Fernando, 1996; McPherron & Randolph, 2014; Moon, 1998). Depending on how one chooses to draw the definitional boundaries, formulaic language typically includes set phrases like (a) greetings and speech formulas (e.g. *what's up?*, *see you later,* and *excuse me*), (b) fillers (e.g. *you know*, *like*, *um*, and *uh*) (c) idioms (e.g. *kick the bucket*, *fat chance*, and *in a nutshell*), (d) proverbs (e.g. *the pen is mightier than the sword*), (e) common collocations (e.g. *black coffee* and *opera house*), (f) expletives (e.g. *oh shoot!*), and (g) multi-word verbs (e.g. *look up*, *get over*, and *put up with*).

Familiarity with these types of vocabulary items and phrases is important for ELLs because spoken language has been shown to be composed largely of formulaic language chunks. As much as 80% of words were "part of a recurrent word-combination" in one corpus analyzed by Altenberg (1990, p. 102). Not only are these items frequent, but some forms of formulaic language are among the hardest items to learn for ELLs because of their uniqueness to English and counterintuitive behavior (McPherron & Randolph, 2014).

Multi-word verbs, especially phrasal verbs like *get over* (i.e. to recover from something), are grouped together with idioms by some researchers (Fernando, 1996; Liu, 2003) and separated by others (McPherron & Randolph, 2014; Moon, 1998; Simpson & Mendis, 2003). In this section, the discussion addresses the various types of and challenges with multi-word verbs and recognizes that many phrasal verbs share several traits with idioms. Therefore, although this study examines multi-word verbs in instructor speech, additional examples which fall under different idiomatic language categories are occasionally used to illustrate the close relationship between phrasal verbs and idioms in particular.

Research on idioms, moreover, shows that they do occur in spoken academic language and can serve important discourse and pragmatic functions such as paraphrasing ideas, emphasizing points, and establishing solidarity in a group (Liu, 2003; Simpson & Mendis, 2003). This study is a complement to earlier research on idioms and aims to add to the overall picture of what university ELLs need to know by addressing the issue of multi-word verbs directly. First, the categories of multi-word verbs are investigated in terms of their syntactic, morphological, semantic, phonological, and pragmatic properties. Then, additional aspects which make multi-word verbs and other idiomatic language difficult for ELLs are explored.

2.3.1 Multi-Word Verbs

Greenbaum and Quirk (1990) divide multi-word verbs into two main categories: *phrasal verbs* (e.g. *get over*) and *prepositional verbs* (e.g. *look at*). Both types contain a main lexical verb (e.g. *get* and *look*) and a particle (e.g. *over* and *at*), which Greenbaum and Quirk (1990) describe as "a neutral designation for the overlapping categories of adverb and preposition" (p. 336). Greenbaum and Quirk (1990) and Biber et al. (1999) distinguish multi-word verbs from *free combinations*, which are verbs followed by a preposition or an adverb that do not share a semantic and syntactic relationship as close as in the case of multi-word verbs (e.g. We *called after* lunch) (p. 388-389). Celce-Murcia and Larsen-Freeman (1999), on the other hand, draw a distinction between phrasal verbs and verb plus preposition combinations but not explicitly between prepositional verbs and free combinations.

In this study, Greenbaum and Quirk's (1990) classification of multi-word verbs is generally followed because I wished to examine phrasal verbs and prepositional verbs but not free combinations. Moreover, this definition of multi-word verbs was chosen for this study because phrasal verbs and prepositional verbs can be difficult to distinguish and are both worthy

of examination by instructors and learners. Additionally, as will be seen, the issue of multiple meanings (i.e. polysemy) makes it necessary to acknowledge that a multi-word verb item can be a phrasal verb, prepositional verb, or free combination depending on the context in which it occurs.

2.3.2 Phrasal Verbs

Phrasal verbs consist of a lexical verb plus an adverbial particle; can be transitive, intransitive, or both; are often opaque in their meaning; and exhibit tendencies toward syntactic cohesion (Greenbaum & Quirk, 1990, p. 336-338). Additionally, single-word verbs can often substitute for phrasal verbs, and the primary stress of phrasal verbs is usually placed on the particle (e.g. *give UP*).

For example, *give up* (i.e. *surrender* or *quit*) can be transitive (e.g. He *gave up* smoking) or intransitive (e.g. The team *gave up*). In both the transitive and intransitive cases, the meaning of *give up* cannot be predicted from its individual parts: what exactly is *given* in the previous examples, and why is it *up* as opposed to *down*, *out*, or some other adverb? *Give* and *up* are not easily separated in the intransitive case (e.g. **gave* slowly *up*), but the object of a transitive phrasal verb can come between the lexical verb and the particle in many cases, especially if the object is a personal pronoun (e.g. He *gave* smoking *up*, He *gave* it *up*, but not *He *gave up* it). Finally, the particle can be fronted in neither the transitive (e.g. **Up* smoking he *gave)* nor the intransitive (**Up* the team *gave*) phrasal verbs (Biber et al., 1999; Celce-Murcia & Larsen-Freeman, 1999; Darwin & Gray, 1999; Greenbaum & Quirk, 1990).

Because some transitive phrasal verbs require separation (e.g. *get* the message *through*, and not **get through* the message) and some do not allow it (e.g. She *came across* it, and not *She *came* it *across*), Celce-Murcia and Larsen-Freeman (1999) suggest a three-way

classification between separable phrasal verbs, inseparable phrasal verbs, and verb plus preposition combinations (p. 430). Although this distinction based on syntactic behavior could be useful for teaching phrasal verbs to advanced learners, this study is more concerned with identifying which multi-word verbs to teach. Because both separable and inseparable phrasal verbs are often opaque in meaning and exhibit overall similar behavior, they are put in a single group here.

Other examples of phrasal verbs include *turn up* (increase), *come up* (appear), *set up* (arrange or prepare), *come back* (return), *pass out* (distribute) and *go on* (happen). As can be seen from these examples, many alternative definitions are possible: *turn up* (appear), *set up* (trick), and *pass out* (faint). Phrasal verbs are highly polysemous, an issue addressed in more detail in section 2.3.4.

Certain phrasal verbs require a specific preposition to follow the adverbial particle (e.g. *come up with*, *look down on*, and *check up on*) (Celce-Murcia & Larsen-Freeman, 1999; p. 427). Greenbaum and Quirk (1990) refer to these combinations as phrasal-prepositional verbs. Phrasal-prepositional verbs always take a prepositional object (e.g. He *came up with* a great idea) and can form passives (e.g. They were *looked down on* by neighbors.) (p. 341-342). As phrasal-prepositional verbs are a type of phrasal verb, they are included in the definition of multi-word verbs for this study.

Additional types of multi-word verb constructions exist in English. Biber et al. (1999) identify a few: verb + noun phrase + preposition (e.g. *take a look at*); verb + prepositional phrase (e.g. *take into account*); and verb + verb (e.g. *make do*) (p. 403). Although these forms were not prioritized in the search for multi-word verbs in this study, two such items were included in the final analysis because they fit with the pedagogical goals of the study. The items included were

(1) *take a look at* and (2) *take into* (account/consideration) and are discussed further in the results section.

2.3.3 Free Combination, Prepositional Verb, or Phrasal Verb?

Prepositional verbs, according to Greenbaum and Quirk (1990), consist of a lexical verb plus a preposition with which the verb shares a close semantic and syntactic relationship (p. 338). Examples include *look at* (examine), *care for* (?help), and *talk about* (discuss). These instances illustrate how prepositional verbs resemble phrasal verbs in some ways but differ in others. They might be replaced by single-word verbs like phrasal verbs; however, prepositional verbs are typically more transparent in meaning and exhibit somewhat different syntactic behavior. For example, prepositional verb objects cannot come before the particle like they can with phrasal verbs (e.g. *look at* the cat, not **look* the cat *at*).

The difference between prepositional verbs and free combinations can be seen in two ways. For prepositional verbs, (1) the prepositional object can become the subject of a passive clause in most cases (e.g. The cat was *looked at*), but not with free combinations (e.g. *He called after lunch* cannot become **Lunch was called after by him*). (2) Questions about prepositional objects are asked with *who(m)* and *what* (e.g. *What/Who* did he *look at*?) for most prepositional verbs. These questions are asked with the adverbials *when* and *where* (e.g. *When/Where* did he *call*?) with most free combinations (Greenbaum & Quirk, 1990, p. 339-340).

Linguists, grammarians, and language teachers have traditionally used several of these kinds of comparisons in grammatical behavior to decide which combinations should be counted as phrasal verbs, which should be considered prepositional verbs, and which are free combinations. Most of these comparisons or tests have already been mentioned in some way in the previous discussion here and are syntactically based. They include (a) substituting the

combination for single-word verbs, (b) fronting the particle/preposition/adverb in phrases and *wh-* questions, (c) forming passives, (d) moving direct objects, (e) determining where primary stress occurs, and more (Biber et al., 1999; Celce-Murcia & Larsen-Freeman, 1999; Darwin & Gray, 1999; Greenbaum & Quirk, 1990).

Table 1 shows nine different tests often used to differentiate phrasal verbs from prepositional verbs and free combinations. The *Rule* column in Table 1 describes what the tendency is for phrasal verbs (PhVs), meaning PhVs should mostly follow these rules while prepositional verbs and free combinations should generally not follow them. Table 1 provides examples of and exceptions to these rules, mainly using the transitive phrasal verb *call up* (e.g. She *called up* her friend), the prepositional verb *call on* (i.e. She *called on* the student), and the free combination *call after* (e.g. We *called after* lunch) (adapted from examples in Greenbaum & Quirk, 1990, p. 339-340). Blank cells in Table 1 indicate that no appropriate example of a rule or exception was able to be identified.

Using these three combinations allows easy comparison of the tests and demonstration of the rules. However, other examples are used to illustrate rules and exceptions where these three could not. The transitive phrasal verb *call up* was chosen over an intransitive example because only a transitive example can illustrate the object movement tests. Many other combinations could be used to exemplify how these tests simultaneously succeed and fail to form clear categories of multi-part verbs in different instances.

In fact, each grammar resource (Biber et al., 1999; Celce-Murcia & Larsen-Freeman, 1999; Firsten & Killian, 2002; Darwin & Gray, 1999; Greenbaum & Quirk, 1990) and empirical study (e.g. Gardner & Davies, 2007; Garnier & Schmitt, 2014; 2016; Liu, 2003; 2011) consulted for the definition of multi-word verbs acknowledges the lack of firm boundaries separating

Table 1: Phrasal Verb Tests, Rules, Examples, and Exceptions

TEST	RULE	EXAMPLES OF THE RULE			EXCEPTIONS TO THE RULE		
	for Phrasal Verbs (PhVs)	**PhVs**	**Prep Verbs**	**Free Combos**	**PhVs**	**Prep Verbs**	**Free Combos**
VERB SUBSTITUTION	Single-word verbs substitute for PhVs	*call up* (to telephone)	*call on* (?to choose)	She *called after* (???) lunch.	*check out of* (?to depart) a hotel	*talk about* (to discuss)	
PHRASE FRONTING	Particles cannot be fronted	**Up* her friend she *called*.	?*On* the student she *called*.	*After* lunch, she *called*.		?*About* politics they *talked*.	
***WH*- FRONTING**	Particles cannot be fronted before relative pronouns or *Wh*- interrogatives	**Up* which friend (whom) did she *call*?	*On* which student (whom) did she *call*?	(After what)When did she *call*?			
OBJECT MOVEMENT	Particles can occur before or after direct objects	She *called up* her friend. She *called* her friend *up*.	*She *called* the student *on*.	*She *called* lunch *after*.	*She *ran* her friend *into.*		
OBJECT PRONOUN PLACEMENT	Particles occur after direct object pronouns	She *called* him *up*.	*She *called* him *on*.	*She *called* it (lunch) *after*.	*She *came* it *across*.		
ADVERB INSERTION	Adverbs do not occur between verb and particle	*She *called* quickly *up* her friend.	She *looked* longingly *at* the pictures.	She *called* quickly *after* lunch.	She *called* him right *up*. She *blew* it all *up*.	?She *called* quickly *on* the student.	
PASSIVIZATION	Transitive PhVs take passive forms	Her friend was *called up*.		*Lunch was *called after*.	*Friends were *come across*.	The student was *called on*.	The field was *played on* yesterday.
ACTION NOMINALIZATION	Transitive PhVs take action nominal forms	The *calling up* of her friend.		?The *calling after* of lunch.	*The *coming across* of friends.	The *calling on* of the student.	The *running up* of the hill.
STRESS PLACEMENT	Particles receive primary stress in PhVs	She *called UP* her friend.	She *CALLED on* the student	She *CALLED after* lunch.	She *CALLED up*, not BALLED up her friend.	Not *called OUT*, she *called ON* the student.	She *called AFTER* lunch, not BEFORE it.

phrasal verbs, prepositional verbs, and free combinations. Rather, these forms should be thought of as a continuum ranging from transparent, loosely connected free combinations to more opaque, cohesive phrasal verb combinations with prepositional verbs somewhere in the middle.

One can immediately see in Table 1 some of the issues with these tests in categorizing different combinations of verbs, particles, prepositions, and adverbs because of numerous exceptions and technically grammatical yet quite clunky forms. First, like phrasal verbs, many prepositional verbs can be replaced by single-word verbs. However, as Celce-Murcia and Larsen-Freeman (1999) describe, the phrasal verb *check out of* (e.g. *check out of* a hotel) might be replaced with *depart* or *leave*, but this does not quite give the full picture of paying, returning keys, and leaving associated with the action (p. 434).

Some may argue that *call on* is a phrasal verb and not a prepositional verb; however, *call on* acts more as a prepositional verb according to the object/object pronoun movement, *wh-* fronting, and possibly the adverb insertion tests. It might also be argued that *choose* or *select* are incomplete substitutes for *call on* because the latter implies a student will speak in some way, either to ask or answer a question or make a comment, while the former single-word verbs do not. These examples demonstrate how multi-word verbs often serve a deeper purpose than simply being an informal form of a single-word verb. Rather, they contribute sometimes very specific and subtle shades of meaning in the language. University ELLs who miss these multi-word verbs are also missing these meanings.

