THE USE OF CORPUS AND NETWORK ANALYSIS IN TEACHING ENGINEERING EAP PHRASES

by

Maria Joy Pritchett

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Linguistics

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Felicia Roberts, Chair

Brian Lamb School of Communication

Dr. April Ginther

Department of English

Dr. Natalie Lambert

School of Medicine, Indiana University

Dr. John Sundquist

School of Languages and Cultures

Approved by:

Dr. Alejandro Cuza-Blanco

Head of the Graduate Program

To my grandparents,

Robert & Constance Prince & Darvin & Winifred Cupery.

From my earliest memories to my academic endeavors,

Your faith and love are a constant source of joy.

ACKNOWLEDGMENTS

My committee, chaired by Felicia Roberts, has provided invaluable assistance. As I wove together the project's disparate threads, each of you contributed your wisdom and expertise. I'm especially indebted to Dr. Natalie Lambert for introducing me to many of the tools used in this work, to Dr. April Ginther for connecting me with the PLaCE classes, and to Dr. John Sundquist for overseeing my literature review. Dr. Matthew Allen graciously allowed me to work with his students and provided insights along the way. Most importantly, Dr. Roberts guided and mentored me at every stage of my graduate career and this dissertation, helping me to become a better writer, thinker, communicator, and scholar.

I'm thankful for the many opportunities afforded by the Purdue Linguistics Program: the chance to teach Introduction to Linguistics, benefiting from Dr. Elaine Francis' example; the community and support provided by my friends in the Purdue Linguistics Association; the chance to present and develop my work at our Ling Lunches and annual conference; and the writing group that helped me through the final stages of dissertating.

I'm grateful to my communities at Covenant EPC and World Welcome for both encouraging me in my endeavors and keeping my faith at the center. Thanks to the pastors and laypeople who make Covenant the place I went for rest and refreshment, truth and direction, when academics were hard. And thanks to my international friends who volunteered to give feedback on my materials, share their perspectives, and help me think like a STEM grad student, not a linguist. To Lily, Becca & Serpil: thanks for sharing the PhD journey. As we've dissertated, we've also thrown parties, crossed disasters, prayed, grown as scholars, visited, shopped, cooked together, and spent hours in coffee shops, making work enjoyable.

And finally, family. Robert, I finished this primarily due to a husband who always took the time to encourage me and applaud my growing page count, even as you wrote your own dissertation. I can't wait to share the title of Dr. Pritchett with you for the rest of our lives. Mom and Dad, your example and wisdom kept me grounded as I went from fresh-faced M.A. student to a triumphant Doctor. Mom and Dad Pritchett & Sam, thanks for lovingly welcoming me (and my dissertation)

into your wonderful family. Lydia and Hanna, you stayed in touch when I was lost in my work, and you made me laugh and you made me think. Cupery family, there are too many of you to name, but my times with each of you were refreshing and humbling, as you each value character over accomplishments and demonstrate that in your lives and your questions to me.

And my dear grandparents: there is some of each of you in me, in my stubbornness, loyalty, curiosity, work ethic, faith, humor, and love of family. I leave it to you to decide who contributed what, but I couldn't have done this without the pieces of you in me and without your encouragement, support and examples.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation is composed of three interlinked studies that pilot new methods for combining corpus linguistics and semantic network analysis (SNA) to understand and teach academic language. Findings indicate that this approach leads to a deeper understanding of technical writing and offers an exciting new avenue for writing curriculum.

The first phase is a corpus study of fixed and variable formulaic language (n-grams and p-frames) in academic engineering writing. The results were analyzed functionally, semantically and rhetorically. While previous n-gram analyses highlighted how engineering writing relies on text-oriented phrases (Hyland 2008a), the p-frame analysis found that variable phrases are often participant-oriented and communicate author stance. The p-frames also demonstrated that prepositional phrase and passive verb constructions were key structures for author stance.

The second phase combined corpus and network analysis tools to create educational materials. Several elements of successful design were highlighted, including how to best combine corpus and SNA tools, the role of a linguistically knowledgeable designer, and the creation of a framework that can visualize rich insights into the rhetorical, semantic, and syntactic nuances of formulaic language.

Given the complexity of engineering writing, students need clear materials that highlight accessible findings and allow them to practice and master formulaic language. Thus, the final phase tested the materials in two classes with fifteen graduate students, finding evidence for the value of this novel approach. The major benefits were that students learned synonyms for overused items; experimented with fresh terms while practicing common syntactic structures and moves; and developed skills in identifying and employing author stance.

CHAPTER ONE: BACKROUND AND RESEARCH QUESTIONS

A Challenge and an Opportunity

Professional academic writing is an important skill and hard to master. One of the hardest areas of academic writing for non-native speakers (NNS) is formulaic language, as this is often a consistent gap between native-speaker (NS) and NNS writing even after NNS writers have acquired syntax and vocabulary at high levels. Many corpus linguistics studies of NNS academic writing and formulaic language analyze this gap, but fail to offer successful pedagogical approaches to addressing it. While the findings of corpus studies of academic writing is full of rich data that can help students improve their own writing, giving the data directly to students seems to only heighten their awareness of the existence of formulaic language rather than improve their writing (see Jones & Haywood, 2004; Cortes, 2006).

However, by drawing on a tool outside of corpus linguistics, it may be possible to present and teach formulaic language in a way that makes it accessible to NNS students and applicable to their writing. This tool is semantic network analysis, an approach to language that seeks to represent the relationships that corpus linguistics can uncover in meaningful ways through knowledge visualization. Teaching formulaic language through visualizations may make the data more accessible to students and help them acquire authentic knowledge on how to use the phrases.

Before considering how this might be done, however, we will review the work done so far and the terminology that will be used in this project. The first section below will look at corpus linguistics approaches to analyzing academic writing; the second section will survey what corpus has shown about academic writing in general and formulaic language in particular; the third section will introduce the field of knowledge visualization and its potential benefits when paired with semantic network analysis; and the fourth section will explain why engineering writing is an ideal case study for combining the tools of corpus linguistics and knowledge visualization to create pedagogical materials. These literature reviews lead to the research questions in the final section.

Corpus Linguistics Methods for Analyzing Academic Writing

Before discussing what corpus linguistics has discovered about the features of academic writing and the pedagogical uses of those results, it is important to understand the methodologies that produce those results. Corpus linguistics has the intriguing ability to pull out patterns from large corpora and using these patterns in writing education can make teaching English more targeted and efficient. However, the findings must be accurate if they are to be useful to learners. The language patterns identified will vary largely depending on how the corpus is constructed, what strategies are used to search for the formulaic language, how the results are categorized, and how much attention is paid to general trends versus specific genres and disciplines.

There is a multitude of corpus analyses of English for Academic Purposes (EAP) corpora that provide models of how to extract meaningful information from large bodies of text. EAP analyses often start by searching for "collocations" – combinations of words that appear together more frequently than expected by chance (Greaves & Warren, 2010). One of the first approaches to unearthing strings longer than two words in length was to extend collocations to find "accumulated collocations" (Hunston, 2010). In this method, collocations are iteratively entered into the search bar to find their most frequent collocates, and continually extended until there are no more common collocates. For example, "the other" could be collocated first with "hand," searched for again, and collocated with "on," to uncover the formulaic sequence, "on the other hand." However, this process is both tedious and leaves no flexibility for inserted lexemes. For example, a search for the phrase 'with our gratitude' will miss occurrences such as 'with our deep gratitude.'

As computing capabilities increased, it became possible to search for 'n-grams,' i.e., *n* number of words that occur together at frequencies above a set cut-off. A recent example of this comes from Simpson-Vlach & Ellis (2010). To find common phrases in an academic English corpus, they searched for 3-, 4-, and 5-grams. Because each sequence length has to be individually searched, this method requires the researcher to manually eliminate the overlapping results (for example, "on the other" and "the other hand" will appear in the 3-gram results, and these will overlap with the occurrences of "on the other hand" in the 4-gram results). One further step was necessary; as with collocation studies, Simpson-Vlach & Ellis believed that an MI (mutual

information) score is a better indicator of internal cohesiveness than raw frequency. MI scores indicate how improbable two words' co-occurrence is, given the frequency of those individual words in the corpus. When teachers were asked to rate the sequences according to their worthiness for teaching, the sequences' MI scores were a much better predictor of a higher teacher rating than their frequency rates (Simpson-Vlach & Ellis, 2010). Thus, to uncover meaningful formulaic sequences, n-gram frequencies must be supplemented with or weighted according to MI scores.

However, the ability to easily pull out n-grams from the corpus does not mean that all patterns have been identified; some formulaic language is in the form of a set sequence with variable slots. Fischer-Starcke's (2012) "p-frames" extend the methodology of n-grams to find formulaic sequences that may be interrupted by inserted words. "P-frames" are phrases of *p* words, where one of the words either varies or is completely absent. It is possible to automate finding these phrases by asking a computer program to identify n-grams that include a wildcard (*) at any of the phrase's slots. This not only allows for non-contiguous sequences to be identified, but also allows more accurate counts, as sequences with inserted or variable words are not lost. For example, a p-frame search would combine the results for "pulling [his/her/my/your/its/our/ their] leg," rather than giving a separate count for each version that occurred frequently enough to make it over the cutoff frequency. Thus, the combination of n-gram and p-frame approaches allow the researcher to extract the most frequent and highly dispersed lexical patterns from the text.

One methodological weakness with the above approaches is that a phrase with a strong frequency and/or MI score may occur in a wide range of texts in a corpus, or it may occur predominantly in only one subsection or text in the corpus, and the search cannot distinguish between these. This has led to the creation of another score for determining a phrase's usefulness: the dispersion score. The dispersion score measures how dispersed the phrase is in the corpus, which can indicate whether it is general enough to be worth teaching, or if it too specific and likely to be useful only in a small percentage of texts. If a term or phrase appears frequently in only one subsection of a corpus, it will likely be defined in that section (Miller &

Biber, 2015), making it unnecessary for students to learn it in advance, whereas the most dispersed items are also the most widely useful ones.

As a result, when creating EAP vocabulary lists, the most common corrective is to weight frequency scores with dispersion scores. However, there are multiple ways to compute a dispersion score. The most common is Julliand's *D* (Juilland & Chang-Rodriguez, 1964), developed for dispersion-weighted word frequency lists for Spanish and French. It is the most widely used measure of lexical dispersion, employed in many frequency dictionaries, like Davies & Gardner's (2010) description of contemporary English. But as it was created for a corpus broken into a few equally sized parts, recent research indicates that it is not fit for corpora divided into many differently sized subsections. In fact, while the results of the statistic are supposed to range from 0 (tightly congregated) to 1 (dispersed) the actual numbers mostly fall between .6 to .99 (Biber, Reppen, Shnur & Ghanem, 2016). Thus, Biber et al. recommend replacing Julliard's *D* with Gries' *DP* (Gries 2013), the sum of differences in absolute expected proportions divided by the observed proportions for each corpus sub-section. Another approach is to simply look at the percentage of texts in a corpus in which a word appears. Whichever method is employed, the consensus is that frequency should be weighted with dispersion to identify the phrases or words most useful for students to learn.

When the aim of corpus work is to identify language patterns for novice learners, there are two ways of using the measures described above. Corpus studies can either identify differences between a student and target corpora (for example, NNS MA theses with NS MA theses, or MA theses with PhD theses); or can seek to accurately characterize a corpus of target writings for the benefit of novice writers. A target corpus is any collection of writings that represent a style of writing that learners wish to master. Rather than focusing on the differences between their current writing and target writing, learners instead observe language patterns in the target corpus and integrate them into their writing (Sripicharn, 2010). The benefits to this approach include apprentice-style learning from master writing instead of "fixing" existing writing; rich information on the context and variation of the language patterns observed; and the learner's active involvement in choosing and implementing language patterns.

Thus, the field of corpus linguistics has much interesting data to offer students, and it also has tested tools and methods for extracting the data. However, as we shall see in the following review of corpus findings on academic writing, the results that corpus linguistics produces are often in the form of pages and pages of texts – data charts, percentages, and, of course, lines of corpus examples (called "concordance lines"). While this information has the potential to benefit students, it is not in a format that is either attractive or easily translated into writing practice. This is where the research in knowledge visualization offers practical insight and solutions.

Features and Challenges of Academic Writing

Corpus analysis of English academic writing has identified the many ways in which it differs from both academic verbal speech and other written genres. Academic writing refers to the variety of texts that are produced in a higher education setting and in the realm of research: master's theses, doctoral dissertations, textbooks, classroom materials, abstracts, research articles and academic papers (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber, Conrad, Reppen, Byrd, & Helt, 2002; Cava, 2011, Coxhead, 2000; Parodi, 2009). This dissertation will focus on the genre of research articles, as for graduate students and other apprentice writers, this genre represents the goal of the other types of academic writing. Peer-reviewed publications indicate a scholar's contributions to the scholarly community and validate the quality and applicability of their work. And the language of most prestigious journals is predominantly English. Thus, both native and non-native speakers are highly motivated to master this genre (Biber et al., 2002; Flowerdew, 2004; Flowerdew, 2012).

Academic research writing is also not a skill that is easily acquired, as it is a compressed, information-dense genre that has several unique syntactic and structural features. Biber & Gray (2010) sum up these difficulties after a syntactic analysis of a 3-million-word corpus of academic research articles covering a spectrum of disciplines. They conclude that English academic writing emphasizes structural compression through the syntactic devices of attributive adjectives, pre-modifying nouns, and adjectival and adverbial prepositional phrases. This is in direct contrast to spoken academic discourse, where elaborated structures (such as finite complement clause and adverbial clauses) appear at much higher rates. These structures lead to higher information density, and to an increase in implicit meanings: the preferred syntactic

constructions in academic writing require the reader's contextual knowledge to decode the correct relationship, unlike spoken academic discourse, where the relationships between words are more explicit. This makes the discourse less accessible, as "students lack the specialist knowledge that would allow them to readily infer the expected meaning of compact, inexplicit constructions" (Biber & Gray, 2010, p. 17). Successful NNS graduate writers increase their use of these implicit, compressed information delivery structures as they advance in their programs (Parkinson & Musgrave, 2014).

Academic writing has other unique linguistic features, such as its vocabulary and its structured rhetorical moves. Coxhead (2000) provides a meticulously researched "academic word list," a set of 570 word families that appeared with significantly higher frequency in a multi-disciplinary corpus of academic writing than in a corpus of general English. Research articles are also known for having discipline-specific genre moves and set structures, which have been analyzed and described by researchers (comprehensively outlined in Swales (1990); see Nwogu, (1997), Skelton (1994), and Maswana, Kanamaru, & Tajino (2015) for specific examples) and codified in books and manuals for novice writers (for example, Swales and Feak's *Academic Writing for Graduate Students*). These lexical and structural features of academic writing are both well-documented and explicitly taught to novice academic writers. Building on this previous research, but extending it beyond lexical and structural features, we will turn our attention to a less-taught linguistic feature that is essential to academic writing. It will be the primary linguistic feature of academic writing considered in this study: formulaic language.

Formulaic Language in Academic Writing

One particularly challenging aspect of academic writing that makes it difficult for NS and NNS writers alike is its dependence on formulaic structures. While vocabulary can often be memorized, and structural norms acquired, the segments of fixed and variable language that make up large portions of academic writing are difficult for both NNS (Jones & Haywood, 2004; Coxhead, 2008) and NS (Cortes, 2006) students to consciously acquire.

Formulaic language has been defined in a variety of ways, but primarily as reoccurring patterns in language use (Wray, 2012). This broad category includes many more specific items; the

definitions tend to depend on the research methodology used to study formulaic sequences. "Collocations" and "lexical bundles" refer to any group of words that appear together in unusually high frequencies, and thus are determined using corpus linguistics methods (Hyland, 2008a). "Collocations" is the older term and can also refer to words that frequently appear together within a window of chosen length (O'Keeffe & McCarthy, 2010). "Idioms," however, are defined as fixed multiword constructions that are institutionalized (conventionalized) and whose meaning is more than the sum of their parts (Fernando, 1996); these are usually more quickly recognized by native speakers, and as they cannot easily be identified through corpus methods, they are usually gathered and selected by human analysis (Simpson & Mendis, 2003).

Academic writing "draws on a much larger stock of prefabricated phrases than either news or fiction" (Hyland, 2008b), and contains a multitude of unique structural patterns (Hyland & Tse, 2005), lexical bundles of varying lengths (Biber & Conrad, 1999) and discipline-specific idioms (Conrad, 2004). These formulaic sequences also vary widely in terms of quantity, type and rhetorical use across the academic disciplines (Hyland 2008b; Conrad & Biber, 2004; Oakey, 2002; Samraj, 2002; Bloch, 2010).

The use of formulaic language is not unique to written discourse, as speakers make use of these sequences in every form of discourse. For example, the studies on the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) emphasize that lexical bundles are less common in written language than in spoken language. Biber & Barbieri (2007) find that lexical bundles are more common in written course management (e.g. course syllabi) and classroom lectures than in academic written registers and, Biber, Conrad & Cortes (2004) find lexical bundles to occur much more frequently in classroom teaching than in textbooks.

However, studies on written academic writing stress the frequency with which formulaic sequences occur and their distinctness from the sequences in written speech; these are not sequences that even a fluent English speaker uses frequently when speaking. Hyland (2012) found that the text-oriented sequences that structure discourse occur much more often in academic writing than in other writing types. He broke the frequencies down by academic genre, and found the highest percentages are in engineering texts, where four-word bundles accounted

for 3.5% of the total words. Similarly, Biber et al. (1999), in their study of an academic corpus, found that three- and four-word lexical bundles occurred at least 65,000 times per million words. Simpson-Vlach & Ellis (2010), who were concerned with how to best teach formulaic sequences, identified a core group of over 200 sequences that occurred frequently in written and spoken academic English. While many of these results come from specially constructed academic corpora, a study on the largest academic corpus available, the academic sub-sections of the British National Corpus and the Corpus of Contemporary American English, similarly asserted the prevalence and importance of formulaic sequence to academic writing (Liu, 2011). Shariari (2017) looked at the spread of formulaic sequences across research article sections and found that the results section was the most densely formulaic, and that each section varied in the rhetorical types of bundles preferred.

Thus, the consensus is that formulaic sequences are an important and frequent feature of academic discourse, and that they differ substantially from the formulaic language used in other genres. Many of these large-scale studies conclude with an appendix of a list of all the formulaic sequences found above a certain frequency. However, there are several substantial challenges that face novice writers seeking to master their use. The following three sections lay out three of these challenges: the variation in formulaic language usage between disciplines; the documented difficulties for NNS writers; and the lack of successful teaching strategies for this area of language fluency.

Variation Across Disciplines

One of the roadblocks to acquisition is that different disciplines, and even sub-disciplines, vary widely on their preferred formulaic chunks and their frequency and rhetorical uses. The best overview of this topic comes from Hyland (2008b). He compares the use of four-word lexical bundles in four distinct disciplines: electrical engineering, microbiology, applied linguistics and business studies. He uses the standard frequency cutoff of 20 occurrences of the bundle per million words. His primary finding is that discipline matters; more than half of the top 50 bundles in each discipline are unique to that discipline, and no more than 30% of the sequences in any discipline are found in two of the other fields. The fields also differ in the functions of these bundles: science and engineering texts employ the most research-oriented bundles, while

the linguistics and business texts use the most text-oriented and participant-oriented bundles. Research-oriented bundles are those that explain experiment design, text-oriented bundles are those used to structure a text, and participant-oriented bundles either directly address a reader or reveal the author's opinions and attitudes. Hyland (2012) uses the same corpus and demonstrates that the frequency of formulaic language also changes dramatically; though the engineering and biology texts were similar in bundle functions, the engineering texts used bundles at almost twice the rate as the biology texts.

Two granular examples of this variation come from Oakey (2002) and Bloch (2010). Oakey studies the phrase "it has been/is (often) + reference verb + that." While he primarily examines its many rhetorical functions, he notes that medical writing prefers the present perfect, while social and technical sciences prefer simple present. Bloch (2010) finds that sciences and engineering articles are more likely to use non-integral quotes, while humanities and social sciences employ more integral quotes. Because many of these patterns are a matter of preference, or small syntactic variation, it is unlikely that the proficient writers in the field will feel the need to explicitly teach these variations to new writers, or that writing instructors who teach a wide variety of disciplines will know of these differences. This is where corpus linguistics is helpful: by aggregating many examples of writing across disciplines, corpus tools can identify what patterns are unique to each discipline.

Challenges for Non-Native Speakers

While there are successful approaches for teaching the vocabulary and grammar necessary for second language academic writing (Coxhead, 2000; Hinkel, 2004), NNS consistently struggle with incorporating formulaic language in their academic writing (Wray, 2002). Multiple studies have shown that NNS tend to have a small stock of formulaic sequences that they overuse, and that they are less likely to be aware of register differences and appropriate contexts when using formulaic sequences (Wray, 2012).

Most corpus studies of this phenomena compare NNS' use of formulaic sequences with reference corpora of native speakers writing in the same genre; these studies map out how NNS handle the challenge of formulaic writing as they advance through undergraduate and graduate

writing and begin publishing. Undergraduate students tend to overuse certain idiosyncratic idioms but underuse them overall (Lee & Chen, 2009). They avoid stance bundles, such as hedging and anticipatory *it* clauses (Chen & Baker, 2010), a trend that will remain persistent across time. On average, only 22% of their commonly used phrases overlap with those of native speakers (Adel & Erman, 2012); their underused sequences include phrases with negatives, hedging, and complex grammatical structures such as the anticipatory *it,* unattended *this* (e.g. "this would suggest,"), and existential *there* (Lee & Chen, 2009).

Graduate NNS writers increase in their use of sequences to the point of overusing them, and their sequences, while not necessarily incorrect, differ from the ones preferred by native speakers. MA & PhD dissertations by NNS writers often have an overabundance of lexical bundles. For example, Öztürk & Köse (2016) compare MA & PhD theses by Turkish NNS writing in English with MA & PhD theses by NS writers. The Turkish writers use twice as many types of lexical bundles as the native speakers and use them at much higher frequencies. Öztürk & Köse examined many of the phrases unique to the Turkish authors and identified them as literal translations of academic Turkish phrases. This overabundance of idioms in post-graduate writing is not limited to Turkish; Wei & Lei (2011) also found it in Chinese NNS writers, and they suggest that Chinese writers' relative underuse of anticipatory *it* and other participant-oriented bundles is because these writers are taught to value impersonality in academic writing.

This overuse settles into a type of hybrid fluency in published academic writing, but still shows marked gaps. Perez-Llantada (2014) found that the overlap between frequent NS and NNS phrases had risen to 31% in published writing, and that many of the phrases unique to NNS were existing English formulaic sequences that they had incorporated at unusually high rates. They use idiomatic English phrases, with minor evidence of cross-linguistic transfer, but they use them at different frequencies. As in the other studies, the NSS still avoided anticipatory *it* clauses and evaluative or probabilistic hedging. Similarly, even experienced NNS writers in English-language telecommunications journals had not adopted the clauses and phrases that were indicative of expert L1 writers (Pan, Reppen & Biber, 2016).

One key area of difficulty for students that appears across studies is that of author stance. According to Biber, Conrad & Cortes, stance phrases are those that "provide a frame that is used for the interpretation of the following proposition," (2003, p 81) and convey the author's attitude or assessment of a proposition (Biber & Barbieri, 2007). Because academic writing in general and engineering writing in particular stress objective authorial voice, several studies comparing NS and NNS academic writing have found that NNS writers underuse stance bundles or do not use them with the same connotations as NS writers (Bloch, 2010, Chen & Baker, 2010) The existential *there* and anticipatory *it* bundles mentioned above provide examples of author stance, as they are often used to provide the author's assessment (with phrases like "it is important to…" or "there is concern that…").

Importantly, these findings do not mean that NNS do not know the same formulaic sequences as the NS writers. Analysis of student bundle use on oral and written portions of the TOEFL exam found that students used formulaic sequences much more frequently in speech than in writing (Staples, Egbert, Biber & McClair, 2013). This indicates that NNS may understand and know these phrases, but that they are unsure how to idiomatically deploy them, especially the participant-oriented phrases whose rhetorical uses are useful when writing academically. Thus, there is room for development, even at higher proficiency levels.

Lack of Successful Pedagogical Approaches

Many of these studies conclude by listing the grammatical or semantic categories that cause NSS writers the most difficulties. This is an essential first step in determining what could be taught; however, fewer studies test effective strategies for helping NNS writers improve in using the formulaic language they avoid. For example, the studies cited show that NNS writers routinely struggle with anticipatory *it* structures (Chen & Baker, 2010; Adel & Erman, 2012; Wei & Lei, 2011; Perez-Llantada, 2014) and evaluative or probabilistic hedging (Perez-Llantada, 2014; Chen & Baker, 2010). But naming the issues has not led to clear solutions.

Studies that focus on teaching formulaic sequences often report low success rates. For example, Cortes (2006) and Jones & Haywood (2004) both explicitly taught students lists of collocations specific to their disciplines, but found that while the students reported raised awareness after the

lessons, they did not use more formulaic sequences in their writing. Li & Schmitt (2009) intensively studied one Chinese master's student's writing over the course of a year, and, though they saw much improvement, they also saw that it was nearly impossible for the student to correct the unidiomatic formulaic sequence usages that had become ingrained in her writing, even with teacher feedback. Pedagogical approaches that teach students to construct and analyze their own corpora and compare their writing with existing corpora of academic writing have been much more successful, but these classes often experience high dropout rates because of the difficulty of the tasks and the time it takes to arrive at the applicable results (Lee & Swales, 2006).

However, several pedagogical successes exist; the best incorporate in-depth review of formulaic language in the students' particular field of study (Eriksson, 2012, Lee & Swales, 2006) and uncontrolled production, where the students implement the formulaic language in their own professional writing (AlHassan & Wood, 2015). Students in technical fields are especially receptive of and interested in pedagogical approaches that are backed with the data a corpus study can provide (Lee & Swales, 2006).

Knowledge Visualization

There is a wealth of information on formulaic sequences available through corpus tools that can benefit NNS and novice writers; but for this information to be useful, it must be effectively communicated. This is where knowledge visualization – specifically through the tools of semantic network visualization – can create the necessary bridge between academic studies and student acquisition.

Potential Benefits and Application to Academic Writing Materials

Many studies show that proper visualization of educational content improves both initial knowledge acquisition and long-term recall. A review of 48 experimental studies found that illustrated texts led to better knowledge retrieval than non-illustrated texts (Levie & Lentz, 1982). Using both text and visual modalities to enforce a concept results in the richest mental representations of that concept (Baggett, 1989). Small adjustments in visualization, such as line

thickness or using similar models to indicate similar problems, can help the audience make intuitive comparisons (Winn, 1989; Gick 1989). Finally, as time between the initial intake of knowledge and testing of its retention increase, visualized knowledge is consistently recalled more easily than textual knowledge (Peeck, 1989). These benefits hold across fields; for instance, the visual exercise of text mapping increased reader comprehension (Geva, 1983), while spatial planning disputes can be solved faster when the plans are visualized (Swaab, Postmes, Neijens, Kiers, & Dumai, 2002), and business students explained novel articles more fluently after visualizing the contents than they did after traditional note taking (Mento, Martinelli, & Jones, 1999). Learners benefit the most when the visual information and the written information are contiguous and complement each other (Mayer et al., 1996) and those with low prior knowledge of the subject material benefit more from visualizations (Mayer and Gallini, 1990). Given the complexity and many variables that arise from combining text with visualizations and measuring its impact on the reader, Salomon (1989) finds that there is not clear theory on the effects of visual stimuli on performance, but posits that successful visuals "overtly model (that is – supplant)" (p.77) the imagery readers may create as they read, or creates that imagery when the reader is unable to do so. One final benefit is that student interaction with effective visualizations improves student ownership of the material (Mondal, Mondal & Das, 2016; Mandl & Levin, Mento, Martinelli & Jones, 1999).

Eppler & Burkhard (2005) sum up the benefits of knowledge visualization in three points: it eases the transfer of knowledge between individuals with different backgrounds, it inspires the creation of new knowledge, and it decreases information overload. These three benefits can potentially address the documented difficulties in teaching formulaic language to NNS writers. The visual aspect forces the materials creator to carefully present and synthesize their knowledge, putting it in a form that is accessible to students who come with a very different perspective and might have trouble noticing those patterns in the raw data. Illustrations that show where formulaic sequences occur in academic writing and the words around them will allow students to visually conceptualize abstract relationships and the role of variable slots in p-frames. Clear diagrams can also help the students to take ownership of the materials and employ the information in their writings in new and creative ways. Finally, most linguistic studies provide a plethora of tiny details and variations that can overwhelm students, so a good visualization will

make the information more intuitively approachable by guiding students to concentrate on the most essential elements.

While the benefits of knowledge visualization seem tailored to this situation, there has been little work relating the two fields. AlHassan & Wood (2015), which has some of the most effective teaching materials based on corpus work to date, still present the information in tables and lists. There have been a few attempts at interpreting other types of corpus research via visualization technology; as summarized in Rayson & Mariani (2009), this includes word clouds and 3D word networks. These techniques have been attempted in discourse analysis (Baker & McEnery, 2015) and historical linguistics (Brezina, McEnery & Wattam, 2015), but not, to my knowledge, in materials development. Thus, we will next turn to the question of what tools for knowledge visualization of corpus findings are available and how they might be used to enable NNS writers to more easily and thoroughly master formulaic academic language.

Corpus and Semantic Network Analysis Tools for Knowledge Visualization

Currently, the primary software packages in which many corpus studies are conducted are AntConc (Anthony, 2018a) and WordSmith Tools (Scott, 2018). Each package can comb through corpora to identify formulaic language and report frequency and dispersion. Each also comes with helpful, elementary visualization tools. AntConc will show the dispersion of a lexeme or phrase within the corpus on a bar graph; this gives an immediate visual clue as to whether it is rare or common, densely gathered or well-dispersed. In addition, the Keyword in Context (KWIC) window in AntConc allows the researcher to scan over many instances of the lexeme or phrase being searched, with common collocates color-coded. The collocation and n-gram search results display simple Excel-compatible charts that can be rearranged by frequency, strength, position in the line, etc. WordSmith Tools has slightly more thorough versions of the same features. These simple visualizations have been used to introduce students to corpus results (Gabrielatos, 2005), but to effectively communicate results, we need software designed for that task.

There are a variety of software that work to identify and visualize the overall topics of and thematic differences between corpora (Kessler, 2017; Singh, Zerr & Siersdorfer, 2017); but

formulaic language requires a fine-grained approach that both identifies individual phrases and aggregates information about their surroundings and typical uses. An interesting and currently underutilized solution is to use graph theory and its visualizations to help linguists interpret patterns in fine-grained language use.

Traditionally, corpus linguistics software has excelled at extracting numerical data and patterns from a corpus, regardless of where the patterns occur. As evidenced by Baker (2016), the next step is a 3D graph-based visualization of the contexts of phrases and patterns in their uses. This is where semantic network analysis (SNA) has potential as an approach to visualization. SNA is a branch of text mining that depicts the relationship between concepts in a text in a network (Doerfel, 1998). While SNA has not been used widely in linguistics research, it is a logical next step to visualize the connections between formulaic sequences and thus better understand their context and relationships. SNA has traditionally used the same graph theory mentioned in Baker (2016) to map the relationships between ideas or social agents (called "concepts") in a text (Diesner & Carley, 2004). With minor adjustments, SNA tools should be able to map the relationships between phrases as easily as the traditional "concepts." There have been efforts to create a corpus tool that uses graph theory to visualize results (see GraphColl in Brezina, McEnery & Wattam (2015)), but as a new tool, its applications are limited.

It is more useful to consider the SNA tools AutoMap (Carley, Columbus & Landwehr, 2013) and NodeXL (Smith et al., 2010). While AutoMap is designed to map concept relationships (as described in the AutoMap guide by Tanenbaum & Brand, 2008), it can also map how any selected items in a text are linked, and what words show up frequently before or after these items. AutoMap goes beyond GraphColl, as it includes all the formulae & visualization strategies of SNA, and thus allows researchers more flexibility in identifying phrases and depicting the type and strength of their relationships. However, AutoMap is best at extracting relationships; and importantly, it can compile that information in a form that is conducive for use in other visualization tools.

Some visualization tools offer slick visuals but little control for the users (such as InfraNodus and texttexure, both developed by Nodus Labs (Paranyushkin, 2011)); others are more concerned

with uncovering social structures than the connections between lexemes, and thus, while effective, they do not take into account data that is relevant from a linguist's perspective; one example of this is VIS, a tool funded by the International Press Institute. One of the more promising tools is NodeXL (Smith et al., 2010). NodeXL is a network visualization tool that gives the user total control over all aspects of the visualization, from the shape of the graph to the color of the nodes. While it is not as visually sleek as some of the other tools, the ability to adjust the information in the graph and personalize visualizations makes it possible to create educational materials where the creator can highlight the parts of the data that are most relevant and interesting to language learners.

Engineering as a Case Study for EAP and Knowledge Visualization

The previous sections have highlighted the challenges NNS face in acquiring academic formulaic sequences, and the potential benefits that a well-visualized corpus analysis of formulaic language can provide. While the proposed methodology should be useful for students working in any field that requires academic writing, it is necessary to choose a specialized discipline to make meaningful and useful materials. As indicated in the section on interdisciplinary variation in formulaic language, there is a great deal of variety between different disciplines in the language patterns they use to communicate their research. Thus, testing the usefulness of combining corpus results with visualization techniques requires a specific field of study.

For this project, engineering academic writing was selected as the discipline best suited to investigation, with NNS engineering graduate students selected as the student population for the materials. There were two reasons for the choice of engineering: the research context of the project and the previous research on engineering writing. This research project was conducted at Purdue University, a Midwestern R1 university known for its engineering programs and history in engineering innovation. The teaching materials will be used with Purdue international graduate students, and 48% of them (nearly 2000 students) are in the College of Engineering, along with an additional 445 international scholars, researchers and faculty in engineering (Purdue Office of International Students & Scholars, 2017).

At the same time, engineering disciplines are underrepresented in the research on academic writing, thus setting up engineering writing as an interesting and useful area of inquiry. The multi-word constructions currently taught in English for Academic Purposes classes are often not representative of the constructions found in engineering writing (Wood & Appel, 2014). Yet, formulaic language occurs frequently in engineering texts. The highest percentages of lexical bundles have been found in engineering texts, where four-word bundles accounted for 3.5% of the total words (Hyland, 2012). As a subset of the research on formulaic language has been conducted on published articles in engineering (Rozycki & Johnson, 2013; Bloch 2010, and Hyland, 2008a, 2008b, and Hyland & Tse, 2005 for electrical engineering specifically), those studies provide starting points and benchmarks to which this study's results can be compared. Give the research on engineering writing and the abundance of graduate engineering students at Purdue, it is a clear choice. Nevertheless, engineering is also a very large discipline; the methodological details in Chapter 2 will explain what branches of engineering were selected for this study and why.

NNS graduate students were chosen as the student demographic because they are most motivated to improve their academic writing and are at a stage when explicit instruction is helpful. Undergraduate students often go into professional fields; at Purdue, only 22% of the 2018 engineering graduates reported continuing education after graduating, while 74% were either employed or seeking employment ("May 2018 First Destination Survey Outcome Report," n.d.). Thus, they are less motivated to improve their academic writing abilities. For the graduate students, however, learning to write academically is essential, as they are usually expected to write research articles and often do not have much background or training in effective academic writing (Campbell & Kennell, 2018). NNS writers especially benefit from explicit instruction, as they are often dealing with both the language barrier and an unfamiliar writing genre (Carter, 2017); the corpus studies above detailed some of specific language difficulties NNS writers face when writing. Thus, explicit corpus-based instruction on academic writing seems most useful for NNS graduate students.

Research Questions

Given the findings from current literature and available tools and methodologies outlined in the sections above, we can now formulate the main questions that this investigation seeks to answer:

(1) What formulaic language is uncovered through a corpus study of engineering research articles?