Only one test in Table 1, *wh-* fronting, shows no exceptions to the rule. While there seem to be no instances where a phrasal verb particle can come before a relative pronoun or a *wh-* interrogative, it is possible to do so with these prepositional verb and free combination examples. Still, although fronting preposition/particle in questions for prepositional verbs might be

technically grammatical in some cases, *On whom did she call?*, *At what pictures did they look?,* they sound oddly formal. Fronting the preposition/particle in statements sounds even more odd and Yoda-like: ?*About politics they talked*; ?*At the pictures they looked*. It is much more likely that university ELLs will hear *Who did she call on?*, *What pictures did they look at?*, and *They talked about politics* in real language experiences. This tendency for the verb and particle to stay together is one reason why it makes sense to consider prepositional verbs a multi-verb unit similar to phrasal verbs.

Free combinations, on the other hand, seem to follow the tests and tendencies fairly well, demonstrating that the verb and preposition/adverb are separate rather than a unit. Only a few of the tests show exceptions for free combinations. Passivization is possible with some free combinations (e.g. The field was *played on*), which might be instances where those examples are closer to prepositional verbs on the continuum. Although action nominalization is possible with some free combinations (e.g. The *running up* of the hill), it is safe to say that these forms are rare in the real world. Finally, depending on the situation, the placement of stress can happen on any word in a sentence or utterance; however, it is still reasonable to expect primary stress to be on the verb for free combinations (e.g. She *CALLED after* lunch) and on the particle for phrasal verbs (She *called UP* her friend). These tests do show a fairly clear distinction between multi-word verbs (i.e. phrasal and prepositional verbs) and free combinations: therefore, free combinations of verbs plus prepositions/adverbs were not included in the definition of multi-word verbs in this study.

Two additional classes of multi-word verbs are described by Greenbaum and Quirk (1990). They identify these as Type II prepositional verbs and Type II phrasal-prepositional verbs. In both cases, these types of verbs are ditransitive, meaning they are followed by two noun

phrases: an indirect and direct object (e.g. Type II prepositional verb - She *thanked* them *for* the meal; Type II phrasal-prepositional verb – He *let* him *in on* a secret). These types of verbs were also included in the definition of multi-word verbs for this study; however, because searching for such explicit combinations in corpus data is very time consuming, the items identified in this study were not sorted into the specific categories described here. Rather, all items are presented under the broad category of multi-word verbs.

As can be seen from the discussion thus far, developing a clear, understandable definition of multi-word verbs is a challenging task. A continued analysis of these verb forms is important, and linguists should continue to debate, develop models for understanding, and address exceptions to linguistic tendencies. However, in their analysis of phrasal verbs, Gardner and Davies (2007) ask the important question, "if even the linguists and grammarians struggle with the nuances of phrasal verb definitions, of what instructional value could such distinctions be for the average second language learner?" (p. 341). Several studies in the literature (Darwin & Gray, 1999; Gardner & Davies, 2007; Garnier & Schmitt, 2014; Liu, 2011) use the term phrasal verb or even just idiom (Liu, 2003) to encompass phrasal verbs and sometimes prepositional verbs. However, this study uses the term multi-word verbs because it recognizes that the real language experiences of university ELLs include both phrasal verbs and prepositional verbs while at the same time acknowledging differences between phrasal verbs and prepositional verbs, such as the tendency for prepositional verbs to be more transparent in their meaning.

2.3.4 Challenges with Multi-Word Verbs

An initial challenge for ELLs, particularly with phrasal verbs, is that they are unique linguistic structures. Phrasal verbs are widespread in English, used in some other Germanic languages, and extremely rare in any other languages. Celce-Murcia and Larsen-Freeman (1999)

mention the possible use of phrasal verbs in some Bantu languages (p. 441). However, it is unclear which languages they might be referring to. I am familiar with the most widely spoken Bantu language, Swahili, and am unaware of any phrasal verb constructions. Therefore, it is highly likely that ELLs coming from a non-Germanic L1 background will not have seen or used multi-word verbs in their native language.

Some of the other challenges ELLs face with multi-word verbs are self-evident in the previous discussion. First, according to McPherron and Randolph (2014), an essential feature of multi-word verbs and idioms is their varying degrees of (1) *noncompositionality*. That is, often their full meaning cannot be deduced by examining the constituent parts. An earlier example, the phrasal verb *get over* means to recover from something (e.g. *get over* an illness). However, the single verb *get* is generally understood as gaining possession of something, and the adverb *over* has seventeen meanings according to Merriam-Webster, most of which indicate the physical positioning of an object. A prepositional verb like *look at* can also mean something like *consider* or *learn about*.

Similarly, the words *in*, *a*, and *nutshell* in no way represent the concepts of simplicity or conciseness individually; however, when combined into the phrase *in a nutshell*, the meaning changes to mean brief and to the point. These examples of noncompositionality show the metaphorical nature of multi-word verbs and idioms. A *nutshell* is something small, representing a reduction to the simplest form. The verb *get* is representing possession or achievement of circumstance, with *over* perhaps indicating moving past a difficult period or beyond an obstacle. Paradoxically, very frequent words like *get* and *over* are typically well-known to ELLs; however, a major challenge for ELLs is that they likely know both the word *get* and the word *over* individually but fail to recognize possible new meanings when they appear together. Similarly,

an imperative to *look at* something might mean to consider it mentally, not to physically examine it with the eyes. It is therefore important that these common multi-word verbs are identified and taught as discrete vocabulary items, rather than separate words.

The metaphorical aspect of multi-word verbs introduces a second challenging aspect for ELLs: what McPherron and Randolph (2014) call (2) *shared knowledge* or *core concept*. Shared knowledge and core concept were examined by Lakoff and Johnson (1980) and relate to the way in which a group of language users see the world, which is often expressed in their language. Lakoff and Johnson (1980) point to metaphorical language embedded in English to exemplify this concept: (a) Time is money – *save time*, *budget time*, *spend time* (b) Relationships are journeys – *we hit a bump in the road*, *smooth sailing*, *on the rocks* (c) Minds are machines – *he blew a gasket*, *see the wheels turning*, *a lightbulb went on/off in her head* (d) Happy is up, sad is down – *my spirits rose, he's feeling down* (p. 15).

Opaque multi-word verbs operate on a similar pattern of a shared understanding of what the parts represent. Celce-Murcia and Larsen-Freeman (1999) report on work from Stauffer (1996) and Pelli (1976) which indicates that native speakers often create their own phrasal verbs because they intuitively understand that many phrasal verbs rely on metaphorical use of the verb and the "literal spatial or aspectual meaning of the particle" (p. 433) . In *Run up* (e.g. to *run up* a bill), for example, the verb *run* suggests motion and change while the particle *up* indicates an increase in something. Many ELLs who have studied English for years are unaware of the metaphorical meaning senses of words and the shared knowledge that underlies them; however, identifying real examples of how speakers are using multi-word verbs can help unlock patterns that can be used in training ELLs to be effective language investigators.

A third challenging aspect of multi-word verbs and idioms is their relative (3) *fixedness*,

meaning these items are usually formulaic and unchangeable. Fernando (1996) describes idioms as "indivisible units whose components cannot be varied or varied only within definable limits" (p. 30). *Kick the bucket*, for example, loses its idiomatic sense if other words are substituted (e.g. *strike the bucket* or *kick the pail*) or is obviously non-native if used in the wrong tense (e.g. *He is kicking the bucket*). On the other hand, some idioms are partially variable as one can *get cold feet*, *have cold feet*, *roll out the red carpet*, or *lay out the red carpet* (Cooper, 1998). Similarly, native speakers sometimes play with fixed language for humor or effect and might talk about *getting out the scarlet rug* or a groom *having chilly toes*.

Multi-word verbs, as has been seen in section 2.3.3 and Table 1, are also relatively fixed syntactically. Similarly, different verbs and particles usually cannot be substituted for most multi-word verbs: *look up* (e.g. *look up* the word) cannot become *see up* or *watch up*, and *look over* and *look down* have completely different meanings. Additionally, the words which function as particles take the same word form as notoriously difficult English prepositions and adverbs.

Still, as with idioms, multi-word verbs sometimes flout this fixedness tendency. *Give up* and *give in* (i.e. surrender or quit) are virtually synonymous, as are *fill out* and *fill in* (e.g. *fill out/in* the form). When Winston Churchill's speech editor changed a line to avoid ending a sentence with a preposition, the prime minister famously violated the syntactical fixedness of the prepositional-phrasal verb *put up with* by replying, "This is the sort of bloody nonsense *up with* which I will not *put*" (Brians, 2018). ELLs, however, would likely meet resistance and confusion if they attempted to alter fixed expressions either lexically or syntactically.

Look up and other multi-word verbs like *work out* and *go on* are good examples of another major challenge for ELLs: the (4) *polysemous nature*, or *multiple meanings*, that multi-word verbs can take. Gardner and Davies (2007) conducted a search for the most frequent two-

word verb plus adverbial particle constructions in the British National Corpus (BNC). The search resulted in a list of one-hundred phrasal verbs in the BNC. The researchers then used the online resource WordNet (2010), developed and maintained by Princeton University, to search for the different meaning senses for each phrasal verb. The average number of meaning senses was 5.6 per phrasal verb (Gardner & Davies, 2007, p. 353).

According to the authors, *break up* has the most meaning senses with nineteen, *go on* (first on the frequency list) has five meaning senses on WordNet (2010), and, surprisingly, *look up* has only one: to consult or refer. This may illustrate some limitations of WordNet (2010) for categorizing multi-word verb meaning senses. In the case of *look up*, it is not difficult to think of additional meanings: (a) to contact someone (e.g. *look* me *up*), (b) to raise one's eyes (e.g. *look up* from the book), or (c) to express that a situation is improving (e.g. things are *looking up*). Conversely, WordNet (2010) sometimes provides an abundance of unnecessary meaning distinctions for a single item. For example, eight meaning senses are provided for *work out*, including two for mathematical calculations.

The challenge for ELLs is that many of these items have multiple meanings and knowing which meaning senses they should spend time learning is not always clear. Phrasal verb and other specialized dictionaries are typically good references because they are comprehensive, but these materials can confuse learners because of the large amount of information they contain, including obscure and archaic definitions of some items.

As we have seen, the issues that formulaic language like multi-word verbs can cause for English language learners are numerous. Perhaps because of these challenges, ELLs have shown a tendency to avoid multi-word verbs in favor of one-word verbs in some studies (Dagut & Laufer, 1985; Liao & Fukuya, 2004). Interestingly, Hulstijn and Marchena (1989) found that, in

their work with Dutch learners of English, the subjects avoided phrasal verbs because the verbs were perceived as idiomatic and too similar to Dutch phrasal verbs. In other words, some ELLs may avoid these verbs because they are too different from their L1, and others may avoid them because they are too similar.

The avoidance of these forms may have negative consequences for language learners. For example, Barekat and Baniasady (2014) studied the level of phrasal verb avoidance and the writing skills of a group of 86 Persian L1 English language learners. Participants completed several tasks with phrasal verbs in order to establish two groups: those who tended to avoid phrasal verbs and those who tended not to avoid them. Next, each group completed a writing task. Scores from the writing task showed better performance from the group who tended not to avoid phrasal verbs. The authors suggest that the avoidance of phrasal verbs may have negatively impacted the participants' writing performance (Barekat & Baniasady, 2014, p. 348). Because multi-word verbs are challenging and may cause negative impacts for learners if avoided, this study attempts to address them directly for a specific population.

Although the issues of multi-word verb vocabulary items and academic listening that we have seen thus far are challenging, university level ELLs and their instructors can approach the identification and prioritization of these items in a systematic way. As mentioned toward the end of section 2.2, language corpora are one way that different categories of vocabulary have been identified and prioritized in recent language learning research (Biber, Conrad & Reppen, 1998; McEnery & Hardie, 2012). However, many studies in the literature address vocabulary items in a very broad way, meaning they attempt to describe the meaning and behavior of items in the whole of the English language or a wide register. This study posits that a focused approach, using a more specific language database may aid learners with specific needs.

2.4 Corpora in Vocabulary Research and Pedagogy

A language corpus, as simply defined by the Merriam-Webster Learner's Dictionary, is "a collection of writings, conversations, speeches, etc., that people use to study and describe a language" (2017). In the modern age, these corpora, or language databases, are stored electronically and can quickly analyze millions of words of text with computer programs. Tools such as key word in context (KWIC) concordance lines, word frequency lists, and tables of grammatical behavior provide unique descriptions of language which can inform the creation of language learning materials.

Furthermore, because the language samples included in corpora are typically authentic (i.e. naturally occurring in communication and not contrived for teaching purposes), they are arguably a good representation of what learners will encounter in the world and can be an effective learning tool (Williams, 2016). Recent research comparing corpus-based teaching materials to more traditional, textbook materials in the classroom suggests that corpus materials can be enjoyable for students and more effective for teaching vocabulary usage (Ashkan & Seyyedrezaei, 2016; Hou, 2014; Paker & Özcan, 2017).

Today, there exist a great number of language corpora that differ by language (e.g. English, Arabic, Dutch, etc.), mode and genre of language (e.g. written, spoken, news, fiction, etc.), and method of collection (Biber, et al., 1998; McEnery & Hardie, 2012). McEnery and Hardie (2012) distinguish several different data collection methods that result in distinct types of corpora. Primary examples include (1) *monitor corpora* – an ever-growing database containing a wide variety of modes, genres, and time-periods such as the Corpus of Contemporary American English (COCA) (Davies, 2008-), (2) *web as corpus* – using texts on the internet as a language corpus such as the massive 19.6 billion-word English Web 2013 (EnTenTen13 Corpus)

downloaded by SpiderLing in December of 2013, (3) *sample corpora* – a specific type of language collected over a certain time frame such as the English Historical Book Collection containing English books published between 1473 and 1820 (Sketch Engine, 2018), and (4) *opportunistic corpora* – a collection of whatever language is available or possible to obtain for a specific task, a useful method for archiving and analyzing minority and endangered languages (McEnery & Hardie, 2012, p. 6-12).

The most relevant corpora in the literature which was reviewed for this study were the British National Corpus (BNC) (Davies, 2004-), the Corpus of Contemporary American English (COCA) (Davies, 2008-), and the Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al., 2002). Many additional corpora, however, have informed grammars and studies also consulted for this project, including the Longman Spoken and Written English Corpus (LSWE Corpus) (Biber et al., 1999), Pearson International Corpus of Academic English (PICAE) (Ackermann & Chen, 2013), corpora compiled to create Dang et al.'s(2017) Academic Spoken Word List (ASWL), and numerous smaller, topic specific corpora (Aluthman, 2017; Hou, 2014; Zhang, 2013).