(a) Do these results agree with or deviate from previous studies of academic writing in general and engineering writing in particular (as summarized in the previous sections)?

(b) Which formulaic sequences should be taught to NNS students seeking to develop their academic writing?

(2) When using SNA tools to design educational corpus-based visualizations of English phrases for NNS learners:

(a) What is the optimal framework for combining corpus results and visualization technology to create accurate and effective materials?

(b) Can SNA programs (particularly AutoMap and NodeXL) create effective, student-friendly visualizations of these corpus findings?

(c) What benefits and challenges are revealed through designer and NNS student feedback on the materials?

(3) Can these visualizations lead to better teaching outcomes than the results documented in Cortes (2006) and Jones & Haywood (2004)?

(a) Will students show improvement in the variety and idiomaticity of their formulaic language from a pretest to a posttest?

(b) Will a delayed posttest confirm the posttest results?

CHAPTER TWO: METHODOLOGY

Introduction: A Three-Phase Project

Given the interdisciplinary nature of the research questions, the methodology for this project involves linking three separate phases with unique methodologies. The first question will be answered through a corpus investigation; the second question through developing a framework for materials creation with corpus and network visualization tools; and the third question through testing the materials in a classroom setting. The following sections outline these three phases of the methodology.

Phase One: Corpus Investigation

Addressing the first research question through a corpus investigation required four steps: corpus creation, corpus pre-processing, identifying formulaic language in the corpus, and categorizing the formulaic language in meaningful ways. The following sections will lay out both the methodology of these four steps and the literature and logic that gave rise to that methodology.

Corpus Creation

The corpus used in this study is a small (almost 2 million word) corpus of published engineering research articles. The following section will lay out the argument for this type of corpus, and the next section will describe the steps involved in constructing the corpus.

The Specialized EAP Corpus

The first methodological question in selecting a corpus was the choice between making use of a large pre-existing corpus of academic writing or developing a new, specialized corpus. Traditionally, some corpus studies have used massive corpora like the British National Corpus (BNC), the Lancaster-Oslo-Bergen Corpus (LOB), and the Corpus of Contemporary American English (COCA), which strive to present as holistic and varied a view of the English language as possible (Lee, 2010). Possibly the largest attempt to create a corpus of specifically academic language is described in Biber et al. (2002). Funded by the TOEFL, this study included samples

of almost all genres of English an international student would likely encounter at an American university. However, most studies of formulaic sequences in academic language are based on smaller, genre-specific corpora. Teachers who use corpora to teach language usage have discovered the benefits of small, specialized corpora (Flowerdew, 2004). For corpora of written language, "small" always means under five million words, and sometimes under 250,000 words (Koester, 2010).

There are several reasons that small corpora have dominated the study of academic English, including feasibility and ease of creation by a single researcher. The primary advantage of a smaller corpus is that it allows for a "closer link between the corpus and the contexts in which the texts in the corpus were produced," (Koester, 2010, p.67). Especially when teaching language usage, context is key. It does a student no good to learn a phrase if they are not also taught where it is used and how it relates to its context. Context is especially important for students specializing in a specific field, as academic disciplines vary widely in which formulaic sequences they use most and their connotations (Hyland, 2008b; Hyland, 2012).

Creating a specialized corpus also means that the compiler and the researcher are generally the same person, so that the corpus is tailored to the researcher's research questions or to the needs of a specific student population (Flowerdew, 2004). Because the compiler has a deep understanding of the materials comprising the corpus, they may intuitively guess which features most merit analysis. In addition, large corpora usually depend on samples of texts to avoid copyright issues. However, for academic writing, the genre pieces are often very structured (e.g., the introduction-methodology-results-discussion format of many research articles), so including whole texts in a corpus helps researchers understand how language use varies from section to section (Shahriari, 2017). Thus, this study will be based on a specialized corpus built by the researcher in order to provide full access to the texts for the researcher and rich contextual information for the materials.

In creating specialized corpora, the main concern is internal representation. Because there are fewer texts, it is essential to choose texts that present an accurate snapshot of the field. Miller & Biber (2015) depict the challenges of creating a truly representative corpus. They created a most

frequent words list for a corpus of 10 undergraduate introductory psychology textbooks, to examine how using different subsections of the corpus affected the results. Nearly 62% of the lemmas in the overall corpora were used in only 3 or fewer of the 10 books; of the over 30,000 lemmas in the books, only about 4,300 occurred in all 10 books. Breaking the corpora down into sub-corpora created significantly different word frequency counts, indicating that including other psychology textbooks could also produce different frequency lists. One of their proposed solutions for this issue of variety in internal representation is to consider more carefully the range of specific topics within an already specialized corpus. Thus, for example, a study of formulaic language in the journal *Science* may want to control for the different sub-disciplines of the articles and determine how changes in topic affect language patterns. In the case of this study, the discipline of engineering is too broad, with too many specialties, to represent each branch equally and meaningfully in a small corpus. Thus, four disciplines within engineering were selected to focus the corpus: aerospace, mechanical, industrial and electrical engineering. These four disciplines ranked in the top five largest engineering programs at Purdue. Computer engineering ranked third, but as it overlapped significantly with electrical, it was not included. This gives the additional benefit that by comparing two or more of the disciplines, the researcher can separate the patterns that occur across engineering disciplines from any discipline-specific patterns and quirks.

Selecting Corpus Materials

Once the disciplines represented in the small specialized corpus are selected, the next two questions are what type of publications represent target writing for engineering graduate students, and what size would allow the corpus to be representative while still being small enough to allow for meaningful contextual analysis.

The first question was answered by surveying two graduate students each in each of the four selected engineering sub-disciplines. They reported that journal articles and conference proceedings are prestigious publications, but that graduate students often publish conference proceedings first and then refine these into journal submissions in the final years of their studies. When asked to name the most prestigious journals in their fields, six of the graduate students answered with a mixture of journals and conference proceedings. However, as most said that the

language in conference proceedings tended to be slightly less formal, and that journal language was generally considered the target style, I decided to include only journal articles and not conference proceedings in the corpus. Databases like Scimago also consistently rank journals above conference proceedings in prestige. Even graduate students who are writing proceedings benefit from a study of journal writing, as the features of journal articles are usually the target in conference proceedings. To maintain consistency, only empirical journal articles with an Introduction-Methods-Results-Discussion format were included.

Some studies do not consider published articles by NNS authors to be ideal examples of target writing, and they only use papers written by English-L1 authors (as guessed by their last names) as target writing samples. However, holding NS writing as the "ideal" target to which NNS writers should aspire is both unnecessary and unhelpful; NNS may use different writing approaches than NS writers and still communicate well (Silva, 1993). In addition, a study of non-canonical grammar in engineering journal articles found that a low number of grammar "errors" or stylistically odd choices by NNS authors did not keep their research from being understood or awarded prestigious awards (Rozycki & Johnson, 2013). Thus, this study did not consider author nationality or native language in choosing target writing samples. If an article with NNS authors is well written enough to attain publication, then it represents target writing in its field.

Once published engineering journal articles were chosen as the type of target writing, the last decisions were which journals and how many articles to include in the corpus. To ensure that the journals were representative of target writing in their field, two graduate students in each discipline were asked to name the top journals for publishing in their fields, both in the field as a whole and in their specialization. Most students were able to easily name the best journals in their specific areas, but three out of the four consulted journal rankings to identify the best in their field. Thus, the all-field journals were generally named in accordance with the rankings on websites like Scimago. The last five available issues of the top two all-field journals and top two specialized journals from each area were included in the corpus, as students seeking to publish are most interested in current writing norms. This combination of general and specialized journals ensures as wide a range as possible of academic writing styles. For ease of access, only journals accessible through the Purdue library in PDF format were considered. If a journal was

not available through Purdue or other public access options such as Google Scholar, then the next journal named by the student was used. Only two journals were eliminated because they were not easily accessible. A list of the final sixteen journals that were used in constructing the corpus, along with their discipline (and specialization, if relevant), their type/token count, and their total word counts can be found in Appendix A. Note that the word counts come from the final cleaned version of the corpus; the steps taken to clean the corpus are explained in the next section.

For the question of size, other specialized EAP corpora were used as a standard. Examples in the literature range from 730,000 words (the RA corpus in Hyland 2008a & 2008b), around 1 million words (Sharhriari, 2017; Lee & Chen, 2009; Pan, Reppen & Biber, 2015) up to 3 million (Öztürk & Gül, 2016). In these studies, corpora with over 1 million words generally consisted of learner and target sub-corpora that were to be compared to each other. Because this corpus does not consist of such subsections, 1.5 million words was chosen as the target size after pre-processing.

Given the average length of published engineering research articles, around 250 articles were needed to achieve a total of 1.5 million words. Since each of the four disciplines were to be represented by four separate journals (to ensure that no one journal's norms dominated the field), this came to 15.6 articles per journal selected. To err on the safe side, 17 articles were collected from each journal. This led to 68 articles from each of the four disciplines of engineering, for a total of 272 articles. They were chosen following the format in Peacock (2015), where the available articles in each journal were numbered and 17 of them selected from a random number generator. As one file was corrupted and could not be transferred to a text file, it was excluded for a total of 271 articles. This process created a corpus with over 2.4 million words before pre-processing. These files were converted from PDF files to text files. A list of the articles in the corpus is available in Appendix B.

Corpus Pre-Processing

To ensure the quality of the corpus, and to get a more accurate word count, several steps were applied to the raw corpus. All numbers were removed. A wordlist of all tokens in the corpus was generated through AutoMap, and every item with over 150 occurrences that was not an intelligible English-language word was deleted from the corpus. Because the research articles included a multitude of mathematical equations, including many with symbols that represented constants and variables, this list of items to delete had 132 items on it which together accounted for 211,929 tokens in the corpus. Next, the text files were manually cleaned of everything that appeared before the opening abstract or summary (generally titles, author and institution information, keyword lists), and of the references, acknowledgements, and author biographies at the end. This led to the final version of the corpus with 1,925,430 tokens, somewhat larger than the 1.5 million words target. The figure headings were left in, as a preliminary analysis with a test corpus where the figure headings were separated out showed that the headings did not contain novel collocations or language patterns, and thus were not worth separating out. Each article was saved in a separate text file, marked for its discipline, and numbered. A separate database with the title, author and webpage of each article was created for documentation purposes; the author and title information can be found in Appendix B.

As every article was in a standard IMRD format (introduction, method, results, discussion), and that the whole text was used rather than a sample, sample size was not considered when building the corpus. Of the four disciplines, the mechanical engineering corpus was the smallest, with 388,961 words, and electrical engineering was the largest, with 536,043 words. Since this was a large range, once the n-grams had been gathered, their dispersion across the four corpora were checked. There was little difference across corpora; only three n-grams had sub-corpora occurrences PMW (per million words) that were more than one standard deviation different than the across-corpus occurrences PMW. Given their similar dispersion across sub-corpora, the n-grams and p-frames were analyzed for their behavior across the whole corpus rather than separately in the sub-corpora.

Once assembled, the corpus was pre-processed using several of the pre-processing settings available in the AutoMap software. All words were converted to lower-case. British spellings

were converted to American spellings. The words were lemmatized using AutoMap's k-stemmer. Lemmatization was used because in the test corpus, the n-gram and p-gram searches were run on both a lemmatized and non-lemmatized version of the corpus. Lemmatizing allowed for a much clearer grouping of plural/singular nouns and past/present verbs, as generally the tense or plural choice was not semantically significant and distinguishing between them made for results with unnecessary and unhelpful detail. After the lemmatizing was complete, there were several words that were incorrectly stemmed (such as "doe" from "does," and "thu" from "thus"). These were corrected to their true stems by creating and running a custom thesaurus within AutoMap. Other AutoMap preprocessing options, such as removing noise verbs and punctuation, or applying preset delete lists, were not used, as these would remove data that is essential to uncovering formulaic language.

N-Gram and P-Frame Identification

Once assembled, several searches were run on the corpus using AntConc (Anthony, 2018a) and AntGram (Anthony, 2018b). These tools were preferred over Wordsmith (Scott, 2018), as they are free of charge, have a user-friendly interface, and could together perform all the desired searches. The first search was similar to much of the previous research, drawing out 3-, 4- and 5-grams. Using the minimum occurrence of 20 times per million words that was used in previous studies (such as Hyland, 2008a; Biber, Conrad, & Cortes, 2003; Cortes, 2004) would have required at least 58 occurrences in the corpus of each item. However, this study wanted to include dispersion as well as frequency as a measure of a phrase's value to learners. Previous literature had suggested .7 or .8 (i.e., appearing in 70% or 80% of texts) as a common cutoff point for dispersion (Biber, Reppen, Shnur & Ghanem, 2016). But this corpus is composed of many shorter documents; the average word count for an article is 11,260. As a result, using a .7 dispersion cutoff would have required the phrases to appear at least 120 times, or at least 62 times per million words. Thus, to compromise between the recommended dispersion and frequency cutoffs for n-grams in this study, a .5 dispersion score was chosen. While this meant that the frequency cutoff had to be set higher than in most previous studies of n-grams (such as Hyland 2008a, 2008b), it meant that students would still access the n-grams that are less dispersed but still common enough to be characteristic of this genre.

This dispersion score meant a minimum frequency of 62 occurrences, or 32 times per million words. This search created a list of 26 n-grams. One of these n-grams was not relevant ('https doi org,' the result of websites that were cited throughout the articles) and was discarded, but the rest were kept. The resulting items were cleaned of duplicates or nested phrases. The n-gram results are discussed in more detail in Chapter 3.

The second search followed Fischer-Starcke's (2012) methodology for finding p-frames, using the AntGram software. As p-frames are much more common in the corpora than n-grams, the suggested dispersion cutoff point of .8 could be used without any issues. The search was set to yield 4-, 5- and 6-frames, with a novel word inserted at any of the inner slots in the p-frame. This search yielded 30 p-frames. Only p-frames whose variable words were semantically homogeneous, as described in the next paragraph, were kept, as ones with broad semantic variation refer to widely disparate concepts and are not conducive to teaching patterns (Fuster-Marquez & Pennock-Speck, 2015).

To be considered semantically homogenous, at least 51% of the tokens in the wildcard slot had to belong to one grammatical category (verb, noun, etc.) have some semantic similarity, and be able to be used across disciplines. For example, with the p-frame *it be * to,* 82% of the slot tokens were adjectives, and they were mostly adjectives of evaluation ('important, sufficient, better, feasible, impractical, appropriate'). These words were also not so discipline-specific as to be useless for writers in other disciplines. Thus, this p-frame passed the semantic homogeneity test and was kept. An example of a discarded p-frame is "on the ___ of," where the variable slot was filled with widely disparate tokens (measurement, plasticity, set, day, basis, position, etc.). While these were mostly nouns, they did not have a semantic similarity to indicate any underlying patterns. This type of p-frames is not useful to students.

Of the 30 p-frames collected, 21 passed the semantic homogeneity test and were kept for further research. For these p-frames, the information on the tokens in the variable slot and the frequency of each token was collected. These results are discussed in Chapter 3.

Next, for each item (n-gram or p-frame), AntConc was used to collect information about the raw frequency, the normalized frequency, and the dispersion scores of that item in the corpus. The dispersion scores have two parts: first, a score representing how dispersed the item is among the four corpora, and secondly, four scores, one for each corpus, representing how dispersed the term is among the 51 articles in the corpus. Dispersion is calculated as the percentage of texts where that term appears. There are multiple ways to calculate dispersion; a current favorite is Gries' *DP* (Gries, 2013), but recent research has shown that simple percentage calculation leads to nearly identical results (Gray, Biber & Geluso, 2015). Like Gries' *DP*, this measure is sensitive to dispersion and can be used to produce a separate dispersion sub-score for each separate section of a corpus (Biber, Reppen, Shnur & Ghanem, 2016). Keyness information was not collected, because that would have required an external reference corpus against which to compare the frequencies of words in this corpus, and since this is a target corpus, there was not a clear choice for comparison. In addition, recent advances in corpus methodology have found that dispersion and frequency information is often a better indicator of a corpus' "aboutness" than keywords, as a term that is used occasionally across many texts is more representative of the writing in the whole corpus than one that is used frequently within a few texts (Egbert & Biber, 2018).

Categorizing Results

Once the quantitative data had been retrieved, the next step was to determine how best to make the data useful to researchers and students. Depending on the goal of the research, there are several different ways to sort the results. Linguistics-oriented studies typically sort the sequences according to two taxonomies; a structural taxonomy described in Biber & Conrad (1999) and refined in further studies (Biber, Conrad & Cortes, 2003; Hyland, 2008a), and a functional taxonomy, usually either the Biber, Conrad & Cortes (2003, 2004) taxonomy or the Hyland (2008a) taxonomy.

The structural taxonomy classifies the phrases by their syntax. Biber & Conrad (1999) created the framework that is employed in most the recent literature. They find that most lexical bundles contain some section of a matrix clause followed by an embedded complement clause (a *to*-clause, *that*-clause, or WH-clause). These lexical bundles can be categorized by which fragment of this combination they represent, such as beginning with a noun phrase, beginning with a

personal pronoun, including the preposition *of* (the most common preposition), other prepositions, or post-nominal clause fragments. Hyland (2008a) modified the structural categorization system to the following categories: noun phrases + *of*, passive + prepositional phrase, other prepositional phrases, prepositional phrases + *of*, noun phrases + other, anticipatory *it* structures, and miscellaneous. Later researchers tend to begin with either Biber & Conrad's categorization or Hyland's, and then focus on the syntactic categories that appear most frequently in their corpora, and, if necessary, add novel categories.

In contrast, a functional taxonomy classifies the phrases by their use and thus requires more time-intensive and subjective sorting. Biber, Conrad & Cortes' (2003, 2004) taxonomy was developed on a corpus of academic speaking and writing and has three major categories: stance expressions, discourse organizers, and referential expressions, along with a small catch-all category of special conversational functions. Hyland (2008a) reworked this classification to apply specifically to academic research writing. Using Halliday's linguistic macrofunctions (Hyland, 2012), he sorted sequences according to if they were research-oriented (describing the field of research), text-oriented (similar to Biber's discourse organizers, referring to the text itself), or participant-oriented (either communicating the writer's attitudes or addressing readers). He found that some structural categories tended to align strongly with certain functional categories (Hyland, 2008a).

In practice, sorting these sequences can be tricky. One example comes from Oakey's (2002) study of the variable phrase *it has been/is (often) * that*, where the wildcard was verbs such as "claimed, asserted, believed." The pragmatic uses of the lexical bundle were much more varied than he originally hypothesized. Thus, researchers must fight the urge to categorize their results simplistically and must make sure to expose their students to the full range of pragmatic uses of the formulaic sequences under consideration.

While much formulaic language research categorizes the phrases by structure as well as function (Hyland, 2008a; Biber, Conrad & Cortes, 2003; Liu, 2011), given the pedagogical focus of this project, the phrases were only sorted functionally, using the schema in Hyland (2008a). Students in previous studies have reported that they prefer functionally organized lessons (Eriksson,

2012). This may be because it is easier to learn the nuances of new phrases when they are presented alongside semantically similar items, and because learning similar phrases together decreases the likelihood that the student will overuse a new functionally novel phrase (AlHassan & Wood, 2015; Lee & Swales, 2006).

I also did not follow the previous methodology of sorting the phrases only according to their most frequent functional use (as in Biber, Conrad & Cortes, 2004); instead, if the bundles served multiple functional purposes, and each purpose occurred above the frequency cutoff score, the item was included in each of the functional categories where it appears. When one or more of the functional uses fell below the cut-off frequency of 30 instances per million words, it was not included on the list.

In addition to Hyland's functional categories, several other syntactic and rhetorical categories have received special attention as being particularly difficult for multilingual speakers to acquire and successfully use in academic writing. Thus, if p-frame and n-grams fit these categories, they were noted as being particularly useful for students. These categories are intensifiers/qualifiers (Ito & Tagliamonte, 2003); if-clauses (Warchal 2010); attribution (Hyland, 2012), passive reporting verb structures (Oakley 2002); anticipatory *it* clauses (Wei & Lei, 2011; Chen & Baker, 2010), evaluative *that* statements (Hyland & Tse, 2005), and hedging phrases (Hu & Cao, 2011; Perez-Llantada, 2014). When the n-gram and p-frame results included phrases from those categories, they were flagged. This information was useful in deciding which formulaic sequences to highlight during the materials construction.

The final step was to use the network visualization tools to see what network the p-frames and n-grams created, and if there were any interesting patterns there that would help teachers and students to understand how the phrases connect. The results are reported in Chapter 3.

Phase Two: Materials Construction

The goal of the corpus investigations above was to provide results that would primarily interest linguists. But as these results are indigestible to most students seeking to apply the findings, we come to the material construction phase. This phase has two goals: selecting and visualizing the

results to make them accessible to students; and developing accompanying exercises to help students identify the context and patterns in formulaic language and apply them to their writing. "Selecting Formulaic Language for Materials" and "Creating Visualizations" address the first goal, while "Developing Exercises" and "Triangulation of the Corpus Results and Materials" address the second goal. It is important to note that in practice, these two steps often occurred simultaneously and informed each other. Of specific interest throughout this process is addressing the second research question of whether SNA visualization can create effective, student-friendly teaching materials.

Given the interdisciplinary combination of corpus linguistics and SNA visualizations in this phase, the methodology is more exploratory and iterative in nature than in the previous phase. While the corpus stage mostly followed established corpus study norms, this approach combines lessons from the literature with experimentation. Thus, the methodology section will outline the tools available and the approaches used in creating the exercises, while the fourth chapter will seek to answer the second research question by describing the optimal framework for materials design that emerged and by evaluating the usefulness of SNA visualization tools in that process.

Selecting Formulaic Language for Materials

The real-world restrictions of the classroom implementation in the third phase meant that only a small section of the results on formulaic language from the first phase could be taught to the students. The implementation consisted of two 75-minute class periods in a six-week short course (see "Phase 3: Classroom Implementation" below for details). Given that the goal of the project was to introduce students to phrases with rich contextual information and rhetorical uses, it was also considered better practice to introduce a few phrases in depth than to skim through everything uncovered through the corpus analysis.

Thus, the first step in creating materials was to select a subset of the formulaic language that would be most relevant and applicable to students. Past research has employed a variety of approaches in selecting phrases, including choosing the most frequent phrases (Cortes, 2004), the most dispersed phrases (Egbert & Biber, 2018), phrases rated by EAP experts as most useful to students (Simpson-Vlach & Ellis, 2010), or selecting a particular rhetorical or functional

category (AlHassan & Wood, 2015). The pedagogical studies named above (Simpson-Vlach & Ellis, 2010; Cortes, 2004; AlHassan & Wood, 2015) all balanced their primary form of selecting formulaic sequences by considering one or more of the other methods as well.

The second question in choosing materials was how much access the students were to have to contextual information, i.e., should concordance lines be included for each n-gram? Should there be an example for every item that appears in a variable slot, or only the ones above a certain frequency? Should the lines be sorted by rhetorical use of the phrases, or should it be left to the student to map the rhetorical functions to the phrases? The tension here is between teacher-directed and data-driven pedagogical uses of corpus (Gabrielatos, 2005). In the teacher-directed approach, the teacher uses corpus data and examples to construct assignments and guide students toward certain findings. For example, if the teacher wants students to learn that one citation verb is often used more positively than another, they pick examples of those two verbs from the corpus that clearly lead students to this conclusion. Data-driven learning, however, would give students either access to the corpus and corpus tools, or to the concordance lines of every example of the verbs, and ask the students to draw their own conclusions on the nuances between the verbs. The benefit of this is that the students are more invested and may be able to spot patterns that the teacher did not see; the cost is that it is time intensive and students can go down rabbit holes. Both approaches, however, offer the benefit of student involvement and consciousness-raising (Gabrielatos, 2005).

The answer for the first question of which sequences to select for the teaching materials was to identify a set of thematically similar n-grams and p-frames that each represented a different aspect of authorial stance. “Stance” is a rhetorical category that covers how “writers present themselves and convey their judgments, opinions and commitments” (Hyland, 2005, p. 176). In academic writing, with its strong emphasis on objective, rational judgements, this is often done obliquely, through the use of evaluative phrases and words (Peacock, 2015). Because it is inexplicit, NNS writers have had trouble acquiring it and have been shown to benefit from corpus-backed explicit instruction (Chang & Schleppegrell, 2016). Thus, this category was chosen because the formulaic sequences in this category were both frequent and well-dispersed; they contained several syntactic elements that tend to appear less frequently in NNS writing than

NS writing; and they represented a rhetorical set that had proven difficult for NNS in previous research (Peacock, 2016; Chen & Baker, 2010; Adel & Erman, 2012; Perez-Llantada, 2014; Wei & Lei, 2011; Öztürk & Köse, 2016). The answers to the second question on teacher-directed or data-driven assignments will be discussed in the fourth chapter, as this tension was addressed in multiple ways while building a constructive framework for materials creation.

Developing Exercises

The second step, which happened concurrently with the first step, was to integrate the visualizations produced in the first step with exercises that would allow students to better understand the sequences and incorporate them in their writing. These two steps were intertwined because the exercises referred frequently to the visualizations, and the visualizations were modified to highlight the parts that were pertinent to the exercises. The first section below summarizes findings from the literature on teaching formulaic sequences that guided the design process, and the second section explains the types of exercises chosen based on that literature.

Lessons from Previous Studies

The first set of relevant literature is that which investigates the challenges of successfully teaching formulaic sequences. Students report that they avoid formulaic language because it is risky; sequences are harder than individual words to implement correctly; they may have been taught them as chunks, without context or deeper understanding; and they are more likely to have grammatical or semantic issues when using them (Coxhead, 2008). Once a phrase is learned a certain way, it can be difficult to incorporate feedback and change its use (Li & Schmitt, 2009).

Several approaches to teaching formulaic language led the students to report raised consciousness, but they did not show improvement in formulaic sequence use in their writing. This was true both for NNS (Jones & Haywood, 2004), and for native-English speaking undergraduates (Cortes, 2006). In the first case, the students were given a series of worksheets on academic phrases, from gap-fill to distinguishing grammatical differences. In the second case, the undergraduates were presented five 20-minute lessons from a targeted corpus created specifically for their discipline (history). While the students reported finding the material helpful

and interesting, in both cases, the posttests of their class writing revealed no significant increase in either the variety or frequency of formulaic language.

However, despite these challenges, the literature offers several successful approaches to overcoming these barriers and encouraging NNS to acquire an accurate command of formulaic sequences. The first is the data-driven learning approach that is recommended in Willis (2003); he encourages teachers to create small example corpora appropriate for their students' levels or interests, and to allow the students to explore the corpus on their own. While this can take considerable time and energy, and searches without adequate direction or understanding can lead to confusion, data-driven learning approaches to vocabulary acquisition leads to higher student investment and better retention than traditional instruction (Soruc & Tekin, 2017, Karras, 2016). More structured approaches are also successful; one particularly useful approach is to take a specific functional category and use corpus concordance lines and directed exercises to help the students understand and internalize their options for that rhetorical category, as this addresses the challenge NNS can have of overusing a phrase because they are not familiar with other alternatives. Bloch (2010) used this approach to help his students understand the nuances among the most frequent reporting verbs in a corpus of academic articles. He selected concordance lined ahead of time that illustrated 27 different verbs and phrases used to report previous research, and he guided his students through understanding the rhetorical connotations and common collocations of the different types via pen & paper exercises. He observed that the students were engaged and understood subtle distinctions quickly.

A similar approach is to take one grammatical feature and teach its functional uses and common collocations, with examples from concordance lines. Hyland & Tse (2005) conclude their corpus study of evaluative *that* statements by recommending this method. They found that while NNS users had no problem with the syntax of this frame, they tended use more affective items in the variable slot than their NS peers, so that their evaluative *that* statements tended to come across as "overstated" or "anxiously persuasive."

Whether data-driven or teacher-directed, students retain more when the exercises involve repetition and ask the students to write novel sentences using the formulaic sequences. Alali &

Schmitt (2012) found that in teaching English idioms to Kuwaiti students, students who reviewed the idioms had better recall and recognition. Written review was more effective than oral repetition. And "uncontrolled production" (AlHassan & Wood, 2015), where students write sentences and paragraphs in their own fields with the sequences, is an essential last step. When the students practiced the material in the context where they were likely to use them, they were both more enthusiastic about using them and more likely to retain them (Eriksson, 2012).

The most successful case studies in teaching academic formulaic language employed a variety of learning exercises that included concordances, worksheets, and uncontrolled production. One was a brief exercise involving two workshops for PhD students (Eriksson, 2012). In Eriksson's workshops, the PhD students engaged in four activities: speculating about the most common bundles & their frequencies in biotechnology, searching the corpora and analyzing concordances, working with prepared worksheets sorted by functional categories, and incorporating new bundles in their writing. While the students reported enjoying concordance activities and functional worksheets the most, the final exercise of writing novel sentences with the bundles was essential in helping student internalize and recall new phrases.

AlHassan & Wood (2015) provide the most thorough and data-supported case study for effective formulaic sequence pedagogy. Twelve students in EAP programs participated in ten weekly, 90-minute classes targeted at improving use of lexical bundles. The participants represented four different L1s and had a wide range of English placement test scores. Their use of lexical bundles was evaluated through pretests, posttests and delayed posttests. The corpora and teaching content were again field-specific, with 80% of class time dedicated to presentation of the material and worksheets, and 20% to uncontrolled production. Both the posttest and the delayed posttest showed significant increase in use of lexical bundles. The students increased in variety, not just number, of lexical bundles used. This study demonstrates the vital importance of discipline-specificity and uncontrolled production to make the information valuable to students.

Exercise Types Used in Materials

The visualizations and exercises that were developed for this project can be found in Appendix E, and their development is explained in Chapter 4. Before creating the materials, however, the

following three broad categories of exercises were chosen as models because of their proven usefulness to students in previous research. As the materials were designed, at least one question from each of these categories was included for each phrase:

a) Comparison exercises that examined and contrasted different phrases used for a specific rhetorical purpose (Hyland, 2008b);
b) Exercises specifically oriented at mastering functional & syntactic aspects of formulaic sequences that have proved difficult for NNS students in previous research;
c) Tasks that require the student to implement the sequences in their professional writing, following the "uncontrolled production" model in AlHassan & Wood (2015).

The (a) exercises help the students to understand and distinguish semantically similar sequences, as several studies have reported that students benefit from this approach (Lee & Swales, 2006; Bloch, 2010; Eriksson, 2012). Comparison helps them to distinguish sequences and can prevent some of the issues of overusing novel phrases. In this case, for the reasons outlined in a previous section, the rhetorical category is authorial stance and the exercises will examine aspects of how authors communicate stance. However, if the students merely read and discuss the comparative data on quantity of and rhetorical distinctions between phrases, they often to do not internalize the information (Eriksson, 2012). Thus, these activities should provide students with the comparative data and ask them to choose the most appropriate sequence(s) and defend their choice in a variety of fill-in-the-blank exercises.

The exercises in (b) fill in the possible application step that has been suggested by many previous studies of NNS syntactic use of formulaic sequences in academic writing. These exercises will give students an additional chance to practice the structures that have been most difficult for their peers. Among the authorial stance phrases are instances of several syntactic constructions that have proven hard for NNS when writing: anticipatory *it* structures (Chen & Baker, 2010; Adel & Erman, 2012; Wei & Lei, 2011; Perez-Llantada, 2014), evaluative or probabilistic hedging (Perez-Llantada, 2014; Chen & Baker, 2010) participant-oriented bundles (Wei & Lei, 2011) and passive voice (Öztürk & Köse, 2016; Wei & Lei, 2011).

Finally, the uncontrolled production in (c) is an essential final step in enabling NNS writers to permanently adopt helpful novel structures (AlHassan & Wood, 2015). These semi-structured activities coach students to write sentences and paragraphs in their fields of research that employ new formulaic sequences. This can be as simple as writing a new sentence with a sequence, or it can involve editing previous writing or comparing the same paragraph written with different, rhetorically similar sequences. Graduate students often take this type of exercise seriously, as it provides a bridge between the instruction and the intended application (Eriksson, 2012).

Creating Visualizations

Once a set of phases had been selected, the next step was to visualize these phrases. The methodological steps necessary to perform this in the AutoMap and NodeXL software will be outlined in this section. In practice, these steps were applied several times in different orders and combinations to produce unique visualizations, and these were shared with NNS for feedback and editing before they were incorporated into the exercises. As a result, while the methodology is laid out below, the fourth chapter will discuss in more detail the best practices of using these methodologies to create effective materials.

In creating the visualizations, the purpose is always to present the linguistic findings in a way that is attractive and accessible to professional NNS writers without a background in linguistics. When students must wade through lists and syntactic information not directly applicable to their writing, they quickly lose interest (Lee & Swales, 2006). Thus, these visualizations concentrate on the most immediately useful and easily explicable findings.

Once AntConc and AntGram had identified the p-frames and n-grams in the materials, the information on the lexemes around the formulaic sequences and the distance and strength of the phrases' connections to the surrounding lexemes could be gathered using AutoMap software (Carley, Columbus & Landwehr, 2014) and then visualized using NodeXL (Smith et al., 2010). To allow this, formulaic sequences had to be marked in a new corpus so that AutoMap and NodeXL would read them as unique items rather than as independent words. Thus, the corpus was converted into a second and third corpus appropriate for AutoMap. The second corpus contained the same material as the first, but the formulaic sequences were joined by an

underscore. For example, the n-gram “on the other hand” became “on_the_other_hand.” P-frames had ‘pf_’ added to the beginning, and n-grams had ‘ng_’ added, for easy identification and ease in sorting the results. The final version of the previous example was “ng_on_the_other_hand.” In the first corpus, the variable slot in the p-frames were replaced with “XXX” so that “be difficult to be,” became “pf_be_XXX_to_be.” This corpus allowed for the p-frames to be studied without consideration for what lexemes appeared in the variable slot. However, because it could be useful for students to also learn how choices about the variable lexeme affected the material around the p-frame, a third corpus was also created. In this corpus, the variable slot in each p-frame was kept, so that “be difficult to be” in the original corpus became “pf_be_difficult_to_be.”

These second and third corpora were uploaded to AutoMap, and two semantic (co-reference) network lists were generated for each corpus. A semantic network is a list of how often each word co-occurs next to another word within a given window size. It is possible to include bidirectional networks (looking to both the right and the left of the word within the window size) or unidirectional networks (looking only ahead of the word). For this project, the semantic lists were unidirectional. As writing is linear, it is essential to know whether the words in question tend to appear before or after the phrases.

For the second and third corpus, a semantic list with both 2- and 3-word-window settings were created. Thus, each list provided the frequency with which any item occurred within either one or two words of each other item, and, for the 3-word-window, it also provided the average distance between the two items. This allowed the researcher to determine whether it was more likely to appear immediately before or after the phrase or a one-word distance away. The information on semantic networks in the second and third corpora were extracted and saved separately. For the second corpus, a 10-word-window semantic list was also generated, but this amount of information was too unwieldy to deal with in the third corpus. AutoMap recorded this data in .csv files that were compatible with NodeXL, the network visualization package.

Finally, a fourth version of the corpus was designed to extract the relationships between just the formulaic sequences. In this corpus, using the settings available in AutoMap, everything that was

not marked with “pf_” or “ng_” was deleted and a placeholder put in its slot. Relationships between the formulaic sequences were measured with a 200-word-window with one paragraph as the stop-unit (thus, when the program arrived at the end of a paragraph, it would close the window, even if it was shorter than 200 words). This data indicates which formulaic sequences tend to occur together within paragraphs, and the average distance between them.

These three datasets were too large for NodeXL to handle without crashing; the first step was to simplify the datasets by removing all relationships that did not involve any formulaic sequences, as these were not relevant to this study. Before removing them, the smallest .csv file had over 400,000 relationships in it and the largest had over 711,000 relationships; after removing them, the smallest .csv file had about 18,000 relationships, and the largest one had about 36,000.

Because the NodeXL software will visualize whatever data is inserted into it, the best approach to making visualizations was to extract all the relationships for one particular formulaic sequence from the .csv files, sort and categorize it in a variety of ways, put the data into NodeXL to see the results, and then work with the visualization tools there to make the information as accessible as possible.

Before inserting it into NodeXL, however, there were a multitude of ways the information could be sorted. Each aspect of interest was mapped in a separate column next to each lexeme (see Figure 1), as NodeXL could treat each column as a variable associated with that vertex and provide many ways to visualize those variables (through color, size, edge length, etc.). The online tool WMatrix (Rayson, 2009) was used to tag each lexeme with a part of speech (POS) tag and semantic category. This allowed the researcher to quickly notice broad syntactic or semantic patterns in the use of the phrase and the surrounding lexemes. For example, some phrases often came directly before verbs or adjectives, while others appeared frequently after transition words or expressions of quantities. These patterns could then be used to group the lexemes when building the visualizations.

There are other tools available for semantic and part of speech tagging, such as WordNet (Fellbaum, 2005). In the future, as this technique for sorting p-frames according to the semantic

features of their variable items is developed, it would be useful to use additional semantic taggers like WordNet and see if they lead to similar results as WMatrix tagging. However, as the WMatrix format was better adapted to providing information in cvs files, it was preferred.

Finally, any relationship with a frequency of two or less was generally removed from the data; these were considered infrequent enough to not be interesting to the students, though if there were any interesting patterns in the data below the cutoff that was not included, this was noted separately so that it could be incorporated in the worksheets. Depending on the formulaic sequence under discussion, often higher frequency cut-offs were employed; this will be discussed in Chapter 4.

	A	B	C	D	E	F	G	H	I	M
1	Vertex	In-Degree	Out-Degree	Freq	one removed	before NG	after NG	before & after	POS	
2	ng_one_of_the	163	204	123	0					
3	be	1	1	67	0	0	0	1	VBI	
4	most	1	0	36	0	0	1	0	RGT	
5	at	0	1	19	1	1	0	0	II	
6	of	1	1	18	1	0	0	1	IO	
7	as	1	1	12	0	0	0	1	CSA	
8	in	1	1	9	0	0	0	1	II	
9	main	1	0	9	0	0	1	0	JJ	
10	the	1	1	9	1	0	0	1	AT	
11	boundary	1	0	8	0	0	1	0	NN1	
12	to	1	1	7	0	0	0	1	II	
13	each	0	1	6	1	1	0	0	DD1	
14	least	0	1	6	0	1	0	0	DAT	
15	reason	1	0	6	0	0	1	0	NN1	
16	higher	0	1	5	0	1	0	0	JJR	
17	and	1	1	4	0	0	0	1	CC	
18	any	0	1	4	0	1	0	0	DD	
19	by	1	1	4	1	0	0	1	II	
20	commonly	1	0	4	1	0	1	0	RR	
21	consider	0	1	4	1	1	0	0	VV0	
22	engine	1	0	4	0	0	1	0	NN1	
23	four	1	0	4	0	0	1	0	MC	
24	from	0	1	4	0	1	0	0	II	
25	present	1	1	4	1	0	0	1	NN1	
26	widely	1	0	4	1	0	1	0	RR	

Figure 1: Example NodeXL Information

Once the selected relationships for one formulaic sequence and its attributes were imported into NodeXL, there were a variety of visualization choices. The core feature of NodeXL is that it takes whatever relationships are inserted and produces a graph showing the connections between vertices, called nodes; in this case, each node is a lexeme or formulaic sequence. It can group

vertices together and tabulate network properties (such as the number of in-degrees and out-degrees and the betweenness and closeness centralities for each vertex). The color, shape, size and label of each vertex can be mapped to any of that vertex's variables (frequency, in-degree/out-degree, semantic tag, etc.). The color, width, style and label of the edges (the lines connecting the vertices) can also be mapped to the edge variables. In addition to the preset layout options, the user can manually move any vertex or alter its appearance. Figure 2 below shows the visualization created when the n-gram *one of the* and all its relationships with a frequency of 2 or higher were imported into NodeXL. Here, the arrows indicate if the node appeared before or after the phrase; the color indicates how close the word is to the phrase (orange being immediately next to, blue being at a 1- or 2-word distance), and the edge width is mapped to the frequency; the fatter the edge, the more frequently that lexeme appeared around the phrase. These are the settings chosen by the software; designing a helpful visual means choosing how and if the data points are visualized, how the nodes should be arranged, and what leads to maximum readability while communicating as much information as possible.

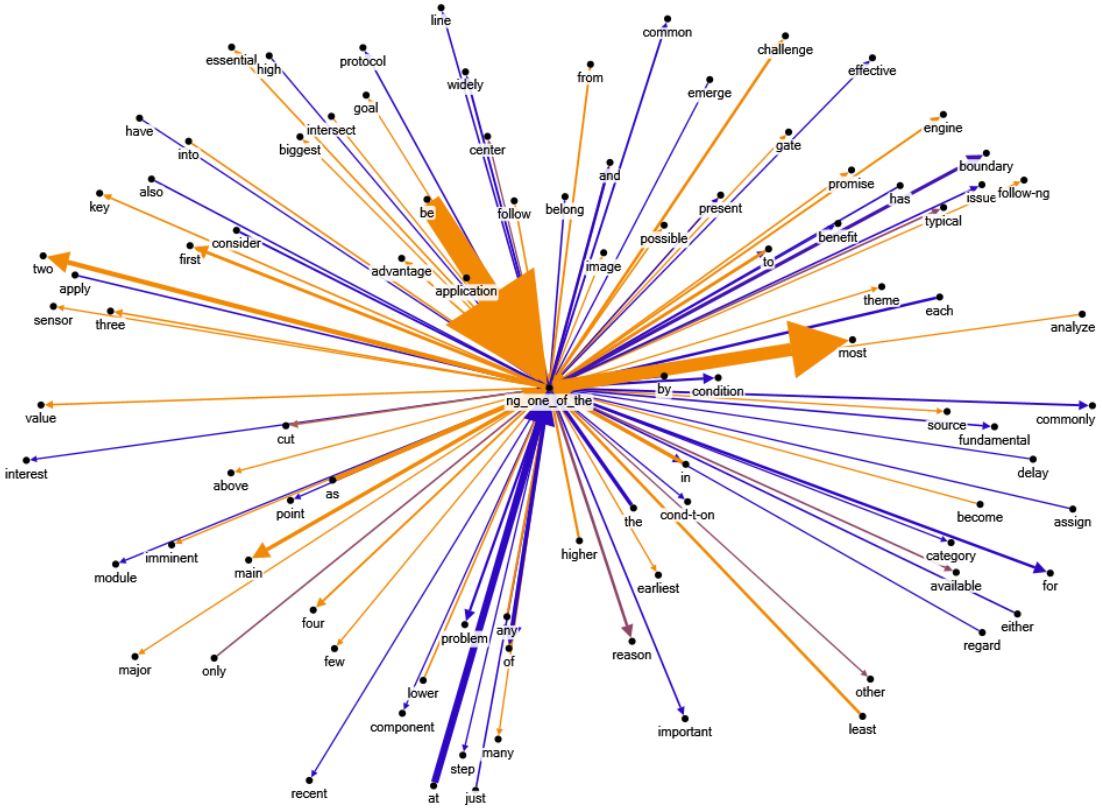


Figure 2: Automatic NodeXL Visualization for *one of the*.

Triangulating the Corpus Results and Materials

The final step was to triangulate the corpus findings by consulting novice and expert writers in each field. While corpus findings can be triangulated through other corpus methods (Baker & Egbert, 2016), in this case it is more helpful to triangulate via outside opinions (as in Jaworska & Themistocleous, 2017). As the formulaic sequences may vary from field to field in ways the researcher is not aware of, input from mature writers in engineering was necessary to ensure that students receive accurate information.

Thus, the materials were piloted with six NNS graduate students to ensure that they were understandable and useful for their target audience, and to answer the second part of the second research question. Each student met with the researcher and went over all the prepared materials, attempting some of the questions and evaluating the information and presentation of the exercises. They were also asked if they see inaccuracies in how the phrases are grouped, what information they find surprising or novel, and if the visualizations are clear or confusing. They were able to provide discipline-specific nuances that I as someone who has not written in the field was unaware of. Their feedback was documented during each session and then incorporated into the materials and into the framework of best practices in making materials.

Phase Three: Implementing Materials in Classroom

The third research question asked if the materials developed in the previous phase could lead to better teaching outcomes than the results documented in Cortes (2006) and Jones & Haywood (2004), and if this could be demonstrated with (a) improvement in the variety and idiomaticity of student formulaic language from a pretest to a posttest and (b) similar results in a delayed posttest. The final methodological phase thus takes the finished teaching materials and tests their usefulness in a classroom setting.

The sections below describes the context and details of the classroom implementation, followed with an outline of the three types of assessment employed to answer the third research question. The first is an ongoing formative assessment of the materials while the class is in progress; the second and third are the summative assessments that seek to answer the questions in 3(a) and

3(b). Formative evaluation is "evaluation used to improve a curriculum during its development," while summative evaluation refers to "the final evaluation of a teaching instrument" that holistically gauges whether the intended effect of the curriculum is visible in student performance (Tyler, Gagné, & Scriven, 1967, p. 87). The goal of the formative evaluation by students and teachers is to gather a qualitative analysis of the strengths and weaknesses of the materials, and to make adjustments in real time when necessary, while the goal of the summative evaluation is to gauge the overall effect of the materials on student learning and writing.

Classroom Implementation of Materials

Ideally, students would study the full range of formulaic language in their disciplines and would be able to spend many weeks learning, reviewing, and incorporating the fixed and variable phrases into their writing. However, the time constraints on graduate students requires that they learn efficiently and quickly; this was why, in part, visualization was judged a helpful tool: it allows participants to understand and apply information quickly and could be used as an immediate reference during future writing. Thus, while it is unfortunate that the implementation could not happen over a longer time period, its brevity has the benefit of testing the materials' effectiveness in a tight timeframe. The situational factors below lay out the context and duration of the implementation, and the implementation describes how class time was used.

Situational Factors

The materials were implemented in Fall 2019 at Purdue University, a large, research-focused university in the Midwest. Purdue is known for its engineering programs, and attracts many international undergraduate and graduate students, primarily for their programs in the sciences and engineering. In recognition of the language needs of the large international student body, Purdue developed PLaCE, the Purdue Language and Cultural Exchange. This program was originally established to create a two-semester program to "improve first-year international students' English language skills" ("About PLaCE," n.d.). However, recently the program has extended to include short courses that cover a variety of language and cultural topics, from academic conversation skills to reading fluency. These classes are short, intensive, non-credit-bearing, and have no additional fees for students. Each course targets a specific language skill

such as prosody, research presentations, academic conversation skills, sentence-level grammar, or common idioms and slang. The classes on academic writing almost exclusively attract graduate students and international scholars, so the materials in this project were implemented in "Essentials of Academic Writing," the introductory short course on academic writing. Information on the student demographics came from current PLaCE short course instructors, who described their learners as highly motivated by intrinsic and extrinsic goals, and under stress to use every minute of their time efficiently. The students were mostly international graduate students, with a minority of them being visiting scholars. Visiting scholars tended to be very committed and make large gains throughout the semester, but they also usually began with weaker speaking and writing skills than the graduate students. The graduate students tended to enroll either because they had received feedback that they need to improve their writing from advisors or instructors, or because they were working on a document or manuscript and wanted to improve the quality of that work. Because of the structure of the engineering programs, the students were highly motivated to learn writing skills and apply them, but they often struggled with not having a scalable way to learn these skills because they had minimal writing assignments outside of the major goals of conference proceedings, articles, theses and dissertations. Their classes rarely required much writing, and lab reports usually received no feedback on structure, syntax or rhetorical effectiveness. Thus, this class offered a valuable opportunity for them to hone their academic writing and learn tools for tackling the high-pressure projects ahead of them. However, given the other pressures on them and the non-credit nature of the class, student dropout throughout the semester was common.

Classroom Time

"Essentials of Academic Writing" was offered twice in Fall 2019, and the materials were implemented in both sections. The course was six weeks long and met twice for 75 minutes every week. In both classes, the fifth week was dedicated to learning and using formulaic language. I guest lectured for he primary instructor for those two 75-minute sessions. The original and revised lesson plans for those two sessions are in Appendix D; their structure and revision are discussed in Chapter 5. To see how the formulaic language section fit into the surrounding coursework, the syllabus for the class can also be found in Appendix F.

Formative Assessment

The formative evaluation combines teacher impressions with student feedback. As in Eriksson (2012), both short courses concluded with a short discussion of the students' opinions on the effectiveness of the class and the content and accessibility of the teaching materials, both via written answers to a questionnaire and oral discussions. After each lesson, the instructor also wrote notes on what content was covered, what questions or concerns arose, and what exercises received the most and least student attention (following Lee & Swales, 2006). Given that previous studies have drawn attention to the difficulty of correcting entrenched inaccurate language patterns (Coxhead, 2008; Li & Schmitt, 2009), special note is taken of inaccurate patterns evident in the students' first writings, and the effect of explicit alteration. This, combined with the student feedback, allowed me to gather the students' more immediate reactions to the materials and reflect on the role the materials played in hindering or furthering the student progress that was measured through the summative evaluation.

Summative Assessment: Pre-, Post- and Delayed Post- Test

Several different evaluative methods have been used to measure if students have increased in the quantity and quality of their use of formulaic language. Students' use of formulaic language can be analyzed in their essays written after the instruction (Cortes, 2004), or in timed paragraphs written in response to a specific prompt (alHassan & Wood, 2015) or via recall and comprehension tests (Alali & Schmitt, 2012).

However, given the time constraints, extended writing would not be feasible. And as the students were mastering the use of a small family of formulaic language and the differences between variants of the variable phrases, it would have been difficult to create a writing prompt that would give them an opportunity to use what they had learned in a scenario close to real life. These instances of academic writing generally appear once or twice every few paragraphs, and the time constraints make it impossible for them to write anything long enough for multiple meaningful uses of authorial stance. A writing exercise also would not accurately judge if the students had met the objectives of the class, as it would only display what formulaic language they chose, not why they chose it or what similar options they considered and discarded. Thus,

for this project, a novel form of evaluation was developed: the re-writing exercise.[1] The re-writing exercises assess the students' abilities to identify the phrases and their uses and produce fluent idiomatic re-phrasing.

In the rewriting exercises, the students are given sentences drawn from the corpus described in Chapter 1, to ensure that they were accurate representations of academic writing. Each sentence contained a highlighted use of one of the phrases covered in class. The prompt asks the student to either identify the function of the phrase or reword part of the sentence, to, for example, indicate a more positive stance, increase the forcefulness of a statement, or indicate a higher degree of certainty or doubt. The total set of questions can be found in Appendix G. The students completed these exercises when the first class started on the first day, at the end of class on the second day, and a week later at the end of the short course.

The students' rewording were rated qualitatively before being quantitatively summarized. In the first round, correct and incorrect responses were noted, and the incorrect responses were labeled with the type of error: no answer, misunderstood question, lack of explanation (if the question called for an explanation of their answer), different part of speech (if, for example, a verb was replaced with an adjective rather than another verb). In the quantitative summary, the types of responses were simplified to three groups: no answer, correct, and incorrect. Partially correct answers were included under correct. Partially correct answers included answers were students and written down multiple synonyms in answer to a question and only some were true synonyms; and when they correctly identified stance or part of speech, but did not provide reasoning. The results from the students' answers on the pre-, post-, and delayed posttest were compared. The results of the tests are reviewed in Chapter 5.

[1] Many thanks to Dr. Matthew Allen, Assistant Director of Curriculum and Instruction at PLaCE, for suggesting this approach to assessment.

CHAPTER THREE: CORPUS RESULTS

Introduction

The first research question was as follows:

(1) What formulaic language is uncovered through a corpus study of engineering research articles?

(a) Do these results agree with or deviate from previous studies of academic writing in general and engineering writing in particular?

(b) Which formulaic sequences are most useful for NNS students seeking to develop their academic writing?

To best address the main question, I will describe the results of the corpus analyses while touching on Research Question 1(a) where relevant. The bulk of the findings consist of classifying and analyzing the 21 p-frames, as they had far more variability and syntactic complexity than the n-grams and fell into three distinct groups. This is followed by a summary of the n-grams and an explanation of the networks between the phrases in the corpus. Previous literature is discussed along with the results that most directly relate to it. For example, the section on "Noun & Preposition P-frames" concludes with a discussion of the connections to Biber & Gray's (2010) research on the increase of nominalizations in academic writing. Following the corpus findings, the connections to literature are briefly summarized to fully answer Research Question 1(a). The final section addresses Research Question 1(b) and leads to the next chapter on materials development.

N-Grams and P-Frames in This Corpus

P-Frames

I will discuss the results of the p-frames first, as the patterns that emerged among them were salient in categorizing the n-grams. Using the methodology laid out in Chapter 2, "N-Gram and P-Frame Identification," 30 p-frames were identified via the corpus search; 22 of these passed

the semantic homogeneity test. Table 1 shows the information for these 22 p-frames, including the syntactic feature of the variable slot that allowed it to pass the semantic homogeneity test.

Table 1: P-Frames in Corpus

P-frame	Occurrences	Range	# variable tokens	Main Type of Variable slot
be * as a	570	199	133	Passivized verb
be * as the	533	201	199	Passivized verb
be * by the	985	249	267	Passivized verb
be * for the	496	194	188	Passivized verb
be * from the	506	193	149	Passivized verb
be * in the	1272	252	383	Passivized verb
be * on the	492	190	165	Passivized verb
be * to be	784	208	110	Passivized verb
be * to the	1162	247	195	Passivized verb
can be * to	444	192	117	Passivized verb
it be * that	597	171	93	Passivized verb
it be * to	568	195	137	Passivized verb
use to * the	565	190	189	Passivized verb
and the * of	819	227	386	Abstract noun
for the * of	695	212	276	Abstract noun
in the * of	1747	250	323	Abstract noun
of the * be	1007	232	449	Abstract noun
on the * of	832	218	317	Abstract noun
that the * of	571	198	281	Abstract noun
the * in the	850	216	387	Abstract noun
to the * of	1083	245	447	Abstract noun
it be * to	568	195	136	Evaluative adjective

This first examination of the p-frames points out several interesting characteristics. Previous analysis of p-frames in business writing found that, using intuitive categorization, the most common p-frames could be split between those used in general English and those specific to a business context. For example, “a * of the” was identified as general English, and “the class *

share" and "net asset value *" were business specific (Fischer-Starcke, 2012). However, none of the 30 p-frames identified in this corpus included engineering-specific fixed words. Not only were the words generic; all words in the p-frames were function (closed-class) words rather than content (open-class) words, with the one exception of the "be" and "use" verbs. As "be" is used to create a passive, it is, arguably, also a function word. The other words are all articles (the, a), prepositions (of, by, from, in, to, with, as), and pronouns (it, that).

Before claiming that this lack of engineering words and predominance of function words was significantly different from Fischer-Starcke's findings, there were two possible methodological reasons for the difference to investigate. The first was that, unlike Fischer-Starcke, I did not include p-frames with variable slots on the external edges, as these are no different than n-grams. Thus, I ran the p-frame search again, allowing for external variable slots. The results overlapped completely with the n-grams discussed in the next section, and none of them had engineering-specific vocabulary. Second, it was possible that the variety of concentrations in the engineering corpus obscured technical p-frames because the technical vocabulary differed across sections; this was checked by analyzing each disciplinary corpus individually, searching for 4- to 6-frames with a dispersion above .8. However, the results in each of the four sub-corpora were nearly identical to the results of the entire corpus; and none of them featured p-frames with technical vocabulary.

This leads to the conclusion that what writing in engineering journals has most in common is not technical idioms and vocabulary; it is phrasal structures. Previous efforts to identify common academic phrases (such as in Simpson-Vlach & Ellis, 2010) have had much lower frequency and dispersion requirements, and thus have discovered idiomatic phrases. But the p-frames approach uncovers the much more common syntactic structures that underlie much of academic writing. Words from grammatical (closed) classes primarily serve syntactic functions, as opposed to the content provided by the open classes such as nouns and verbs (Harley, 2006: 118-119). Essentially, what the p-frame search uncovered was a set of grammatical structures that are used heavily in academic engineering writing. These structures could be neatly divided into two groups, which we will call *passive verb* p-frames (with 12 p-frames) and *noun + preposition* p-frames (with 8 p-frames). Corpus examples of each of the p-frames within both groups are

shown in Table 1Table 2. The following sections will discuss the *passive verb* p-frames, the *noun + preposition* p-frames, and finally, the one outlier is "*it be * to.*"

Table 2: Corpus Examples of *Passive Verb* and *Noun + Preposition* P-Frames

Phrase	**Examples from Corpus**
*be * as a*	"etoclopramide *is used as a* short-term treatment of heartburn" "The in-person training *was suggested as a* way to help lower barriers"
*be * as the*	"where the current disturbance *is chosen as the* band-limited white noise" "Although the latest test results *were described as the* most important,"
*be * by the*	"the other quantities *are calculated by the* solver" "Evolution *was measured by the* Hitachi S4 scanning electron microscope"
*be * for the*	"alleviated concentrations… *were observed for the* neutrophil-NP" "the SPH *are utilized for the* deformation behavior analysis"
*be * from the*	"The model of the vibrational dynamics *is borrowed from the* work of Esbrook et al" "The plate angle *is derived from the* height measurement"
*be * in the*	"Although various distribution network issues *are addressed in the* above literature," "When T1 and the TOI *are located in the* abnormal regions,"
*be * on the*	"SIMS measurements *were started on the* randomly selected analysis areas" "all systems *are trained on the* corresponding train set"
*be * to be*	"the coupling ratio *is estimated to be* 0.24%" "given the long time span of the data, the bias *is expected to be* marginal"
*be * to the*	"Both of our methods can *be related to the* surface registration approaches" "the plasma flow *is introduced to the* domain from the left-hand side boundary."
*can be * to*	"MUT can *be implemented to* use either capactive (CMUT) or piezoelectric" "Algorithm 2 can *be shown to* exhibit the same desired properties"
*it be * that*	"*it is envisioned that* a kinetic impactor would be jettisoned" "*It was observed that* increasing the aspect ratio decreased the length of the core;"
*use to * the*	"the classifiers are *used to classify the* unseen data" "It is *used to determine the* inflow conditions for the second sphere"
and the # of	"H corresponds to the proton consumption *and the amount of* hydroxo linkages" "they alter the interpretation *and the generalizability of* the experimental results"

Table 2 continued

for the # of	"a new sufficient condition *for the stability of* the peak covariance was established." "*For the sake of* simplicity, we merge the quarterly reports of every three months"
in the # of	"#4 is dominated by the cost of the SVD involved *in the computation of* SVT" "(See definition 8.14 *in the work of* Bohner and Peterson)"
of the # be	"The concept *of the method is* illustrated in Fig. 1. "a distributed point-mass model *of the spacecraft is* developed to reflect this"
on the # of	"reduce the influence of DRIE process variations *on the accuracy of the* frequency" "this problem formulation relies *on the validity of* Assumption 3"
that the # of	"Raman spectroscopy confirms *that the presence of* strain coupling…is strongest" "Information processing theory suggests *that the value of* information systems increases"
the # in the	"one must consider *the uncertainty in the* estimate" "Fig. 21 illustrates *the difference in the* temperature modeling"
to the # of	"low rank assumption is difficult to be satisfied due *to the existence of* tail labels" 'the average thickness was unchanged compared *to the thickness of* the original"

Passive Verb P-Frames

Role of the Passive

The use of "be" to indicate a passivized verb accounted for 11 of the 12 instances of p-frames with *be*: for example, *be * as a* most commonly occurred with the variables, "used, defined, considered," and *be * to be* most commonly occurred with "assumed, found, considered." Since the p-frame *use to * the* also primarily had passive verbs in the variable slot, it was included in this group. Table 3 gives the percentage of the total occurrences of each p-frame where the variable slot was a passive verb. It be * to is included to show that, despite having a similar appearance to the other p-frames, it did not have enough verbs in the variable slot to belong. The verb count was created by first tagging all the items in the variable slot with WMatrix's Part of Speech Tagger, and then manually resorting the verbs that it had tagged as nouns (WMatrix chooses Noun before Verb when tagging ambiguous words like 'train, design, convert', so the verbs did not have to be checked).

Table 3 gives the percentage of the total occurrences of each p-frame where the variable slot was a passive verb. *It be * to* is included to show that, despite having a similar appearance to the

other p-frames, it did not have enough verbs in the variable slot to belong. The verb count was created by first tagging all the items in the variable slot with WMatrix's Part of Speech Tagger, and then manually resorting the verbs that it had tagged as nouns (WMatrix chooses Noun before Verb when tagging ambiguous words like 'train, design, convert', so the verbs did not have to be checked).

Table 3 gives the percentage of the total occurrences of each p-frame where the variable slot was a passive verb. It be * to is included to show that, despite having a similar appearance to the other p-frames, it did not have enough verbs in the variable slot to belong. The verb count was created by first tagging all the items in the variable slot with WMatrix's Part of Speech Tagger, and then manually resorting the verbs that it had tagged as nouns (WMatrix chooses Noun before Verb when tagging ambiguous words like 'train, design, convert', so the verbs did not have to be checked).

Table 3: Passive Verb P-Frames

P-frame	% passive verb
be * as a	90%
be * as the	69%
be * by the	93%
be * for the	67%
be * from the	75%
be * in the	77%
be * on the	86%
be * to be	88%
be * to the	55%
can be * to	94%
it be * that	62%
it be * to	17%
use to * the	98%

This predominance of passive verb constructions in academic writing is a commonly highlighted feature of academic writing and is known to cause difficulties for NNS writers. This is especially true of engineering and science writing, which employ "about four times more passive bundles [than linguistics and business texts], often followed by a prepositional phrase marking a locative or logical relation," according to Hyland's 2012 corpus study (p. 164). This predominance is difficult for NNS, not because of the grammatical construction, but because of the difficulty in identifying its appropriate rhetorical functions. Öztürk & Köse (2016) find that Turkish writers import awkward-sounding passive constructions from their L1s. Both Conrad (2018) and Wei & Lei (2011) found that advanced Chinese EFL learners overused passives in their writing and hypothesized that this stemmed from an attempt to sound impersonal while writing. Oakley (2002) emphasized the difficulty NNS students have in mastering these forms when he identified the six different rhetorical functions that one passive phrase had in his corpus. Thus, considered pedagogically, the existence of a high number of passive constructions in academic writing is not a barrier to students; rather, the rhetorical functions of these constructions are the barrier. To better understand their rhetorical functions, we will now discuss the semantic categorization of the verbs in the variable slots.

Semantic Trends in the Verbs in Passive Verb P-frames

WMatrix provided semantic as well as part-of-speech tags for each verb, but its categorization needed amendment; for example, key verbs like 'design' and 'depict' were labeled "Arts and Crafts.' Thus, where necessary, WMatrix categories were merged into my categories, which were based on the most common academic use of those items (see Appendix C for full details on which categories were combined and renamed, and examples of the verbs in each category). Any category with less than then 10 total instances in the corpus was deleted; for example, there the six instances of 'suture, inject' that were categorized as "Medical," and these are not included; this removed 306 of the 6,297 tokens. The 5,991 tokens remaining were split between 26 semantic categories, shown in Figure 3.

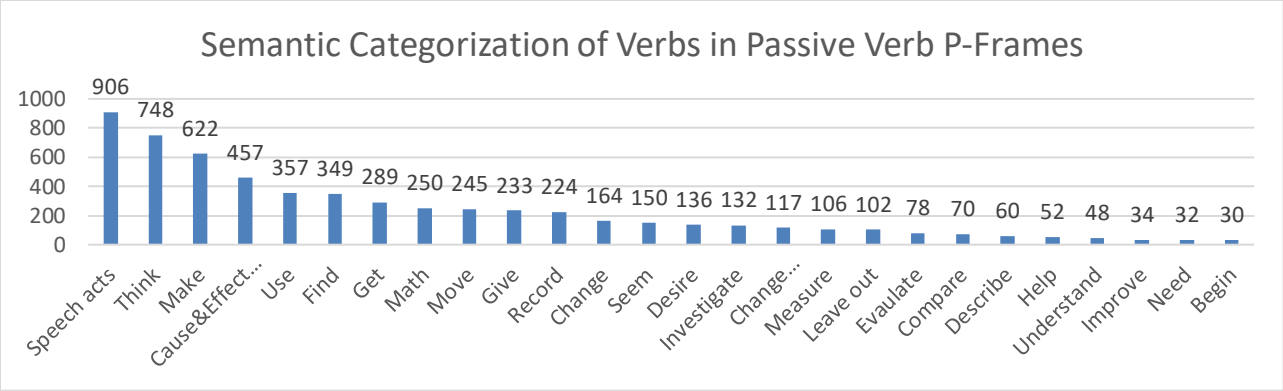


Figure 3: Semantic Categorization of Verbs in *Passive Verb* P-Frames

Surprisingly, as with the fixed words in the p-frames, the words in the variable slot were not predominantly technical engineering vocabulary. The technical vocabulary was broadly represented by the Make, Math, Measure, Move, Change, and Change Quantity categories, as these included verbs such as "solder, calibrate, filter, transpose." However, combined, these categories represent only 20% of the verbs. Instead, the largest theme among the categories is that of knowing and transmitting knowledge. The categories of "Speech Acts, Think, Find, Record, Investigate, and Understand" together accounted for 40% of the variable slot verbs. The five most frequent verbs in each category, along with their frequencies and additional examples for each category, are represented in

Table 4.

Table 4: Knowing Verbs in Passive Verb P-frames

Category	Verbs	Count
Speech Acts	define	234
	apply	150
	describe	94
	attribute	64
	introduce	47
	refer, discuss, say, report, propose, explain, confirm, guarantee, interest, predict, summarize, specify, suggest, mention	
Think	assume	210
	consider	188
	base	114
	expect	76
	estimate	70
	regard, formulate, suppose, concentrate, believe, deem, visualize, think, extrapolate, generalize, conclude, resolve, reconstruct,	
Find	find	181
	see	90
	observe	83
	know	66
	identify	22
	Others: detect, reflect, indicate, demonstrate, display, recognize, corroborate, perceive, discover, reveal	
Record	show	202
	represent	63
	observe	42
	write	29
	denote	22
	map, rewrite, record, note, notice, document, encode, highlight, address, list, register	
Investigate	investigate	27
	focus	24
	study	23
	analyze	21
	test	18
	Others: assess, review, seek, check, examine, inspect	
Understand	interpret	24
	understand	11
	infer	6
	realize	5
	deduce	2

Together, the occurrences of the passive knowing verbs in these p-frames account for 2,874 occurrences. While these p-frames indicate the most common fixed surroundings of this set of verbs, these verbs appear, preceded by a passive *be*, a total of 11,039 times in the corpus, meaning that passivized verbs of knowing and understanding are used on average 41 times in every article.

This use of passive verbs of knowing also paints a fuller picture of engineering writing. Previous studies of n-grams in engineering writing have emphasized that they are used to connect the

visuals to the text and organize the text (Hyland, 2012); examples include "are shown in Table X" or "are summarized in Figure Z". (Hyland, 2008b). Hyland categorizes these as text-oriented phrases and point out that engineering uses fewer participant-oriented bundles than the linguistics and business texts (Hyland, 2008b).

In contrast, the *passive verb* p-frames provide thousands of examples of formulaic language that is not text-oriented. Examples (1)-(4) below are all participant-oriented, as the writers lay claim to previous knowledge and guide their reader to the desired conclusions. (5) links the text to a visual, but even this is still participant-oriented, as it asks the reader to reach a certain conclusion.

(1) *It is thought that* the jetting force serves as the mechanical removal mechanism.
(2) *It is known that* the main limitations of the SRS based pharmacovigilance include under-reporting, duplicate reporting and lack of specific control group for comparisons.
(3) It can *be concluded from the* stress analysis… that the tube is not only subjected to the vertical upward bending force, but also the axial thrust force.
(4) This idea proves to be particularly effective when FLQR is adopted within the MPC framework, in which case our approach can *be interpreted as a* tracking strategy for the optimal final time while solving standard LQR problems.
(5) This also has a detrimental effect on performance as it can *be noted on the* oscillatory behavior in Figure 12.

While most of these *passive verb* p-frames are too infrequent to appear on an n-gram search, the p-frame search allows us to group these functionally similar verbs together and realize how often engineering writing actively constructs a discourse of knowing and learning. These are not verbs used to report experiments; instead, these verbs have complex rhetorical functions – establishing authority, building a niche, linking visuals to text, appropriately making and hedging new claims, creating arguments. Especially important are the participant-oriented functions: building a shared knowledge base with the reader, highlighting important aspects of one's own findings, and using the work of others to create a background for one's own work. We will discuss these functional uses more in the functional categorization section.

As mentioned earlier, this semantic analysis explains why the passive construction is difficult for NNS to use. The rhetorical functions described above are the very ones that have been found to be difficult for NNS writers (Bloch, 2010; Wei & Lei, 2011; Hyland & Tse, 2015). NNS students are taught that scientific writing should be objective and impersonal, but they are not taught how expert writers use objective-sounding passive constructions to establish their authorial stance, highlight interesting points, and engage the audience.

Noun & Preposition P-Frames

The use of *be* + passive verb was not the only similarity shared by many of the p-frames. Eleven of the 30 original p-frames, and eight of the 20 p-frames that passed the semantic homogeneity test belong to the *noun + preposition* group. Six of these include "*the * of"* within their four-frame. "*With the * of,*" and "*the * of the*," are examples of discarded p-frames, while "*for the * of,*" and "*that the * of*" were kept. The other two members of the *noun + preposition group* are *of the * be* and *the * in the*.

While previous studies of engineering writing have found a "greater real-world, laboratory-focused sense of writing," (Hyland 2008b), this is not entirely true of this set of p-frames. Examination of the nouns that filled these slots revealed several key features: first, that they are predominantly abstract nouns; second, that they feature many nominalizations, and third, that they nouns fall into several broad semantic categories that again are essential in incorporating author stance and author opinion into objective-sounding prose.

Abstraction and Nominalization

We will look at abstractions and nominalization first, as these semantic and morphological features are linked. Notably, 73% of the categories, referred to abstract items. To determine abstractness, every semantic category uncovered by WMatrix was classified as abstract or concrete, following the definition in Abrams & Harpham (2011), where concrete "denotes a particular person or physical object" and abstract "denotes either a class of things or else… qualities that exist only as attributes." (p. 60) There were a few categories with both abstract and concrete words; these were labeled as "Mixed," and a rough estimate of the abstract words was

used to determine what percentage of the category would count towards the count of abstract nouns. Table 5 shows the highest frequency categories and their classification as concrete or abstract. This serves as a reminder that though engineering writing is focused on explaining laboratory studies, much of the writing is abstract and, correspondingly, difficult.

Table 5: Abstraction in *Noun + Preposition* P-frame Variable Slots

Top Categories	Abstract or Concrete	Examples	Number of Items
Kinds, groups, examples	Abstract	case, form, instance, kind, sample	436
Location and direction	Abstract	accessibility, orientation, position, trajectory, transfer	420
Technical	Mixed	bevel, cardinality, cathode, conductivity, emitter, etiology	400
Language, speech and grammar	Abstract	abbreviation, term, terminology, usage, work	304
Cause & effect/connection	Abstract	basis, effect, impact, relation, result	286
Quantities	Abstract	amount, extent, portion, set, sum	257
Objects Generally	Concrete	ball, basin, hammer, instrument, spring	249

This abstraction is closely linked with a second trend in the data: verbal and adjectival nominalizations, which were tabulated because they are discussed extensively in Biber & Gray (2010). While abstraction is a semantic categorization, nominalization refers to morphologically deriving nouns from verbs and adjectives, and these words form a subset of the abstract words. Biber & Gray find a significant increase over time in intangible nouns derived from verbs or attributes in academic writing, so I looked for evidence of similar intangibility in these nouns by searching for common morphological noun endings that denote adjectival and verbal roots (-ion, -ment, -nce, -y, and -ness) and manually adding the others nominalizations in the dataset. This returned 487 nouns, which collectively accounted for 34% of the slots (Table 6).