A major advantage of using large corpora to analyze language is that they offer the ability to extract a vast amount of quantitative data not easily accessible without computer programs. For example, a query using Brigham Young University's tools for searching COCA can reveal the most common adjective-noun collocations in the corpus in a matter of seconds (Davies, 2008). Such data is extremely valuable for teaching and understanding a language. Biber and Conrad (2001) posit that although quantitative analysis using corpora might seem like "elaborate bean counting," frequency data has informed two ideas vital to language teaching today: patterns of language use across registers and the unreliability of intuition about language (p. 332).

(1) Patterns of use in language are strongly linked to register. Although this notion makes intuitive sense, providing empirical evidence has become much easier through the use of computers with language corpora. For example, Biber and Conrad (2001) report that the twelve most common verbs in the conversational register in the LSWE Corpus (i.e. *say*, *get*, *go, know, think, see, make, come, take, want, give,* and *mean*) account for 45% of all lexical verbs in that register, but the same verbs only account for 11% of the verbs in the academic register in the LSWE Corpus (p. 332-333). Therefore, what students want to do with language needs to be a primary consideration for curriculum development. Preparing for casual conversation or academic study in English are very different animals.

(2) Intuition about language use is unreliable. Using the same corpus data as the previous example, Biber and Conrad (2001) challenge the commonly held belief that progressive aspect is the unmarked, or most common, choice of English speakers in conversation. This belief is evidenced by textbooks often covering the present-progressive in the first unit. The authors' analysis found that verb phrases in conversation were more than twenty times as likely to be in simple-aspect than progressive aspect. Similarly, progressive aspect was extremely rare in academic writing (p. 333). Teachers and materials writers who use intuition to prioritize vocabulary items and structures may be making costly choices for learners. Certain multi-word verbs, for example, might be considered by teachers as frequent, important, or rare and therefore unimportant in the academic register; however, evidence of their prominence and usage should inform pedagogy, not speculation based on personal experience or conventional wisdom.

2.4.1 Multi-Word Verb Frequency and Academic Language

As discussed in section 2.1 on the importance of vocabulary in SLA, corpora have long been used to identify the frequency of vocabulary items in English in order to inform teaching

priorities. Although many lists have been created, particularly over the last two decades, this section focuses mainly on literature related to examining multi-word verb frequency, the creation of teaching lists, and the academic register because these are also the focus of this study. However, other vocabulary or phrasal categories such as academic discourse markers (e.g. *moreover*; *on the other hand*) or adjective-noun collocations (e.g. *brief overview*) would be worthy of exploration.

Biber et al. (1999) investigated multi-word verbs across registers in the LSWE Corpus, a database of forty-million spoken and written words of American and British English. Overall, they found that phrasal verbs and prepositional-phrasal verbs were about as common as other lexical verbs (i.e. one-word verbs) in informal conversational and fictional registers, making them an important part of everyday communication. On the other hand, phrasal verbs were much less common in academic prose. Prepositional verbs were also found to be as frequent as lexical verbs across most registers (i.e. conversation, news, and academic) and proved to be even more frequent in fiction (p. 411-424).

Judging from this data, it seems ELLs studying for academic purposes would get the most out of learning prepositional verbs first. However, the information on the academic register is misleading because a major piece missing from the corpus used in the study was spoken academic language. In fact, the entire 5.3 million words of American and British academic English in the corpus was taken from books and research articles (Biber et al., 1999, p. 33-34). It may well be the case that prepositional verbs should take the priority in general English language learning and for academic writing. Nevertheless, the analysis from this study fails to give teachers and learners a sense of the prominence and use of any multi-word verbs in spoken academic English.

Although several studies have built off the work with multi-word verbs conducted by Biber et al. (1999), authors have tended to focus more on phrasal verbs than prepositional and phrasal-prepositional verbs (Gardner & Davies, 2007; Liu, 2011; Garnier & Schmitt, 2014). This seems to be due to the idea, discussed in section 2.3.4, that phrasal verbs are often more opaque in meaning than prepositional verbs and therefore more challenging for learners. However, these authors have acknowledged that their definitions of phrasal verbs are not always clear cut, and that some of what this study calls prepositional verbs are included in their results because of the polysemous nature of multi-word verbs.

For example, in their Phrasal Verb Pedagogical List (PHaVE List), Garnier & Schmitt (2014) only included the meaning sense *raise one's eyes* for the multi-word verb *look up* because it accounted for 88% of its uses in their analysis (p. 657-658). This is an important idea because, as discussed in the first section, both difficulty and frequency should be considered in prioritizing vocabulary. If the opaque meaning of *look up* is very seldom used, the more common transparent meaning should be prioritized in teaching or possibly left out of the curriculum if it is easy for learners to grasp.

Gardner and Davies (2007) searched the BNC for the most frequent phrasal verb combinations by using the part of speech (POS) annotation, or tagging, present in the BNC. In the BNC, and most corpora today, each *token* (i.e. smallest unit that the corpus divides into, typically each word form and punctuation) is designated a POS, such as verb, noun, adverb, etc. by a computer program and/or human annotators. Therefore, the researchers were able to transfer the BNC text into Microsoft SQL Server and search the corpus for each instance of a lexical verb which was followed by an adverbial particle, simplified here as [verb] + [particle]. Additionally, the search was lemmatized, meaning the program searched for verb *lemmas* (i.e. a verb in all its

inflected forms – *go*, *goes*, *going*, *went*, *gone*). Therefore, a single search was able to identify all the items the researchers were looking for, such as *go over*, *went over*, *take off*, *took off*, etc.

Gardner and Davies (2007) also closely analyzed each lexical verb and adverbial particle in their search. Findings indicated that the top-twenty lexical verbs found accounted for 53.7% of all the phrasal verbs in the BNC, meaning those twenty verbs were highly productive. They also discovered that *out*, *up*, *down*, and *back* occurred very frequently as adverbial particles in phrasal verbs and posit that phrasal verbs comprise a major grammatical class. By their calculations, learners will encounter a phrasal verb in every 150 words they are exposed to, roughly two per page in a 300-word per page written text (Gardner & Davies, 2007, p. 346-349). Their findings resulted in a list of 100 phrasal verbs consisting of the 20 most productive lexical verbs in the BNC. The authors concede, however, that these results relate to the entire BNC and that the frequency of usage likely varies across registers and in different data sets (p. 347).

Liu (2011) followed Biber et al. (1999) and Gardner & Davies (2007) in investigating phrasal verbs in English by conducting a comparison between American English and British English using COCA and the BNC. The researcher first combined the phrasal verb lists created in the two previous studies, which resulted in a list of 104 items once overlap was accounted for. Next, Liu (2011) searched COCA and the BNC for those 104 items as well as several thousand others from dictionaries, totaling 8,847 phrasal verbs. This number was then narrowed to 150 by eliminating false positives and only including items deemed highly frequent, determined by the items occurring at least ten times per million words in either corpus (p. 666-667). The *words per million* concept is important in corpus linguists as it allows researchers to compare the frequency of items in corpora of different sizes.

Liu (2011) used statistical analyses to compare the frequency patterns of these 150 items

in COCA and the BNC. He claims that although a two-way chi-square analysis showed a significant difference between the items' frequencies in COCA and the BNC, the difference is accounted for by a small effect size score, Cramer's V of 0.0032, and because of the large size of the two corpora (p. 670). Put another way, frequency of phrasal verb items in the BNC and COCA are somewhat different but follow a generally similar trend. For example, many of the items have the same ranking on the list in COCA and the BNC: 1st – *go on*, 14th – *come in*, 19th – *get back*, 44th – *bring back*, and 94th – *turn down* (p. 670).

A close analysis of individual items was also conducted, which resulted in about 20 items that appeared significantly more in American English and about 10 items that appeared significantly more in British English (Liu, 2011, p. 671). On the whole, Liu (2011) claims that although the general distribution of these 150 phrasal verbs in COCA and the BNC are fairly similar, there are usage differences between the American and British varieties, notably an increased use of phrasal verbs in American English (p. 672).

One issue clouding the data from this study, which Liu (2011) does address, is that the BNC and COCA cover different time periods. The BNC contains language from the 1980s to 1993, while COCA covered 1990 until 2011 at the time of the study. Analysis showed that certain items, such as *check out*, *hang out*, *show up*, and *come up*, have increased in use across the 1990 to 2011 timeframe in COCA (Liu, 2011, p. 672). Evidence that American English uses multi-word verbs in a distinct way from other English varieties and that usage of these items changes over time suggests the need for collection of more regional specific and current corpora. In other words, university ELLs in the United States who are using textbooks or dictionaries which include other national varieties or that are outdated might be missing important vocabulary for their time and place of study.

Liu (2011) additionally examined the 150 phrasal verbs in terms of their distribution across registers in COCA. As with Biber et al. (1999), Liu (2011) found phrasal verbs to be much less frequent in the academic register. Liu (2011) did identify just one example, *carry out*, which showed a very high frequency in academic writing. However, as with Biber et al.'s (1999) analysis, COCA's academic register is entirely written language, limiting the findings to academic writing. The generalizability of these findings to academic oral language remains to be determined.

Still, the research reviewed here has tremendous value to English language learners in general. The 150 item list which resulted from the work of Biber et al. (1999), Gardner and Davies (2007), and Liu (2011) makes a great phrasal verb vocabulary list for general English learners. Garnier and Schmitt (2014), furthermore, addressed the issue of polysemy with this list by creating the previously mentioned PHaVE List, a pedagogical list with the top meaning senses of each of these 150 items.

To create this teaching tool, the researchers took two random samples of 100 concordance lines for each item in COCA and recorded occurrences of different definitions by going through them line by line. Inter-rater reliability was also included for a small sample (5 items), which showed very close judgement of meaning senses by the researchers and an outside rater. Because the list is pedagogically based, and very long lists can become cumbersome, the researchers only included meaning senses which accounted for at least 10% of all sample item occurrences.

For example, if an item was identified as having three meaning senses, where meaning-one accounted for 65% of all uses, meaning-two accounted for 30% of all uses, and meaning-three accounted for 5% of all uses, only meaning-one and -two were included on the list because

they are more likely to be encountered by learners (Garnier & Schmitt, 2014, p. 652). In addition to definitions, each meaning sense for an item has an example sentence. Items are listed in order of frequency in COCA. For example, number twenty-eight on the list is *take off* and displayed in the following way:

> **28. TAKE OFF**
> 1. Remove STH (esp. piece of clothing or jewellery from one's body) (41%)
> I **took off** my shirt and went to bed.
> 2. Leave a place, especially suddenly (28.5%)
> They jumped into the car and **took off**.
> 3. Leave the ground and rise into the air (14%)
> The plane **took off** at 7am. (Garnier & Schmitt, 2014, p. 658).

Although a useful teaching tool, the PHaVE List does not address the spoken academic register. Where the academic register has been addressed, multi-word verbs have not been central to any analysis. Biber (2006) describes in his book *University Language: A corpus-based study of spoken and written registers* how vocabulary in the T2K-SWAL Corpus, a proprietary corpus created to inform iBT TOEFL exam study material, demonstrates vast differences in classroom teaching language and academic writing. For example, the classroom teaching segment of the corpus revealed the use of only 14,500 unique words by speakers while textbooks used some 27,000 unique words (p. 36).

Furthermore, words such as *get*, *say*, *think*, *want*, *see*, and *thing* were extremely common in classroom teaching (e.g. between 2,000 and 6,000 words per million) while they were much less common in textbooks (e.g. at or below 1,000 words per million for each). Even common textbook words such as *occur*, *include*, *control*, and *analysis* did not reach the 1,000 words per million mark, suggesting the use of a wider vocabulary. These items also occurred much less frequently in classroom teaching (Biber, 2006, p. 37).

The findings also suggest additional differences between classroom teaching and written

academic vocabulary. Biber (2006) states, "the dense use of complex noun phrase constructions" (e.g. *the frequent word in the corpus which is used only in one academic discipline*) was common in academic writing while classroom teaching tended to avoid these constructions (p. 173). However, no analysis of multi-word verbs was found in this extensive examination of the academic register.

Analyses using a previously mentioned academic corpus, MICASE (Simpson, et al., 2002), could offer better evidence of the use of multi-word verbs in the spoken academic register. MICASE is an academic spoken corpus which was collected over a period of five years (1997-2001) and is comprised of lectures, small-group study sessions, dissertation defenses, office-hour discussions, and more at the University of Michigan. Simpson and Mendis (2003) analyzed the occurrence and function of idioms in MICASE; however, as no multi-word verbs were included in their definition of idiom, none were included in their analysis.

Similarly, Liu (2003) investigated the most frequently spoken American idioms by triangulating results from three corpora, one of which was MICASE. The author did include phrasal verbs in his definition of idioms for the study and created individual lists for each corpus used. Therefore, one of the products of the study was a list of 289 frequent idioms in MICASE, including phrasal verbs. Notable multi-word verbs high on the list included *go on*, *figure out*, *come up with,* and *find out*.

However, the methods for searching the corpora differed from the part of speech methods used by Biber et al., (1999) and Gardner and Davies (2007). As with Liu's (2011) study on phrasal verbs in COCA and the BNC, each item had to be searched individually because idioms do not necessarily follow a part of speech pattern like multi-word verbs. For example, phrasal verbs are consistently verb plus particle, while idioms like *fat chance*, *go hand in hand*, and *take*

a stab at are all constructed from different parts of speech. Therefore, they are quite difficult to search with current corpus tools.

Although Liu (2003) made thorough use of textbooks, dictionaries, and wildcard features to conduct the search, this method only found items which were searched for, perhaps missing others. The results offer valuable insight into idioms in academic speech and American English in general. Nevertheless, because the search of MICASE created by Liu (2003) was focused on idioms, it is unlikely that the phrasal verbs included in the findings are a solid representation of multi-word verbs in the corpus or the academic register in general.

2.4.2 Focus on Local Language

The mantra in corpus linguistics seems to be that bigger is better. According to Biber et al. (1998), even a one-million word corpus is inadequate to make generalizations about the meaning and use of words (p. 30). The BNC contains over 100 million words (Davies, 2004), while MICASE has over 1.8 million (Simpson et al., 2002), and COCA is over 560 million words and growing (Davies, 2008-). As computing power and the internet proliferate, the number of words in language corpora are starting to reach the tens of billions. The website Sketch Engine houses dozens of corpora for over 85 languages. Out of the fifteen corpora on the website's *Featured Corpora* page, only three fall below one billion words (Sketch Engine, 2018).

It is logical that making generalizations about an entire national language, such as Liu's (2011) comparisons of American and British English, would require a colossal sample size. However, when investigating the meaning and behavior of vocabulary, such as multi-word verbs, we must consider two questions: (1) Is it possible to make generalizations about these items in the context of such a large population of language users? (2) Are generalizations of this kind

helpful to specific language learning populations, such as ELLs at a university in the Midwestern United States?