Table 6: Nominalization in *Noun + Preposition* P-Frame Variable Slots

Percentage	Type	Examples	% of Words
Verbal Nominalizations	"-ion verbal nominalization"	production, interaction, separation, generation	15.7%
	-ment verbal nominalization"	development, improvement, management, treatment	2.5%
	-nce verbal nominalization"	existence, dependence, governance	4.0%
	other verbal nominalizations	choice, removal, recovery, release	4.4%
Adjectival Nominalizations	-y adjectival nominalizations	variability, mobility, convexity	4.4%
	-nce adjectival nominalizations	absence, distance, convenience	4.0%
	-ness adjectival nominalizations	thickness, skewness, hardness, inertness	0.7%
Total			**34.7%**

The data develops and extends several of the points in Biber & Gray, 2010. They argue from a variety of data points that nominalization has increased dramatically in pre-modifying nouns (nouns that modify other nouns), and that a rise in prepositional phrases has contributed to "the loss of explicit meaning" and the decrease of verbs in academic prose. What this family of p-frames reveals is a combination of these two trends. The nominalizations are not unique to pre-modifying nouns; they are also present in the increasing use of prepositional phrases.
Biber & Gray's overall argument that pre-modifying nominalizations reduce the use of verbs and obscure the relationships between elements holds true in these prepositional phrases.

First, the predominance of the verbal nouns over the adjectival nouns (25.6% of slots as opposed to 9.1%) point to the way these nominalizations are replacing verbs. The authors of these engineering articles are replacing verbs with verbal nominalizations more than they are replacing adjectives with adjectival nominalizations.

Second, Biber & Gray argue that pre-modifying nominalizations (such as "reprisal raids") completely obscure the relationship between the two nouns; but prepositional phrases ("raids of

reprisal") are scarcely clearer. Sentence (1) below, from an aerospace text, demonstrates how these embedded relationships lead to highly compact, inexplicit meaning in the same way as pre-modifying nominalizations. It includes both the pre-modifying nominalizations described in Biber & Gray ("separation bubble, "pressure gradient"), and several italicized nominalizations embedded in prepositional phrases, which have been indicated with brackets.

(1) It appears that the adverse pressure gradient that is produced [by the *addition* [of the flow [through the rocket ejector]]] is primarily responsible [for the *separation* [of the boundary layer [on the inner wall [of the duct]]]] and thus the *production* [of the *separation* bubble]."

In this sentence, the author has embedded actions into prepositional phrases four times, thus stringing a great deal of information and action into a small space. While this sentence may be a more extreme example, it is neither a poorly written example, nor was it hard to find.
Thus, before exploring the semantic categories of the variable slots in this set of p-frames, we have found further support for the arguments that academic writing is abstract, dense, and inexplicit, and requires expertise to read and, particularly, to write. The emphasis on "hard science" in engineering writing does not exempt it from these challenges.

Semantic Categories

Since the nouns are predominantly abstract and highly nominalized, we expect the semantic categories to be broadly abstract as well, and they were. As WMatrix identified 232 categories, there was much more variety in than in the *passive verb* p-frames. However, some of these were inappropriately sorted; for example, the category "Calm" consisted of the word, "rest," but in this corpus it is used in the sense of "remainder." These items were manually resorted, and categories with fewer than ten occurrences were removed, leaving 62 categories. The majority of these categories could be combined in 4 broader groups that together accounted for 79% of the variable slots. The full details on the categorization adaptions are in Appendix C, with example words, and this data is summarized in Table 7.

Table 7: Semantic Categorization of *Noun + Preposition* P-frame Variable Slots

My Category	Percentage	WMatrix Categories	Examples
Research	29%	Investigate, examine, test	research, assessment, analyze
		Knowledge	database, identification, information
		Language, speech, grammar	terminology, vocabulary, abbreviation
		Mental Objects	system, method, target, pattern
		Paper documents and writing	paper, flowchart, graph, text
Evaluation	22%	Likelihood	potential, viability, prospect
		Quantity	increase, reduction, peak, scarcity
		Cause & Effect /Connection	result, impact, effect
		Comparing	norm, variance, comparison
		Evaluation	quality, accuracy, error, validity
		Helping/Hindering	recovery, benefit, drawback, obstacle
Engineering	16%	Anatomy and physiology	eye, pore, tissue, optic
		Business	vendor, office, retailer
		Electricity	anode, radar, voltage
		Medical	hospital, physician, diagnosis, injection
		Abbreviations	CSA, RSE, ATP
		Objects generally	bundle, pipelines, pendulum
		Science & technology	engineer, topography, wavelength
		Substances & materials	chemical atom, granite, substrate
		Technical	flowfield, accelerometer, ferrofluid
		Flying & aircraft	trajectory, asteroid, Jacobean
		Light	light, beam, laser
		Vehicles and transport	bumper, cart, transport
Math	12%	Mathematics	ratio, computation, equation, sum
		Measurement	radius, diameter, gauge
		Shape	sphere, spiral, geometry
Total	**79%**		

As with the p-frames, the slots are remarkable for the degree to which they do *not* refer to technical information. The two groups "math" and "engineering" accounted for 28% of the slots and covered the categories that one would expect to find in unusually high frequencies in engineering research articles. Some of them clearly correspond to the topics of the sub-disciplines; for example, "flying & aircraft" with aerospace engineering, "electricity" and "light" with electrical engineering, "medical" with the biomedical emphasis in both industrial and mechanical engineering. While the mathematical variable slots were most common in aerospace articles, again, their overall presence in all four disciplines is expected.

However, these together accounted for less than a third of the variable slots. 51% were occupied with two distinct yet related categories: the first, accounting for 29% of the nouns, were general words related to research, as in the following sentences.

(1) most *of the literature is* dedicated to measurements under stagnant flow configurations.
(2) Algorithm 2 can… get better inner estimates of domains of attraction than the method *in the work of* Luk and Chesi.
(3) An edge in a directed graph denotes that the node VJ has access *to the information of* node VI.
(4) the components obtained by employing the JIVE *in the analysis of* grossly corrupted data may be arbitrarily away from the true ones.
(5) If the owners of all the other entries *of the database are* colluding with the enemy…

These broad terms, which appear in academic research regardless of topic, account for a greater percentage of the variable slots than the technical math and engineering words. They serve two primary rhetorical uses: first, as in (1) and (2), they refer to the works of others and thus set up the niche and purpose of the authors' work. Second, as in (3) through (5), they are ways to discuss the authors' own work. The functional categorization of these two purposes will be discussed in the section on functional categorization, but for now it is enough to note that one of the central uses of these prepositional phrases is reference to others' and own work.

The evaluation words, which account for 22% of the words, are words which judge or measure an aspect of something. The words include primarily positive attributes (applicability,

effectiveness, importance, fidelity) and negative attributes (negligence, drawback, difficulty), but also evaluative terms for likelihood (possibility, capacity) and quantity (scarcity, minimum, maximum) and effect (impact, implication, result). In the context of these p-frames, they allow the author to express their opinion.

(1) Access to space would yield a giant leap *in the viability of* many proposals
(2) These discrepancies can be especially problematic… given the multidisciplinary nature of their field *and the lack of* unified theories.
(3) The input design would depend *on the quality of* the chosen estimate.
(4) The event-triggered parameters… are all constants… which leads *to the difficulty of* selecting optimized initial parameters.
(5) Fig. 9… does not account *for the impact of* these errors on the control system.

These evaluative words are thus a powerful way that authors can communicate authorial stance towards the subject. Academic writers are generally expected to maintain impersonality and objectivity in evaluating their own work and the work of others, and these nominalized, abstract evaluations provide an objective-sounding way of arguing for the value of their work and commenting upon the work of others.

As with the *passive verb* p-frames, we find that a large portion of the structured phrases used in engineering writing reference previous work and evaluate one's own work. Both are done in large part through abstract and nominalized nouns embedded in prepositional phrases. These verb-less constructions are not only frequent (and frequently interlinked); they serve essential, if sometimes hidden roles in building author credibility and explaining the importance of one's work.

It is worth noting that there was not much variation between the *noun + preposition* p-frames in terms of which semantic categories filled their slots. Table 8 shows the variation between the eight p-frames, including the standard deviation for each category; light red and green indicate one standard deviation above or below the average frequency, while the two dark green blocks indicate more than two standard deviations above the average frequency. *The * in the* had much

more engineering items in the variable slot, and *of the * be* featured more math items. The engineering items in *the * in the* tended to denote a spatial relationship, with examples like, "*the spacecrafts in the* asteroid-fixed frame" "*the electrons in the* current collector," or "*the flowfield in the* flame duct." *Of the * be* had an unusually high percentage of math variable slots, as it was used to refer to mathematical processes, for example: 'most *of the computation is* integrated," and "the rotation *of the constraint is* only applied in this window."

Table 8: Semantic Categories in Noun + Prep P-frames

Category	***of the * be***	***to the * of***	***that the * of***	***the * in the***	***on the * of***	***for the * of***	***in the * of***	***and the * of***	**Total**	**SD**
Research	14%	32%	51%	22%	38%	14%	16%	19%	**29%**	12%
Evaluation	32%	24%	12%	8%	26%	35%	19%	34%	**22%**	9%
Engineering	12%	8%	9%	47%	9%	9%	30%	9%	**16%**	13%
Math	20%	11%	11%	10%	9%	14%	11%	13%	**12%**	3%
Other	22%	25%	17%	12%	17%	27%	25%	25%	**21%**	5%
Total	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*	***100%***	

*The Final P-frame: It be * to*

The final p-frame to consider is "*it be * to*." This p-frame is not only remarkably different from the other ones it the contents of its variable slot; it also contains a syntactic structure that is difficult for NNS, and has several specific, distinct rhetorical functions.

The variable slot differs from the other p-frames with *be*: 87.5% of the variable slots are filled with adjectives (12.5% of the slots contain passive verbs). These adjectives fit neatly into two categories: adjectives of possibility and adjectives of importance. In Table 9, the examples of each category have been listed on a semantic cline, from most possible to least possible and from most important to least important.

Table 9: Semantic Categorization of *it be * to* Adjectives

Type	Examples	% of slots
Adjectives of Possibility	easy, convenient, simple, straightforward, feasible, possible, hard, difficult, suboptimal, impractical, inappropriate, infeasible, unlikely, impossible	41%
Adjectives of Importance	essential, critical, crucial, imperative, vital, necessary, optimal, important, worthwhile, useful, advantageous, beneficial, helpful, interesting, reasonable, meaningful, noteworthy, instructive, meaningless, trivial, unnecessary, unwise	39%
Other Adjectives	approximate, close, equivalent, likely, sufficient	7%
Passive Verbs	used, applied, designed, set, aimed, expected, required, known, referred, restricted	13%

The limited slot options available to this p-frame hints at its rhetorical functions. This particular p-frame is not a broad syntactic structure that enables a variety of academic language, as with the others. Rather, it is usually used to perform three organizational and meta-textual roles. The first is to introduce a topic, where "*it is helpful to* introduce the estimation error" means, "I will introduce the estimation error next." The second is to draw attention to a claim or point, usually with an adjective of importance, as in "*it is important to* point out that since the fluorescence-based technique is the only available approach…" The third use is to defend or explain methodological choices with adjectives of importance or possibility, as in "*it is necessary to* create a metric in a functional space," or "*it is reasonable to* ignore their influence to derive an approximate analytical solution."

This structure is particularly of interest to us because NNS speakers struggle with the anticipatory *it* construction in general, avoiding it because of the difficulties in employing it correctly (Chen & Baker, 2010, Lee & Chen, 2019, Wei & Lei, 2011). This extends even to published research articles (Perez-Llantada, 2014). The syntatic difficulties overlap neatly with difficulties with the structure's pragmatics, as NNS writers tend to avoid phrases that would communicate opinion or personal interest (Conrad, 2018; Chen & Baker, 2010, Wei, 2007). Just as with the evaluative nouns in the *noun + preposition* p-frames, however, this is a common

construction used in academic writing to argue for a work's worthiness or interest while maintaining an impersonal authorial voice. Thus, it is a valuable structure for NNS to master. It falls clearly under the participant-oriented (interpersonal) categorization (Hyland 2008b, 2012), so it will be discussed further under the section on functional categorization.

N-Grams

The n-grams results were much simpler than the p-frames results, but they both emphasized some of the points made in the p-frame analysis and served to validate Hyland's previous research on lexical bundles in electrical engineering writing (2008b, 2012) to three new engineering disciplines. The n-gram search with the parameters described in Chapter 2 resulted in 25 n-grams, 22 of which were kept; their information is shown in

Table 10. *Show in fig* and *be show in* were partially nested within *be show in fig*. They both had a frequency count that was higher than the 4-gram frequency count; this meant that every instance of *be show in fig* was included in the 3-grams, so *be show in fig* was removed. Two of the n-grams, *where be the* and *and is the,* were actually p-frames derived from common mathematical expressions, and thus were removed. Because all numbers and most of the single characters had been deleted from the corpus, the original mathematical use of, for example "where *y* is the,' showed up as the n-gram "where be the," and "and L is the…" showed up as the ng-ram "and is the" in the pre-processed corpus. *Be use for* and *be use to* partially appeared within the p-frames *be * for the* and *can be * to*, but their p-frame uses counted for only a small portion of their total appearances, so they were kept in the n-gram results. The final 22 n-grams are shown in Table 10, with their occurrences per million words (PMW), the percentage of articles in which they were found (range), and their occurrences PMW in each of the four branches of engineering writing. The final column gives the standard deviations among the four corpora's occurrences PMW.

Table 10: N-gram Results

N-gram	PMW	Range	AE PMW	EE PMW	IE PMW	ME PMW	SD
depend on the	151	53%	177	170	108	149	27
one of the	166	52%	185	153	194	121	29
show that the	206	61%	243	229	129	231	46
as well as	339	68%	262	392	352	342	47
according to the	204	55%	268	170	223	147	47
in term of	244	54%	239	321	187	221	49
there be no	132	51%	29	28	140	167	63
as show in	307	56%	343	211	312	391	66
be use for	141	50%	173	127	15	198	70
due to the	413	76%	314	338	354	494	70
in this paper	304	62%	275	360	146	231	77
use in the	155	51%	235	17	142	149	78
base on the	374	77%	304	310	112	288	82
note that the	190	53%	293	229	25	149	100
be show in	367	74%	354	267	194	473	104
be use to	414	83%	297	271	379	553	110
in order to	340	54%	133	321	443	244	113
with respect to	279	52%	441	125	231	193	118
in fig the	269	57%	408	129	192	391	122
the number of	394	66%	150	539	429	288	147
the effect of	380	68%	260	166	566	350	148
show in fig	667	69%	701	394	512	1031	241

Disciplines whose PMW are lower than the mean by more than one standard deviation are light yellow; values more than two standard deviations are dark yellow. PMW above the mean by more than one standard deviation are light green.

We will discuss the functional categorization of these n-grams together with the p-frames. However, we can note here that several of Hyland's claims about lexical bundles in electrical engineering have been validated in the fields of aerospace, industrial and mechanical engineering as well. Hyland found that "formulas and graphs are linked in routinely patterned, almost formulaic ways" (Hyland, 2012), and this is just as true here, where six of the n-grams are commonly used in this way. Table 11 gives examples from the corpus of these six n-grams.

Table 11: N-grams Linking Visuals in the Corpora

show that the	"The results in Table *7 show that the* network can exploit more complex lighting"
as show in	"Greedy search, *as shown in* [1], can achieve very good results in some cases"
note that the	"*Note that the* condition in item (1) follows since every maximal solution…"
be show in	"simulated crater geometries PAGOSA *are shown in* blue triangles"
in fig the	As shown *in Fig. 3a, the* orbit does not repeat"
show in fig	"a line through the time history of 0osc via least squares *as shown in Fig.* 4a."

Hyland based his claims on an analysis of electrical engineering texts; but in this corpus, electrical engineering lags behind other disciplines in the use of language that links visuals to text. When the use of these six visual-linking phrases is totaled for each of the four corpora, electrical engineering uses it significantly less often than aerospace and mechanical. Industrial has 1,486 uses PMW, slightly below electrical engineering's 1617. Aerospace and mechanical almost double that, with 2,729 occurrences and 3,031 occurrences PMW respectively. It is not surprising that the n-gram search should provide results similar to his lexical bundle search; it is the p-frame results that give new information and may lead us to adjust his claims about engineering writing.

Relations among Phrases in the Corpus

The primary goal in applying network analysis tools to the corpus linguistics results is to create better ways to visualize and present those results to non-linguists. However, network analysis has also developed measures for evaluating the relationships between nodes, and these can help us to

spot interesting patterns that are not usually available to traditional corpus methods (Baker, 2016). NodeXL, the tool used to create the visualizations, also provided statistical information about each "node" (the p-frames and n-grams) and each "edge" (the relationship between each p-frame and n-gram). These measures were originally developed to identify key players, potential communication breakdowns, influencers, etc., in social networks, so we must be careful in drawing too much from them. However, there are a couple salient points worth noting.

As discussed in the methods section, these statistics come from a corpus with only the p-frames and n-grams mentioned in the previous section where every other word had been replaced by a placeholder. Thus, it was possible to accurately measure the distance between the phrases without documenting their relationships to every word in the corpus. The information about their relationships was gathered twice: once with a window of 50 words, and once with a window of 200 words, thus looking for phrases that would be within a sentence or two of each other and phrases that would span paragraphs. The phrases in the first corpus composed 51% of the phrases in the corpus with the larger window, meaning there is substantial overlap in the findings. For the measure discussed below, there was no significant difference between the two analyses.

Both analyses of the corpora communicate the same picture. The phrases did not form dense networks; rather, they were spread across many scattered and often reciprocal relationships. This is validated by their high levels of correlation between the different measures for centralizations, as reciprocal relationships and low density have been found to correlate with high correlation between centralization measures (Valente, Coronges, Lakon & Costenbader, 2008). The four main measures of centrality used here (in-degree, out-degree, betweenness centrality, and eigenvector centrality) measure how central a node is to the graph in four different ways, but the top 50 vertices in each category for each analysis demonstrated almost perfect overlap. The top 50 on each list were primarily the n-grams, followed by the p-frames with the highest frequency variables, such as *in the* case *of*, *be* derived *in the*, and *in the* work *of*. This is also correlated with raw frequency; thus, the more common a phrase is, the more nodes it connects to.

Just as there were no phrases that were unexpectedly central, so there are not groups of phrases that co-occur at unusually high rates. This was tabulated through two measures: clustering

coefficients and closeness centrality. The clustering coefficient measures the degree to which a node's neighbors create a clique by measuring how many of the possible interconnections between a node's neighbors exist (Watts & Strogatz, 1998). The nodes with the highest clustering coefficients in the two analyses were primarily middle-frequency phrases that were connected only to the most-frequent n-grams. For example, *be* connect *to the* was linked to "*in order to, be use for, be show in*" (see Figure 4). Because these are such high frequency items, it is not surprising that they all linked to each other; but as each have 50+ other relationships, it is clear that "*be* connect *to the*," is not in any way part of a tight cluster. This trend holds for the other items with high clustering coefficients.

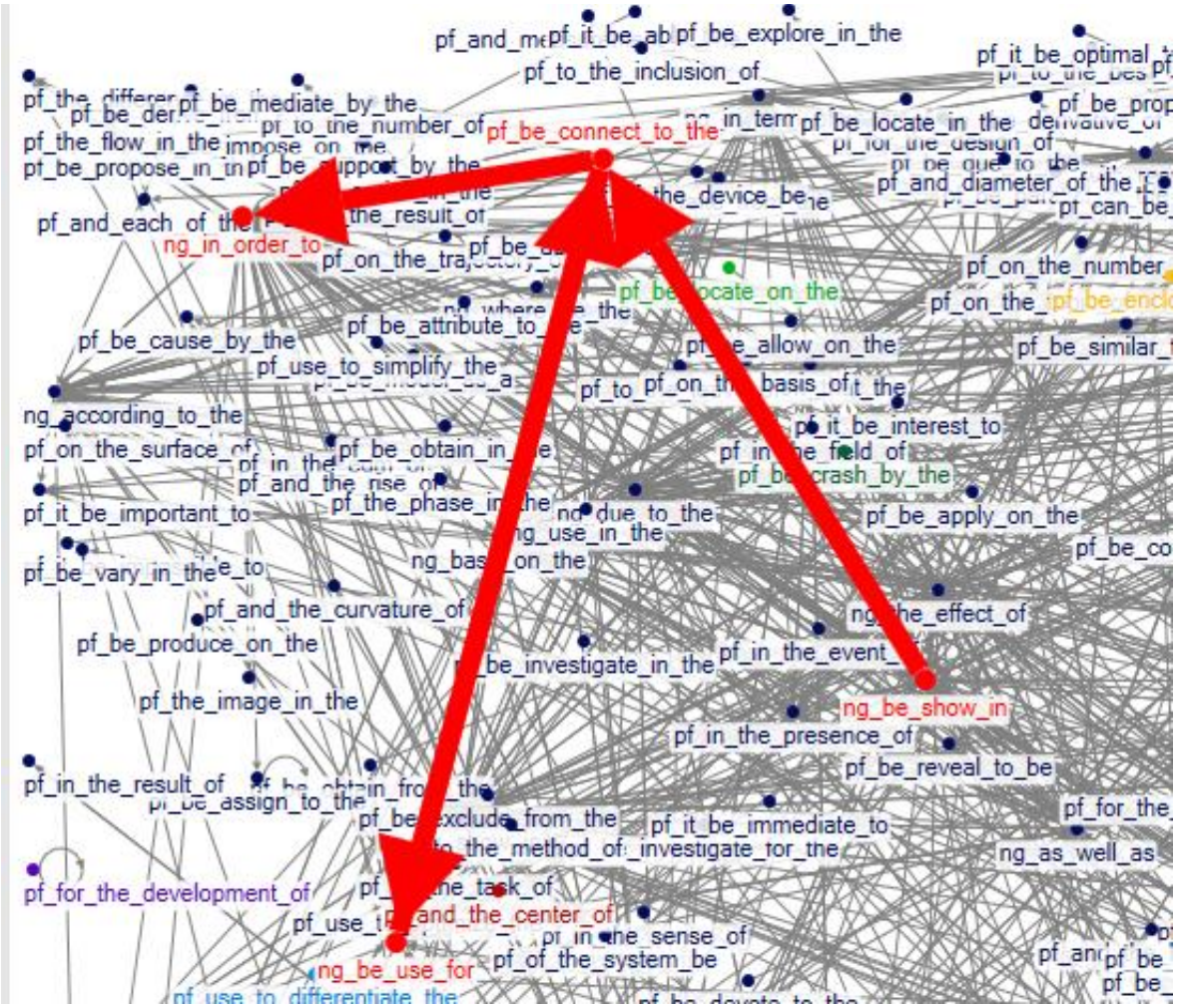


Figure 4: High Clustering Coefficient Node and Relationships

The closeness centrality, which measures the total distance between that node and all its neighbors, identified a few nodes in each corpus that were exclusively connected with only one another; but this revealed only a few pairs and triangles of nodes, such as the exclusive relationship of "*the* center *of the*" with "*be* tangent *to the*" and "*be* propagate *in the*" with "*be* detect *in the*." A corpus search confirmed that these cliques were the result of one author using them several times within a stretch of writing that did not have any of the other marked p-frames or n-grams.

The picture of the corpus that thus emerges is one of very even dispersal – the more high-frequency an item is, the more likely it is to have reciprocal relationships with other items and be interconnected. There does not seem to be strong cliques or other patterns, such as phrases that come primarily before or after other phrases. This has one primary implication for choosing phrases to teach. As the phrases seem to appear indiscriminately near and around each other, it does not make sense to teach them in groups based on the proximity. Rather, the syntactic, semantic and rhetorical patterns identified in other approaches will be more useful to NNS writers hoping to understand and use these phrases. The usefulness of network analysis is not in identifying connections between the phrases, but in identifying connections between unique phrases and their environments, as will be shown in the next chapter.

Functional Categorization

The Phrases Categorized

The final step was to categorize the p-frames and n-grams according to Hyland's functional categories (as described in Hyland 2008a, 2008b, 2012) and compare the results with his observations on lexical bundles in electrical engineering. While the n-gram results patterned nicely with Hyland's findings, the p-frame results provided novel information and revealed that, while engineering might have fewer partcipant-oriented n-grams, it makes up for this through frequent use of participant-oriented p-frames.

It is important to note that for the n-grams, my cutoff frequencies are much higher than for Hylands' lexical bundles. I required a dispersion score of .5 (i.e., that the n-gram appear in at

least 50% of the articles), and this meant that it had to occur a minimum of 60 times per million words (PMW) instead of Hyland's 30. This higher dispersion score was used to identify n-grams that were not only frequent, but also appeared in at least half of the total articles in the corpus. Though this meant I was classifying only 21 n-grams, instead of the 50 top bundles Hyland classified, the overall results were similar. Table 12 represents the results; of the 12,392 instances of n-grams, 31% were research-oriented, 58% were text-oriented, and 11% were participant oriented. Research-oriented n-grams are split evenly between procedural phrases and quantifiying phrases. The majority of text-oriented n-grams are concerned with linking visuals with the text, which was discussed above. Finally, there are only three participant-oriented phrases where the author states their own opinion or explicitly addresses the reader. This is similar to the findings in Hyland.

Table 12: Functional Categorization of N-grams

Research-Oriented	**3832**	**Text-Oriented**	**7146**	**Participant-Oriented**	**1414**
Procedure	***2023***	***Structuring Signals***	***4082***	***Writer's Stance***	***1040***
be use for	272	as show in	592	there be no	255
be use to	797	be show in	706	due to the	795
use in the	299	in fig the	517		
in order to	655	show in fig	1284	***Address Reader***	**366**
		show that the	397	note that the	366
Quantification	***1809***	in this paper	586		
one of the	319				
the number of	758	***Transition Signals***	**652**		
the effect of	732	as well as	652		
		Framing Signals	**2122**		
		base on the	721		
		according to the	393		
		in term of	470		
		with respect to	538		
		Resultative Signals	**290**		
		depend on the	290		

We cannot directly compare these percentages with Hyland's results, as he never gives precise numbers for electrical engineering research articles. This is because in one study he compares research articles across four disicplines to the genres of masters and doctoral theses, and in

another he compares three genres of electrical engineering (research articles, masters theses, and doctoral dissertations) to the same genres in business, science and linguistics (see Table 13). However, as he notes, research-oriented bundles across all disciplines decrease and text-oriented bundles increase as writers move from masters theses to dissertation to research articles as "this is the most discursively crafted and rhetorically machined genre of the three" (Hyland, 2008a, p. 58). Thus, the functional categorization of n-grams in ths corpus validates Hyland's claims about engineering writing by extending the corpus to include three new engineering disciplines.

Table 13: N-gram Results Compared to Hylands' Results

	Research-oriented	**Text-oriented**	**Participant-oriented**
This corpus	31%	58%	11%
All RAs in Hyland's corpus (Hyland, 2008a)	26%	60%	14%
All EE genres in Hyland's corpus (Hyland, 2008b)	48%	44%	8%

When Hyland's functional categorization is adapted to the p-frames, however, the data tells a markedly different story. As the functions of the p-frames varied by the type of word in their variable slot, the *be passive* and the *noun + prepositions* p-frames were split across functional categorizations according to their uses. Table 14 shows the totals.

Table 14: Functional Categorization of P-frames

Research-Oriented	**5385**	**Text-Oriented**	**2205**	**Participant-Oriented**	**6070**
Procedure	***4000***	***Structuring Signals***	***2205***	***Writer's Stance***	***3196***
PV with procedural verbs	2784	N+P with research nouns	2205	*It be * to*	568
N+P with engineering nouns	1216			PV with evaluative verbs	182
Quantification	1385			PV with evaluative adjectives	773
N+P with math nouns	912			N+P with evaluative nouns	1673
PV with quantifying verbs	473			***Sharing Knowledge***	***2874***
				PV with knowing verbs	2874

PV refers to the passive verb p-frames, while N+P refers to the noun + preposition p-frames.

For the *noun + preposition* p-frames, the ones with engineering variables belonged to the Procedure sub-category of research-oriented, the math nouns came under Quantification, the phrases with research nouns were used were Structuring Signals, as they referred to graphs, findings and methods, and the phrases with evaluative nouns were counted under Writer's Stance, as they revealed the writer's evaluations.

For the *passive verb* p-frames, the verbs that were semantically categorized as "Use, Make, Move, Change, Cause & Effect, Get, Give," belonged to Procedure, as they generally described experiments. The verbs under "Measure, Math, Change Quantity" came under Quantification under Hyland's scheme. The largest group of the *passive be* verbs, however, did not fit under any of Hyland's categories. These were the verbs of knowing (categorized under Speech Acts, Think, Find, Record, Investigate, and Understand by WMatrix) discussed in the previous section. Hyland described the text-oriented category as including "engaging with the literature, providing warrants, establishing backgrounds" (2008a, p. 58) Bit his top 50 results for each discipline and

genre in his three studies did not include any of these functions, so he did not provide a subcategory of text-oriented phrases to which they can belong.

These phrases with passivized verbs of knowing are also different in that the majority of them are *not* engaging with previous literature or using specific sources to explain the background or justify a procedure, which is the reason Hyland classifies them as text-oriented. Instead, as described in the section on these verbs, they are generally used to lay claim to general knowledge that the author presumes to share with the audience, to engage the reader, or to discuss the author's own work. Some verbs often have a primary use; for example, *know* usually refers to uncited shared knowledge (Example 1 below), and *note* often emphasizes a point the author wants the audience to pay attention to (Example 2 below). But verbs like *find* are generally split between the author's research and outside research, and arguably, is also used for drawing attention to points the author wants to highlight (Examples 3-4 below).

(1) Markovian jump systems *are known as a* special family of hybrid systems and stochastic systems. (shared knowledge)
(2) *It is noted that* these shape models were readily available [29], and no attempt was made to create a more optimal lower-fidelity model. (drawing attention and own results)
(3) The optimal value of Q maximizing Equation (60) *is found by the* critical-fractile approach. (own results)
(4) This *was found to be* the case in Ozer et al.'s (2014) evaluation of trust and trustworthiness across U.S. and Chinese cultures. (reporting specific finding)

Due to the passive nature of the verb, the knower is often not named or even immediately clear from context. Thus, rather than attempt to separate each verb out by their individual uses, it is most useful to classify them under the new category of Sharing Knowledge under participant-oriented phrases. These verbs are how authors both construct their own credibility and invite their reader to agree with their conclusions, thus making these phrases much more about projecting stance than structuring a text.

Finally, *it be * to* was the one p-frame that only belonged in one of the three categories, as the adjective inserted in the variable slot is usually a way for the author to express their opinion on a subject. It is primarily used to indicate importance and probability, and thus is used to defend methodological choices, state the value of a project, and draw reader attention to the author's main points.

These p-frames thus provide essential complementary data to previous work on the role of lexical bundles in academic writing. While many authors have noted that engineering writing values objectivity and thus features fewer overt bundles of author stance and audience engagement, the way that skilled engineering authors engage their audience and declare their opinions through a variety of p-frames has not been explored. These are valuable structures that are important for NNS writers seeking to establish their credibility and the worthiness of their findings.

Connections to Literature

In summary, this analysis of p-frames and n-grams in this corpus has connected to many points from previous corpus studies of academic writing, both engineering and general. The primary phrasal patterns in this corpus were syntactic structures which could be divided into passive verb structures and preposition and noun combinations. The passive verbs were frequently verbs concerning knowing and learning, serving interpersonal purposes. This helped to explain the findings that NNS often use passives awkwardly, even though they understand the syntax well (Wei & Lei, 2011; Oakley 2002; Öztürk & Köse, 2016; Conrad, 2018). Both the levels of abstraction and nominalization in the *noun + preposition* phrases provided evidence for Biber & Gray's claims that use of nominalizations and prepositional phrases is on the rise, that this is partially explained by a decrease in verbs, and that this contributes to denser, inexplicit prose. *It be * to* and *it be * that* are especially interesting p-frames for NNS because they use the dummy *it*, a syntactically difficult construction to acquire (Chen & Baker, 2010, Wei & Lei, 2011; Perez-Llantada, 2014; Lee & Chen, 2019). *It be * to* also conveys author stance, another aspect of academic writing that can be difficult for NNS (Conrad, 2018; Chen & Baker, 2010, Wei, 2007).

The functional categorization of the n-grams revealed an emphasis on phrases that link visuals to text, and on text-structuring phrases in general, a finding that was in line with previous comparative work that found engineering writing to be more text-focused and less participant-oriented than other genres (Hyland, 2008b, 2012). However, the functional categorization of the p-frames revealed a different pattern, where participant-oriented phrases that conveyed author stance predominated. Given that this is a feature of engineering writing that has not been much marked upon or explicitly taught, it is a helpful addition to the literature and points the way toward what elements of this linguistic analysis are helpful for NNS engineering graduate students who want to improve their writing.

Which Should Be Taught?

One recent study of the textual changes made to NNS writers' research articles by NS editors demonstrates the difficulties NNS face in balancing confusing and sometimes contradictory advice on how to write for publication. Yli-Jokipii & Jorgensen (2014) found that the editorial changes tended to be in the direction of increasing the structural explicitness (more transitions and language that indicated the flow of the article) while simultaneously editing on the sentence level to create more dense, implicit prose (Yli-Jokipii & Jorgensen, 2014). Given these at times contradictory directives, what phrases can we teach NNS graduate students to help them write both professionally and well?

Because this project seeks to present phrases with rich contextual information gleaned from network analysis, it is important to not just pick the most common phrases, but to choose ones where students will benefit from analyzing the language that surrounds the phrase. In particular, this means picking phrases where there are distinct, varied language patterns centered on the phrase. In the final phase of this project, the time allotted to teaching these formulaic expressions to a class of NNS graduate students was two 75-minute lessons. Because of the emphasis on the phrases' contexts, it seemed best to use each class period to go in-depth on a few phrases, rather than attempt to cover all the phrases briefly. Given the time constraints, it was only possible to discuss two phrases per class. Two n-grams could be taught in the first class and two p-frames taught in the second, so that students began with a concept they were more familiar with and then

went on to the more difficult phrases with variable slots. The question, then, was which two n-grams and p-frames to select for teaching.

Of the n-grams, the ones that link visuals to the text are so frequent and formulaic that students will probably have already acquired them before reaching the level where they are writing their own papers. The research-oriented phrases are also similarly common, and do not seem to represent the type of syntactic structures that gives NNS writers trouble. However, the framing signals and the participant-oriented phrases offer a variety of phrases that are both syntactically interesting and have distinct and potentially difficult rhetorical functions. Thus, *due to the* and *there be no* were chosen for the classroom exercises. *Due to the* provides an approachable introduction to the phrases, as it was a set phrase that students were likely familiar with. However, its context had many hedging adverbs before it and a variety of abstract nouns and evaluative nouns after it. This provided a way to introduce author stance and some of the semantic and syntactic concepts that were found to be important in this chapter; these concepts were then further developed in the next phrases taught.

The second n-gram was *there be no*. This one was selected because it uses the dummy *there*, a structure that NNS writers often avoid. Its context contained a wide selection of discourse organizers and a variety of verb phrases that were used to build author credibility. As it occurred in four different positions in the sentences, it also provided an ideal set of corpus lines for practicing rewriting sentences and moving sentence components around to create more powerful sentences. Due to time constraints, these materials were not used in the classroom teaching, but they were fully developed in the materials development phase.

Of the p-frames, *it be * to* stands out as a unique structure that combines known syntactic and rhetorical difficulties for NNS writers, and thus is certainly worth teaching. The analysis of its context provided in the next chapter also shows that its three distinct rhetorical functions: citing outside sources, organizing the text and depicting authorial stance. All are valuable skills to practice. The gradations of possibility and importance expressed by the near-synonyms in the variable slot are also valuable for equipping NNS students to increase their vocabulary.