Most language users know that the English spoken in Boston is different in many ways from that spoken in Alabama, the United Kingdom, or South Africa. The same is true for nearly all languages in the world. The feature we typically notice first is pronunciation differences. However, vocabulary also varies regionally. Ask someone outside of Wisconsin where the *bubbler* is, and you will likely get some quizzical looks. Using the term *water* or *drinking fountain* would be more effective. It is these language variations that have spurred projects such as the creation of the *Dictionary of American Regional English* (DARE) led by researchers at the University of Wisconsin-Madison (2018). Furthermore, it is not a stretch to think that multi-word verbs may occur in different patterns or even carry different meanings across different language communities.

Research using corpora in linguistics has started to narrow its focus in some ways in recent decades. COCA and the BNC, for example, are searchable by mode (i.e. written or spoken), year (e.g. 1980-1985), and certain contexts (e.g. broadcast news, parliamentary minutes, social science research, etc.). Additionally, researchers and even teachers have compiled corpora on specific topics such as the oil and gas industry (Aluthman, 2017), the study of medicine, and specific registers like spoken academic English (Dang et al., 2017; Simpson et al., 2002). MICASE (Simpson, et al., 2002) is a prime example of an attempt to focus corpus research on a more specific register in a specific location. It is with this idea of specificity to register, location, and time that motivated this study.

2.5 Research Questions

Several central themes in the literature reviewed here have contributed to the need for analysis of spoken academic language at the local level. First, prioritizing vocabulary which is frequent, challenging, and useful to ELLs' academic needs is essential for their success in a university setting. Second, because many ELLs studying at universities have spent more time with academic reading and writing skills, academic listening in the classroom poses a unique challenge to them.

Furthermore, the literature on academic vocabulary and multi-word verbs in academic writing is plentiful, while the research on spoken academic vocabulary, especially multi-word verbs, is less so. Additionally, the complexity of multi-word verbs and the challenges they pose to ELLs makes them an interesting and important vocabulary item to investigate in academic discourse. Finally, although corpus linguistics has provided tremendously valuable and powerful tools to analyze humungous collections of language, closer analysis of specific types of language are needed to provide evidence about vocabulary and other linguistic trends in those communities.

Accordingly, this study had two main aims. One aim was to investigate if it would be possible to compile a spoken academic corpus of instructor speech on a local level, with lecturers that currently teach or will likely teach university ELLs at their institution. The second aim was to explore the use of multi-word verbs in this created local corpus. These aims of the study prompted the following research questions:

1. **Which of the multi-word verbs identified in the local corpus occur most frequently?**
2. **How do the most frequent multi-word verbs in the local corpus compare to phrasal verb frequency lists from large English corpora?**

3. **What proportion of the local corpus is comprised of multi-word verbs?**
4. **Does multi-word verb use differ between general academic contexts and ESL contexts in the local corpus? If so, in what ways?**

3. Method

To address the research questions, instructors at the University of Wisconsin – River Falls (UWRF) were recruited to provide existing speech samples in order to create a local spoken academic corpus. This section first describes the university community at which the study was conducted and then the corpus creation process in terms of the departments and instructors who contributed speech samples, the speech collection process, descriptions of the types of media included in the corpus, transcription, the corpus compiling tool, and finally the methods for gathering data from the corpus and answering the research questions.

3.1 University of Wisconsin – River Falls

UWRF is a public university located in River Falls, WI, a town of about 15,000 residents located approximately 30 miles east of Minneapolis-St. Paul in the United States. The university is a member of the broader University of Wisconsin system and offers 70+ majors from four colleges: College of Agriculture, Food and Environmental Sciences (CAFES); College of Arts and Sciences (CAS); College of Business and Economics (CBE); and College of Education and Professional Studies (CEPS). According to the Office of Institutional Research Enrollment Report for Fall 2016, the student population was 5,931 students (5,482 undergraduate; 449 graduate) with over 120 international students from more than twenty countries. During the Fall 2016 semester, according to the UWRF International Education Office, the top five countries of origin for international students at UWRF were China, India, South Korea, Japan, and Taiwan,

and the top five majors for international students were Other (59), Business (41), Elementary Education (15), TESOL (12), and Computer Science (10) (personal communication, 2017).

For ELLs at UWRF looking to take classes to improve English language skills, the university offers two programs: English Language Transition Program (ELT) and Pathways Program. According to the UWRF website, the mission of the ELT Program is "to assist non-native English learners to develop and strengthen the knowledge and language skills necessary to achieve their academic, professional, and personal goals" (2018). In spring 2018, the ELT Program offered ESL courses at the intermediate and advanced levels and in all four skill areas as well as business English, academic language and classroom culture, grammar, vocabulary, and pronunciation. In previous semesters, beginning level courses have also been offered (Petree, R., personal communication, 2018)

The Pathways Program is a two-semester program for first-year domestic students for whom English is not their first language, according to the UWRF website. Students in Pathways take general education courses (e.g. communications or psychology) along with academic support courses focused on academic language, reading, writing, and oral communication (2018). Both the ELT and Pathways Programs are aimed at helping students succeed in using English for university study while addressing their many needs as international or domestic ELL students.

This study was aimed at analyzing vocabulary trends, mainly multi-word verbs, in the UWRF classroom by creating a local spoken academic corpus from classroom instructors in a range of academic disciplines including general education, major-specific, and ESL contexts. Additionally, it is hoped the corpus can serve as a research and language learning tool moving

forward. The project was approved by UWRF's Institutional Review Board (IRB) and assigned protocol number H2018 - T057.

3.2 The Local Corpus: FISC – Participants and Collection Procedures

The Falcon Instructor Speech Corpus (FISC) is a collection of UWRF instructor language, transcribed into text from pre-recorded video and screen capture lectures, tutorials, and task instructions. FISC was gathered and compiled over four months, from November 2017 to February 2018, and contains 52,725 running words.

3.2.1 Instructors, Academic Departments, and Speech Collection

Before contacting specific departments or instructors, I searched for courses offered online during the Fall 2017 and Spring 2018 semesters using the *electronic Student Information System* (eSIS). The thinking was that courses offered online would have a higher likelihood of using instructor-created media; however, media created for online courses, face-to-face courses, and recorded instructions for placement testing were accepted. Additionally, I spoke with colleagues, friends, and classmates to narrow the search for instructors who were known to have used original media in the past. Once a list of courses and instructors was compiled, appointments were made with instructors to explain the project and request media which included instructor speech. Instructors were provided with several avenues for electronically sharing media (see Appendix A) and asked to sign a consent form (see Appendix B). Once an instructor granted media access and had signed the consent form, I uploaded media to a cloud-based account connected to university email for secure storage. No material inducements were offered to instructors who shared media.

Attempts were made to collect media from a variety of departments and courses,

including content intended for ELLs specifically, content for the general education courses, and upper-level undergraduate and graduate courses. The end goal was to create a corpus that was as representative of UWRF instructor speech as possible. However, time and logistical constraints made creating a corpus of prerecorded speech a realistic and useful choice. Because of these time constraints, a goal of ten instructors (speakers) from various academic disciplines was set. At the end of data collection, a total of fifteen speakers from thirteen different departments contributed media for the creation of FISC: Agricultural Engineering (AGEN), Communication Studies (COMS), Computer Science and Information Systems (CSIS), Crop and Soil Science (CROP), Economics (ECON), English (TESL and ELT), Geology (GEOL), History (HIST), Psychology (PSYC), Teacher Education (TED), Music (MUS), and the UWRF Library (LIB).

Table 2 shows information about the instructors who shared media including their college, gender, and if they are a native speaker (NS) or non-native speaker (NNS) of English.

Table 2: Speaker Demographics in the Falcon Instructor Speech Corpus

College and Instructors		**Instructor Type**		**Gender**		**NS/NNS**	
CAFES	3 (20%)	ESL only	1 (7%)	Female	5 (33%)	NS	13 (87%)
CAS	8 (53%)	ESL + GC	3 (20%)	Male	10 (66%)	NNS	2 (13%)
CBE	2 (13%)	GC	11 (73%)				
CEPS	2 (13%)						
Total Colleges: 4		Total Instructors: 15					

Note: CAFES = College of Agriculture, Food and Environmental Sciences; CAS = College of Arts and Sciences; CBE = College of Business and Economics; CEPS = College of Education and Professional Studies; ESL = English as a Second Language; GC = General Content; NS = Native English Speaker; NNS = Non-Native English Speaker

Instructor Type indicates if the instructor teaches only ESL courses, general content (GC) courses (e.g. psychology), or both ESL and GC courses (e.g. ESL and TESOL teacher training courses). One NS in the corpus speaks a British English variety while all others, including all ESL instructors, speak various American English varieties.

3.2.2 Speech Event (Video) Descriptions

One of the corpora which inspired this study was MICASE (Simpson et al., 2002), described in section 2.4. While MICASE represents a wide variety of university campus language (e.g. classroom lecture, advisor meeting, and study sessions), FISC represents a more focused sample of a specific genre: prerecorded, academic classroom speech. In addition to having a more specific focus than MICASE, FISC was able to be compiled in just four months by a single researcher and provides a rich representation of the language that instructors at UWRF are using currently while teaching, especially for online and flipped classes. All speech events in FISC are monologues which were prerecorded by the instructor with students as the intended audience. Because all the speech events were recorded videos, the terms *speech event*, *media*, and *video* are all used interchangeably to refer to the videos collected from instructors.

The media which were used to create the transcripts that comprise FISC are typical of online courses and as supplements for face-to-face classes at UWRF. They are often shared with students through an online course management tool: Desire2Learn (D2L). All prerecorded videos used to compile FISC were created by instructors prior to the idea of using them for an academic corpus was conceived and for real instructional purposes; therefore, the language used in the media was not influenced by the project. For the purposes of this study, I have grouped the speech events used to create FISC into three categories: *prerecorded lectures*, *task instructions*, and *screen capture tutorials*.

Three main characteristics of the *prerecorded lectures* in FISC are that they include (a) an audio recording of the instructor's voice, (b) visual aids (e.g. text and pictures), and (c) are aimed at providing students with content knowledge (e.g. history of trade routes in the Middle East). The most common example of this type of lecture is the PowerPoint (PPT) voice-over: an

instructor lectures while moving through a PPT presentation. Twenty-two out of thirty-two (68.8%) speech events in FISC are prerecorded lectures.

Task instructions also include an audio recording of the instructor's voice and visual aids; however, the purpose of these videos is to give directions or instructions on how to complete a specific task. For example, FISC includes transcripts from instructions on procedures to take a computerized English placement test, how to complete a specific research project for a course, and the solutions for problems assigned as homework in an engineering course. Five task instructions (15.6% of the total speech events) are included in FISC.

Finally, although *screen capture tutorials* share the voice audio and visual aid aspects of prerecorded lectures and task instructions, their purpose is to teach a student how to use, typically, an online or computer tool, such as search for scholarly articles on the university library webpage or create a podcast using the software program Audacity. Five screen capture tutorials (15.6% of the total speech events) are present in FISC

There is undoubtedly some overlap with these categories. Recorded lectures may include task instructions for homework, usually at the beginning or end of the video, and the line between task instruction and tutorial can get blurry. For example, a tutorial could be aimed at learning how to use the tool to later complete a specific task for a course. Nevertheless, these categories help conceptualize the context in which the language in FISC is occurring. Media which was not created by UWRF instructors or did not include UWRF instructors' voices were not included in the creation of FISC.

In total, thirty-two speech events made up the entire 52,725 words of FISC, making an average of 1,647 words per speech event. However, the number of words in each speech event ranged from just 369 words to 3,477 words. Individual speech events ranged from one minute

and twenty-one seconds (1:21) to thirty-six minutes and thirty-four seconds (36:34). However, as discussed in the next section, some transcripts provided by instructors were from videos which were about one hour. The fifteen speakers averaged 2.13 speech events each, with four speakers providing just a single speech event and two speakers providing four. Some speakers provided multiple types of speech events (e.g. one lecture video and one tutorial video).

Table 3 provides a simple breakdown of the categories of speech events in terms of the number of videos, speakers, and words. As can be seen, the bulk of FISC is made of prerecorded lectures.

Table 3: Media Type Breakdown in FISC

	Prerecorded Lectures	Task Instructions	Screen Capture Tutorials	Total
Videos	22	5	5	32
Speakers	11	4	2	*15
Words	42,871	4,213	5,587	52,725

* Note: Two speakers contributed two types of speech events (e.g. 1 lecture and 1 tutorial); FISC=Falcon Instructor Speech Corpus

Although instructors may use other media, such as TED Talks or videos available through YouTube, the aim of this project was to analyze local language that students are meeting most often. The speakers represented in FISC are the instructors with whom UWRF ELL students have class or are likely to have class, and therefore an excellent sample to draw from for informing vocabulary pedagogy for ESL courses at the university.

3.3 Instruments: Transcription, Reliability, and Data Gathering

This section first defines the method used for transcribing the spoken language from instructor videos into text form using both automated and manual procedures. Second, it addresses the process of testing the reliability of this method. Finally, it describes the web-based

program used for compiling the corpus and the tools used to collect frequency and other information on multi-word verbs in the corpus.

3.3.1 Transcription

Transcription of audio from videos was primarily a two-step process. First, all speech was transcribed using the voice-typing tool on Google Docs and an audio program which enables sound to be sent from one application to another on a single computer. After adjusting sound settings, the media would be played on the cloud storage software on my laptop, and the voice typing tool on Google Docs would be activated in another window. Through this process, a transcript of the audio was created in the same time it took to play the video in its entirety (e.g. a ten-minute video took ten minutes to transcribe). This technique was extremely useful for creating transcripts and could be used in the future for making subtitles for instructor videos or further language research; however, depending on the audio quality and speaking style, the transcripts contained various mistakes and no punctuation.

The second step in transcription was to fix the errors made by the voice typing tool and add capitalization and punctuation that reflected the speakers' pauses and thought groups. For example, some homophones (e.g. *for*, *four*), technical words (e.g. *datum*), and proper nouns (e.g. *Chalmer Davee Library*) needed to be fixed as they were transcribed incorrectly by the program. Punctuation and some metalinguistic information was added by adapting the linguistic transcription guidelines as described by Brezina (2013, p. 7-8). This process consisted of listening to the media while simultaneously reading along and manually editing the transcripts. The speaking speed made it necessary to frequently pause videos while editing of transcripts took place, which made this a time-consuming process. Editing a transcript from a ten-minute video could take anywhere from one to two hours to complete. After editing was concluded, I

watched each video at least one more time in entirety while reading along with the transcript to check for errors. A few minor typos were also discovered while testing the corpus search tools and rectified at that time.

Several goals were kept in mind during the transcription process. First, conventions were followed to keep consistency between documents (See Appendix C for Transcription Guidelines for FISC). Also, as the primary aim of this study was to examine vocabulary, special attention was paid to accurately reflect the words used by instructors. Next, punctuation was added for two reasons. (1) Punctuated text facilitates the computerized tagging of parts of speech better than unpunctuated text, a topic discussed further in the next section. (2) Punctuated text aids the researcher, instructor, or student who is reading excerpts from the corpus by adding context, especially the thought groups used by the speaker. For example, compare the two concordance lines taken from FISC below:

Example A1: and expected outcome we've already talked about the importance of intrinsic motivators

Example A2: and expected outcome. We've already talked about the importance of intrinsic motivators,

It is unclear in Example A1 if the phrase *we've already talked about* belongs with the text that precedes it or that follows it. The punctuation in Example A2 makes the speaking boundaries clear for any readers and the program's part of speech tagging program.

Conventions in regard to capitalization, spelling, and punctuation aimed at making the corpus a useable teaching tool. In Examples A1 and A2 above, for instance, it would be easier to make and understand presentation slides or handouts from lines which closely follow writing conventions. However, FISC was also meant to represent real spoken discourse. Therefore, commas were used primarily to mark pauses in the speech that were not the end of a thought group and did not necessarily follow writing conventions. As illustrated in Example B, the

speaker pauses after saying *so* at the beginning of an utterance while the speaker in Example C does not pause between *so* and *that's*.

Example B: Sorry, I had to uh, clear my throat there. **So**, now we have the sketch set up. I'd like to rotate

Example C: offset in this, uh box here. We'll select okay. **So** that's our point. Um, then step six, create a

Of course, adding commas to the text in odd places can make it harder to read and might also confuse the part of speech tagging system. Ultimately, the end goal was to maintain an accurate representation of the speech as it occurred while simultaneously making the text understandable for myself and the computer program, a balance I believe was achieved.

As a final note on transcription, it should be stated that three of the speakers contacted for the project provided existing transcripts to be included in FISC. One of these transcripts was created by an instructor after the lecture had been recorded, while five others were scripts from which the instructors read while recording their videos. I read through these six transcripts and edited them to match the formatting of the created transcripts and fix typos. However, I did not request to view the videos connected to these transcripts; therefore, punctuation was left as it was or clarified by following writing conventions. Although this presents possible limitations in terms of the accuracy of these transcripts, there is no reason to believe the vocabulary used by the instructors in the videos is any different from what is represented in the transcripts.

3.3.2 Transcription Reliability

To ensure the transcription process described above was reliable, meaning others following the conventions would create similar transcripts, an undergraduate research assistant replicated transcription for six of the videos, including ESL contexts, general contexts, and one NNS. Excluding the existing transcripts provided by instructors, the sample that was replicated included roughly 42% of the speakers (five out of twelve that submitted videos), 23% of the

videos (six out of twenty-six transcribed), and 12% of the total running time of videos transcribed (about thirty-four minutes out of 4.5 hours).

After about one hour of discussion and training, the research assistant closely followed the transcription procedures outlined in the previous section to create six replication transcripts. Next, the researcher and research assistant each made a document containing the six transcripts selected for reliability testing. The researcher's document contained the original transcripts which are included in FISC, and the research assistant's document contained the six replication transcripts.

Both documents were then uploaded to a sandbox course management account. To measure how similar the documents were, both documents were submitted to Turnitin.com for originality checking. The idea behind this method was that if the transcripts were the same, they should each return a high unoriginality score and report each other as the matching documents. The replication document was submitted first and generated a first report indicating 7% unoriginality, which stemmed from a website on instructions for a computerized English placement test: the same placement test for which directions were given in one of the reliability transcripts. Next, the original document was submitted to Turnitin.com, which generated a report indicating that it contained 95% unoriginal content (2018). The document which generated the 95% match was the replication transcript document submitted by the research assistant, indicating the researcher's transcripts and the research assistant's transcripts were about 95% the same.

The 5% difference between the documents was accounted for by observing that some reduced forms were transcribed differently (e.g. *going to* versus *gonna*), some compound nouns were hyphenated differently (e.g. *offset* versus *off-set*), and fillers such as *um*, *uh*, and *oops* did

not always match. Furthermore, a few minor typos in the replication transcripts were observed. Overall, a 95% match suggests strong reliability of the transcription method. This is especially true for the vocabulary forms this study is examining because the issues accounting for the 5% difference between the documents would not have an impact on the multi-word verb forms being studied in the original transcripts.

3.3.3 Compiling the Corpus

After the transcripts were created and edited, they were uploaded one at a time to the website *Sketch Engine*, a web-based corpus management interface tool (Kilgarriff, 2014). As files were uploaded and then compiled by the program, it would automatically annotate, or morphologically tag, parts of speech (POS) in the text using a program called TreeTagger (Marcus, 1993). The program assigns each word in the text a POS (e.g. noun, verb, preposition, etc.), which allows the program to build sketch grammars (i.e. displays of the grammatical and collocational behavior of specific words) and makes searching for specific words and structures much more efficient. According to research conducted on the accuracy of TreeTagger for automated POS tagging, an error rate of 2-6% is typical (Marcus, 1993). A two-step searching process, which is detailed in the next section, was used to mitigate inaccurate data as a result of the errors in POS tagging.

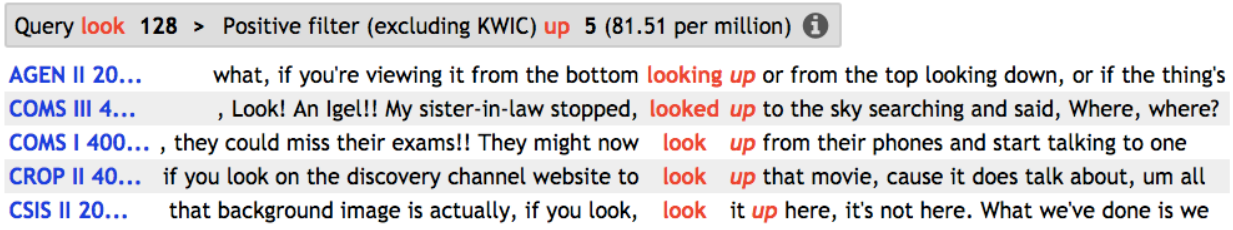
Transcripts were uploaded one at a time to create individual files for each speech event. This makes it possible for a researcher, teacher, or learner to have more control when interacting with the corpus. For example, using Sketch Engine to explore FISC, it is possible to search and analyze the entire database of thirty-two speech events, the language in an individual lecture about crops, all the transcripts from the UWRF Music Department, or only materials used in an ESL context. Users are also able to create sub-corpora including any combination of files they

wish. This specificity is useful because of the differences in discourse structures and vocabulary among different academic disciplines discussed in the literature review.

3.3.4 Data Gathering

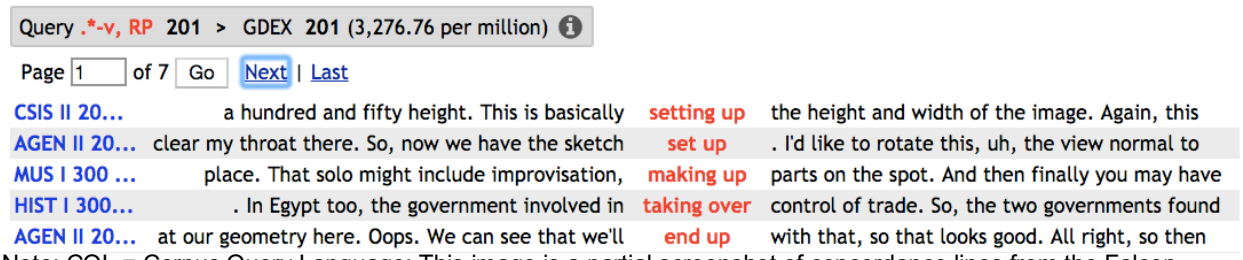
Automated searching of FISC for multi-word verbs was completed using a search tool called CQL (Corpus Query Language). According to the Sketch Engine website, CQL is a code used for complex searches that cannot be completed using the standard concordance search tools (2018). For example, standard searching tools allow a user to search for a specific multi-word verb by entering a lemma (e.g. *look*) and context information about what occurs near that lemma (e.g. *up* within three tokens to the right of *look*). CQL, on the other hand, allows the user to search for multiple combinations of POS with a single query, such as all verbs which are followed by a particle. Figure 1 shows the search results for a standard search of *look up* in FISC, and Figure 2 shows five results of a verb plus particle CQL search.

Figure 1: Lemma + Context Search Example Using Sketch Engine



Note: This image is a partial screenshot of concordance lines from the Falcon Instructor Speech Corpus in Sketch Engine (2018)

Figure 2: CQL Search for Verb + Particle Occurrences Using Sketch Engine



Note: CQL = Corpus Query Language; This image is a partial screenshot of concordance lines from the Falcon Instructor Speech Corpus in Sketch Engine (2018)

To search for multi-word verbs, CQL was used to conduct three searches: (1) verb plus particle, (2) verb plus preposition, and (3) verb plus adverb. Particles, prepositions, and adverbs were searched within three tokens of the verb in order to include multi-word verbs separated by objects or adverbs and three-part phrasal-prepositional verbs (e.g. *look* it *up*, *talk* excitedly *about*, or *come up with*). The CQL codes used for the searches are as follows:

verb + particle = [lempos=".*-v"]([tag="RP"]|[][tag="RP"]|[][][tag="RP"])

verb + preposition = [lempos=".*-v"]([lempos=".*-i"]|[][lempos=".*-i"]|[][][lempos=".*-i"])

verb + adverb = [lempos=".*-v"]([lempos=".*-a"]|[] [lempos=".*-a"]|[][][lempos=".*-a"])

However, after identifying some examples in the corpus which were separated by more than three tokens (e.g. *making* all of that *up*), individual items were searched with up to five tokens between verb and particle/preposition/adverb. Several reasons motivated conducting the multi-word verb search in this way.

First, in addition to verb plus particle combinations, searching for verb plus preposition and verb plus adverb combinations was thought necessary due to limitations in POS tagging, the ambiguity of defining multi-word verbs, and to include polysemous occurrences. For example, Figure 1 illustrates how, in one case, *look up* means to raise one's eyes (*look up* from their phones), while in another case it means to search for (*look up* that movie). As previously discussed, even linguists sometimes disagree whether *up* is functioning as a particle, preposition, or adverb in these cases, and the tagging program had trouble accurately marking the POS for some of these items. Therefore, all possibilities were searched for. Additionally, including polysemous occurrences of these items, as was seen with the PHaVE List (Garnier & Schmitt, 2014), is important from the perspective of informing pedagogy.

The reason for searching with CQL was that it was much faster than searching for specific combinations one at a time. Additionally, if only specific combinations were searched for, only those combinations would have been found, possibly missing others. Finally, separating the search into three categories created three lists of combinations which could be prioritized for analysis. For example, verb plus particle combinations were addressed first because this category would likely contain the most idiomatic phrasal verbs, while the other search categories would contain many free combinations that could be eliminated from the final list. If analyzing all three lists proved impossible for this study's timeline, verb plus particle combinations would have already been separated and singled out for analysis. The same searches were performed on multiple days to test the reliability of the search program. Searches returned the same results each time.

After each search, a frequency tool was used to create a list of the lemmas found by the program. Figure 3 shows the frequency tool display for the five most frequent verb plus particle combinations found using CQL in FISC. The display indicates that *come up* occurred thirteen times, *go up* occurred nine times, and so on.

Figure 3: Frequency of Top 5 Verb + Particle Constructions in FISC in Sketch Engine

Note: FISC = Falcon Instructor Speech Corpus; This image is a partial screenshot of concordance lines from FISC in Sketch Engine (2018)

The three lists created by the frequency tool were saved, loaded into a spreadsheet, and used to search for each item individually. For example, each item in Figure 3 was searched individually

using the standard lemma plus context search tool described in Figure 1 (e.g. lemma *come* followed by *up* within five tokens to the right).

These additional searches were conducted for two reasons. (1) Because the lemma plus context tool searches for lemmas, not tagged POS, it was another way to put all of the instances of an item together, regardless of the POS. For example, some instances of *talk about* are tagged as verb plus preposition, and other instances are tagged as verb plus adverb. Having all the concordance lines of *talk about* in one place made it easier to search each line for false identifications, usable examples, and other data. This type of search also displayed other phrasal items like those discussed in section 2.3.2 on phrasal verbs. For example, instances of *look at* where *look* is a noun, such as in *take a look at*, were counted in the total occurrences for that item because the function is essentially the same as the verb *to look at*. (2) Each line was analyzed to gather additional data such as how many different speakers used each item (identifiable by the file name) and to identify false positives, to sort occurrences into general academic and ESL contexts, and judge meaning sense occurrences for some examples.

Although all of the items on the three lists were searched and analyzed, only items that occurred a minimum of three times were included in the final list and analyses because less frequent items are less likely to be important to learners and would have resulted in a very long word list with many items occurring only once. Moreover, an item which occurred three times in FISC had a frequency rate of 48.91 words per million, demonstrating the items' frequent nature. Biber et al, (1999) and Liu (2011), for example, set the bar for item inclusion at only 10 words per million, meaning even the items at the bottom of this studies' list are relevant. As a result, the final FISC Multi-Word Verb List (FISC List) includes a manageable sixty-eight unique items with high frequencies. The items are lemmatized, meaning an item like *go on*, for example,

includes occurrences like *went on*, *going on*, etc. (See Appendix D for the entire FISC List). Although items occurring only one or two times in FISC, such as *sit down*, *turn down*, *list out*, and *extrude up*, were not included, they might be worthy of further of analysis where time allows.

4. Results

RQ1: Which of the multi-word verbs identified in the local corpus occur most frequently?

The purpose of the first research question was to create a list of frequent multi-word verbs from FISC to offer UWRF ESL instructors evidence regarding which items might occur often. To answer RQ1, the sixty-eight items included in the final multi-word verb list were first ranked according to their total number of occurrences. However, a simple occurrence ranking was insufficient for this study because the corpus was limited to only fifteen speakers, and items used numerous times by a single speaker appeared relatively high on the frequency ranking list. For example, although *breed for* occurred eleven times and ranked as 19th out of sixty-eight on the frequency list, it was only used by a single speaker and in a single video. In this case, *breed for* was used only in a lecture about disease resistance in crops and is likely specialized vocabulary for specific academic disciplines.