The second p-frame selected was *it is * that*. This one appears similar to the first, yet its variable slot is much different. As discussed above, it is primarily filled with passive verbs, and the majority of these fell into the verbs of knowing category. It thus extended the explanation of the dummy *it* and provided an opportunity to work with the passive, to learn the nuances of the knowledge verbs, and to explore the qualifying phrases and degrees of emphasis provided by the context. All of these are elements of author stance that are usually not discussed in materials on engineering writing, but which occurred frequently in the analyses above.

The next chapter will explain how the corpus information and network analyses of these four phrases were converted into classroom materials, and the framework for future materials creation that was developed through this process.

CHAPTER FOUR: MATERIALS DEVELOPMENT

Introduction

This project combines information and tools from two disciplines – corpus linguistics and social network analysis – to create uniquely helpful materials for teaching English academic phrases to NNS graduate engineering students. Thus, the second research question, re-stated below, aims to find the best framework for combining tools to create classroom materials. In doing so, the project will evaluate the role of Social Network Analysis (SNA) in developing these pedagogical materials. Designing the materials involved both discovering best practices for doing the task at hand and evaluating if the chosen tools were the best tools for this endeavor. Thus, in this chapter, the project produces a framework for design that incorporates the best practices from materials creation and evaluates the usefulness of the SNA tools in that process. Finally, having developed and critiqued this approach to materials design, I will summarize its benefits and challenges.

The second research question is as follows:

(2) When using SNA tools to design educational corpus-based visualizations of English phrases for NNS learners:

> (a) What is the optimal framework for combining corpus results and visualization technology to create accurate and effective materials?
> (b) Can SNA programs (particularly AutoMap and NodeXL) create effective, student-friendly visualizations of these corpus findings?
> (c) What benefits and challenges are revealed through designer and NNS student feedback on the materials?

The answers to these questions, elaborated in this chapter, come from my experience designing pilot materials and receiving NNS student feedback on those materials. I learned best practices as I went along, because, for each successful approach, there were usually one or two unsuccessful attempts; and building on these experiences, I developed a framework for simpler and more

successful materials development in the future. I then had six graduate NNS students from a variety of engineering backgrounds review the materials. Based on their feedback, I restructured the materials and incorporated both minor and major revisions. The answers to the three components of this research question emerged through this cyclical process of developing, refining and redeveloping. The end result of this process was fourfold: four in-class worksheets on the four phrases chosen at the end of the last chapter; a one-page summary for students for each phrase; a lesson plan for two 75-minute classes on the four phrases; and a summary of learning objectives. As the lesson plan was designed for the specific circumstances of the classroom research described in Chapter 5, it will be discussed in more detail there; this chapter will primarily discuss the development of the in-class exercises, phrase summaries, and learning objectives. These class materials can be found in Appendix E, and the learning objectives are in Table 15 on page 87.

This chapter draws on the experience designing the end products to address all three components of RQ 2. The next section addresses RQ 2(a) by elaborating the four principles that create an optimal framework and finishing with a proposed strategy for future materials development. The four principles are to maintain simplicity, incorporate corpus insights, structure to align with clear learning objectives, and design attractive materials. The proposed framework incorporates these principles while offering step by step suggestions for how to organize and order the work. It is hoped that the material in this section will be useful to other materials designers who want to incorporate SNA and corpus insights in creating advanced language learning materials.

The following section addresses RQ 2(b). There has been extensive research on the pedagogical value of corpus-based materials for language learners (see Kettemann & Marko, 2002 for a recent compilation), and there are even guides that explain how corpus tools like AntConc or WordSmith can be used in writing classes (Anthony, Wulff & Boettger, 2016). However, SNA tools are primarily designed to investigate social networks (Carley, Columbus, & Landwehr, 2013), not explain language to students. Thus, there is not much research on the usefulness of these tools for designing pedagogical language materials. This section specifically evaluates the challenges and advantages of using NodeXL and AutoMap to create pedagogical materials, as these were the primary SNA tools used. This section will be useful to materials designers who

are debating which tools to use. While I cannot meaningfully evaluate the tools I did not use, my experience with these specific ones will provide a helpful starting point.

The final section incorporates points from RQ 2(a) and RQ 2(b) to answer the holistic question posed in RQ 2(c). Having established the best practices for this material design process, are the materials produced worthwhile? The process of materials development, student evaluation, and redevelopment, as well as the advantages and disadvantages of the SNA software, revealed several important challenges and benefits to this approach. The challenges included the time-consuming nature of materials construction, the necessity for teacher-driven outcomes, and the difficulty in aligning the materials with student levels. The benefits, however, are that this approach gives students an essential opportunity to understand nuances in vocabulary choice; it combines and visualizes sentence variety and vocabulary variety in a novel way; it is easier for the researcher to spot patterns in data; and it is easier for students to visually understand and work with the patterns.

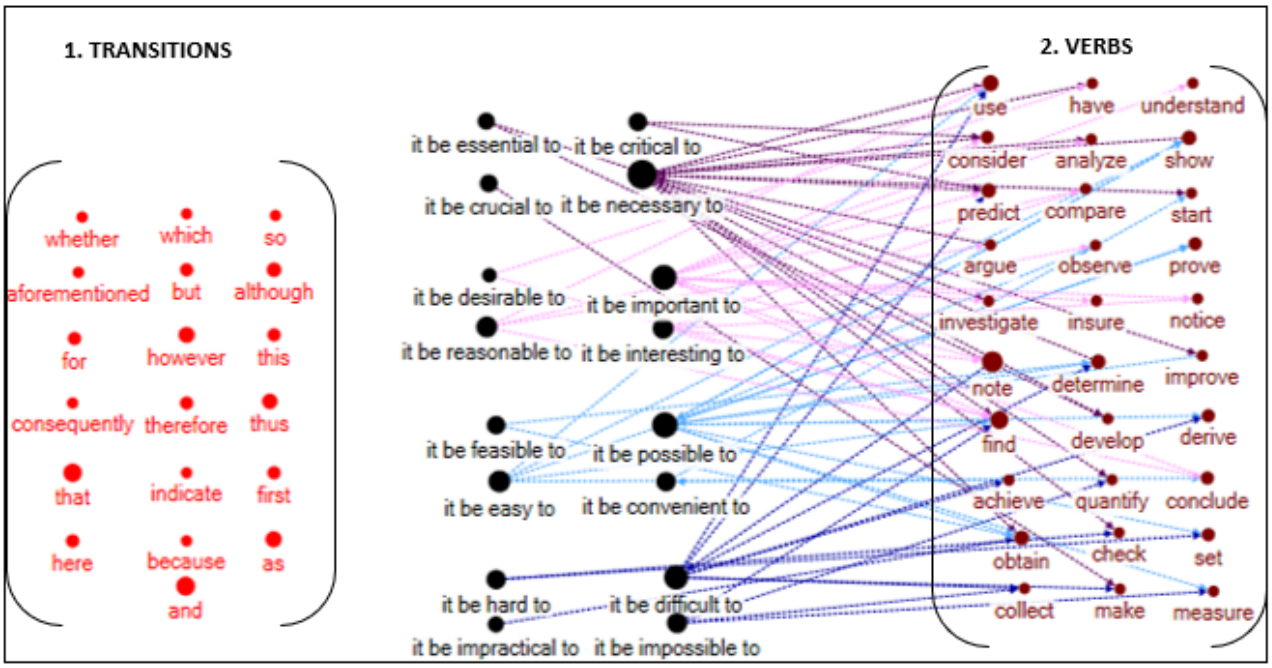
While every pedagogical approach has distinct shortcomings and advantages, this approach of marrying corpus and SNA produces novel materials that can effectively address the needs of advanced level NNS writing students. Thus, as we shall see, the answers to RQ2 are cautiously positive. The final test of its usefulness, however, is discussed in the classroom research in Chapter Five.

Creating a Framework for Material Development

The main endeavor of RQ 2(a) was to identify best practices for using existing corpus and SNA tools to create advanced language learning materials. Incorporating this technology meant that there were often many creative options available for materials design, but that a large amount of time and trial and error was required to properly make use of the tools. As I became acquainted with the tools and their limits, tried a variety of materials designs, and incorporated student feedback, I took notes about what succeeded and what failed. As the materials neared their final state, four major principles of materials design emerged. These principles are not unique to my situation; previous researchers have emphasized the importance of simplicity (Dondis, 1973), corpus contributions (Gabrielatos, 2005), structure and learning objectives (Richards, 2001) and

visual design (Hall, Dansereau & Skaggs, 1992) in creating educational materials. However, this section examines the four principles' specific implications for corpus-based pedagogical visualizations and activities for language learners. These lessons are then incorporated into a final template for future materials design using corpus and SNA tools. Figure 5 contains one of the worksheets in its final form; the other final worksheets can be found in Appendix E. This worksheet was accompanied by a selection of corpus lines that showed the variable phrase in sentences from the corpus. While the following sections will include specific examples of how each principle was applied, it may be useful to return to Figure 5 occasionally to visualize how these principles were applied in the actual worksheets.

Variable Phrase 1: "it BE * to"



Exercises

Adjectives in this phrase not included above that occurred 2+ times: *advantageous, applicable, appropriate, beneficial, challenging, common, helpful, imperative, inappropriate, indispensable, infeasible, instructive, meaningful, meaningless, noteworthy, optimal, plausible, proven, prudent, rare, remarkable, robust, safe, sensible, simple, straightforward, suboptimal, tedious, trivial, unavoidable, unlikely, unnecessary, unreasonable, unwise, useful, vital, worthwhile*

1. a) Which of these adjectives can be used as synonyms for the adjectives in the four groups?

   b) Which are used for separate purposes?

2. Rewrite five of the sentences from the Corpus Examples with different adjectives. How does the new adjective change the author stance?

3. *It BE * to* has three rhetorical uses. Identify a few sentences in the Corpus Examples where you think the author was employing each of those uses.

4. Write one sentence for each of the three rhetorical uses of *it BE * to*.

5. Write four to six sentences on your research using some of the less-common adjectives in the "It is X to" format. Can you employ each of the three uses above?

Figure 5: Worksheet for *it be * to*

Maintain Simplicity

In creating the materials, I substantially simplified the information and visualizations that students would receive. Some of this simplification was easy, such as removing nodes with two or fewer occurrences from NodeXL graphs. Some were more driven by teacher instinct; I categorized the main trends around the phrases and left out words that did not seem to contribute to student understanding of how to use the phrases. Still, one of the main pieces of feedback on the draft materials from five of the six students was to simplify the sheets to make it even clearer what they should focus on and what to take away. This simplification occurred in three ways: selecting main points, avoiding technical language, and maintaining consistent color choices across the worksheets.

Select Main Points

The main form of simplification was to decrease the amount of information on each worksheet and the number of aspects students discussed for each phrase. Of course, the act of choosing four phrases to teach out of the 45 possible p-frames and n-grams was already a major act of simplification; but each phrase has a sprawling amount of contextual information that could be categorized in a number of ways. There is the frequency data about what comes before and after; clustering data that shows which words co-occur in the context; all the rich information that can be gleaned through going through the corpus lines; and of course, a wide variety of possible pedagogical exercises on the semantic, syntactic or rhetorical functions of the phrase.
Thus, when first creating the visualizations, I began by lowering the number of nodes per graph. The first step was to remove words that had occurred two or fewer times in the vicinity of the phrase. The next step, which relied on designer intuition, was to remove the high-frequency words that did not contribute to student learning. These were primarily closed class lexemes: determiners, pronouns, and prepositions that tended to follow certain nouns or verbs. While I kept the pronouns as high-frequency examples of the start of a new clause in the *it be * that* visual, everywhere else I removed the closed class words.

Selecting main points also meant not discussing every aspect of each phrase. My first draft of the worksheets gave the students open-ended questions and allowed them to decide for themselves

what information was helpful and what they wanted to learn. However, the primary critique that the graduate students provided was that the questions were too open-ended. They felt lost in the corpus lines when they did not have a specific phrase or syntactic pattern to look for. They also did not like not knowing the learning goals of each exercise.

Figure 6 and Figure 7 demonstrate the reviewers' felt need for exercises with clear, predetermined answers on two questions. In Figure 5, Question 1 connected to the *it be * to* diagram, where the adjectives had already been split into four groups and the student was asked to identify variations within two sub-groups; this structured question with a clear goal received positive feedback. In Figure 7, Questions 4(a) and (b), which asked students to make their own categories for the verbs in the variable slot of *it be * that*, received negative feedback: "can try – need more experienced opinions," as the student told me that they could try to answer the question, but they would prefer for someone "experienced" to guide them in grouping the verbs.

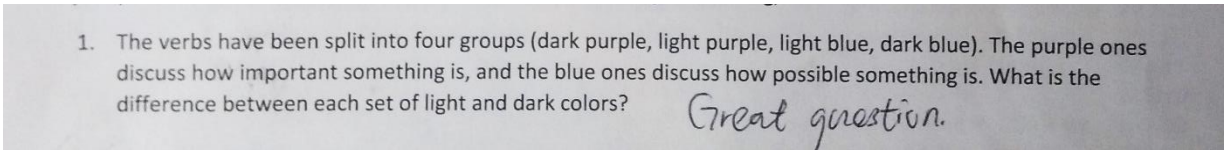


Figure 6: Positive Feedback on Specific Question

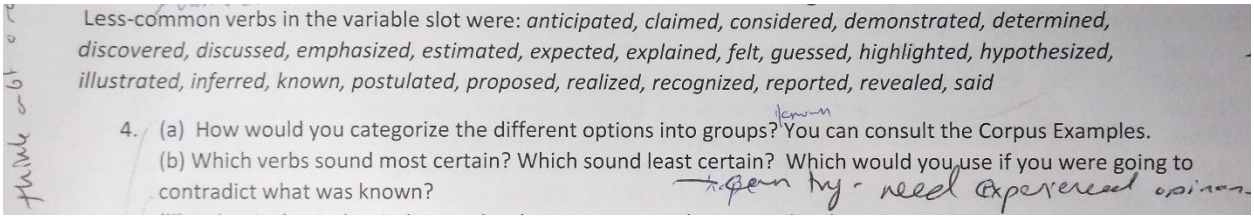


Figure 7: Negative Feedback on Open-Ended Question

Thus, when revising the materials, I chose learning objectives for each phrase (discussed in the section on structuring materials) and edited the exercises so that each question connected to one of those goals. This eliminated many of the questions that students had originally objected to as being too open-ended.

Avoid Technical Language

Another strategy for simplification was to avoid technical language, specifically linguistic jargon and unfamiliar word forms. Students were unfamiliar with the terms "strengthen" and "hedge," and wrote question marks next to these. One student also pointed out that I employed synonyms like "highlight or diminish" elsewhere in the sheets, and he was unsure if they were the same. I kept the terms "strengthen," "hedge," and "author stance," but made a note to explain them in the lesson plan, and then edited the sheets so that they were used consistently throughout. I also removed linguistic terms such as "collocation" or "agent of passive sentence" from the worksheets and designed an introduction sheet that can be found at the beginning of Appendix E to explain where the data came from and why it was useful without using corpus terms. I referred to the n-grams and p-frames as "fixed phrases" and "variable phrases" and explained verb stems and corpus lines with language that the average NNS graduate student would recognize.

The other way to reduce technical language was to return the words from the pre-processed corpus in the visuals to their original forms in the un-processed corpus. This meant editing the labels on the nodes to their most common form, usually by restoring verb roots to inflected forms and turning singular nouns into plurals. For example, for *it be * that*, I returned the verbs stems in the variable slots to past participles ("known" for "know," "seen," for "see," etc.). This was not possible for words like BE, which occurred in multiple tenses, so they were written with uppercase letters and explained in the introduction sheet. I also removed the ng_ and pf_ tags from the phrases and restored their original spacing instead of underscores. When importing corpus lines into the exercises, I used examples from the un-processed corpus. These changes meant that students accessed the materials in the most idiomatic form, without having to distinguish between the phrases that had been altered by preprocessing and authentic writing examples.

Maintain Consistent Color Choices

The final method of simplification was to add color coordination. In the original draft, there was no clear reason for the color choices in the visuals; they were simply as distinct from each other as possible to help with ease of reading. However, based on student feedback, the colors in both

the network visualizations and the corpus lines were redesigned to be consistent both within the sheets and across them. Within the sheets, the colors used in the visuals for categories like Transitions and Hedging Adverbs were the same as the colors used to highlight examples of these categories in the corpus lines. Originally the corpus lines were not color coded at all, as the students were to pick out examples. However, several students said that they would like examples highlighted, both to save time and to help them better understand the categories through real examples. Second, the colors were changed to be consistent across the sheets: where possible, categories like "Transition," "Noun," and "Verb" were colored the same so that the students could make meaningful comparisons across the sheets.

By reducing the content of the visuals and the questions, avoiding technical terms, and color coordinating the exercises, it is possible to use student time efficiently and to highlight the most important data and takeaways. Students were initially overwhelmed by the amount of information in the visualizations and the corpus lines. As they are used to reading for content, not style, it was hard for them to concentrate on the use of phrases and syntax in the corpus lines and not the content of the words. Thus, simplifying the content and the visuals reduced the clutter and helped them concentrate on main points.

For an example of the ramifications of this principle, compare Figure 8 and Figure 9 below. The first visualization is my first attempt at drafting a visual for *due to the*; the second visualization is the final form used in the worksheet. Figure 8 contains every lexeme that occurred 3 or more times within a two-word distance of the phrase, colored in accordance with syntactic and rhetorical categories. This is still the result of simplification; the low-frequency nodes have been removed and the remaining nodes categorized.

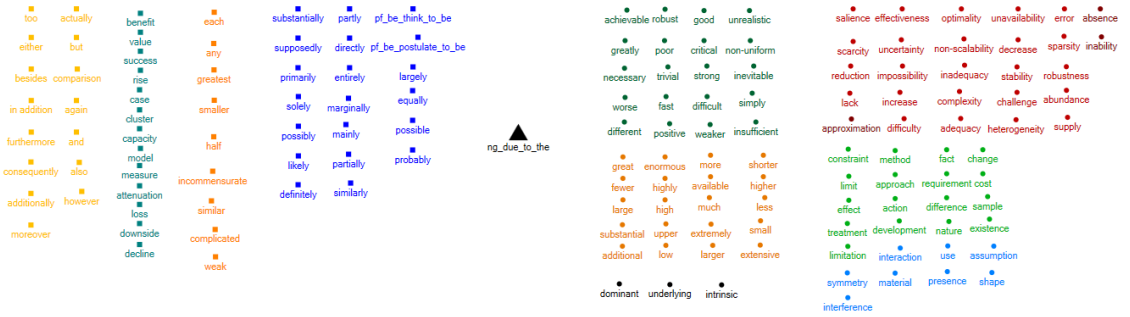


Figure 8: Original *due to the* Visualization

However, the n-grams and p-frames are still shown with their corpus markings, and there is too much information for a student to absorb quickly. There are four categories before the phrase (transition, noun, adjective, and adverb) and three categories of adjective and three of noun after the phrase. The visual contains so many nodes that it cannot easily be read. Thus, the content was simplified again, and is shown in Figure 9.

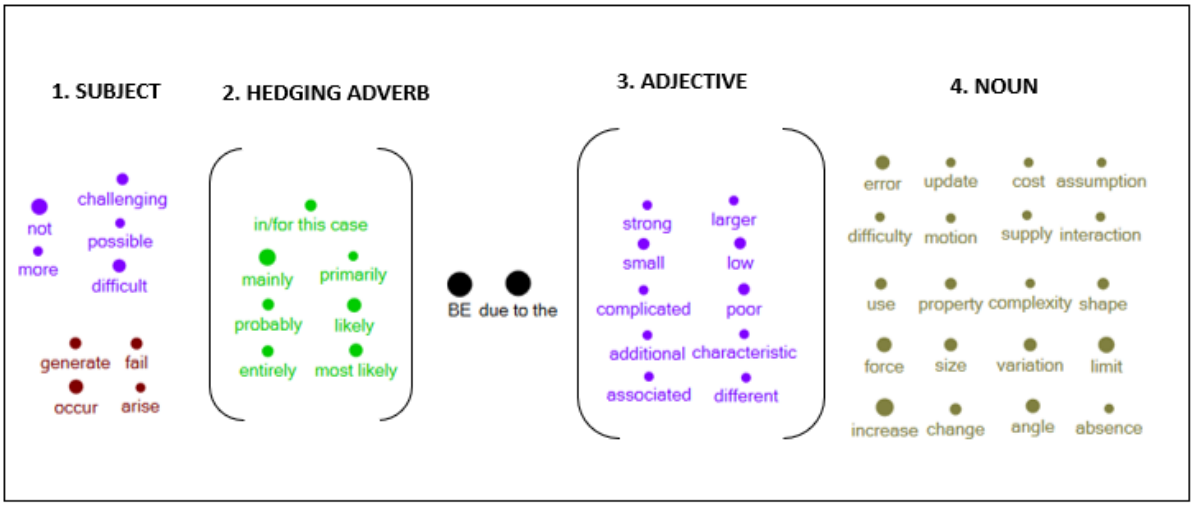


Figure 9: Revised *due to the* Visualization

In Figure 9, the material before the phrase has been reduced to the subject of the phrase (with the verbs and adjectives in separate colors and discussed in the worksheet) and the hedging adverbs, as the syntax of the subject and rhetorical purpose of the adverb are explicitly discussed in the worksheet. The words afterwards have been similarly simplified. Each category has been

numbered and given a title so that students can easily refer back to it when it is discussed in the worksheet, as the material in the visual is now closely connected to specific learning objectives. The one downside to this simplification is that the students do not get to self-select which areas are most interesting to them, an approach that is widely cited as useful in engaging students and giving them ownership of the material (Gabrielatos, 2005). However, simplicity and structure were elements students asked for; and we will return more to the justification for simplicity over choice in the section on structuring the materials to align with learning objectives.

Incorporate Corpus Insights

To create useful worksheets for the students, the visualizations produced by NodeXL were not enough; they had to be refined through analysis of the corpus lines. The corpus lines were essential for accurately interpreting the information in the graphs and for modifying the nodes where necessary. They also made it possible to identify syntactic trends and rhetorical trends in how the phrases were being used. Thus, I discovered that before I designed each sheet, it was useful to open AntConc and read through the corpus lines for that phrase. As I sorted the nodes into semantic categories and used frequency and distance information to decide where to place them in relation to other categories, I referred back to the lines to make sure that I was correctly sorting them. I also skimmed the lines to see if there were any obvious trends that were not clearly shown in the NodeXL networks.

The corpus lines are as useful for a student working with the phrases as for the researcher uncovering points worth teaching. Once the sheets were finished, I pulled 30-40 corpus lines into each worksheet under a section entitled "Corpus Examples," ensuring that every takeaway in the phrase summaries was illustrated by multiple lines in the Corpus Examples. Incorporating corpus insights helped identify four key types of information: low-frequency synonyms in or near a phrase, idioms near a phrase, and syntactic and rhetorical patterns in phrase usage.

Incorporate Low-Frequency Synonyms

The corpus lines were especially helpful in creating complete lists of the items that appeared in the p-frames' variable slots. For example, *it be * to* mostly had adjectives of importance and

possibility in the variable slot. The items that made it past the frequency filter into the network visualizations were words students were already familiar with and could use confidently, words like *important, critical, necessary, easy*, and *hard*. The less common ones, however, were valuable synonyms: *plausible, unreasonable, vital, suboptimal, imperative*. When two of the students first learned how the data were gathered, their concern was, rightfully, that the most frequent expressions are not necessarily the best expressions; one of them had been warned by her advisor to avoid clichés and the other one to never use the same word twice in a sentence. For more advanced students who are already comfortable with the most common forms of *it be * to,* incorporating the less frequent synonyms meant the worksheets were still valuable.

Additionally, primarily in the case of the transitions, some items that appeared at low frequency in the immediate vicinity of the phrase appeared much more often a few words farther before or after the phrase. For example, 'consequently' appeared only twice in the two-word vicinity of *there be no*, but reading through the corpus lines, it appeared several more times just outside that boundary. In the case of these low-frequency synonyms or words that appeared more often in a larger window, adding them to the graph would have contradicted the previous principle of maintaining simplicity. Thus, this information was added into the exercises and summaries, as students were asked to identify the nuances between synonyms and practice using them in place of their more cliché counterparts.

Incorporate Idioms

Another way of incorporating corpus insights was to change the labeling on individual nodes from a single word to a full phrase when the corpus lines indicated that the words were usually part of longer phrases or idioms. For example, before *due to the*, the word *most* always appeared before the word *likely*, but *likely* also appeared on its own, so the label for *most* was changed to *most likely*. The words *in*, *case*, and *this* also appeared before *due to the*, and the corpus lines revealed that they primarily appeared as *in this case* or *for this case*, so I combined the nodes into one node labeled *in/for this case*. With *there be no*, the corpus lines revealed that whenever the phrase was followed by *significant*, the word *difference* also followed, so the *significant* node was relabeled as *significant difference*. Small changes like these ensured that students had a more holistic picture of how the phrases were used, and enabled exposure to common academic

phrases outside of the four in the lesson plans. It also ensured that students were studying functionally significant chunks rather than lexemes that were not fully meaningful on their own; the importance of this is discussed in Simpson-Vlach & Ellis (2010).

Incorporate Syntactic Trends

The other two contributions of the corpus lines were to identify syntactic and rhetorical trends in phrase use. These data were essential in choosing how to order the pieces of the sentences linearly and identify the rhetorical functions of each phrase. The corpus lines were useful in making sense of clusters of nodes around the phrases. In general, this step was carried out by first noting a pattern or unexpected node in the NodeXL networks and then searching for it in the corpus in the vicinity of the phrase. One example of this is the cluster of verbs that appeared shortly before *there be no*. By arranging the corpus lines to sort alphabetically according to the word to the left of the phrase, it was quickly apparent that there was often a clause before the phrase that indicated on whose authority the claim was made, and that the verb of the clause was almost invariably followed by *that* (see Figure 10). Thus, these verbs were grouped before the phrase and both the syntax and rhetorical purpose of the clause explained and practiced in the worksheets. In a similar way, the corpus lines demonstrated that *it be * that* was often followed by a new clause starting with pronouns, and these could be grouped into a New Clause category.

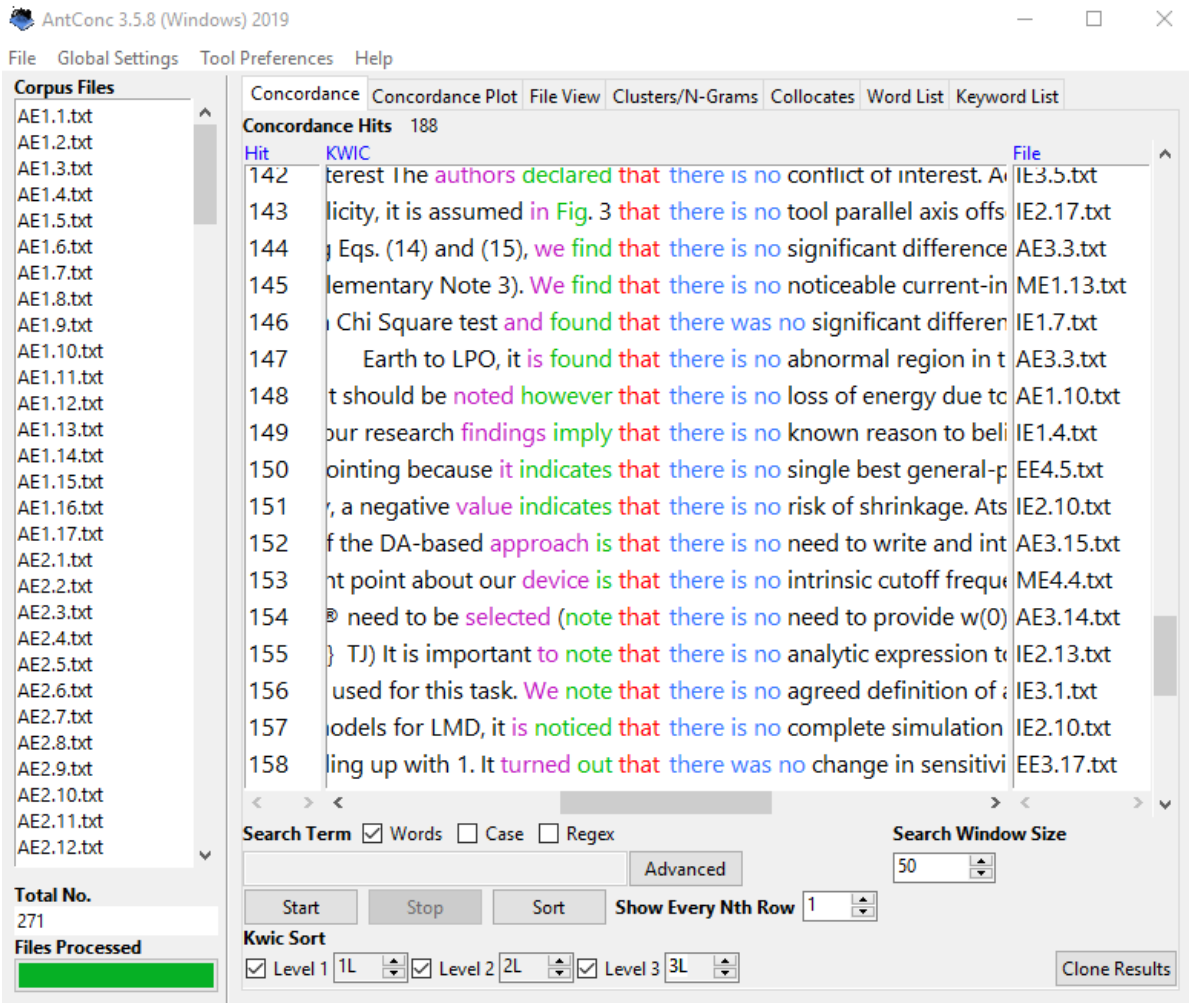


Figure 10: Identifying Syntactic Trends before *there be no*

AntConc was also useful for identifying the phrase's position in sentences, using both the punctuation information in the original corpus and the corpus lines. Some phrases, like *it be * that*, appeared almost always after the start of a sentence or a comma. Other phrases, like *due to the* appeared is a variety of positions: sometimes as a phrase attached by comma, sometimes following a linking verb, and both before and after whatever was caused by the object of the prepositional phrase. The placements were much more diverse than what I would have imagined if I had not referred to the corpus lines, and by identifying and practicing the full range of syntactic placements, the students are able use the phrase more fluently.

Incorporate Rhetorical Trends

In addition to syntactic trends, the corpus lines made it possible to pull meaningful rhetorical trends from the graphs that NodeXL produced. For example, for *it be * that*, there were a cluster of words *(for, at, with, as, when, where)* that appeared primarily after the phrase. Reading the corpus lines revealed that these started qualifying phrases that restricted the domain of the claim made by the phrase, as in this line from the corpus: "It is noted that, *with a higher-resolution instrument*, the differences may be more pronounced." While these words were not syntactically identical, they shared the rhetorical purpose of introducing phrases or clauses to explain in what conditions the claim was true. The corpus lines allowed me to see that these phrases also occurred, though much more rarely, before the phrase, so I could highlight them in the corpus lines and encourage students to consider the rhetorical function and the placement of these qualifying clauses.

One rhetorical trend that occurred frequently throughout the material was the use of adverbs and adjectives to either hedge or strengthen main claims. For example, in *due to the*, hedging primarily occurred through a variety of hedging adverbs (*probably, possibly, largely, mainly*) before the phrase, but with *it be * that*, hedging occurred through qualifying phrases that followed the phrase. On the other hand, the clauses with verbs of knowing before *due to the* usually served to strengthen the claim, and the adjectives in the variable slot of *it be * that* could either strengthen or hedge the following clause. Thus, the idea of strengthening and hedging claims was clearly a point worth teaching students, especially as authorial stance towards ideas has been shown to be difficult for NNS writers (Perez-Llantada, 2014; Wei & Lei, 2011; Öztürk & Köse, 2016) and easier to understand when presented through corpus materials (Chang & Schleppegrell, 2016). Thus, I was able to identify the different types of hedging and strengthening nodes throughout the phrases and teach them cohesively.

Ultimately, the wealth of information gained by incorporating corpus analysis created more useful materials that provide a fuller picture of the phrases' context. Students could access a holistic picture of the synonyms, idioms, syntax and rhetoric that surrounds these phrases and use the information to add variety and idiomaticity to their writing. The details provided by the

corpus make the visuals more interesting and give NNS writers the tools to address their desire for greater variety and idiomaticity in their writing (Wray, 2012).

Incorporating this corpus information, however, leads to the need for the third principle: with such a variety of insights and information, the material had to be carefully structured and the takeaways clearly identified. It also increased the difficulty of applying the fourth principle, as combining corpus information into the existing exercises led to unattractively dense worksheets where information and questions were confusingly intermingled. The best solution for this was to keep the worksheets for questions and design a separate summary sheet for each phrase; this step will be explained under both the third and fourth principles.

Structure to Align with Learning Objectives

While the previous two principles were ones that I as a designer consciously incorporated before receiving student feedback, the third principle of structuring the materials to align with clear learning objectives developed primarily as I considered the feedback from the students who reviewed my materials. As it partially contradicted some of the previous recommendations on how to use corpus-based materials with language students, I will explain the justification for this principle before explaining how it was carried out through choosing learning objectives, designing summary sheets, and encouraging student participation.

Justification

Much research on corpus-based pedagogy has argued that students should be given the ability, as much as possible, to explore the corpus findings on their own, study what interests them, and apply that information to developing their own writing (Bernardini, 2002). However, student feedback on the materials consistently emphasized a desire for more structure and clear takeaways rather than more freedom in exploring the corpus lines. In fact, several students suggested that the corpus lines be removed, or that only the most pertinent examples be worked into the sheets. While I chose to leave the corpus lines in for students who wanted the contextual data, I did take the students' requests for teacher-driven findings seriously. I did so both because of the nature of corpus-based visuals and because of the real-world constraints the students faced.

As noted many times in the principles above, the step of visualizing corpus findings means that the author must select and highlight certain parts of the findings over others. The acts of creating syntactic and rhetorical categories, of color-coding the data, of choosing which nodes should be deleted, which labels should be changed, which low-frequency synonyms should be included, all involve decisions about what information will most benefit students. Visualizations simply cannot hold all the information the corpus contains and still be easily read and understood. Thus, by creating specific learning objectives for each phrase, I was not greatly restricting the students' ability to decide what to study; this had already been restricted by the visualizations, and it made sense to make the questions as goal-oriented as the visualizations.

In addition, the students' request made sense given the realities of their context. International graduate students seeking to improve their academic writing face serious time constraints, which meant that they preferred mastering specific learning objectives over an exploratory approach. This was as true of the students who used the materials in the classroom setting in Chapter Five as for the students who reviewed the materials. They were making time in their busy schedules for a no-credit class; their main motivation in signing up was primarily that someone with authority had told them that they needed to develop their academic writing to meet their career goals. Thus, they were most willing to engage with pedagogical materials when they could clearly see how it would improve their writing.

Given the needs and desires of this audience, incorporating clear goals and takeaways was necessary. To best reach this goal of creating specific learning outcomes, it still is essential for the materials designer to ask open-ended questions and then allow clear goals develop; it is useful to create summary sheets for student; and finally, it is important to find ways to encourage student participation and peer-to-peer instruction within the instructor-led framework.

Start Complicated, Then Choose Learning Objectives

From a materials development point of view, developing focused, structured worksheets still means starting with as much data as possible and pursuing many points that may not make it into the final activities. As I explored the NodeXL visuals and corpus lines, I noted several trends and patterns in the phrase contexts that I excluded in the final versions. For example, *due to the* was

the only phrase that was not preceded by a group of transition words, and in the first draft, I asked students to explain why there was this difference. However, as it was an open-ended question, and overlapped with the following questions on sentence order, I removed it. In the original worksheets, there were also more questions about categorizing the nouns and verbs that followed the phrase, but as the content of these words often varied by the discipline, students would have been categorizing and discussing words that they would not use in their personal field of study. While this information was linguistically interesting, and perhaps useful to students, it was not universally applicable and did not directly improve fluency of phrase usage, so I removed it and decreased the amount of space in the visuals given to those verbs and nouns. Similarly, the *it be * to* worksheet originally noted the preponderance of prepositional phrases before the phrase and asked students to examine the corpus lines and decide what rhetorical functions these phrases served. As there was no one specific answer (they served a variety of overlapping tasks), I removed this question.