Therefore, to take into account how widespread an item was among multiple speakers, a score was calculated for each item based on both the raw number of occurrences and the number of speakers who used the item. These *item scores* were calculated by dividing the number of speakers who used the item by the total number of speakers in the corpus (15) and multiplying that result by the raw number of occurrences. Basically, the percentage of speakers who used the item multiplied by the number of occurrences. For example, the item score for *talk about* was calculated in the following way: [number of speakers (11)/total speakers in FISC (15)] * number

of occurrences (96) = 70.40 or (11/15) * 96 = 70.40. Items were then re-ranked according to item score.

Ranking by item score decreased the ranking of items which were used many times but by few speakers and increased the ranking of items which were used fewer times but by a wider range of speakers in the corpus. In other words, higher item scores suggest higher salience for the learner because a higher score indicated frequent as well as widespread use of the item. Table 4 shows the top ten multi-word verbs in FISC ranked by raw occurrence in the far left column, and the top ten ranked by the calculated item score in the far right column.

Table 4: Top 10 Multi-Word Verbs in FISC

Raw Occurrence Ranking	Number of Occurrences	Number of Speakers	Item Score	New Ranking by Item Score
1. talk about	96	11	70.40	1. talk about
2. look at	64	11	46.93	2. look at
3. think about	47	10	31.33	3. think about
4. listen to	36	4	9.60	4. go on
5. focus on	29	5	9.67	5. refer to
6. go on	25	10	16.67	6. focus on
7. refer to	25	7	11.67	7. listen to
8. get to	19	4	5.07	8. look for
9. work with	18	6	7.20	9. deal with
10. deal with	17	7	7.93	10. start with

Note: FISC = Falcon Instructor Speech Corpus

As can be seen, high-ranking items with few speakers, such as *listen to* and *get to*, dropped down the list, while items with more speakers (e.g. *go on*) moved up the list. The previous example with only one speaker, *breed for*, dropped from 19th to 51st on the list, demonstrating a more appropriate ranking for a specialized item.

Thus, the answer to RQ1 in terms of raw frequency is seen in the far left column of Table 4; however, perhaps a more valuable answer for teachers to RQ1 is in the far right column of Table 4. *Talk about*, *look at*, and *think about* were the most frequent by raw occurrence as well

as by item score, demonstrating frequent and pervasive use. *Go on*, *refer to*, and *focus on* were also in the top ten by both raw occurrence and item score. Other items in the top ten are *listen to*, *look for*, *deal with*, *get to*, and *start with*. See Appendix D for the full list of sixty-eight items of the FISC List ranked by raw occurrence and by item score.

RQ2: How do the most frequent multi-word verbs in the local corpus compare to phrasal verb frequency lists from large English corpora?

The second research question examined if the FISC List identified similar items as the previously mentioned PHaVE List (Garnier & Schmitt, 2014). The PHaVE List was selected for comparison to the FISC List in this study for several reasons. First, the PHaVE List contains all the phrasal verb items identified as most frequent in the LSWE Corpus by Biber et al. (1999), the most productive phrasal verb combinations in the BNC according to Gardner and Davies (2007), and additional frequent items added from COCA by Liu (2011). In other words, the PHaVE List is the most current pedagogical list of frequent and productive phrasal verbs created from large English corpora available in the TESOL literature.

Liu's (2003) list of spoken American idioms also includes some phrasal verbs; however, it includes many formulaic items which do not resemble multi-word verbs in any way (e.g. *as a matter of fact*) and is nearly 300 items long. The PHaVE list includes only phrasal verbs as well as some prepositional verbs due to the authors including polysemy examples (e.g. *look up* – search for; *look up* – raise one's eyes), and is a more manageable 150 items long.

Additionally, because the PHaVE List was created from very large corpora (i.e. BNC and COCA), a comparison with a list created from a smaller, more focused corpus (i.e. FISC) can help shed light on if vocabulary patterns across a very large, non-specific sample of language

occur in a similar way when a smaller, more specific sample. We should keep in mind, however, the research the PHaVE List was created from did not search for prepositional verbs explicitly. Therefore, a comparison of these lists is not a comparison of items with exactly the same definition, which greatly limits findings from the comparison. It could be argued that the research that informed the PHaVE List excluded prepositional verbs because they are typically more transparent and known to learners.

Still, both the FISC List and the PHaVE List are pedagogical lists, meaning they aim to help instructors and learners know which vocabulary items to address first in language learning. This study assumes that frequent prepositional verbs are also important to include in vocabulary curriculum for university ELLs because they are not always transparent in meaning and known to learners. If a large number of items from the FISC List are covered by the PHaVE List, this would provide some evidence that the PHaVE List might be sufficient for teaching multi-word verbs in this specific academic context. Conversely, if many items from the FISC List do not appear on the PHaVE List, it would provide some evidence that perhaps more prepositional verbs should be included in spoken academic vocabulary research, especially if those items are opaque or idiomatic. Finally, a lack of coverage by the PHaVE List could also suggest that more specific and local corpora may be needed to address the specific language learning needs of ELLs.

To begin to answer RQ2, a side-by-side comparison of the two lists was made. Table 5 displays the top ten items from both lists. The item score ranking was used for the FISC List because it more accurately reflects item salience in FISC as described in the previous section. Similarly, the main source of the PHaVE List, Garndner and Davies (2007), did not rely on

frequency alone to prioritize items. As can be seen, only one item, *go on*, occurred in the top ten items of both lists.

Table 5: Top 10 Item Comparison Between the FISC List and the PHaVE List

FISC List	PHaVE List
1. talk about	**1. go on**
2. look at	2. pick up
3. think about	3. come back
4. go on	4. come up
5. refer to	5. go back
6. focus on	6. find out
7. listen to	7. come out
8. look for	8. go out
9. deal with	9. point out
10. start with	10. grow up

Note: FISC = Falcon Instructor Speech Corpus; PHaVE = Phrasal Verb Pedagogical List (Garnier & Schmitt, 2014)

To compare the lists more thoroughly, the entire PHaVE List was searched for items from the FISC List. The search identified that twenty-seven items out of the total sixty-eight on the FISC List are also included in the PHaVE List, meaning overlapping items account for about 40% (27/68) of the FISC List and about 18% of the PHaVE List (27/150).

However, a more useful measurement might be to examine how much coverage of multi-word verb occurrences in FISC these twenty-seven items provided. If many of the twenty-seven items were highly frequent in FISC, such as *go on*, the percentage of coverage would be relatively high. On the other hand, if many of the twenty-seven items are the less frequent in FISC, the percentage of coverage would be low. In other words, if a student had learned the PHaVE List, and therefore these twenty-seven items, what percentage of the multi-word verb occurrences in FISC would they have understood, assuming an ideal situation where all

occurrences were understood?

These twenty-seven overlapping items occurred in FISC a total of 204 times, an average of 7.6 times per item. Dividing these 204 occurrences by the total number of multi-word verb occurrences (811) results in a figure of approximately 25%. Therefore, a student who had learned the PHaVE List would have understood about a quarter of the multi-word verb occurrences identified FISC by this study, assuming all other items were unknown.

From these analyses, we can answer RQ2 by saying that the FISC List and the PHaVE List are quite different in terms of the top-ranked items, only one out of ten items occurred in both lists. Moreover, knowledge of the items on the PHaVE List would provide only 25% coverage of the multi-word verb items which occurred in FISC. On the other hand, it might be reasonable to assume that some frequent items from the FISC List (e.g. *talk about* and *think about*) are fairly transparent and therefore known to many learners. Furthermore, the research that informed the PHaVE List was not searching for prepositional verbs, so we should not expect overlap of those items. Appendix E displays all twenty-seven items which occurred on both lists and their rankings in their respective lists.

RQ3: What proportion of the local corpus is comprised of multi-word verbs?

The third research question aimed to address how urgent this issue of multi-word verbs was in terms of the impact on total input received by students. Put another way, are multi-word verbs a large percentage of what is heard by students in the classroom? To answer RQ3, the number of multi-word verb occurrences in FISC needed to be compared to the total number of words in the corpus. In other words, what percentage of the full-text of the corpus consisted of multi-word verbs?

This percentage was calculated using the sixty-eight items from the FISC List. These sixty-eight multi-word verb items occurred a total of 811 times in FISC. Because the tagging program counted each one of these 811 occurrences as at least two words (e.g. verb and particle), the number of occurrences (811) was doubled to make a total of 1,622 words. Therefore, by dividing these 1,622 words by the total 52,725 words, we can see that about 3.1% of the words in FISC were part of a multi-word verb.

However, it can safely be assumed that this 3.1% estimate was lower than the actual percentage of multi-word verbs in the corpus for two reasons. (1) The FISC List only contains items that occurred at least three times. If all of the low-frequency items were included, the percentage would increase, although probably by a small amount. (2) The tagging program counted phrasal-prepositional verbs as three words (e.g. verb + particle + preposition), meaning 1,622 was a low estimate of the total number of words which were part of a multi-word verb. A closer analysis would need to be conducted to determine a more exact number; however, the data collected for this study suggests that at least 3.1% of the corpus was made up of multi-word verbs.

RQ4: Did multi-word verb use differ between general academic contexts and ESL contexts in the local corpus? If so, in what ways?

The fourth research question aimed to address the differences between multi-word verb use in the ESL classroom and other classrooms at the university. It was hypothesized that general academic contexts would contain a higher rate of multi-word verbs than ESL contexts because ESL instructors sometimes simplify their vocabulary by avoiding idiomatic formulas such as multi-word verbs or intentionally use academic vocabulary to increase learner exposure to those

forms. General academic instructors, on the other hand, are probably less likely to be as intentional about vocabulary choices as language instructors and mix colloquial and academic forms while teaching.

Answering RQ4 was approached in terms of the percentage of multi-word verbs that occurred in general academic contexts compared to ESL contexts. To compare the two categories, the corpus and the occurrences of multi-word verb items were separated into *general contexts* and *ESL contexts*. As seen in Table 6, the majority of the total words in FISC were from general contexts at 50,147 words (95.1% of FISC), while only 2,578 words (4.9% of FISC) were from ESL contexts.

Table 6: General Context and ESL Context in FISC

	Total	General Context	ESL Context
Words in FISC	52,725	50,147	2,578
MWV Words	1,622	1,590	32
Percentage	3.1%	3.2%	1.2%

Note: FISC = Falcon Instructor Speech Corpus; ESL = English as a Second Language; MWV Words = multi-word verb words (e.g. *look up* = 2 words)

Of the roughly 1,622 words (811 items * 2 words each) that made up multi-word verbs in FISC, 1,590 occurred in general contexts, while only 32 occurred in ESL contexts. That means that only 1.2% of the ESL context words were part of multi-word verb items, while 3.2% of the general context words were part of multi-word verbs. Looked at another way, if the sample of words which occurred in a general academic context and an ESL context were the same size (e.g. 50,147 words), the number of words that were part of a multi-word verb item in a general context would have remained at 1,590 and would have been 623 for the ESL context, assuming a linear relationship between multi-word verb items and other language in the ESL context.

Whether we compare the percentages of items in the two contexts (i.e. 3.2% to 1.2%) or examine a proportional comparison of the words in the two different contexts (i.e. 1,590 to 623), the numbers suggest the rate of multi-word verb items in ESL contexts was less than half of that in general academic contexts. Therefore, according to this analysis, the answer to RQ4 was yes, multi-word verb use did differ between general academic contexts and ESL contexts in terms of rate of use. Multi-word verb items were used more often in general academic contexts than in ESL contexts.

5. Discussion

Taking a glance at the top ten multi-word verbs from the first research question, we might say these are mostly prepositional verbs with fairly transparent meanings. However, the items that follow in bold do appear in the *Cambridge Phrasal Verbs Dictionary* (2006), and the underlined items appear in a student phrasal verb workbook, *Zero In! Phrasal Verbs in Context* (Root & Blanchard, 2003): *talk about*, ***look at***, *think about*, ***go on***, ***refer to***, ***focus on***, *listen to*, ***look for***, ***deal with***, and *start with*. As can be seen, seven of the top ten do appear in resources available to ELLs at the university library. However, the specialized dictionary, though a good reference, is limited in helping teachers and learners select words or meanings for study because of its comprehensive coverage of items and meanings. Additionally, Root and Blanchard's (2003) workbook contains only two of the top ten items. Using the FISC List and the context provided in the corpus can help prioritize multi-word verb items and their uses in a way that might be more useful for ELLs at UWRF.

Not visible from the list itself is that the items did not all have only a single, transparent meaning in FISC. For example, number two on the list, *look at*, had two general meaning senses in FISC, one of which occurred twice as often as the other. The following example is a

representation of the item in a style similar to the PHaVE List (Garnier & Schmitt, 2014), with the definitions, occurrence numbers, and examples with subject areas taken directly from FISC:

2. look at – 64 total occurrences
a. Consider, examine, or study in some way – 43
In the next chapter eight video, we'll **look** more closely **at** how to demonstrate professional excellence when using email at work. (*Communications Lecture*)
b. Examine with the eyes – 21
And if we **look at** it, from the top, we can see that we have one hook between these two legs and one hook between those two legs. (*Agricultural Engineering Tutorial*)

This type of information could be helpful to learners at UWRF because it provides not only items and definitions, but also contextualized examples from real speakers at the university. The fact that the examples are from real instructors at the students' university could motivate learners to engage with the material in a thoughtful way.

Number four on the list, *go on*, is an unsurprising item as it is near or at the top of all the phrasal verb frequency lists in other studies reviewed in this paper. We see a similar split in meanings sense with this item as well:

4. go on – 25 total occurrences
a. Happen, take place – 17
I think it'll give you some great background to some of the major issues, themes, and research that's **going on** in the field right now. (*TESOL Lecture*)
b. Continue, proceed – 8
So for example we now have a form that is ABCDEF, and it could **go on** and on. (*Music Lecture*)

One example of *go on*, however, is not clear cut. Figure 4 shows expanded context of *went on* in a history lecture in FISC where it could mean either *to happen* or *to continue*.

Figure 4: "Went on" in Context of a History Lecture in FISC using Sketch Engine

< previous , and the Italian merchants who controlled their export-import economy. Along its western end, Muslim and Christian traders carried out a profitable trade. There was frequent violence, military confrontations, and trading embargoes. Still trade **went** on. It was hinged on Baghdad. If the Italians were the hinge of Europe, Baghdad was the hinge of the Middle East. During the tenth to thirteenth Centuries it was a great and very prosperous metropolis next >

Note: FISC = Falcon Instructor Speech Corpus; This image is a partial screenshot of context from concordance lines from FISC in Sketch Engine (2018)

I have counted it in definition *b. Continue, proceed* in the occurrences counted above, but

this is only a guess at what the speaker meant. Interestingly, the context shows two additional multi-word verb items, *carried out* and *hinged on*. *Hinge* is also used in a metaphor by the instructor.

This polysemy information is important because instructors and learners might focus attention on the most straightforward and literal definition of these combinations. But, as we can see, the more frequent uses of these two examples, *look at* and *go on*, are figurative and idiomatic in FISC. Evidence that the items are both frequent and likely difficult for learners due to non-literal meanings makes them good candidates for being prioritized for vocabulary study. Furthermore, the abundance of authentic examples provided by even a modestly sized corpus could be extremely useful to instructors.

However, it is difficult to say how challenging on the whole the multi-word verb items in the FISC List are for the ELLs at UWRF. Some of the items are idiomatic or figurative (e.g. *look at*, *go on*, *come up with*, *point out*), while others are fairly literal (e.g. *think about, depend on*, *click on*). Ultimately, the FISC List and examples from the corpus are tools that instructors and learners can add to their store of resources. The choice of which of these items to address, if any, falls to individual instructors and learners.

Another interesting observation about the top ten items is that they seem to represent some of the language of a lecture with words which are usually not identified as academic. During university courses, things are *talked about*, *looked at*, and *thought about*, and that is reflected by these items being used a total of 207 times in FISC. In the pursuit of addressing academic language for learners, ELL instructors might be tempted to assume that other university instructors avoid these colloquial multi-word verbs and choose academic vocabulary alternatives such as *discuss*, *examine*, or *consider*; however, these items only occurred a

cumulative 41 times in FISC. The academic terms are important because they are less likely to be in everyday conversation and perhaps less likely to be known to ELLs.

Still, evidence from this study suggests these multi-word verbs are used by instructors while teaching. University ESL instructors should confirm that their students understand how these items can be used in academic contexts, especially that they are used with literal and figurative meanings. Furthermore, instructors and learners should be disabused of the idea that classroom lectures are a realm where academic language dominates. The idea that colloquial language is useful for university study might also encourage learners to engage in more casual language experiences in their host communities because the value of such experiences can be connected to their academic goals.

Rather than focus on the results of research question two in terms of the differences between the FISC List and the PHaVE List, it might be more productive to consider the twenty-seven items which occurred on both lists. These items are likely important for ELLs at UWRF, especially items which rank highly on both lists. For example, *figure out*, *go back (to)*, *go down*, *go on*, *make up*, *set up*, and others all occurred in the top fifty items on both lists. This means they have been identified as productive and frequent across registers in the LSWE Corpus, the BNC, COCA, and now in a local, spoken academic corpus at UWRF. Accordingly, these items might serve as all around very useful items for ELLs at UWRF. A closer examination of the meaning senses in FISC, however, is needed before a more meaningful comparison of these twenty-seven items can be made. We can see that *go on*, for example, follows similar meaning patterns in both FISC and the PHaVE List: *to happen* or *take place* accounts for about 60-70% of the meanings while *to continue* accounts for most of the remaining occurrences in both analyses

(Garnier & Schmitt, 2014). Other items from these two lists might follow similar or different meaning sense patterns.

The results for research question three, at least 3.1% of the FISC being multi-word verbs, suggests that they are an important part of the language in FISC and possibly in instructor speech across campus. Although a small percentage, we should remember that because these verbs are content words, they carry important information which greatly impacts the meaning of sentences.

The results of research question four did suggest differences in the way multi-word verbs are occurring in general academic contexts and ESL contexts. This should raise some questions about the vocabulary trends in these different contexts and about what steps should be taken to ensure these and other vocabulary items are being addressed to prepare students for future lectures in their general education and major specific course. However, given that the portion of the corpus which was made up of ESL context material was very small (only three instructors, five videos, and 2,578 words) it is difficult to make definitive generalizations.

Another factor that may have influenced the occurrence of multi-word verb items in ESL contexts in FISC was that the ESL context videos all fell under the categories of task instructions and screen capture tutorials. In other words, none of the ESL videos were lectures. This is understandable because English language classes are often more about learners doing tasks with language than instructors lecturing about a topic. Furthermore, all of the videos from ESL contexts were supplements for face-to-face classes, while general content videos were from a mix of online, flipped, and face-to-face classes. Comparable sample sizes and speech events which are more alike in purpose might provide better insight into differences between the two contexts. It may be the case that ESL instructors at UWRF are exposing students to lecture-style

speaking in the classroom. If not, it may be helpful to learners to experience listening to such lectures in their preparation for general academic classes.

6. Limitations

Although this study has shown it is possible to compile a local, spoken academic corpus in a relatively short amount of time, the methods of compiling and searching the corpus, as well as the results generated in this study have several limitations. First, the type of prerecorded language in FISC is undoubtedly somewhat different from live classroom lectures and discussions. Some of the videos collected for this project had the feel of extemporaneous speech, while others were clearly scripted as evidenced by the instructors saying as much. Although this local corpus gives us a glimpse into how instructors are likely speaking in the classroom, samples of live lectures and classroom activities would be a valuable addition to what has already been collected.

A second limitation was the size of the corpus. Because this study was not attempting to generalize findings to an entire language or register as other large corpora studies often do, millions or billions of words were probably not needed to provide evidence of vocabulary patterns. However, a larger number of words, speakers, academic topics, and speech events would make a stronger case for any findings.

In terms of searching methods, issues with the computerized part of speech annotation, or tagging, were mitigated in this study by searching each item identified in CQL searches as individual lemmas; however, it is possible that items were missed due to some inaccurate tagging. Although it is possible to edit part of speech tagging with Sketch Engine after the process has been completed by the software program, this is a time consuming task. Solving the

issue of accurate corpus tagging requires either more advances in the technology or cumbersome human editing.

These methodological limitations suggest the results for the four research questions are similarly limited. As should be stressed with all corpus research, findings from corpora analyses apply to the database of language from which they were derived. Therefore, ELLs and others at UWRF should be aware that the items on the FISC List and other information provided through answering the research questions apply to the thirty-two videos from the fifteen instructors in the corpus, with possible, but not concrete, evidence for vocabulary patterns in classrooms across the university.

Regarding the results from research question four, larger sample sizes and a more thorough statistical analyses using a two-way chi-square and effect size measures similar to those used by Liu (2011) would provide more compelling evidence of general context versus ESL context differences. Additionally, the FISC List is a helpful starting point in identifying the multi-word verbs which were frequent in the corpus, but meaning sense information such as that provided for *look at* and *go on* is needed to make the list a more useful teaching tool. Furthermore, the meaning sense information provided in the discussion of this study is limited in that I alone tallied the meanings for different occurrences. Multiple readers and inter-rater reliability similar to that used by Garnier and Schmitt (2014) would make a stronger case for the meaning trends.

Finally, automated transcribing with Google Docs saved a considerable amount of time. However, although not a limitation of the methods or results of this study, editing and cleaning transcripts manually was very time consuming. This study demonstrated it is possible to create a localized, spoken corpus in a relatively short amount of time. Still, for the moment, the creation

of spoken corpora has to rely, at least in part, on good old fashioned manual transcription. This reality makes creating spoken corpora a cumbersome task. However, speech recognition technology is improving rapidly, and it may not be long until an instructor simply wearing a microphone in class can provide an accurate transcription document.

7. Future Research and Conclusions

This study achieved its first aim by demonstrating it is possible to create a localized, spoken academic corpus to analyze vocabulary trends. Continued investigations of the academic register with localized corpora can add to informed language teaching and learning methods for university ELLs, especially in terms of vocabulary and discourse patterns. More of these types of corpora are needed to investigate if claims from large general corpora about items such as phrasal verbs hold true in specific registers and language communities. This study has provided possible methods for creating such localized corpora and a solid foundation for the local university to build from. As technology progresses, it will likely become easier and easier to build these spoken corpora and quickly collect evidence of vocabulary frequencies and patterns, which can be used to inform pedagogy.

The second aim of the study, exploring the use of multi-word verbs in the localized corpus, has provided some interesting findings about frequency and meaning. Although some evidence suggests multi-word verbs are less frequent in written academic corpora, this preliminary study suggests they may be more important in classroom language than intuition would tell us. As American university classrooms shift from traditional lecture models to more informal and interactive models, more and more conversational vocabulary will likely mix with the academic vocabulary which we think of as typical to university study. Now that tools for investigating different vocabulary patterns locally are available, researchers, teachers, and

students should allocate available time and resources to investigating what important vocabulary items are common at their institutions. Creation of such databases would not only be useful for the local English learning community, but also for comparison between varying language communities.

References

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12(4),* 235-247.

Altenberg, B. (1990). On the phraseology of spoken English: The evidence of recurrent word-combinations. In Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications* (pp. 101-122). Oxford: Clarendon Press.

Aluthman, E. S. (2017). Compiling an OPEC word list: A corpus-informed lexical analysis. *International Journal of Applied Linguistics and English Literature, 6(2),* 78

Ashkan, L., & Seyyedrezaei, S. H. (2016). The effect of corpus-based language teaching on Iranian EFL learners' vocabulary learning and retention. *International Journal of English Linguistics, 6(4),* 190.

Barekat, B., & Baniasady, B. (2014). The impact of phrasal verb avoidance on the writing ability of the university EFL learners. *Procedia - Social and Behavioral Sciences, 98*, 343-352.

Barker, L., Gladney, K., Edwards, R., Holley, F., & Gaines, C. (1980). An investigation of proportional time spent in various communication activities by college students. *Journal of Applied Communication Research, 8(2),* 101-109.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Philadelphia: John Benjamins Publishing Company.

Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly, 35(2),* 331.

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

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V… Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. RM-04-03, TOEFL-MS-25, ETS Research Memorandum, 374 pages. Retrieved from https://www.ets.org/research/policy_research_reports/publications/report/2004/ibyq

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

Brezina, V. (2013). *Practical workshop: Compiling and analysing a spoken academic corpus*. Lancaster University. Retrieved from http://www.lknol.com/Docs/booklet.pdf

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 36(1),* 1-22

Brians, P. (2018). "Churchill" on prepositions. The Website of Prof. Paul Brians. Washington State University. Retrieved from https://brians.wsu.edu/2016/11/14/churchill-on-prepositions/

Brown, H. D. (2007). *Teaching by principles: an interactive approach to language pedagogy*. White Plains, NY: Pearson Education.

Cambridge phrasal verbs dictionary (2nd ed.). (2006). Cambridge, UK: Cambridge University Press.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Boston: Heinle.

Celce-Murcia, M., Brinton, D., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide (2nd ed.)*. New York, NY: Cambridge University Press.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10(2),* 157-187

Cooper, T. C. (1998). Teaching idioms. *Foreign Language Annals, 31(2),* 255-266.

Corpus Query Language (2018). Sketch Engine Glossary. Retrieved from https://www.sketchengine.eu/user-guide/glossary/?letter=C

Corpus. (2018). In Merriam-Webster's Learners Dictionary online. Retrieved from http://learnersdictionary.com/definition/corpus

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford: Clarendon Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34(2),* 213.

Cummins, J. (1980). Psychological assessment of immigrant children: Logic or intuition? *Journal of Multilingual and Multicultural Development*, *1*, 97-111.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition, 7(01),* 73.

Dang, & Webb. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*, 66-76.

Dang, T., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning, 67(4),* 959-997.

Darwin, C. M., & Gray, L. S. (1999). Going after the phrasal verb: An alternative approach to classification. *TESOL Quarterly, 33(1),* 65.

Davies, M. (2004-) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Retrieved from https://corpus.byu.edu/bnc/.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Retrieved from https://corpus.byu.edu/coca/.

Dictionary of American Regional English (2018). University of Wisconsin – Madison. Retrieved from http://dare.wisc.edu/

Dudley-Evans, T. (1994). Variations in the discourse patterns favoured by different disciplines and their pedagogical implications. In J. Flowerdew (Eds.), *Academic listening: Research perspectives* (pp. 146 - 158). Cambridge: Cambridge University Press.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24(2),* 143-188.

English Historical Book Collection Corpus (2018). Sketch Engine: Historical collection: EEBO Phase I, ECCO, Readex's Evans. Retrieved from https://www.sketchengine.eu/historical-collection-eebo-ecco-evans/

English Language Transition Program and Pathways Program (2018). Retrieved from uwrf.edu

English Web 2013 Corpus (2018). *Sketch Engine: TenTen Corpus Family*. Retrieved from https://www.sketchengine.eu/documentation/tenten-corpora/#toggle-id-4

Fernando, C. (1996). *Idioms and idiomaticity*. Oxford: Oxford University Press.

Ferris, D., & Hedgcock, J. (2014). *Teaching L2 composition: Purpose, process, and practice (3rd ed.)*. New York, NY: Routledge.

Ferris, D., & Tagg, T. (1996). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly, 30(1),* 31.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10(2),* 102-112.

Firsten, R., & Killian, P. (2002). *The ELT grammar book: A teacher-friendly reference guide*. Burlingame, CA: Alta Book Center.

Flowerdew, J. (1994). *Academic listening: Research perspectives*. Cambridge: Cambridge University Press.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41(2),* 339-359.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35(3),* 305-327.

Garnier, M., & Schmitt, N. (2014). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 645-666.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System, 59*, 29-44.

Graves, M. F., August, D., & Mancilla-Martinez, J. (2013). *Teaching vocabulary to English language learners*. New York, NY: Teachers College Press.

Greenbaum, S., & Quirk, R. (1990). *A student's grammar of the English language*. Harlow: Longman.

Hansen, C. (1994). Topic identification in lecture discourse. In J. Flowerdew (Eds.), *Academic listening: Research perspectives* (pp. 131 - 145). Cambridge: Cambridge University Press.

Harada, T. (1998). Mishearings of content words by ESL learners. *The CATESOL Journal, 10(1),* 51-70.

Hou, H. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching, 7(5),* 26-37.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes. *Studies in Second Language Acquisition, 11(3),* 241-55.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41(2),* 235-253.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7-36. Retrieved from http://www.sketchengine.co.uk

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: Univ. of Chicago Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension and production of new words? In Laurén, C., Nordman, Marianne, & European Symposium on Language for Special Purposes. (eds.), *Special language: From humans thinking to thinking machines* / edited by Christer Laurén and Marianne Nordman.