In addition to these minor points that were removed, I designed entire visuals that did not make it into the final edition of the worksheets. At various points I created visuals for the p-frames that retained all the edges so that students could see what context words linked most with which variable slots in the p-frames; I created separate networks for the adjectives of possibility and the adjectives of importance in the *it be * to* variable slots; and I used WMatrix to semantically tag the nodes and sort them accordingly. At some points during the development, I had two pages of questions and exercises for some of the worksheets. It was only after thoroughly exploring the variety of visuals and objectives for each phrase and receiving student evaluation on what points where helpful and which were too easy or too abstract that I was able to confidently select the three or four most useful objectives for each phrase.

Create Summary Sheets

As mentioned above, the process of incorporating corpus data and creating specific learning objectives meant that there was too much information to incorporate easily into an in-class activity sheet and summarize efficiently for future student use. When reviewing the first draft of the exercises, two students separately asked if I had the data and "answers" on another sheet that they could use as a reference. A third student pointed out the alternation between the corpus

information and questions were confusing, and that after working through the sheet, he wanted a separate clean paper with the information that he could refer to in the future.

The solution to this was to create a summary sheet for each phrase that highlighted the corpus information and learning objectives for that phrase. The summary sheets had the following categories: relevant grammatical information, sentence order information, relevant linguistic concepts, the phrase's rhetorical uses, and synonyms. Not every phrase summary had every category, only the ones relevant to that phrase and its objectives. The summary sheets are in Appendix E, along with the worksheets.

The summary sheets facilitated designing three or four learning objectives for each phrase, shown in Table 15. It was much harder to draw the learning objectives from the draft worksheets with the open-ended questions, but after creating the summaries with their terms and descriptions, it was simple to select learning objectives. The process was cyclical, as I also tweaked the summary sheets to better connect to the learning objectives, and this often meant changing a question on the worksheet as well. But in the end, the three sets of materials were clearly aligned, and this meant I could also align each of the pre- and posttest questions with a learning objective. Table 15 shows how the objectives for each phrase matched a certain section of the summary and specific questions in the worksheet for each phrase.

Table 15: Alignment Between Learning Objectives, Summary and Worksheet

Phrase	Learning Objectives	Section in Summary	Worksheet Questions
due to the	1. Identify the antecedent of the phrase	*Pieces of the Sentence*	2
	2.Change or add a hedging/strengthening adverb	*Strengthening & Hedging*	5, 6
	3. Rewrite sentences to change phrase order and rhetorical impact	*Place in Sentence*	3, 4
there be no	1. Change and add transitions before sentence	*Transitions*	1
	2. Change and add hedging adjectives	*Adjectives that Strengthen/Hedge*	4
	3. Adjust preceding verb to hedge or strengthen	*Verbs that Strengthen/Hedge*	2, 3
*it is * to*	1. Change adjective of possibility	*Adjectives of Possibility*	1, 2
	2. Change adjective of importance	*Adjectives of Importance*	1, 2
	3. Identify and use the phrase's three rhetorical functions	*The "Dummy It"* and *Uses*	3, 4
*it is * that*	Identify qualifying phrases	*Position in Sentece*	1
	Identify and use the phrase's three functions	*The "Dummy It"* and *Uses*	2, 3
	Identify the passive agent of the verb	*The "Knower" of the Phrase*	4

It was useful as a researcher to be able to pinpoint what students should learn and how to evaluate if they had acquired that knowledge. But more importantly, these summary sheets meant that students could grasp the most salient points about each phrase and refer back to them if they were struggling to use the phrase fluently in their future writing.These objectives also allow students to see how the knowledge they gathered from a particular phrase is transferable to other phrases. Learning to distinguish when an author is hedging or strengthening a claim, how a

phrase's impact is changed by its placement in a sentence, how to replace clichés with more original synonyms, or to notice when an adjective reflects author stance are all transferable skills.

Encourage Student Participation/Sharing

Despite the emphasis on instructor-directed learning outcomes and summary sheets, true learning requires student involvement and peer-to-peer sharing. International graduate students are by necessity multilingual, gifted, motivated learners who are becoming experts in their specific fields. Those who have had to acquire a language as adults have helpful insights into its use that are often not available to NS writers (Llurda, 20015). As I reviewed the drafts with the six students, I was continually impressed by the acuity of their observations and the amount of thinking and talking they had done about their own English academic writing. They knew which words and phrases their advisors employed frequently, and which writing pet peeves to avoid. They also knew field-specific connotations that I could not gather by scanning corpus lines. For example, one student in chemical engineering informed me that the type of experiment performed to uncover the chemical makeup of a material constrained whether the researcher used "obtained" or "determined" in documenting the results.

Thus, as a NS teacher who had not done writing in any of the students' fields, it was essential to create space in the exercises and the lesson plan to ensure that students were encouraged to share their insights with each other. One student recommended that I clearly delineate which questions were for individuals to answer on their own and which were group work. Thus, in the final sheets, the questions were ordered so that students had some time to familiarize themselves with the phrase and answer a few basic questions before devoting the majority of their time to going over the questions in groups. There are still a few open-ended questions for which the teacher does not have a clear answer, and what is learned will be a result of what the students know or discover. The lesson plan (discussed in the next chapter) also included times for the groups to share their insights with the whole class. As the main learning objectives were clearly fixed, it was important to balance that by making room for peer-to-peer learning.

Design Attractive Materials

The final key principle is to make the materials attractive. While this goal is not often discussed in research on corpus-based pedagogical materials, the addition of visualizations highlighted the focus on the appearance. It was also something students regularly gave feedback on – from quantity of white space to color choices to suggestions for bullet points, they wanted a design that made sense and was easy to navigate.

For this particular set of worksheets and summaries, that meant being selective about what was included in the visuals, labeling clearly, and separating the materials into exercises and summaries. It also required listening to students' comments and incorporating some small suggestions, some as simple as not stapling the sheets with the corpus lines to the worksheets, as that made it difficult to flip back and forth between the examples and the questions.

Visualize Only Select Elements

This application connects directly to the first principle on maintaining simplicity. In this case however, the simplicity serves to enhance the appeal of the visuals rather than direct the students' attention. In particular, it means deciding what information gets visualized and what is included as text or left out completely. Since the visuals are what makes these materials unique, it was tempting to visualize everything possible. However, in the final version of the materials, only a few main points were visualized in the phrase networks. The relative frequency of a word is indicated by the size of its node, as more frequent words have larger nodes. The colors correspond with syntactic or rhetorical categories. And the linear order, where possible, represents the order in which the elements of the sentence appear.

There was a great deal more network and corpus information that could have been visualized. The average distance of the word from the phrase is not included; there is no indication of the raw frequency of each node's occurrences; and in most cases, the edges between nodes have been deleted. This means that with the p-frame exercises, the students do not get to see which words before and after varied in accordance with the word in the variable slot. Even when the edges were kept after the phrase in the *it be * that* worksheet, I did not set the edge widths to

reflect frequency information. While I designed versions that visualized all these pieces of information, it was quickly apparent that more than 3 or 4 types of information per node cannot be simultaneously visualized in a way that fits attractively onto one page.

Selecting elements to visualize also meant that I did not use many of the visual bells and whistles that NodeXL offers. I could have made the nodes different shapes; I could have employed directional arrows between nodes; I could have differentiated between dotted and solid lines; I could have used more complex graph shapes. However, I settled on simple left to right visuals because this most closely followed the linear structure of writing. I retained the arrows only on the right side of *it be * to*, so that students could see which adjectives in the variable slot appeared with which verbs after the phrase. For the rest, I let the linear structure convey the order. In the original graphs, bidirectional arrows were useful for demonstrating which nodes appeared both before and after the phrase; however, for the sake of clean visuals, I used the edge count information to determine whether a node occurred most often before or after the phrase, and placed the node on the side where it occurred most frequently. Again, this simplified the information; but it was more attractive to have a linear graph with less information than to be able to identify the handful of words that appeared both before and after the phrase.

Label Clearly

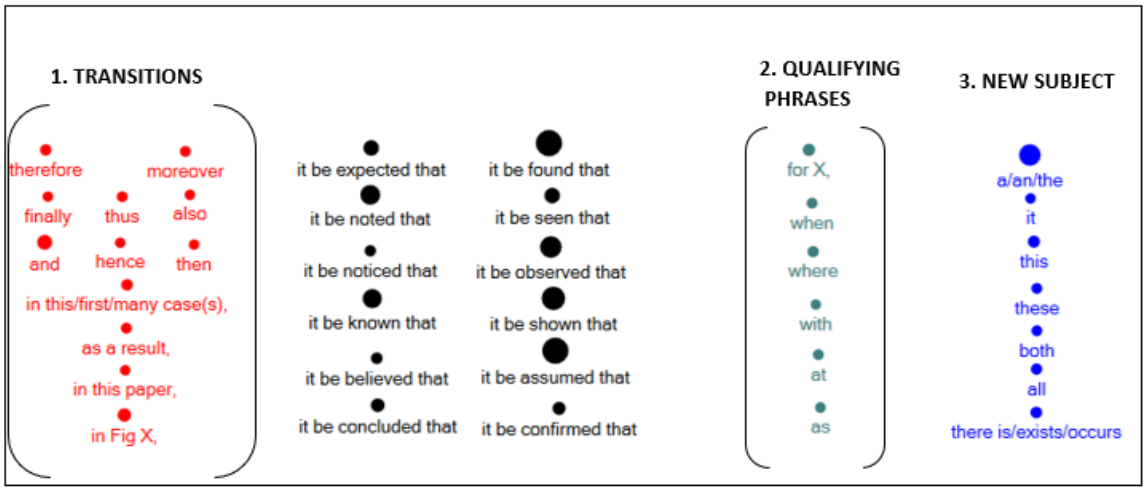
One improvement to the visuals that several students requested was to label each category of items before and after the phrase. Numbering these categories made it easy for students to refer back to the visual when working on the questions. The very first draft of the worksheets only had numbers, not labels, because, following Gabrielatos (2005), there are pedagogical benefits to asking the students to identify the trends in the data and name the categories themselves. However, as mentioned before, the students preferred to work with and learn from existing categories, and once created, the categories were much easier to work with when labeled.

Separate Information from Exercises

Three of the students asked for more modern-looking, spacious worksheets. When there were lists in the text, one student recommended bullet points. This was the second reason, in addition

to including more corpus information, that I redesigned the original single worksheets into a worksheet and a separate summary sheet. With the new design, the important terms and corpus information can be presented orally by the teacher in class, the students can work through the worksheets and write on them if they wish and mark up the corpus lines, and then they can take the summary sheets home. This change made the in-class exercise much less cluttered.

A before-and-after example of this is found in comparing Figure 11 and Figure 12. In Figure 11, we have the original introductory paragraph on the *it be * to* worksheet. The categories are explained in full sentences, without referring to the category numbers, and examples from corpus lines are mixed in; the questions do not start until halfway down the page. In Figure 12, however, all that content has been removed from the worksheet and placed in a separate summary sheet. In addition to making more space on the worksheet, the information moved to the summary sheet is also more attractive. The categories are explained in a numbered list that matches the visuals, and headings now separate the pieces of information. This creates more white space, making the page easier to navigate. The examples have been removed, since they will be discussed in class. Finally, rhetorical uses have been added, so that students can easily return to this summary to recall the important points.



Again, we have a "dummy it" phrase – but this time, the variable word is a passive verb about knowing. Many transitions come before, and this is followed by two more complex categories. The first one is optional: qualifying phrases that restrict the following clause, i.e., "It is noted that, *with a higher-resolution instrument*, the differences may be more pronounced." These qualifying phrases may also appear before the variable phrase. The new subject then introduces what it is that the author says has is known.

Figure 11: Original Paragraph for *it be * that* Worksheet

Order

(1) Transition
(2) Qualifying Phrase that restricts the claim. e.g., "It is noted that, *with a higher-resolution instrument*, the differences may be more pronounced." This can also come after the phrase.

it BE (past participle verb) that

(3) Beginning of a new clause

The Dummy "It"

The *it* doesn't refer to anything specific. Instead, here it is a holding structure for a passive verb of knowing.

Uses

(1) to explain the author's expectations, assumptions, or thoughts without using personal pronouns
(2) to state general truths or background information
(3) to summarize and highlight certain research results

Figure 12: New Summary for *it be * that*

To summarize, clean visual design of the materials is not something that either corpus linguistics or social network analysis tools can manage. It takes a designer to decide what information goes into the visuals, how data should be sorted and labeled, and how to lay out the information on the page. Both expert knowledge of what linguistic data is useful to writers and user feedback on what makes sense or feels clunky are essential elements. Only by combining the insights from corpus linguistics and pedagogical knowledge can a designer choose what elements to visualize. But when creating materials that are attractive to the eye and easy to work with and learn from, the most important element is student feedback. The designer must understand what a student sees when they pick up a sheet and redesign until the students' perceptions align with the pedagogical goals.

Template for the Future

The four main principles that led to successful materials design were to maintain simplicity, incorporate corpus insights, structure materials to align with learning objectives, and attend to design aesthetics. However, as is clear from the examples above, these principles were applied iteratively throughout the design process; sometimes they conflicted with each other or conflicted with student feedback. It took hours of trial and error to settle upon the final form of the visuals, worksheets and summary sheets, and there is still room for improvement. However,

given the lessons learned, this is the linear process recommended for developing language learning materials on academic phrases that effectively use corpus-based visuals. The nine steps are listed in Figure 13, along with the tools used in each step. The graphic flow begins in the upper left.

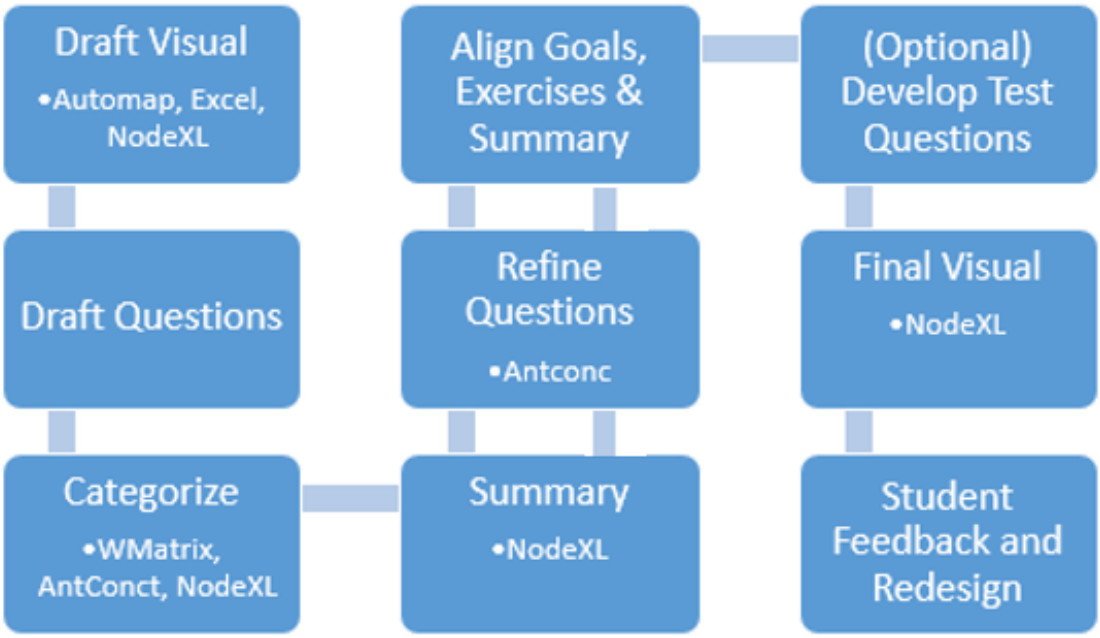


Figure 13: Template for Materials Design

These steps start with the assumption that the materials designer has access to a cleaned corpus of relevant material with the phrases/lexemes of interest already marked, and that the designer is familiar with linguistic terms and language acquisition pedagogy.

1. *Draft the visual*. This is the most technologically time-consuming step. To begin, input the corpus texts into AutoMap. Window size of two words (i.e., two words to the right and left of the phrase) is recommended; widening the window did not help with identifying useful information. From the union semantic list, extract all edges that contain the phrase (this is done in Excel). Place this information in a new Excel document, and clean out whatever information is clearly not needed (such as edges with a frequency below 3 and scrambled scientific abbreviations). It is much easier to do this in Excel than in NodeXL. NodeXL can be buggy and has built-in commands and formatting that means

it occasionally reverses your decisions about which nodes to hide and show, so put as little excess data as possible into NodeXL. Once cleaned in Excel, the data is inserted into NodeXL to see what information emerges from the visual. If there are too many nodes to see clearly, reduce the nodes in Excel and try again in NodeXL; it can be helpful to alternate between the two to immediately picture the results of changes. Keep careful documentation of what is excluded and why.

2. *Draft potential questions, both for the exercises and the researcher*. This means that initial good ideas will not be lost; note idea for a useful activity, or questions raised by the graphs. For example, in creating the *there be no* graph, one question I wrote is, "why are there so many reporting verbs before this phrase?"

3. *Develop categories for the nodes in the visuals.* This means working with AntConc, NodeXL, and, optionally, WMatrix's semantic tags. WMatrix can be used to semantically tag the nodes and highlight clusters of similar words. The AntConc lines are essential for deciding the linear order of the categories, answering questions like, "when a phrase is preceded by both qualifying adverbs and transitions, which occurs first?" AntConc corpus lines will also indicate if a category appears frequently outside the window-size distance, and thus if it is worth including. As categories are chosen, arrange the nodes in the visual accordingly. When a node does not fit into a category, search for it in the corpus lines to pinpoint its role and decide whether to keep it.

4. *Draft the summary.* As with the second step, pause in drafting the visual to draft the summary, noting down linguistic terms that would be useful in explaining the patterns, synonyms in the variable slots or in the vicinity of a fixed n-gram; and rhetorical and syntactic patterns identified in the AntConc lines. This step relies most heavily on the designers' research and pedagogical knowledge, as it is appropriate to highlight the patterns that have been shown to be most difficult for language learners.

5. *In light of the summary, redevelop the questions and add corpus lines*. As salient patterns are identified, it is possible to change open-ended questions into directed questions. With

the question mentioned above, I identified three rhetorical uses to having a reporting verb before the claim and asked students to identify and practice those uses. This is also an ideal time to pull in corpus lines for the Corpus Examples section, as you can pick the most relevant lines and use them to illustrate questions and concepts.

6. *Pick 3-4 learning objectives for each phrase and align them with the content*. This is the point where the many interesting points and patterns are narrowed down and structure created. The objectives may involve rhetoric, syntax, sentence development, vocabulary and synonyms, author stance or other useful points, but make sure that they are not all syntax or all vocabulary. Rotate working on the objectives, exercise questions and phrase summary till they are all clearly aligned.

7. *Optionally, develop test questions to measure if students have acquired the concepts in the learning objectives*. The design of the test questions will be discussed in the next chapter; here, they are mostly valuable as a way of checking the alignment of the goals, exercises and summaries.

8. *Finalize visual design*. This is the time to go back into NodeXL and make the graph as attractive and useful as possible. Color-coordinate the categories. Rearrange the nodes so their labels are easily readable and they fit the span of one page. This should be saved until the very last, since NodeXL can have trouble saving visuals; node position can be locked, but color and shape choices made on a graph will be overridden by the Excel lines when it is reopened. Thus, the final visual should be designed in one sitting and exported. On the exercise sheets, label and reference each category correctly, color coordinate corpus line examples, and make sure the layout on the page is attractive.

9. *Invite and incorporate feedback from the target audience.* Finally, ask members of the intended audience to evaluate your materials, and redesign the sheets as appropriate. As not every suggestion can be incorporated, pay attention to those which align with the four principles and to points raised by more than one reviewer.

As noted in the beginning and evident throughout the steps, this process only works when the designer is educated in linguistics and pedagogy. While the centrality of the designer's knowledge is clearest in step 4, their own experiences and leanings will influence the decisions made in other steps as well. Figure 13 also highlights the necessity of access to an array of corpus and SNA tools. This approach combines researcher expertise with five separate tools, all of which inform the final choices made. Thus, it is appropriate to next evaluate the role of the SNA programs in the framework.

Evaluation of SNA Programs

The two main social network analysis programs used in this project were AutoMap and NodeXL. They are separately evaluated below for their usefulness in the task of designing visualizations for corpus-based teaching materials. AutoMap was used to clean the corpus and extract relationship information; NodeXL was used to visualize that information. While both played an important role in corpus design, the past sections have made clear that they were not adequate on their own. In order to use these tools successfully, the materials designer needs to be able to successfully use AntConc or other corpus tools, WMatrix or other forms of POS and semantic tagging, and Excel.

AutoMap

One small challenge of this program is that it was not designed for a corpus analysis where the researcher is most interested in the words themselves. It was designed for researchers interested in network analysis, who want to map out relationships between key people and places for purposes like counterterrorism and understanding social networks (Carley, Columbus, & Landwehr, 2013). But this does not stop Automap from being highly effective at the two key steps of corpus pre-processing and extracting basic relationship information.

For pre-processing, AutoMap has the key benefit of being able to quickly make changes to a large set of text files. Features such as stemming, merging hyphenated word at line ends, converting to lower case, and removing any word or numbers that the designer asks for all make it easy to efficiently clean a corpus. While there are other files that do similar work (e.g.

Notepad++ (Ho, 2019)), AutoMap comes with a variety of presets that are specific to text editing, the interface is easy to work with, and it is possible to reverse changes to the text and watch the effects of doing them in a different order. AutoMap also made it easy to fix its own errors. For example, there were a handful of words ending in /s/, like 'thus, analysis, this,' which it misidentified as plurals and transformed into 'thu, analysi, thi,' when stemming. But thanks to its excellent documentation, I could look through the list of transformations, identify the incorrect ones, and run a thesaurus that changed them back to their original form.
In addition to corpus cleaning, AutoMap was used to extract relationship information from a corpus, and it did so perfectly. The task is a simple one (to count how often words co-appear within a certain window), but AutoMap did this work efficiently and gave a multitude of options in addition to window size. It was possible to select textual cues like sentence end and paragraph end when setting window size and to choose to incorporate directionality (which is essential for this type of analysis, where whether the word comes before or after the phrase matters). It produces data files that are compatible with a variety of file formats, and, importantly for NodeXL, can be easily imported into Excel files.

Thus, the way AutoMap was employed in this project is not a fair assessment of its true capabilities; it can do much more than what it was used for here. However, its preprocessing presets and ability to extract accurate relationship data from large corpora make it a helpful tool for a linguist.

NodeXL

NodeXL played an essential role in creating the visualizations. While these visualizations were received favorably by the students in the piloting and have a number of clear benefits that we will discuss later, as a software NodeXL has both serious advantages and disadvantages.
The main concerns for future materials designers to consider when selecting data visualization software is that NodeXL is not always intuitive, that it occasionally crashes when faced with large quantities of data, and that there are a few key features missing.

Because it is built as an extension to Excel, experienced Excel users may enjoy the familiar features and find it easier than starting from scratch on a new platform; however, the downside is

that the goals of NodeXL do not sync perfectly with the Excel interface. The graph is displayed in Excel's Document Actions Pane, so the user is limited in the size and shape of the space. Selecting and managing data points in the pane is not intuitive; the NodeXL controls are spread out between a special tab at the top of the document, the graph pane, and the rows of data. The rows of data are all easy to edit, but because of all the presets in a NodeXL template, Excel formulas often do not work, and editing the data rows does not always update the visual, even after refreshing the graph.
It is also worth noting that there were a couple points in my design process where NodeXL could not handle the large amount of data inserted into it; the data had to be filtered in a separate Excel file to a smaller size before being imported. This is not an insurmountable problem, but it means that the researcher or materials designer cannot benefit from the visualizations until after paring down the data.

Finally, while it is excellent at providing full control to the designer, there are some shortcomings that affect the quality of the visualizations. There is no setting that groups nodes together so that the labels do not overlap. Aside from pre-set graph shapes that did not fit with the goals of this project, there is no way of telling the software that you want a certain group of nodes in a certain part of the graph without manually selecting and moving them. The nodes cannot be lined up in neat rows, as the option that should do this does not consider label width, so the labels overlap. Even for directed graphs, there is no way of arranging nodes in accordance with the direction of the arrows. Finally, exporting the images is clunky; the only way to do so with a Basic membership is to upload it to their website and download it from there. Even with the Full membership, the graphs have to be emailed to someone and downloaded from the email.

However, despite the sometimes clunky or unreliable interface, there are clear benefits to this software; it enables the researcher total control over all the variables, provides immediate visual understanding of key ideas, and does not require coding abilities or special training to use effectively.

The aspect of total control is this software's strongest selling point. Other visualization software take inserted data and produce visuals, but the transformative work between the data and the

visualization is often under the hood, and it is difficult to tweak small aspects of the visualization (such as colors, sizing, placement). This is where incorporating Excel becomes NodeXL's strength rather than its weakness. The researcher can insert as many additional variables as they wish for each node and for each edge, and change graph qualities in coordination with those variables. For example, I inserted a column for parts of speech and for semantic tags for the nodes on some exercises, and then asked the program to color each POS or semantic category a different color. Equally important was the ability to remove or add individual data points. As discussed above, reviewing the corpus lines sometimes brought out important information that was not apparent in the NodeXL images, and I could manually add a lexeme or phrase, choose the appropriate size, and connect it to the appropriate existing nodes.

The visualizations are also intuitively helpful to both the researcher and the students. As a researcher, it was helpful to take all the relationships identified in AutoMap and visualize them before choosing points to highlight and teach. The initial visualization was helpful, for example, in noticing the preponderance of transition words before some of the phrases and in differentiating between the frequent and infrequent nouns that followed *due to the*. The students who reviewed the materials were similarly drawn to the NodeXL graphs. Being able to visualize the parts of sentences as puzzle pieces, some optional, that could all be fit together around a central phrase, was a new and interesting way think about phrases and sentences.

The final draw of NodeXL is that it is not intimidating for materials designers with little knowledge of programming. All that it requires is a basic knowledge of Excel and a willingness to learn the quirks and rules of the program. A couple hours clicking through the features and reading through the Help or online forums will give the materials designer enough information to get started, and additional skills can be learned as needed. Thus, despite its flaws, its shorter learning curve and the control and abilities it gives the designer are major benefits.

Conclusion

In conclusion, SNA provides useful tools that can take massive files of data and turn them into easily understood visuals with a lot of information in an attractive package. However, they are far from perfect, and require patience with their small oddities and limitations. They also require

a materials designer with a solid grasp of the corpus data, who can use corpus tools to interpret the findings and choose what elements to highlight. Without some sort of informed filter, these programs will merely highlight the most common phrases and the connections between them, which means that they would highlight the fixed class words (prepositions, articles, and the like), rather than the less frequent but more unique or important phrases and words. A designer who understands what the measures and numbers represent, however, can work with the data to highlight the linguistic features that are useful for students. Thus, these tools do not replace the need for an intelligent and trained linguist to interpret the data; but they can make it possible to portray the information in unique ways.

Evaluation of Worksheet Value

Given the work necessary to learn the tools, rotate between them, and build cohesive material, it is clear that the goal of combining SNA and corpus tools to create worksheets with corpus-based network visualizations is no small effort. Is it ultimately a worthwhile effort? The next section addresses this question by explaining the challenges and benefits of the approach.

Challenges

The primary challenge of this approach is the time-consuming nature of materials construction. Designing materials from scratch is never a speedy process and will always involve trial and error. However, this approach is particularly slow because there is so much information to work with and because of the number of tools involved. The corpus and semantic network approaches provide a large variety of rhetorical, syntactic and semantic information about the context of each phrase, and it takes a while to determine which information is most useful. Even once the information has been selected, the process of rotating between the corpus and SNA tools, including some that were not made with pedagogical materials development in mind, is a lengthy commitment.

Secondly, the necessity for teacher-driven content and learning outcomes may or may not be a challenge, depending on the pedagogical perspective of the teacher and the needs and desires of the students. Creating visuals from corpus content means highlighting certain pieces of

information and relationships at the expense of others. Students are not able to freely explore corpus lines, but rather work to identify and practice selected patterns. In this case, this approach fit the expressed needs of the students and thus was more of a benefit than a challenge. In addition, it is possible to keep the visualizations but create more exploratory exercises that gave students greater freedom. However, designing the visualizations puts the power of interpretation in the hands of the designer, and a poor or inaccurate visualization can easily give students a false impression, as visuals can speak more powerfully than text. It is easier for a student to trust a graph that shows a phrase always being used for a certain purpose than to scan through fifty corpus lines and note how else it is used.

Finally, the emphasis on visualizations and directed questions may make it difficult to align the materials with student levels. NNS graduate students represent a diversity of language backgrounds and needs; some excel at writing while others excel at speaking; some struggle with grammar while easily handling idioms and phrases; others have excellent command of grammar but lack vocabulary. A less structured corpus exercise lends itself to working with such a varied population, as students can draw out the information that is helpful to them. Because these materials are so specific, emphasizing certain vocabulary, syntax or rhetoric points, it is more likely for some of the content to be too simple or to be irrelevant. This challenge is slightly offset by including the Corpus Examples section in the worksheets for each phrase, so that more advanced students can explore the corpus lines and draw their own conclusions.

Benefits

Despite these challenges, there are distinct advantages to this approach. The first two have to do with the benefits of visualization; the second two are how this approach meets two specific needs for advanced NNS learners, giving them important transferrable skills beyond fluency with these select phrases.

First, the researcher benefits from the process of semantic network visualization because it brings to light patterns that they might not otherwise see. Categorizing the nodes, sorting through frequency information, and observing the clusters identified by NodeXL means that very little

can get overlooked. For example, when first categorizing the rhetorical uses of *it be * to*, I assumed that this phrase was primarily used to justify methodological decisions and highlight key points. However, as I sorted the nodes into categories and checked them against the corpus lines, I noticed a third prominent use: the phrase was used as a type of topic sentence. For example, the sentence "it is critical to consider scale-up and manufacturing issues," led to a discussion of those issues. It was only through sorting nodes and corpus lines that I was able to see this function and thus to include it in the materials.

The students also benefit from working with the visualizations. The proven benefits of visualizations were discussed in the first chapter. Reinforcing text with visuals increases reader comprehension (Baggett, 1989), leads to better long-term retention (Peeck, 1989), is especially helpful for introducing readers to unfamiliar concepts and new material (Mayer and Gallini, 1990), and improves student ownership of the material (Mondal, Mondal & Das, 2016). The students who reviewed the materials were able to quickly understand the visuals; in fact, while they had many suggestions on improving the questions, none of them had complaints about the visuals, instead stating that they could understand the layout and work with the information easily. As the goal of the materials is for students to use the information (rather than simply understand it), the benefit of quickly communicating information made it much easier for the students to jump from understanding to applying.

The combination of corpus and network visualizations also met two needs of academic NNS writers. It develops learner vocabulary, specifically by asking students to identify the nuances between word families and synonyms and practice using the synonyms in sentences. NNS have been found to use more ambiguous or indirect language (Hinkel, 1997) or to rely on the words they know well, rather than risk unfamiliar words (Coxhead, 2000). These exercises present words and terms in semantic families, and the summaries give additional low-frequency synonyms found in the corpus lines. Because students work through the choices on their own, discuss the nuances and place words on clines as a class, and practice making sentences, they learn nuances that are otherwise inaccessible.

And finally, this approach holistically combines sentence variety with vocabulary variety. NNS writers can suffer from lack of sentence structure variety (Hinkel, 2003). By portraying the

different categories of words and phrases surrounding a certain phrase, students develop their abilities to identify and move around the pieces of a sentence. Each of the phrases had questions that required constructing and modifying sentences, as the student is asked to vary the placement of the phrase in the sentence and determine the syntactic and rhetorical effects of the change. The students also interact with their peers' sentences, thus benefiting from observing how their peers write and the sentence structures they employ. While it is possible to have these types of exercises without the visuals, the visuals emphasize the puzzle-piece quality of good sentences and make it easier for students to decide what pieces they want to add, remove or move.

Conclusion

This chapter sought to answer three primary questions: What is the most effective framework for combining corpus and social network tools to create teaching materials? Are the SNA tools capable of creating student-friendly visualizations? And, once an optimal method of creating materials has been found, are the end results worth all that work?

It is useful to answer the second question first, as the usability of the SNA tools contributed to the challenge of building an efficient framework. While AutoMap was not designed for creating teaching materials, it was an efficient and effective method of extracting network information from texts. However, the majority of the work was done in NodeXL, and this was a more mixed experience. While NodeXL gives the materials designer complete control over the visual and is easy to learn, it can be clunky. The benefits of detailed control are balanced by the difficulty of manipulating more than one data point at once, and the possibility of the program not saving visualization choices. Over the course of materials developing, I was developed best practices that protected against NodeXL's weaknesses. These practices includes as saving copies of files often, always exporting the visuals before leaving the program, and documenting design choices carefully so that they could be quickly replicated if lost. NodeXL made it possible to develop complex, polished visuals without extensive technical skills, but it also required careful handling. Thus, while the SNA tools can be used to create effective, student-friendly visualizations of corpus findings, they can only do this within a well-developed design framework that provides safeguards for their flaws and best practices for how to use them for purposes outside of their original intent.

The optimal framework that answers the first question highlights a tension that is common in educational materials development. For the materials to be well-received and effective, the designers need to incorporate both general design principles and in-depth discipline-specific knowledge. Of the four principles of design that emerged through the process, three apply to any endeavor to create educational materials, and one required technical linguistic knowledge. The principles of establishing clear learning objectives, maintaining simplicity, and designing attractively are none of them specific to corpus materials or teaching academic English. As I approached this project with the skills of a linguist rather than a curriculum developer, I had to learn to apply these principles by trial and error. For example, when I thought I had adequately simplified the material, the students who reviewed my material told me it was confusing and they were overwhelmed, and so I simplified further. Similarly, given the arguments that corpus is excellent for student-driven learning (Bernardini, 2002), I originally created too many open-ended questions and learned later in the process that my student audience reacted best to and were most willing to learn from more targeted questions that had clear answers. Thus, some of the key elements of developing corpus-driven exercises on phrases in academic English required deepening my knowledge of curriculum development and materials design, fields that were outside my personal experience with corpus linguistics and SNA.

However, the fourth principle points to the key role my linguistic knowledge played in developing these materials. Students were most interested in the rich contextual corpus information I included in the worksheets and which they would have had difficulty accessing on their own. The synonymous terms in the variable slot of the variable phrases; the clear lists of the different rhetorical purposes of the two variable phrases; the many syntactic options for the placement of *due to the* in sentences: all of this was valuable information. The learning objectives I eventually developed all built on these corpus insights, as I translated the technical terms into explanations and exercises that students could work with.

Thus, the final design framework had to weave together SNA software, corpus software, and other tools; consistently focus on the proven design principles of simplicity, attractiveness, and clear goals; and translate technical corpus-based information into approachable exercises. Figure

12 summarizes my working solution to this three-pronged challenge. The main takeaway is that the process is iterative, and that student feedback is key. Without student feedback, the designer cannot know if they are meeting all the goals and still producing useful materials. While the designer does not need to take every feedback comment into consideration, the patterns in the responses will highlight the materials' problem areas and strengths.