Lee, S. (2007). Revelations from three consecutive studies on extensive reading. *RELC Journal, 38(2),* 150-170

Lems, K., Miller, L. D., & Soro, T. M. (2010). *Teaching reading to English language learners: Insights from linguistics*. New York: Guilford Press.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning, 54(2),* 193-226.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly, 37(4),* 671.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly, 45(4),* 661-688.

Lynch, T. (2011). Academic listening in the 21st century: Reviewing a decade of research. *Journal of English for Academic Purposes, 10(2),* 79-88.

Marcus, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19 (2),* 313-330.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics, 33(3),* 299-320.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Theory, method and practice*. Cambridge: Cambridge University Press.

McPherron, P., & Randolph, P. T. (2014). *Cat got your tongue: Recent research and classroom practices for teaching idioms to English learners around the world*. Alexandria, VA: Tesol Press.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines 5,* 12-25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesi, H. & Thompson, P. (2006) *The British Academic Spoken Corpus Manual*. Retrieved from https://warwick.ac.uk/fac/soc/al/research/collections/base/history/base_manual.pdf

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.

Office of International Research (2016). *University of Wisconsin – River Falls Enrollment Report (CDR Data, Fall 2016)*. Retrieved from https://www.uwrf.edu/Research/upload/Enrollment-Report-by-Acad-Level-4.pdf

Over. (2018). In Merriam-Webster's Dictionary online. Retrieved from http://learnersdictionary.com/definition/over

Paker, T., & Özcan, Y. E. (2017). The effectiveness of using corpus-based materials in vocabulary teaching. *International Journal of Language Academy, 5(1),* 62-81.

Root, C., & Blanchard, Karen Lourie. (2003). *Zero in!: Phrasal verbs in context*. Ann Arbor: The University of Michigan Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Cobb, T. Horst, M., & Schmitt, D. (2017). How Much Vocabulary Is Needed to Use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50(2),* 212-226.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18(1)*, 55-88.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31(4),* 487-512.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37(3),* 419.

Simpson, R., Briggs, S., Ovens, J., & Swales, J. (2002). *Michigan corpus of academic spoken English*. Available from University of Michigan Web site, https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase

Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL - International Review of Applied Linguistics in Language Teaching, 45(2),* 119-139.

Sketch Engine (2018). Feature Corpora. Retrieved from https://the.sketchengine.co.uk/auth/corpora/

Turnitin.com (2018). *Originality check report* Retrieved from https://ev.turnitin.com/app/carta/en_us/?o=939496715&s=1&session-id=089b23add11567eb7f31595ce85a4ba2&lang=en_us&u=1072038308

Van Zeeland, H., & Schmitt, N. (2013). Lexical Coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457-479.

Waring, R., & Nation, P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English Speaking World, 4*, 97-110.

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, 44(3), 263-277.

West, M. (1953). *A general service list of English words*. London: Longmans, Green.

Williams, D. (2016). Teaching collocation through dictionaries and corpus-based resources. *Modern English Teacher, 25(4),* 9.

Wilson, M. (2003). Discovery listening--improving perceptual processing. *ELT Journal*, 57(4), 335-343.

WordNet (2010) *Princeton University about WordNet*. Princeton University. Retrieved from https://wordnet.princeton.edu/

Zhang, M. (2013). The application of corpus tools in the teaching of discipline-specific academic vocabulary. *International Journal of Computer-Assisted Language Learning and Teaching, 3(4),* 33-47.

Appendix A

Ways for subjects to share their media files
How to share your media files

Option 1: Collaboration on Kaltura MediaSpace (preferred)

Step 1: Login to Kaltura Media Space - https://uwrf.mediaspace.kaltura.com/

Step 2: From the *Home* page, find your name in the upper right-hand corner and click it. Next, click *My Media* to display media you have created.

Step 3: Select what you would like to share by clicking *Edit* to the right of the media.

Step 4: From the *Edit* screen, click on *Collaboration*. Next, click on the blue *+Collaborator* button to the bottom-right.

Step 5: Enter *Tyler Theyerl*, into the *Enter a Collaborator* field and select what pops up *w3108545 (Tyler Theyerl)*. Finally, select the *Co-Editor* option and click *Add*.

Option 2: Sharing via OneDrive

Step 1: Login to your UWRF email account and access your OneDrive via the square dropdown menu located in the top-left next to *Office 365*.

Step 2: Select the folders or files you would like to share by clicking the small circle with the checkmark in it to the left of file names (Note: you can only *Share* one folder or one file at a time. The easiest way to share multiple files is to have them in a single folder) Next, click the *Share* button on the top toolbar.

Step 3: Make sure you have selected the *Allow Editing* option in the *Send Link* menu. Enter the email address tyler.theyerl@uwrf.edu into the *Enter a name or email address* field. Finally, click *Send*.

Option 3: Your preferred method

Step 1: Contact me at tyler.theyerl@uwrf.edu or 715.220.8646 to inform me of your preferred method.

Appendix B

Subject Consent Form

Project Title: Idiomatic Phrases in the Spoken Language of the UWRF Online Classroom
Researcher: Tyler Theyerl - English Dept. - TESOL Program – tyler.theyerl@uwrf.edu
715.220.8646 - 132 W. Johnson St. #3 River Falls, WI 54022

1. **Purpose**: The purpose of this study is to identify idiomatic phrases used in recorded media (e.g. video lectures) designed for student consumption to help facilitate communication between instructors and English language learners at UWRF.

2. **Procedure**: Identifying these phrases will be accomplished by transcribing the spoken language in media, such as online lectures created by UWRF instructors, to create a small language database of spoken academic English. Very few of us enjoy our recorded voices. I'd like to stress that the focus of this study is about identifying words and phrases used in recorded media in a general sense. No identifying information about instructors or courses will be shared. Furthermore, no visual elements of any media will be analyzed or shared.

   I'm asking you to participate in this project by sharing media files, such as video lectures, you have already created for your courses. This can be accomplished one of three ways.
   a. Add me, Tyler Theyerl, as a collaborator to media files on your Kaltura MediaSpace account. (once downloaded, you will be contacted to disable access)
   b. Through your Outlook OneDrive account connected to your university email by sharing with tyler.theyerl@uwrf.edu. (once downloaded, you will be contacted to disable access)
   c. If you have a preferred, alternative method for sharing media (e.g. YouTube or flash drive), please let me know.

3. **Time Required**: Your participation involves only the time it takes to share media files.

4. **Risks**: Sharing media for this project involves risks commensurate with sharing media with students in your courses.

5. **Your Rights as a Subject**:
   a. Participation is completely voluntary, and no identifying information about any instructors or courses will shared. Data and results will be presented in aggregate form and will not be released in a way that could identify you.
   b. If you wish to withdraw from the study at any time, you may do so without penalty. Any data you submitted will be destroyed if you so desire.
   c. If you wish, results will be made available to you once the study is completed.

6. If you have concerns about how you were treated in this study, please contact:
   Dianne Bennett, Director of Grants and Research,
   101 North Hall, UWRF, 715/425-3195.

This project has been approved by the UW-River Falls Institutional Research Board for the Protection of Human Subjects, protocol # H2018 – T057.

I have read the above information and willingly consent to participate in this experiment.

Signed: _____ Date: _____

Appendix C

Transcription Guidelines for FISC

General Tips

1. Transcribe all words spoken in the video, including hesitation and fillers (um, you know, etc.).
2. Produce the Google Docs automated transcript first, then focus on details later.
3. When editing a transcript for details, pause the video frequently to make edits to the transcript. Having to move the video back to listen again can become very time-consuming.
4. If the video contains a new word that you are unsure about, use visual information from the video, surrounding context, and the Internet to aid you.
5. Follow the **Transcription Conventions** table below.
6. Because this study focuses on vocabulary, transcribing the vocabulary accurately in terms of spelling and word order is the most important aspect of the transcription.
7. Be consistent within a video and from video to video.
8. If you have any questions, please email me! Tyler.theyerl@uwrf.edu

Transcription Conventions		**Examples**
General	Transcribe all spoken words, even if the speaker misspeaks or is ungrammatical	ain't no problem y'all
General Spelling	Use standard American English spelling for most words, even if they are not fully pronounced, pronounced with an accent, etc.	Speakers says "dis," you transcribe "this" if this is what was meant in your judgement. Speaker says "k," write "okay"
Repairs	Place a hyphen where the pronunciation of a word stops if the speaker repairs it	This is a trans- transcript…
Reduced Forms and Contractions	Transcribe common reduced forms and contractions if they are used by the speaker	wanna, gonna, cuz, kinda, sorta, alright, okay, oops, uh oh, yeah, yup, I'm, there's, etc.
Hesitations and Exclamations	Use these standard hesitation and exclamation sound spellings, regardless of how long the sound is or insertion of other sounds	um, uh, ah, oops, oh, uh oh
Unintelligible Speech	Place the word *unintelligible* in brackets if the speech is not recognizable	I said I would <unintelligible> after…
Names	Anonymize speakers' names by placing the word *name* in brackets	My name is <name>. Today…
Pauses and Punctuation	Place a comma for pauses within thought groups. Place a period for the end of a thought group, usually indicated by a final falling intonation. Use a question mark for rising intonation typical of a question.	It's a simple, pretty simple issue. Let's move now to our next topic, group communication. So what makes a group? *when in doubt, follow writing punctuation conventions

Capitalization	Follow normal written capitalization conventions. Capitalize proper nouns, names, acronyms, and first words after a period.	Pavlov's dog was named Beck. He was a mutt and I think PETA wouldn't have liked the situation.
Numbers, Time, and Websites	Spell out numbers fully as words as the speakers says them. Use lowercase a.m. and p.m. for times. Spell out website names and "dots"	Numbers: three, twenty-five, a hundred and fifty, one-hundred fifty, two thousand. Time: five-thirty p.m. Website/Email: tyler dot theyerl at my dot uwrf dot edu
Metalinguistic Information	Put context information in <brackets>, if absolutely necessary for understanding when reading the transcript	Instructor plays a recording of another voice = <plays audio clip> transcribe what the audio clip says <end audio clip>

Appendix D

FISC Multi-Word Verb List (FISC List)

Items in Order of Raw Occurrences	Number of Raw Occurrences	Number of Speakers	Item Score	Items Ranked by Item Score
1. talk about	96	11	70.40	**1. talk about**
2. look at	64	11	46.93	**2. look at**
3. think about	47	10	31.33	**3. think about**
4. listen to	36	4	9.60	**4. go on**
5. focus on	29	5	9.67	**5. refer to**
6. go on	25	10	16.67	**6. focus on**
7. refer to	25	7	11.67	**7. listen to**
8. get to	19	4	5.07	**8. look for**
9. work with	18	6	7.20	**9. deal with**
10. deal with	17	7	7.93	**10. start with**
11. think of	17	6	6.80	**11. work with**
12. click on	16	3	3.20	**12. go up (to)**
13. look for	16	9	9.60	**13. go back (to)**
14. come up (with)	13	4	3.47	**14. think of**
15. put in/into	13	6	5.20	**15. put in/into**
16. start with	13	9	7.80	**16. depend on**
17. go up (to)	12	9	7.20	**17. get to**
18. go back (to)	12	9	7.20	**18. come to**
19. breed for	11	1	0.73	**19. set up**
20. depend on	11	7	5.13	**20. look like**
21. engage in	11	4	2.93	**21. put on**
22. set up	11	6	4.40	**22. come up (with)**
23. divide into	10	1	0.67	**23. click on**
24. put on	10	6	4.00	**24. get into**
25. come to	9	8	4.80	**25. relate to**
26. get into	9	5	3.00	**26. engage in**
27. look like	9	7	4.20	**27. go down**
28. make up	9	4	2.40	**28. talk to**
29. relate to	9	5	3.00	**29. go into**
30. go down	8	5	2.67	**30. make up**
31. go over	8	4	2.13	**31. go over**
32. go through	8	4	2.13	**32. go through**
33. interact with	8	3	1.60	**33. involve in**
34. involve in	8	4	2.13	**34. figure out**
35. know about	7	3	1.40	**35. show up**
36. result in	8	2	1.07	**36. interact with**
37. talk to	8	5	2.67	**37. look up**
38. advocate for	7	2	0.93	**38. know about**
39. lead to	7	2	0.93	**39. check out**

40. point out	7	2	0.93	**40. begin with**
41. expose to	7	1	0.47	**41. result in**
42. respond to	7	2	0.93	**42. move on**
43. show up	7	4	1.87	**43. start out**
44. begin with	6	3	1.20	**44. take into**
45. contribute to	6	2	0.80	**45. write down**
46. figure out	6	5	2.00	**46. advocate for**
47. go into	6	6	2.40	**47. lead to**
48. look up	6	4	1.60	**48. point out**
49. scroll down	6	1	0.40	**49. respond to**
50. check out	5	4	1.33	**50. contribute to**
51. move on	5	3	1.00	**51. end up (with)**
52. start out	5	3	1.00	**52. breed for**
53. take into	5	3	1.00	**53. divide into**
54. take over	5	1	0.33	**54. work for**
55. work for	5	2	0.67	**55. build up**
56. write down	5	3	1.00	**56. flow through**
57. end up (with)	4	3	0.80	**57. take out**
58. speed up	4	1	0.27	**58. expose to**
59. build up	3	3	0.60	**59. scroll down**
60. check off	3	1	0.20	**60. continue on**
61. continue on	3	2	0.40	**61. get rid (of)**
62. flow through	3	3	0.60	**62. go in**
63. get in	3	1	0.20	**63. look down**
64. get rid (of)	3	2	0.40	**64. wrap up ***
65. go in	3	2	0.40	**65. take over**
66. look down	3	2	0.40	**66. speed up**
67. take out	3	3	0.60	**67. check off**
68. wrap up *	3	2	0.40	**68. get in**

* one occurrence of *wrap up* was mistakenly transcribed as *rap up* in the corpus

Appendix E

27 Items Which Appear on both the FISC List and the PHaVE List

FISC List Ranking (out of 68)	Item	PHaVE List Ranking (out of 150)
55	build up	84
39	**check out**	49
22	**come up (with)**	4
50	**end up (with)**	18
34	**figure out**	21
68	get in	98
13	**go back (to)**	5
27	**go down**	26
62	go in	54
4	**go on**	1
31	go over	74
32	go through	76
12	**go up (to)**	33
63	look down	42
37	**look up**	20
30	**make up**	17
42	**move on**	50
47	**point out**	9
15	put in/into	149
21	put on	87
19	**set up**	11
35	**show up**	27
43	start out	91
44	take in(to)	134
57	take out	24
65	take over	37
44	write down	119

Note: **Bold** items are in the top 50 items on both the FISC List and the PHaVE List