The final question was, once this framework has been developed, do the results merit the work? The full answer to this question is left to the next chapter, as the NNS student feedback focused on the material's accessibility and usefulness, not on the measurable impact it made on student writing skills. However, the preliminary indications from the student feedback touch upon a few key benefits. First, developing the visuals assists the designer in identifying patterns, and it assists the students in conceptualizing the syntactic elements of a sentence and their options in assembling the pieces. Second, the use of SNA analysis leads to an emphasis on developing vocabulary variety and sentence variety, two key areas of growth for NNS writers (Coxhead, 2000; Hinkel, 2003). The next chapter will explore how these potential benefits were explored and confirmed in a classroom setting.

CHAPTER FIVE: CLASSROOM IMPLEMENATATION

Introduction

At the end of the second phase of the research project, four worksheets and summary sheets had been created to be used in an advanced English for Academic Purposes (EAP) classroom, ideally with graduate students in STEM fields. The third and final phase of the dissertation was thus to use these materials in two separate classes to see if, first, they facilitated student gains in writing fluently with a variety of the phrases, and second, how students and teacher assessed the usefulness of the materials in classroom use.

The third research question guided the process of classroom implementation, assessment, and evaluation. The research question was as follows:

(3) Can these visualizations lead to better teaching outcomes than the results documented in Cortes (2006) and Jones & Haywood (2004)?

(a) Will students show improvement in the variety and idiomaticity of their formulaic language from a pretest to a posttest?

(b) Will a delayed posttest confirm the posttest results?

Although the third research question is predominantly concerned with the test results, the implementation also created rich opportunities for student and teacher feedback and materials evaluation, so this chapter will first lay out the methodology used in implementing the materials, then overview the results (separated into student test results, student evaluation of materials, and teacher evaluation of materials), and finally discuss, based on these results, the efficacy of the visualizations specifically and the approach as a whole.

Overall, both the test results and the student and teacher evaluation were positive, indicating that the corpus analysis plus network analysis depicted through network visualizations provided students with valuable and accessible information on language patterns that they were able to grasp and apply. However, while the corpus network visualizations played an important role in

bridging the gap between technical linguistic information and graduate students' ability to improve their technical writing, it is still an imperfect bridge. The improvement in the posttest and delayed posttest scores demonstrates the value of delivering this information to NNS graduate students, and how they were able to immediately incorporate the linguistic information to enrich their writing. However, their suggestions for improvement indicate that there were still spaces where the gap between their knowledge and the technical information presented was daunting. Thus, while this method is one with great potential, it can yet be refined to facilitate easier understanding and acquisition of the linguistic material.

It is important to acknowledge that the data in this chapter comes from a small sample of students; a total of 14 students participated across the two sections of "Essentials of Academic Writing" in which I lectured, but not all the students who came to the first day of each class returned for the second day or the delayed post-test. Thus, I will rely on descriptive statistics and student feedback in evaluating the materials. Given the new approach to materials design, this small sample still represents a valuable field test. The implementation provided valuable insights on honing the design process of the materials and on the usefulness of the materials to the students. Although I cannot generalize the findings to future uses of these materials, they indicate the teaching potential of combining corpus insights with network visualizations.

Classroom Methodology

Context

As mentioned in Chapter 2, the final implementation of the materials developed in Chapter 4 took place in the Fall of 2019 in two sections of the "Essentials of Academic Writing" short course that was offered to international students at Purdue; the primary target audience of that course is graduate students. This non-credit, six-week course, whose syllabus is found in Appendix F, introduces students to academic genres, rhetorical moves and sentence structure. I taught the two 75-minute lessons in the fifth week of the course. The first ten minutes of the first class were devoted to the pretest, and the final fifteen minutes of the last class were dedicated to the posttest, the feedback questionnaire, and student discussion. The course concluded the next week, so the posttest was performed exactly one week after the second lesson.

The first section took place in the first half of the semester, allowing time for revision of materials before implementing again in the second section. For the first implementation, eight students participated in the first class; however, as one student arrived late, only seven took the pretest. Six returned for the second lesson and took the posttest and evaluated the materials; only four were present in the final class a week later when the posttest was administered Three were graduate students in civil engineering, and one each in aeronautical engineering, statistics, animal sciences, and physics. The eighth student (who was one of the two to not return) was an undergraduate.

In the second half, the first class had seven students, so again there were seven pretests, for a total of 14 pretests across the two classes. Four returned for the second class and took the posttest, for a total of 10 posttest results. When the delayed posttest was administered a week later, five students took the posttest. However, only three of them had attended both classes. One student did not answer most of the test questions, and their responses indicated that they had not attended either instructional day, so their results were not included. The fifth student marked that they had attended the first day. As every other posttest and delayed posttest had been taken by students who had attended both days, this delayed posttest was not included in the results. Of the seven students in the first class, three were in engineering programs, one in molecular biology one in information systems, and the last one in urban planning. Two of the engineering students and the information systems and urban planning students returned to the second class. In terms of L1 background, of the fourteen total students who participated, 10 were Chinese speakers, one was an Arabic speaker, one was a Spanish speaker, and the last student did not volunteer their native language.

Redesigned Lesson Plan

Given that students' evaluation of the materials was built into the process of classroom implementation, it is impossible to discuss the lesson plan and supplemental materials that guided the implementation without mentioning how it was revised by teacher evaluation and student evaluation. Thus, the following description of the lesson plan and the supplemental materials will include a preview of some of the points raised in the student feedback on the materials and the teacher assessment of their usefulness and efficacy.

Originally, after the materials for the four phrases were finished, I created a lesson plan to teach all four phrases during the two 75-minute classes; it can be found under "Original Lesson Plan" in Appendix D. However, the lesson plan was redesigned twice; the first redesign before the first implementation, when the materials covered in class were reduced. The process of developing the original lesson plan made it clear that it was not realistic to fit four phrases with three learning goals each into two 75-minute classes, especially as 25 minutes of the 150 minutes had to be devoted to the pretest, posttest and student feedback on the materials. Thus, the first redesign eliminated the second fixed phrase (*there be no*). This one was eliminated because the first phrase (*due to the*) did not have any grammatical variation in it, unlike *there be no,* and thus was the easiest starting place for introducing the fixed and variable phrases. This change allowed for more of the class time to be spent on the worksheets and student writing.

The second redesign happened between the first and second day of the first implementation and was extended when preparing for the second implementation. After the first day in class with the students, it was clear that they needed more scaffolding and direction in understanding the material and applying it, in part because their English levels were lower than expected, and lower than that of the students who had provided most the feedback on the materials in development stage. Thus, the second lesson plan revision focused on dedicating more of the lesson time to scaffolding, especially in defining key terms and in conducting group exercises that used corpus examples to strengthen concept comprehension and application before writing practice. In Appendix D, the final revised version used in the second implementation can be found under "Revised Lesson Plan." The second redesign of the materials meant that PowerPoint slides were developed to guide the class through the concept; the role of the PowerPoints in the classroom are discussed below, and the PowerPoint slides used in the second section of the class can be found in Appendix E.

Supplemental Teaching Materials and Related Revisions

Originally, the supplemental materials used in class were a handout that explained the corpus and the basis for the materials in class, and a set of PowerPoint slides that defined important vocabulary on the worksheets and provided a visual version of the lesson plan (directing the order in which questions were to be worked on, and the time allotted to each one).

However, along with the lesson plan, the supplemental materials were revised. In the first revision before the implementation, the handout was totally removed. In the second revision (which applied to both the second day of the first implementation and the entire second implementation) the role of the PowerPoints was expanded, and corpus examples incorporated slightly differently.

Rather than hand out the Introductory Sheet (found in Appendix E.I) that explained what the phrases were, why they mattered, and where they came from, I chose to include that information on the slides. The reason for this was that between the pretest, the worksheet for the first phrase, and the summary sheet for the first phrase, there were already several handouts in the first fifteen minutes. A fourth handout would have been overwhelming. In addition, it was a more efficient use of time to summarize the main points verbally than to have students distracted by a sheet of information that they would read and ingest at very different speeds.

The PowerPoints were expanded after the first day of the first implementation, as it was clear that students would be able to participate more fully if the slides provided more scaffolding for terms and had more activities to strengthen understanding. Thus, the original slides were expanded, and new group exercise slides were added.

The first addition was a set of slides explaining how to read and work with the diagrams. Figure 14 is an example of one of these slides. This slide helped students to understand the type of sentences that led to the collection of the data, and to visualize the puzzle piece nature of sentence construction from the diagrams. This approach to explaining the visuals was successful. While several students from the first section expressed that they struggled to understand or apply the illustrations, no one in the second section reported this issue. In fact, on the feedback form for the second section, all four students answered the question “How useful were the pictures?” with a 5 on the 1-to-5 Likert scale (‘very useful’).

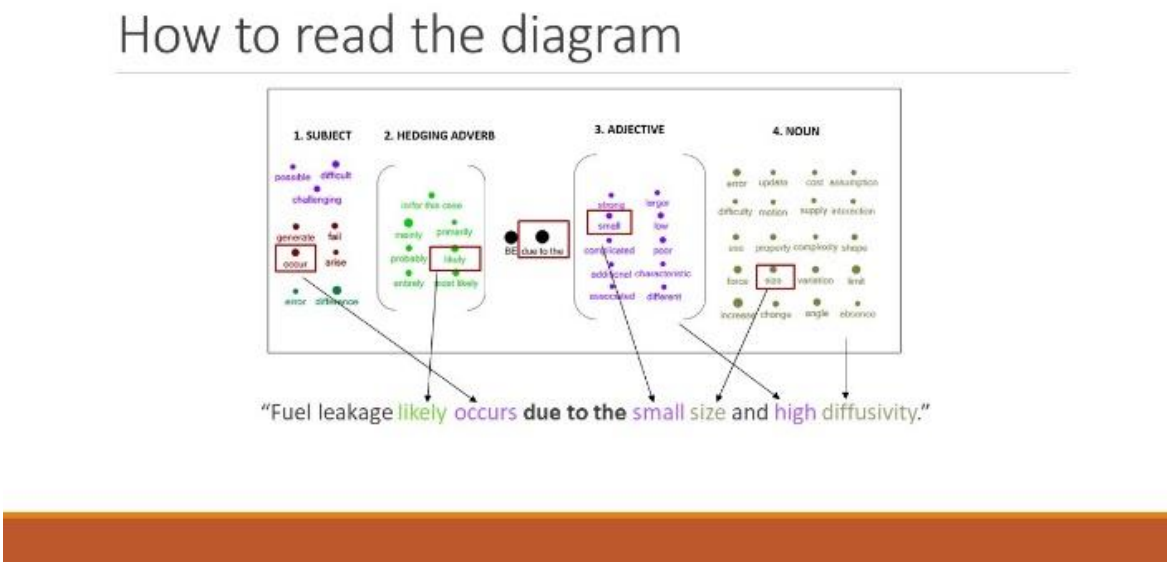


Figure 14: Slide Illustrating Worksheet Visuals

In addition to explaining how to use the sheets, the existing content on the PowerPoints was deepened. This was primarily done by adding group exercises where students reviewed and practiced the information about the phrases before we transitioned to writing sentences with those phrases. Figure 15 and 16 show what this expansion looked like for the *due to the* slides. In Figure 15, the original introduction to hedging and strengthening had the terms briefly defined, and then asked students to use the hedging and strengthening words in their own writing. In the redeveloped slide, shown in Figure 16, hedging & strengthening has been expanded into a full slide that includes an exercise where the students identify the strengthening and hedging words in multiple complete sentences from the corpus.

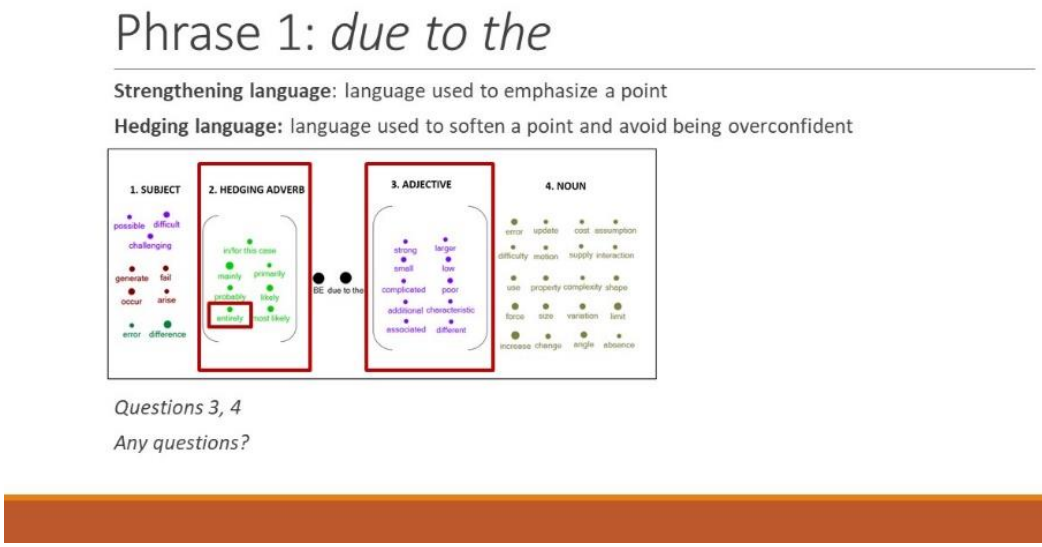


Figure 15: Original Hedging & Strengthening Slide

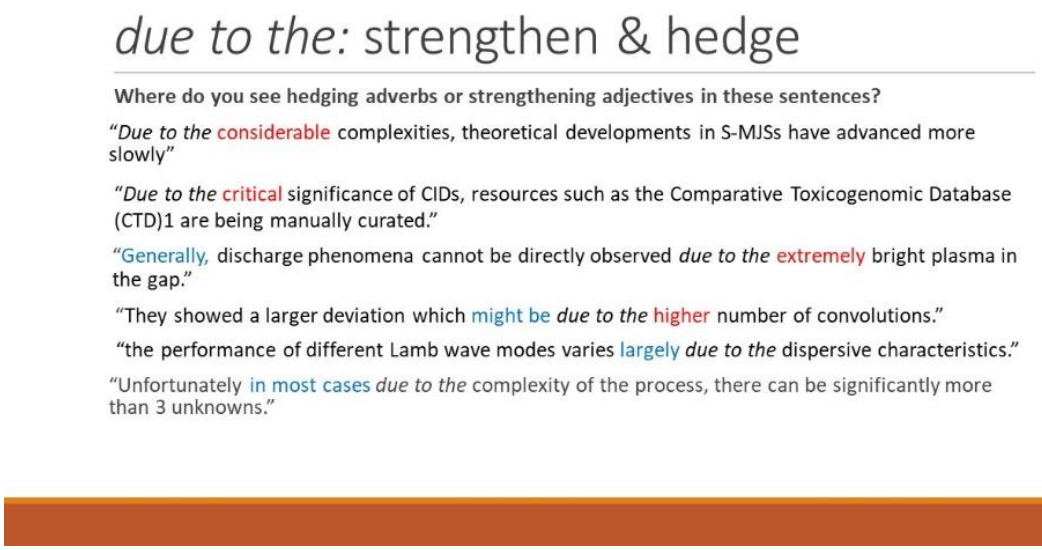


Figure 16: Revised Hedging & Strengthening Slide

Similarly, in the revision, additional activities outside the exercise sheets were developed. Figure 17 is an example of an activity that provided a review of the adjectives in *it be * to*. The adjectives in the center of the slide did not have markers that indicated which category they belonged to until after the students classified the adjectives in groups. At that point, I showed

them my categorization. Figure 18 shows a comparable activity for practicing the primary rhetorical uses of *it be * that*: the students again as a class discussed how they would classify each sentence before I showed them how I had matched the sentences to the three uses.

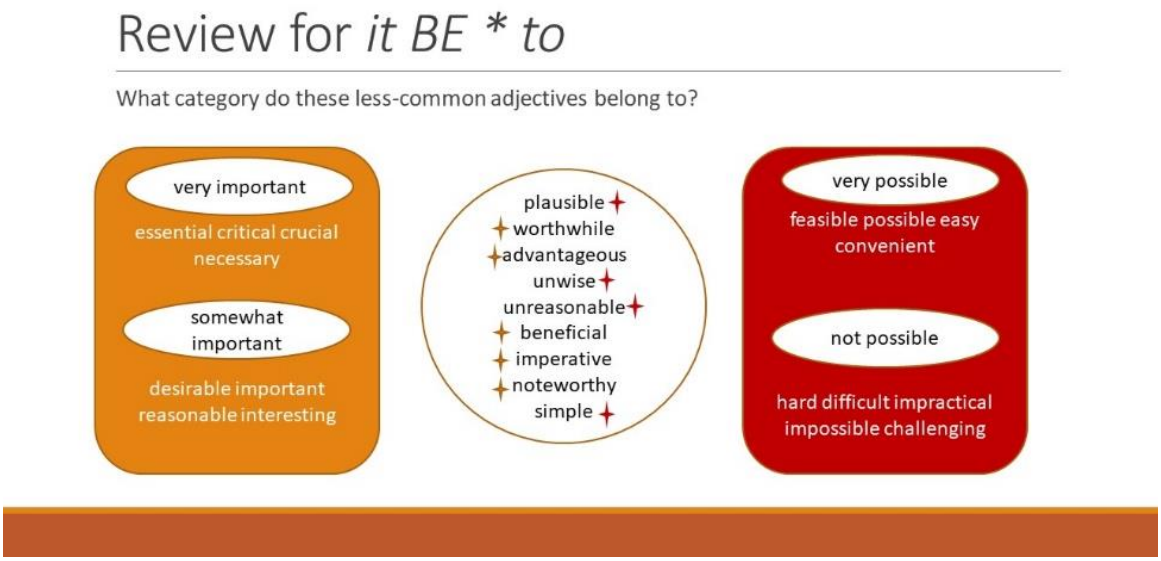


Figure 17: Scaffolding Exercise Developed for *it BE * to*

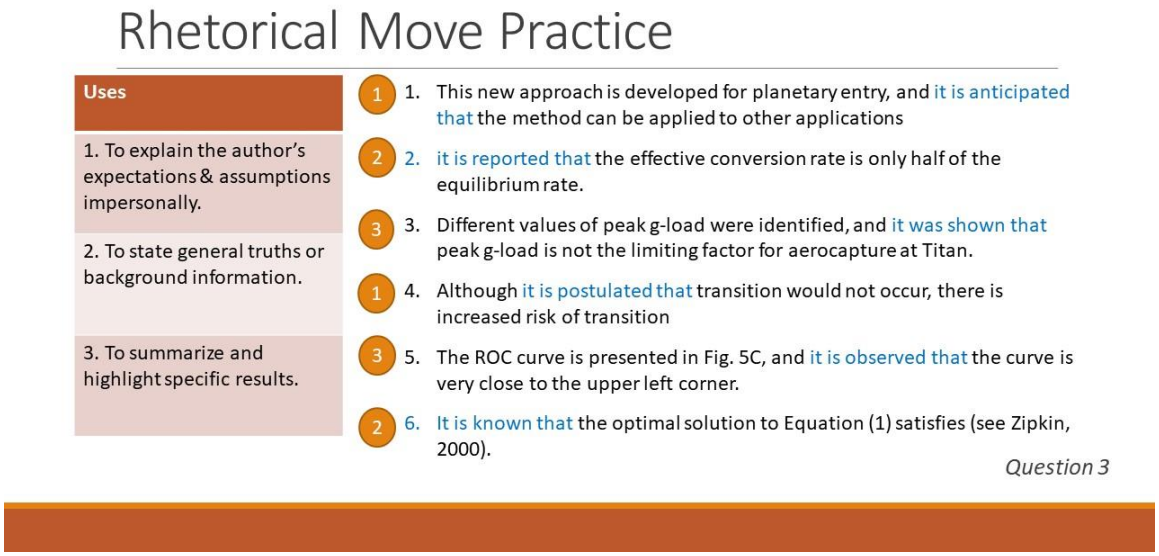


Figure 18: Scaffolding Exercise Developed for *it BE * that*

One crucial element of these redeveloped slides is that they show full sentences with the academic phrases, rather than the one-line snippets with twelve to twenty words that were included under "Corpus Examples" on each worksheet. While I had included the Corpus Examples format because it allowed students to see a large quantity of examples on one sheet and to look for overarching patterns, their in-class feedback on the first day of the first implementation was that they had a hard time understanding the snippets, and thus were not sure how the phrases fit into the full sentences. Putting full-sentence examples on the slides provided an intermediary step where they could understand the phrases and their connection to the rest of the sentence better before they scanned the snippets for patterns to answer the exercise questions. This final set of better-developed PowerPoint slides can be found at the end of Appendix E.

Classroom Teaching Results

This approach used both formative and summative evaluation (Tyler, Gagné, & Scriven, 1967) to evaluate the effectiveness of the materials and the usefulness of a corpus-based network visualization approach to academic phrases. The first set of results is the summative data from the pre-, post- and delayed posttests. It demonstrates that the students did improve in their performance on the nine teaching objectives developed in Chapter 4. This is followed by a summary of the major themes from the student responses on the feedback form and their in-class discussion of the materials' usefulness. The students' formative evaluation of their experience also provides valuable insights into what made the materials effective and how they could be improved in the future. This is concluded with a review of the teacher assessment of the materials and class based on their use in the four class sessions.

Pre-, Post- and Delayed Posttest Results

As explained in the methodology section, a ten-minute pretest was given before the first class started, and a similar post-test was given at the end of the second class. A delayed post-test was given a week later. In all cases, the tests consisted of two questions for each of the nine learning objectives, for a total of 18 questions. All three tests can be found in Appendix G. The test results were first reviewed qualitatively, and the source of errors marked. Then the student answers were divided up into the three categories of no answer, correct, and incorrect for

analysis. Between the two classes, 14 students took the pretest, 10 students took the posttest, and seven students took the delayed posttest.

Given the lower proficiency of many of the students, most students were unable to complete all questions. Every student started on the first question and answered as many questions as they could before the time ran out. Some students skipped questions they could not answer, but all students proceeded linearly. This means that an increase in the number of correctly completed question is also evidence of a learning gain, as their learning allowed them to rewrite the sentences and explain their answers more quickly. Table 16 below summarizes the average number of unanswered, correct, and incorrect answers across the three tests, and the percent of answered questions that were correct.

Table 16: Overview of Test Results

	Pretest	Posttest	Delayed Posttest
Average # of Questions Unanswered	10.5	5.5	1.9
Average # of Questions Correct	5.4	10.5	12.4
Average # of Questions Wrong	2.1	2	3.6
Average % of All Questions Correct	30%	58%	69%

From the results, there is evidence of learning across the test. On average, students went from leaving 10.5 questions blank, to 5.5 on the post-test and fewer than two on the delayed posttest. From the pre-test to the post-test, students nearly doubled the amount of questions they answered correctly in the ten minutes (from 5.4 to 10.5, on average). The delayed post-tests do not provide as much evidence, as they were only taken by seven students. In both cases, they were given on the last day of class, which had low attendance. However, the students who did take the delayed posttest retained their learning gains.

When the student answers are separated by the learning objectives of each of the three phrases, we can observe which areas had the most growth. Below is a list of the three target phrases and their related goals (Table 17), and a breakdown of test data by target phrase and goal (Table 18).

Table 17: Learning Objectives by Target Phrase

Phrase	Learning Objectives
due to the	1. Identify the antecedent of the phrase
	2.Change or add a hedging/strengthening adverb
	3. Rewrite sentences to change phrase order and rhetorical impact
*it is * to*	1. Change adjective of possibility
	2. Change adjective of importance
	3. Identify and use the phrase's three rhetorical functions
*it is * that*	1. Identify qualifying phrases
	2. Identify and use the phrase's three rhetorical functions
	3. Identify the passive agent of the verb

Table 18: Test Results by Target Phrase and Related Goals

	Pre-test			**Post-Test**			**Delayed Post-Test**		
	No answer	**Correct**	**Incorrect**	**No answer**	**Correct**	**Incorrect**	**No Answer**	**Correct**	**Incorrect**
due to the									
Goal 1	64%	21%	14%	45%	50%	5%	14%	86%	0%
Goal 2	93%	4%	4%	5%	70%	25%	0%	86%	14%
Goal 3	64%	18%	18%	30%	65%	5%	21%	71%	7%
*it is * to*									
Goal 1	4%	82%	14%	5%	75%	20%	7%	64%	29%
Goal 2	71%	21%	7%	65%	25%	10%	29%	36%	36%
Goal 3	93%	7%	0%	40%	50%	10%	7%	50%	43%
*it is * that*									
Goal 1	29%	50%	21%	60%	40%	0%	14%	71%	14%
Goal 2	32%	46%	21%	10%	75%	15%	0%	93%	7%
Goal 3	75%	21%	4%	15%	75%	10%	0%	71%	29%

It is impossible to make strong claims based on the descriptive statistics in Table 18, as part of the variance is due to whether the students were able to reach both of the questions for each particular goal. Some of the 'no answers' are due to the students not reaching the questions, and the others are due to the student skipping the question because they were not confident in their ability to answer.

However, there are a few notable points. First, the no answer rate decreases substantially for all but Goal 1 of *it is * to*, showing that the gain in speed is distributed across all the learning goals. Second, two of the largest gains are in similar goals; both Goal 3 of *it is * to* and Goal 2 of *it is * that* have to do with identifying and using the phrases' rhetorical functions. And finally, while the goals that emphasized finding synonyms (Goal 2 of *due to the* and Goals 1 & 2 of *it is * to*)

do not show dramatic improvement, it is important to note that the quality of the correct answers improved. For example, in the pre-test, students primarily wrote 'hard' or 'impossible' when asked to replace an adjective of possibility with a word that meant less possible. In the post-tests, the answers included 'infeasible, unwise, unpractical, inappropriate, plausible,' in addition to 'impossible, hard.' Several students wrote down multiple words. The same pattern of increasing quantity and quality of correct synonyms was seen in the reporting verbs for *it is * that*. Thus, there is solid indication that at least some of the students acquired new rhetorical understanding and deepened their knowledge of synonyms.

Student Feedback

At the end of both implementations, students received a feedback form and were asked to fill it in; I sat away from the students and collected the forms once they were turned over to provide some anonymity, though the students may have been more likely to respond positively since I was still in the class. The forms had three questions where student selected their answer from a five-point scale, and three open-ended questions. The feedback form can be found in Appendix G. 10 students filled out the form: six from the first class and four from the second. While the responses to the scaled questions were generally positive in the first class, there was somewhat conflicting feedback on the open-ended questions that addressed the unique role of the visualizations and some ideas for improvement. After I implemented many of the suggestions, the feedback from the second class was overwhelmingly positive. Both are discussed in turn below.

Feedback on Scaled Questions

The three questions with five-point scales asked the students to rate how useful the visuals were, how useful the information was, and how much of the content they understood. The responses from the first implementation are summarized in Table 19, and the responses from the second implementation are summarized in Table 20.

Table 19: Scaled Feedback Question Responses from First Implementation

Questions with Five Point Scale	S1	S2	S3	S4	S5	S6	Average
Q1. How useful were the pictures?	5	5	3	5	3	5	**4.3**
Q2. How much of the information was useful?	5	5	4	5	3	5	**4.5**
Q3. How much did you understand?	4	4	4	4	2	4	**3.7**

Table 20: Scaled Feedback Question Responses from Second Implementation

Questions with Five Point Scale	S1	S2	S3	S4	Average
Q1. How useful were the pictures?	5	5	5	5	**5.0**
Q2. How much of the information was useful?	5	4	5	5	**4.8**
Q3. How much did you understand?	5	4	4	3	**4.0**

In the first implementation, as reported In Table 19 for questions 2 and 3, the students rated the "usefulness" of the material at 4.5 on average, which was higher than their rating of how much they understood (3.7 on average). This indicates that the material was slightly too difficult. They appreciated the material and thought it was important, but that the lack of scaffolding and/or manner of presentation was confusing to them. The second implementation, reported in Table 20, received higher scores for both these questions; students rated the usefulness at 4.8 and how much they understood at 4.0. Thus, the changes between the implementations may have increased student ability to understand the materials. The increase in understanding also led to higher usefulness ratings for the visuals, from 4.3 to 5.0.

Feedback on Open-Ended Questions

The three open-ended questions were: "What was most helpful about the materials?" "What was most confusing or unhelpful about the materials?" and "Other comments or questions?" The students in the first implementation provided conflicting feedback. One student mentioned the visuals as the most helpful element; two students mentioned them as confusing. Of those two, one found it hard to distinguish between the different visuals, and the other felt like the visuals

were not adequately self-explanatory. In addition to the visuals, the students named the rhetorical uses of each phrase and the teaching about variable phrases and structures as the most helpful parts of the materials. This was unsurprising, as they were visibly most engaged in class when we discussed synonyms for the most frequently used adjectives and verbs, and when we covered the rhetorical uses of the phrases.

For the second implementation, only four students filled out the feedback form, and their comments were brief. They answered the first open-ended question about what was most helpful with all-encompassing answers like "phrases & their function" and "three common and useful phrases were introduced." None of them answered the question about what was confusing or unhelpful, and the only comment on the last question was one student asking to learn more about corpora work. As they answered the Likert scale questions with more positivity than the first group, this lack of issues is not surprising. Both the generality of the positive answers and the absence of negative answers could indicate that the adjustments I made to the content of the second set of classes was useful for students.

Teacher Assessment

As the teacher, I followed the example of Lee & Swales (2006) and took copious notes on both classes as they happened and immediately afterward, primarily documenting student responses to activities, including positive responses and students' questions or areas of misunderstanding. My primary takeaway from the very first class was that the students had a lower level of comfort with academic English than I had anticipated, and this manifested in three ways. First, most of the students were hesitant to write sentences on their own, even after we had discussed the phrase and several examples. As several of the exercises included a prompt to write an original sentence with the phrase, this meant they just waited or worked on the other exercises when I gave them time for sentence writing. Second, it took longer than expected to go through and understand the example sentences. Finally, the students were most interested in the synonym options, especially in the words that were unfamiliar to them. Building vocabulary was still a key concern for them.

As mentioned above, to address the first concern I built much more scaffolding into the lesson plan. I included slides that labeled every part of the example sentences, and for almost every learning goal, I created a PowerPoint slide with a related exercise, so that they spent more time reviewing all the information and playing with existing sentences before writing their own sentence. While the first class was still hesitant to write on the second day, the second class responded more easily to the sentence-writing prompts. The thoroughness with which we covered the examples seemed to give them the confidence to mimic the structures while writing in their own discipline.

As a whole, my notes from the second day of the first class and from the second class were much more positive about student participation. Both classes asked thoughtful questions about how to use patterns best. A student in the first class mentioned that the most common phrase in his discipline was "it is obvious that," and he found this annoying, as he found that it was often followed by a statement that was not obvious to him. This led to a discussion of when to use the most common patterns, and when to seek out more creative expressions. The students were especially interested in accurately acquiring and using the less common synonyms for the variable phrases, so we discussed the subtle connotations of the unfamiliar synonyms and when they could best be used.

Finally, both classes were most engaged when we discussed how the variable phrases could be used to express author stance and voice while avoiding personal pronouns. Again, we discussed that these popular workarounds do not always lead to the best or clearest writing, but the students found it valuable to be able the recognize and manipulate the tools used for inexplicit author stance. The whole class was also engaged when we did PowerPoint exercises (such as sorting adjectives into adjectives of possibility vs. importance, identifying the knower behind *it be * that*, and naming the rhetorical uses of example sentences for both variable phrases).
Thus, my assessment from the first lesson led to several adjustments that seemed to substantially improve student attention and buy-in. By providing more scaffolding exercises and making space for discussion of points that caught their attention, I could accommodate the students who were not comfortable jumping into writing and ensure that they still learned information and skills that they found valuable.

Discussion

This chapter summarized the classroom teaching and its results, including how the material was adapted to the circumstances, student test results, and student and teacher feedback on the materials. All this work sought to answer one primary question: is teaching from corpus findings, using network visualizations, an effective approach to aiding advanced NNS students improve their use of formulaic language in academic writing? The answer rests first, in the effectiveness of the visualizations, and second, in the usefulness of the approach as a whole.

Efficacy of Visualizations

The graduate students who reviewed the materials when I was designing them were overwhelmingly positive about the visualizations. The direct feedback on the visualizations from students in the first class was more mixed, however. The difficulties with the visualizations came primarily from the density and unfamiliarity of the information. The visualizations combine part of speech color-coding, rhetorical information, frequency information, synonyms, and ordering information. While some students in the first class appreciated this, others found it overwhelming or confusing. I was able to improve student understanding in the second class by including several slides that walked students through the information in the visualizations and how to read and use them. In my in-class observations, the students were more willing to use the visualizations after that introduction, and they responded more positively on the surveys. This indicates two possible avenues of improvement: either the visualizations could be kept as is and better introduced and demonstrated, or they could be simplified or redesigned to be more intuitive.

Either way, the visualizations offered a novel avenue through which the students could acquire rich contextual information based on real data and employ that in their own writing. The generally positive student feedback on the visualizations shows it had value to them. They frequently referenced the visuals in class when writing or rewriting sentences, demonstrating its value as a tool. The students' gains from the pre- to the post-tests demonstrate that much of their new skills stayed with them, and that could be due in part to the visuals. A final point in favor of these visualizations is that the learning objectives that were best received by the students were developed as I designed the visualizations and wrestled with which patterns to highlight and

what contextual information to include in the visuals. Not only did the visuals aid the students in learning; the development of those visuals led to the designer having a deeper understanding of the phrase and being able to select thoughtful and authentic learning objectives that engaged students.

Efficacy of Approach

One key component of the third research question was if incorporating network visualizations into the teaching process could lead to better results than previous attempts to teach corpus-based phrases to students. Specifically, the research question asked if the teaching led to better results than those experienced in Cortes (2006) and Jones & Haywood (2004). As each of these three corpus studies targets different populations, uses different strategies for both identifying and teaching formulaic language, and measures student success differently, there is the danger of comparing apples to oranges. However, there are some comparisons worth noting. Cortes (2006) found that while her students demonstrated increased awareness of lexical bundles, there was no change between pre- and post-instruction production of lexical bundles in their classroom writing. Jones & Haywood (2004) found a slight improvement in students' ability to fill in the missing word of a formulaic sequence in a sentence context, but no change in their use of formulaic sequences in their essays. Cortes suggests at the end of her study that "students might need more exposure to examples in their contexts" so they can see "how these expressions are used by published authors." (2006, p 401). This study did exactly that, and also included rich information on each phrase's context. The hope was that the visualizations, increase in contextual information, and deep focus on a few phrases rather than a brief introduction to many phrases would lead to better outcomes.

My use of re-writing exercises as the avenue for student evaluation is a half-way point between Cortes' tests of student awareness of lexical bundles and their spontaneous production in writing. Through the re-writing exercises, I was able to demonstrate that the students did improve in their ability to recall, manipulate, and correctly use the phrases. Thus, the students in this study were not only able to recognize the patterned language they studied; they were also able to manipulate that patterned language, select synonyms and antonyms, and name the rhetorical functions of the language they used. Given the limitations of the class time, we were unable to measure if the training had impact on their writing outside the class. Thus, while they went a step beyond

Cortes' and Jones & Haywood's students' ability to recognize the phrases and were able to accurately and meaningfully use the phrases when asked, it is not known if this will transfer to their academic writing. Thus, there is preliminary data showing that this approach has a whole may have pedagogical value; but it would take larger, longer-term studies to conclusively demonstrate that.

Conclusion

In summary, this study created and tested a new method for developing and implementing teaching materials using corpus-based findings for instruction about formulaic academic language to advanced NNS graduate students. This method emphasized the rich contextual information around the phrases and visual representation of complex linguistic data. Through test data and student feedback, though the sample was small, it was clear that this approach was both well-received and led to an increase in student reading and writing skills. Students showed improvement on almost all of the nine learning objectives. There was significant improvement in their comprehension, ability to rearrange the sentences, and understand and use different rhetorical nuances. They were especially interested in the less-frequent synonyms for common expressions and improved in the quality of their synonyms in the re-writing exercises.
One major impediment to the effectiveness of this approach was the difficulty in bridging the gap between the complex and technical syntactic and rhetoric content and the varied but consistently lower writing level of the students who were in the classes. While they were all intelligent and involved students eager to improve their writing, they were daunted by writing tasks and not always familiar with rhetoric or grammatical terms. To address this, I incorporated clearer scaffolding and exercises for each learning objective, additional explanations of the visuals, and more warm-ups and model sentences before asking students to write. This led to higher student engagement during class and more positive student feedback after the second class. The nature of the exercises accommodated this flexibility, as with more or less scaffolding and explanations, they could be immediately useful to both more advanced and less advanced students.

All in all, the results from the classroom study were positive. The process of teaching with the materials offered several clear avenues for improvement at this specific level, and when the

improvements were implemented, the materials were effective and useful. This indicates that the materials could be usefully implemented at a variety of levels, and that they are helpful resources to advanced NNS students who are improving their academic writing. Given the small sample size of this study, and the fact that the materials were redeveloped as they were implemented, it would be valuable to conduct further studies with larger student samples.

CHAPTER SIX: DISCUSSION

The previous chapters have laid out the background and methodology of this study and the results of the three phases of research. Below, the key findings from each phase are highlighted, followed by a discussion of future steps. The first phase, which analyzed the most frequent and dispersed n-grams and p-frames functionally, semantically and rhetorically, led to insights that both complemented and contrasted with previous studies of engineering writing, and showed the usefulness of p-frame analysis of academic writing. The second phase experimented with a novel combination of corpus and network analysis tools to create educational materials, and thus both evaluated the usefulness of the tools and created a design framework for this type of materials design. The final phase tested these materials in a classroom setting, finding evidence both for the value of this novel approach and possibilities for improvement. These three sets of findings point to several avenues for future research.

Findings from Corpus Analysis of Engineering Writing

While the final two phases of this project were concerned with the pedagogical application of corpus findings, the first phase was most interested in identifying characteristics of engineering writing via corpus methods. In doing so, the research built upon the work of previous linguists who had studied academic engineering writing and contrasted it with other disciplines. This had two main outcomes: first, it showcased the importance of p-frame analysis, as it can lead to findings not discoverable via n-gram analysis. Second, it provided a more nuanced picture of engineering writing, highlighting new aspects of author stance and rhetorical use of language.

The Value of P-Frame Analysis

One of the primary findings of the first phase that is of interest to other researchers is that the work of describing and understanding a corpus is not done when we have finished n-gram analysis; p-frame analysis is an essential second step. N-gram analysis has been the traditional approach to corpus analysis of academic writing, and the work on engineering writing has used this approach to identify the most frequent fixed expressions and generalize from them about patterns in engineering writing. Hyland 2008a, 2008b & 2012 are the most in-depth example of

this approach, as Hyland compared the findings on engineering writing with findings from three disparate disciplines. Biber & Gray (2010) also effectively use n-gram research to create compelling arguments about the nature of academic writing as a whole. However, Chapter 3 showed how p-frame analysis complements previous n-gram analyses in several unique ways. It allows a deeper understanding of syntactic patterns in engineering writing; it identifies variation and useful low-frequency language that complements high-frequency terms; and it can impact how we conduct writing education.

First, semantic analysis of the variable words in p-frames highlights syntactic patterns in the corpus that we cannot find with n-gram analysis. Because English is an analytic language, our grammatical patterns tend to be spread across many small words rather than carried through the addition of word endings. Thus, in English, an approach that looks for a general trend across words, allowing for variation in selected slots, will most likely identify closed class words such as prepositions, articles, and verbs that have more syntactic than semantic roles. For example, in this corpus, it highlighted the use of "be" and "use" to create passives, and the frequency of prepositional structures. While similar trends can be observed in some of the lexical bundle findings, this approach was able to identify the systematic semantic patterns associated with syntactic patterns. The predominance of passive verbs of knowing, and of nouns of research and evaluation, are two examples of this. Semantic analysis of the variable slots in p-frames in other genres may similarly be able to uncover trends that are widespread but are not visible among n-gram results.

In addition to providing a richer understanding of syntactic patterns, the p-frame analysis also complements the n-gram approach in that it identifies useful low-frequency synonyms and synonymic phrases for the most common expressions in academic writing. These aid both the researcher and the student of writing. For the researcher, this approach allows us to see not only what structures authors use often, but the variety within those structures. For example, the *passive verb* p-frames differed in the semantic category of their variable slot. Syntactically, while all were over 50% in containing verbs, some had higher percentages of adjectives. For example, 75% of the items in the variable slot in *be * from the* were passive verbs like *seen*, but 22% of the slots were adjectives like *separate, independent, obvious*). Semantically, some phrases were

more had higher percentages of verbs of knowing, while others were more likely to have technical or math-related verbs. Although this study centered on identifying the information most useful to students, it is easy to see how the p-frame approach makes it possible for the researcher to study variety within patterns. For example, a study of the most commonly employed verbs of knowing can be deepened by also investigating the less-frequent synonyms used by novice or established authors. Inter- and intra-author variation in the variable slots, as well as genre-based variation, would be excellent avenues for further research for synonyms in p-frames.

For the student writer, the benefits are also clear. Students want to learn how their published peers write, but if they merely depend on the top fifty or hundred phrases in their discipline, they run the risk of parroting clichés. The students I worked with expressed a rightful desire to avoid this trap. While academic writing will always require a large number of similar moves (citing sources, creating a niche, defending methodological choices, etc.), the p-frame approach allows students to experiment with fresh ways of performing repetitive tasks. It addresses the dual issues of lack of vocabulary (Wray, 2012) and avoidance of unfamiliar syntax (Lee & Chen, 2009). By attempting to write with the common p-frames, but using their less frequent variable slot options, students can expand their vocabulary while practicing unfamiliar or daunting syntactic patterns.

Thus, p-frame analysis and semantic analysis of the variable slots has a lot to offer to the fields of writing and education, as it provides a fresh approach that supplements and sometimes even contradicts the established wisdom of n-gram analyses. As described in the section below, thee deeper understanding of rhetorical functions in engineering that p-frames provides should also impact how we teach rhetorical functions to students. This study conducted p-frame analysis with semantic analysis for engineering writing. However, there are many other types of academic and professional writing that have been primarily studied for fixed phrases; they could benefit from p-frame analysis, as this would allow us to better understand the role of the phrases, the variety within them, and how to teach them and their functions to students.

The Characteristics of Engineering Writing Uncovered

In addition to demonstrating the benefits of p-frame analysis, this study uncovered new aspects of engineering writing that deepen our understanding of what it is and how it can be taught. The n-gram analysis complemented Hyland's (2012, 2008b) work on electrical engineering writing by validating its findings in academic research articles from three new branches of engineering (aerospace, mechanical, and industrial). Hyland categorized formulaic language according to its function; text-oriented language organizes the text, research-oriented language structures and reports real-world activities, and participant-oriented language conveys the writer's stance or engages the reader directly.

Similar to Hyland's findings, the n-grams in this corpus were functionally primarily text-oriented, with research-oriented coming in second and participant-oriented representing a small minority. In fact, in this corpus, electrical engineering lags behind other disciplines in the use of text-oriented language. While electrical engineering had 1617 occurrences per million words (PMW), aerospace had 2,729 occurrences PMW, and mechanical had 3,031 occurrences PMW. Thus, the n-gram analyses strengthened Hyland's claim that the majority of fixed bundles in engineering writing are text-oriented, either linking visuals to the text or organizing the text.

The syntactic patterns in the p-frames also provided further evidence for claims made in previous research. Biber & Gray (2010) documented the rise of prepositional phrases and pre-modifying nominalizations in academic writing. The predominance of prepositional phrases and the widespread use of nominalizations in the *noun + preposition* p-frames' variable slots confirm that engineering academic writing relies just as heavily on these constructions as the rest of academic writing. These features were not only widespread; the prepositional syntactic constructions were a key structure through which authors embedded their stance in writing while maintaining an objective-sounding voice. This analysis of the rhetorical function of these widespread prepositional structures furthers Biber & Gray's argument that academic writing relies on compact, inexplicit language.

However, the p-frame analysis also found novel characteristics that contradicted previous n-gram findings. Semantic analysis of variable slots found that the most common functional use of both

the *noun + preposition* p-frames and the *passive verb* p-frames was participant-oriented. The verbs of knowledge (e.g., *find, see, show, note, investigate, identify*) in the *passive verb* p-frames were used for a complex array of participant-oriented activities: building a shared knowledge base with the reader, directing reader attention, and using others' work as background for one's own work. Similarly, 51% of the variable slots in the *noun + preposition* phrases were research nouns (e.g. *paper, method, result,* data) or evaluative nouns (e.g. *performance, quality, effectiveness, purpose*). The research nouns were used for similar rhetorical purposes as the verbs of knowing; and the evaluative nouns provided objective-sound avenues for author stance. In this corpus, text-oriented and research-oriented variable phrases accounted for only 33% and 18% of the variable phrases respectively; this is a neat reversal of the spread of functional uses that Hyland found across n-grams.

Hyland (2012) contrasted the lack of participant-oriented n-grams in electrical engineering and biology (9.2% and 8.4% of all n-grams, respectively) with the abundance in applied linguistics and business studies (18.6% and 16.6% of all n-grams, respectively). In this study, 49% of the p-frames were participant-oriented. Since I did not look at other disciplines, it is an open question if other disciplines use p-frames for participant engagement at equally high, or higher rates. An analysis of p-frames in different disciplines will reveal if all disciplines tend to use variable phrases at higher rates than fixed phrases for participant-oriented rhetorical moves. If not, engineering's relatively low use of participant-oriented fixed phrases (e.g. *as can be seen, it should be noted*) is balanced by its higher use of participant-oriented variable phrases (e.g. *it be * to* with adjectives of importance in the variable slot*)*, and it would be worthwhile to revisit the claims that other disciplines use more participant-oriented rhetorical moves than engineering writing.

Finally, the last important characteristic of engineering writing documented in the first phase of research was that the p-frames in engineering consisted primarily of function words, and the majority of the slots were not filled with technical or discipline-specific vocabulary. According to WMatrix's semantic categorization of the variable slots, the majority of the words fell into categories like "Speech Acts, Thinking, Cause & Effect, Research, Evaluation." While it is essential that student writers learn the precise technical language necessary for their disciplines,

this technical language is not the primary vocabulary in the variable phrasal structures. The variable phrases are most used for cross-discipline rhetorical moves such as structuring the text, evaluating results, drawing attention and arguing, rather than for communicating technical findings. Again, p-frame analysis of other disciplines' academic writing will reveal if this pattern holds true for other branches of knowledge. It could be that formulaic language is predominantly used for rhetorical moves across disciplines, or that this usage of formulaic language is specific to engineering.

These characteristics of engineering writing have important implications for the teaching of engineering writing. If teaching materials are based on Hyland's n-gram findings only, then students will concentrate primarily on learning formulaic language for linking illustrations to the text and structuring the text. While these are important skills, they are not the only or even the primary linguistic skills students need to acquire. Arguably, the participant-oriented structures are more important, because they are more hidden; they provide an avenue for authors to express their stances while still adhering to the norms of objective scientific writing. These are not just linguistic skills; they are also rhetorical skills. Since the p-frames seems to play an important role in constructing persuasive arguments in engineering writing, students would benefit from being able to both recognize how this happens when they read, and practice building these arguments themselves.

Finally, the p-frame findings highlight that academic writing should be taught not as expressions to memorize, but as variable structures with rich internal options. All the p-frames discussed in Chapter 3 had clear semantic patterns in their variable slots, such as the adjectives of possibility and importance in *it be * to* (e.g, *is it essential to, it is unnecessary to*) and the division of the knowing verbs into four semantic families of 'see, say, expect, confirm,' where each group has both high-frequency items (*hypothesize, believe, expect*) and less common synonyms (*postulate, propose*). Teaching these semantic families to students allow them to use set expressions creatively, using common devices but avoiding overused idioms.

Contributions to Curriculum Design

Having demonstrated the usefulness of p-frame analysis and expanded our understanding of published engineering writing, the purpose of the second step was to develop materials and an approach to teaching formulaic language via visualizations to advanced NNS writers. While there are other studies on teaching formulaic language, this study was the first to use network analysis to both understand the linguistic context of the phrases and to visualize that information in a way that would be attractive and accessible to advanced NNS students. Thus, the key takeaways from this phase were lessons on the usability of the network analysis and corpus tools; the creation of a materials design framework; and the potential educational benefits of this approach as discovered through the materials design process.

Usability of Network Analysis & Corpus Tools

This phase required cobbling together and transferring data across a variety of corpus and network programs that were not designed to interface with each other. There were no insurmountable problems; but there were a variety of benefits and disadvantages to the various programs which are important to highlight for anyone who is considering these programs for future design work.

On the positive side, AntConc (Anthony, 2018a) and AntGram (Anthony, 2018b) worked exceedingly well at handling the corpus tasks. AntGram was used only for the specific tasks of extracting the n-grams and p-frames, for which it was designed. AntConc was essential in selecting what information to highlight when designing the sheets and in providing rich contextual information and selecting the most useful corpus examples for the worksheets. While the interfaces are basic, they are also straightforward and effective.

In addition, Automap is an excellent example of a program that was designed for network analysis, but which was also highly effective at corpus pre-processing and extracting basic relationship information for analysis in other programs. It interfaced well, as it was possible to transfer data from AntConc to AutoMap, and from AutoMap to NodeXL.

The bulk of the difficulties, as well as major benefits, came from NodeXL. The difficulties came from its inability to efficiently and reliably handle large amounts of data, and from the fact that while it offered the user total control of every data point and visualization, it also did not have helpful ways to automate many of the tasks the visualizations required. The first issue is simply a lack on the part of the software, and one that future designers should be aware of. The second issue comes from using the software to create educational visualizations, for which the system was not designed. For educational visualizations, it is essential that the arrangements make clear sense and that every point is labeled and explained. However, the system was created to depict broad trends; thus, many of my issues came from prioritizing abilities that were not the main strengths of the program. However, given the total control over visuals that it provided, it was worth developing strategies to overcome the difficulties and continue to produce the visuals in NodeXL. While NodeXL is not the easiest program, it offers materials designers capabilities and control that to my knowledge are currently not available elsewhere.

Lessons from the Design Framework

While it is useful to evaluate the tools on their own, the major outcome from the second phase was a design framework for future construction of educational materials that marry corpus findings with the visualization benefits of network analysis. The many iterations of design possibilities and incorporating feedback from students led to a final design process suggestion that connects the disparate pieces as smoothly as possible, allowing the findings from one step to benefit the other steps. The design framework from Chapter 4 is reproduced below in Figure 19.

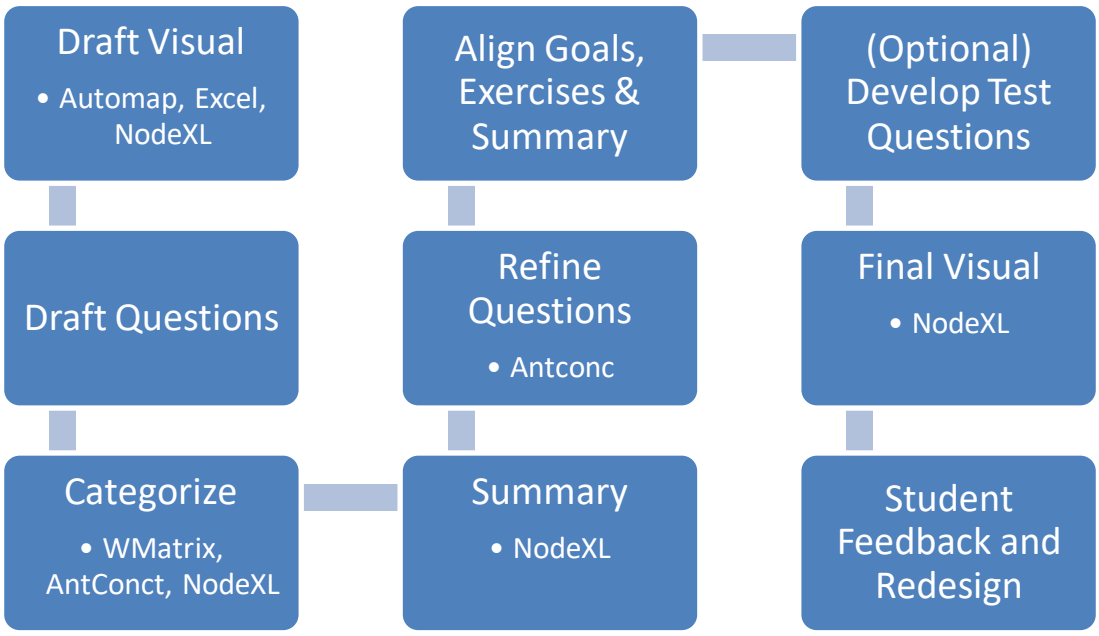


Figure 19: Template for Materials Design Framework

The first key element of a successful design framework is to allow the information from the corpus tools, semantic categorization, and NodeXL to interact with each other. Interesting patterns in NodeXL can become the starting point for semantic categorization or analysis that uses AntConc lines. The semantic data can identify word families in NodeXL, which will affect how the designer chooses to sort and color the data points in the visuals. The visuals have to be developed and redeveloped as the learning objectives and exercise questions are refined to ensure that all three are aligned.

The second essential component is a linguistically knowledgeable designer. Without expert input, the tools will simply spit out the most frequent phrases, items in the variable slots, and contextual words without categorizing or explaining them in a way that this helpful to NNS students. The materials design process requires a designer who is familiar with research on second language writing and academic writing, has extensive syntactic knowledge, can translate sometimes complex syntactic patterns into accessible terminology and learning goals for

students, and can identify rhetorical patterns and points of linguistic interest that will vary from phrase to phrase. When the design process is iterative and the designer knowledgeable, however, it is possible to access rich insights on everything from rhetorical to syntactic patterns and convey them efficiently to students.

Educational Benefits of Network Approach

While the framework for materials is complex, there were several educational advantages of this approach that were made clear throughout the design process. The complex, hands-on nature of the design process meant that the materials themselves provided rich, unique insights into language use that could be immediately applied by the students. And the network visualizations fit the task well, as they highlighted the puzzle piece nature of building academic sentences, allowed large amounts of information to be succinctly conveyed, and emphasized variation in patterns.

The first benefit is that the design process forces the designer to sort through the contextual data point by point, noticing and searching for patterns. If the software were more automated and identified the patterns with less guidance from the designer, important information would be lost. Given the necessity of pairing down the data before visualizing it, the researcher or teacher directs the learning, as students spend most of their time discussing and studying the points the researcher selected. Applied corpus studies often encourage student-directed learning (Gabrielatos, 2005), but this study found that for this particular audience, teacher-directed learning was desired and led to positive outcomes. In both reviewing the materials and using them in class, students tended to disengage when the questions were open-ended, and responded well when I explained specific findings. For example, they remained silent when I asked what rhetorical purposes they saw in a set of corpus examples, but they engaged when I named three common rhetorical functions and asked them to identify the rhetorical functions in a set of corpus examples. The detailed, phrase-specific findings that this process produces empowers students with the details they want as they practice new constructions.

The second benefit is how well network visualizations convey much of the important information from corpus findings. Previous attempts at teaching corpus linguistics have usually

begun with handing lists of phrases to students or having them search for phrases in a corpus. It can be daunting for NNS to start with these blocks of texts. In contrast, the visualizations allow them to access information about frequency, part of speech, and place in the sentence for a great number of lexemes with relative ease. This gives students the choice to concentrate on the data that is most helpful or interesting to them, while still having access to other data. It is also an ideal setup for visualizing the less-frequent synonyms in variable slots, which is data that students particularly appreciated. And representing each lexeme as a point on a graph, grouped near similar points in its most common sentence position, emphasized the puzzle piece nature of sentence building. The visuals indicated which sections could appear at other points in the sentence and which were fixed, and the exercises guided students in rearranging pieces and deciding which arrangements worked best for their purposes and why.

In conclusion, the process of designing materials provided many valuable insights into the tools themselves, the best practices for a design framework that combines corpus findings with network visualizations, and the potential educational benefits of this approach. These findings will be especially relevant to future educators teaching formulaic language at an advanced level.

Lessons from Classroom Application

The final phase of this project consisted of refining and testing the educational materials via two sections of a writing course for international graduate students and visiting scholars. In both cases, the students took a pre-test on the first day of class that measured their performance on the three learning objectives for each of the three phrases. Then I taught, using the materials developed in Chapter 4, for two 75-minute class periods. The students took a post-test at the end of the second period and gave their oral and written feedback on the materials; seven of the sixteen students also took a delayed post-test a week later. While the study sample was small, it provided several key insights into how to make the materials accessible to the target audience, what educational content was most useful to the students, and the overall value of this novel educational approach.

Lessons for Student-Friendly Design

The primary takeaway from the class for creating student-friendly materials design was that careful and explicit scaffolding greatly facilitated student engagement and willingness to practice writing with the phrases. This was in part because the English level of the students in the classes was lower than that of the international graduate students who had reviewed my learning materials. The classroom students benefited from more explicit instructions on how to read the visualizations and they also appreciated more intermediate exercises before writing their own sentences.

Two students in the first section had difficulty reading the visualizations on the worksheets. After I included several slides that explained the information in the visualization step by step at the beginning of the second section, no students reported problems understanding them, and all students gave the visualizations top scores for usefulness (5 out of 5 on a Likert scale). The students were also quicker to use the visualizations as a reference tool when working with the sentences if they felt like they understood them. Thus, the visualizations were a helpful tool, but like all new approaches, students were much more likely to embrace it when it was explained, and its value made clear.

The second type of scaffolding that was developed as a response to student engagement on the first day was the inclusion of intermediate activities that asked students to categorize or manipulate the data on the worksheets before using it to write sentences. Originally, the worksheets introduced a concept (such as strengthening adverbs, or a phrase's rhetorical function) and asked students to write a sentence where they employed that. However, as most the students were hesitant to begin writing, I developed more intermediate activities. Examples of these include the exercises where students sorted synonyms into semantic families, identified prepositional phrases in the sentence, found the agents behind passive verbs, decided if adverbs in corpus sentences strengthened or hedged the sentence's claim, and classified corpus sentences according to the rhetorical function of the variable phrase. Students fully participated in these activities when we did them as a class or as individuals, and they also asked more questions and engaged with the material more deeply than when just asked to write sentences. While this meant

less time writing in class, it also meant that when we did switch to writing, students were more willing to try writing sentences.

In addition to the benefits of scaffolding, another classroom finding was that the students preferred to work with full example sentences from the corpus, rather than from the page of corpus lines that was attached to each phrase worksheet. Again, this was partially because of the students' comfort with English. Gleaning useful information from a snippet and being able to generalize patterns from corpus lines requires both a high level of vocabulary and the ability to guess at what the remainder of the sentence would look like. The students in my pilot study were not comfortable doing this, as they wanted to understand every part of the full sentence before making guesses about the role of the phrase. As a result, I reworked the lesson plan so that we looked at more example of phrases within full sentences before I asked them to spot patterns in the corpus lines. I also simplified the corpus line tasks; asking them just to find hedging adverbs and other one-word items rather than identifying prepositional phrases or rhetorical moves within the snippets. The addition of scaffolding and more full-length example sentences were the two primary changes to materials design developed through the classroom teaching.

Most Valuable Content

Another benefit of the classroom sessions was that it became clear what content the students most valued and were eager to discuss and apply in their own writing. These concepts were the tension between learning accepted patterns and parroting, extending their vocabulary via synonyms, identifying and employing rhetorical moves, and identifying author stance in their reading and employing it in their writing.

The students were interested in the frequency information and appreciated knowing which phrases appeared most often and what the most common variable words were. However, both sections asked if they should be using or avoiding these phrases; would it make their prose stronger or weaker to use the most common idioms? They found the list of less-common synonyms particularly helpful, because it enabled them to use common constructions in novel ways. Thus, in both sections we discussed the synonyms in detail and discussed the nuances that the less familiar words could provide.

The students were also particularly interested in identifying rhetorical moves and author stance. As most of the students hoped to write academic publications in English but had not yet done so, they were interested in how the findings could improve both their reading and their writing. When we discussed the three most common rhetorical moves for each of the two variable phrases, they found it helpful to practice identifying these moves in existing sentences before attempting it in their own sentences. They also found the information and examples on strengthening and hedging useful, especially the nuances that certain adverbs provided. The tension between sounding objective and writing persuasive arguments is one that international students often struggle with (Chen & Baker, 2010; Wei & Lei, 2011; Conrad, 2018), so corpus findings that show how published authors navigate this tension are particularly useful.

Overall Value of Approach

The work in Chapter 5 was both a continuation of and an evaluation of the materials development in Chapter 4. While it is hard to definitively evaluate the materials as they were still being improved, the test results and student feedback from the two class sections were mostly positive. The study did not measure if the students used the phrases in their academic writing after the class, but the test results did indicate substantial gains in the students' ability to manipulate patterned languages, select diverse synonyms and antonyms, and recognize the rhetorical functions of the phrases. Students also improved the speed at which they were able to do these tasks, as they went from answering an average of 7.5 questions in 10 minutes on the pre-test to answering an average of 12.5 questions on the post-test. In addition to an improvement in the quantity of answers and quantity of correct answers, there was a jump in quality. Students were able to name more synonyms and less common synonyms when asked to rewrite a sentence, and they supplied a great diversity of adverbs and word order variations.

While the small nature and unique conditions of studies of teaching formulaic language make it hard to compare results across studies, these are positive indications that this approach has potential to assist students in overcoming common barriers to acquiring academic phrases. Both Cortes (2006) and Jones & Haywood (2004) found that acquainting students with lists of formulaic phrases did not improve the likelihood of using them in their own writing, and they recommended more contextualized teaching. This study chose to concentrate deeply on three

phrases rather than give students the whole list, and to employ visualization to help students understand the contexts and uses of the phrases they were learning. While we cannot say if the students in this study used the phrases more in the writing after the classes, they did clearly learn to understand and manipulate these phrases.

Limitations

Even though the corpus analysis, teaching materials and classroom implementation offer unique insights, there are several limitations in the methodology and pilot study that must be acknowledged.

In the methodology for the corpus study, the semantic categorization of the p-frames was developed specifically for this dissertation, and thus was less-tested than the methodologies used in n-gram analysis. While WMatrix is a powerful tool, it is impossible for it to reach complete accuracy when it identifies the semantic categories of words. This is true especially for technical engineering language, where the less-common meanings of words may be employed more often than the primary dictionary meanings. For example, "field" was classified under 'Farming and Agriculture,' but in this corpus, it primarily referred to branches of learning (as in "the medical field"), or to electric, magnetic, and gravitational fields. As I manually re-categorized words that were inaccurate, I used their corpus concordance lines to guide my decision-making. However, semantic categorization is still an approximate process, as minority uses of the words (such as the 2 instances were 'Field' was the last name of a quoted source) make it impossible to identify the exact semantic distributions. In the future, it would be best to check WMatrix's categories and my corrections against another semantic tagger like WordNet (Felbaum, 2005).

Second, because Hyland's functional categories were developed to classify n-grams, his three-way distinction may not perfectly apply to p-frames. When classifying the n-grams, it was easy to check this classification by identifying how Hyland had sorted identical or equivalent n-grams. However, for the p-frames, they were sorted according to the most common functional uses I identified in the corpus. While I followed Hyland's definitions carefully, another researcher might have classified some phrases differently. Thus, in the future, it would be useful to test

these categorization methods on p-frames from other corpora, and perhaps amend the categories to capture the full range of functional uses of p-frames.

There were also several limitations in the classroom implementation. Because this was a pilot study, there can be no strong claims about what the pre- and post-tests proved. The first major limitation was that the students represented a variety of English levels and academic backgrounds. For example, the international STEM graduate students who reviewed the classroom materials were slightly more comfortable with reading and writing academic texts than most of the students in the PLaCE class. To cover the gap, more of classroom time was spent on explaining terminology and on scaffolding exercises. However, if the students' levels had been clearer in advance, I might have structured the exercises differently and reduced the terminology. In addition, not all the students came from engineering backgrounds, and thus were less familiar with the engineering-specific language in the examples from the corpus. This may have presented a barrier to their acquisition of the formulaic language.

The second major limitation that affects the interpretation of the test scores is the difficulty in distinguishing between an increase in student writing skills and an increase in student meta-language about writing, i.e., their ability use writing terms like *hedge*, *strengthen*, *rearrange*, *clause*, or the parts of speech. Across the test scores, there is a notable increase in speed, as students went from answering only 7.5 questions on average on the pre-test to answering 12.5 questions on average on the post-test. It is possible that this gain is not only because they had developed the skills necessary to answer the questions; it could be that it was because they had been exposed to the meta-language in class and were able to more quickly understand what was being asked of them. While both the ability to perform a skill and the ability to name and discuss that skill are important gains, further research is necessary to untangle the two elements.

Future Steps

Given the results of this study, there are a variety of logical next steps that would extend this work to answer questions about the nature of published academic writing and to improve the instruction of academic writing.

The p-frame and n-gram analyses conducted in Chapter 3 supported previous findings on academic writing in general and on engineering writing in particular. However, the functional categorizations of the p-frames had a markedly distinct distribution from the functional distributions found in Hyland (2008b, 2012) for n-grams. In particular, the p-frames were used for participant-oriented functions such as establishing authority and persuading the reader at much higher percentages than the n-grams were. It remains to be seen if the predominance of participant-oriented p-frames is specific to engineering writing or if other disciplines have a similar functional distribution to their p-frames. As it has been claimed that the lack of participant-oriented phrases is evidence that engineering writing strives for a more objective tone (Hyland, 2012), it is important to see if engineering is more participant-oriented than other disciplines when it comes to p-frames. If it is, then students should not be taught that engineering writing is less participant-oriented overall, but rather how to perform participant-oriented acts through the subtler tools offered by p-frames.

It would also be interesting to do a deeper analysis of the types and kinds of functional categorization of the p-frames. The analysis in this paper did not go in depth; it relied primarily on WMatrix's automatic categorization of the variable slots with manual analysis where WMatrix was clearly wrong. The analysis also adopted all of Hyland's functional categories, despite the fact that it was designed for n-grams there were, nonetheless, some p-frames that did not fit perfectly in his categories. This was adequate for identifying major trends and choosing which phrases to teach students. However, in the future, it would be worthwhile to revisit the categorizations and see if Hyland's functional categorization could be refined to more accurately and sensitively capture the p-frames' rhetorical uses.

Other analyses of lexical bundles have categorized them syntactically as well as functionally (see Biber & Barbieri (2007) and Hyland (2008b) for examples). As this was beyond the scope of this project, there was little syntactic analysis beyond the two main categories of *noun + preposition* and *passive verb*, and that categorization was based primarily on the semantic characteristics of the variable words. Thus, the syntactic categorization of p-frames across academic disciplines is an open avenue of research for future investigators. It would be particularly interesting to see how it overlaps or diverges from the syntactic trends found in the n-grams.

From a pedagogical perspective, it would be helpful to create more educational materials to better refine the framework for materials development developed in Chapter 4 and to test those educational materials on more students. At the end of the first section, one student asked if I had worksheets on all the other phrases and was disappointed when I did not. He asked why, and I explained that the process of creating the existing worksheets was complex and had taken weeks. The framework was created through four processes to create learning materials for four phrases and following it in the future would decrease the production time. But while the framework shows the best way to create materials for those four phrases, work on other phrases might require a slightly different process. Applying it to phrases from another discipline would also be an interesting exercise, as it is an open question if the formulaic language in other disciplines is substantially different and if the same types of visualization, exercises and goals would work for them.

Finally, this project did not isolate the contributions of the network visualization elements to student learning. Given how tightly interconnected the exercises and visuals were, as well as the small number of students, it was impossible conduct a controlled study where students received the same pre- and post-tests but only learned from corpus findings without the aid of the visualizations. If there were a separate set of educational materials that included the same information but did not present it visually, it would be possible to compare across sections to see if including the network visualizations had a significant impact on student acquisition of the phrases. As the visualization creation is a complex task, it would be valuable to know if it its impact makes it a worthwhile investment.

REFERENCES

Abrams, M., Harpham, G. (2011). *A Glossary of Literary Terms*. Boston, MA: Cengage Learning.

About PLaCE. (n.d.) Retrieved from https://www.purdue.edu/place/about/index.html

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes 31*, 81-92.

Alali, F., & Schmitt, N. (2012). Teaching Formulaic Sequences: The Same as or Different from Teaching Single Words? *TESOL Journal, 3*(2), 153-180.

AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes, 17*, 51-62.

Anthony, L. (2018a). AntConc (Version 3.5.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Anthony, L. (2018b). AntGram (Version 1.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Anthony, L., Wulff, S., & Boettger, R. (2016). Workshop: Integrating data-driven learning into the technical writing classroom. *2016 IEEE International Professional Communication Conference (IPCC), 2016*, 1-2.

Baggett, P. (1989). Understanding visual and verbal messages. In H. Mandl & J.R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 101-124). Amsterdam: North-Holland.

Baker, P. 2016. The shapes of collocation. *International Journal of Corpus Linguistics, 21*(2), 139-164.

Baker, P., & Egbert, J. (2016). *Triangulating methodological approaches in corpus-linguistic research* (Routledge advances in corpus linguistics; 17). New York: Routledge.

Baker, P., & McEnery, T. (2015). Who benefits when discourse gets democratised? Analysing a Twitter corpus around the British Benefits Street debate. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 244-265). Basingstoke, UK: Palgrave.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis*. Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July 2000 (pp. 165-182). Amsterdam: Rodopi.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversations and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

Biber, D., Conrad, S., Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A Wilson, P. Rayson, T. McEnery (Eds.), *Corpus linguistics by the lune*, (pp. 71-93). Pieterlen: Peter Lang

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., & Gray, B. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), p.2-20.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics, 21*(4), 439-464.

Bloch, J. (2010). A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers. *Journal of Writing Research*, *2*(2), 219-244.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173.

Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Chang, P., & Schleppegrell, M. (2016). Explicit Learning of Authorial Stance-taking by L2 Doctoral Students. *Journal of Writing Research, 8*(1), 49-80.

Campbell, M., & Kennell, V. Working with graduate student writers: Faculty guide. West Lafayette, IN: Purdue Writing Lab.

Carley, K., Columbus, D., & Landwehr, P. (2013). AutoMap User's Guide 2013. *Technical Report, CMU-ISRI-13-105*. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Institute for Software Research International.

Carter, T. Working with multilingual student writers: Faculty guide. West Lafayette, IN: Purdue Writing Lab

Cava, A. (2011). Abstracting science: A corpus-based approach to research article abstracts. *International Journal of Language Studies, 5*(3), 75-98.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Conrad, S. (2004). Corpus variety: Corpus linguistics, language variation, and language teaching. In J.M. Sinclair (Ed.), *How to use corpora in language teaching (Studies in Corpus Linguistics, 12)*. Amsterdam, Netherlands: John Benjamin.

Conrad, S. (2018). The Use of Passives and Impersonal Style in Civil Engineering Writing. *Journal of Business and Technical Communication, 32*(1), 38-76.

Conrad, S., & Biber, D. (2004). The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica: International Annual for Lexicography, 20*, 56-71.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, *17*, 391-406.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ, 8*(4), A-1. [42]

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*, 213-238.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 149-162). Amsterdam, Netherlands: John Benjamin.

Davies, M., & Gardner, D. (2010). A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists. London, New York: Routledge.

Diesner, J., & Carley, K. M. (2004**).** AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts**.** *Technical Report, CMU-ISRI-04-100*. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Institute for Software Research International.

Doerfel, M. (1998). What constitutes semantic network analysis? A comparison of research and methodologies. *Connections, 21*(2), 16-26.

Dondis, D. (1973). *A primer of visual literacy*. Cambridge, Mass.: MIT Press.

Egbert, J., & Biber, D. (2018, September). *Incorporating text dispersion into keyword analyses.* Paper presented at the meeting of the Association of American Corpus Linguistics, Atlanta GA.

Eppler, M. J., & Burkhard, R. A. (2005). Knowledge visualization. In J. Edward (Ed.), *Encyclopedia of Knowledge Management,* (pp. 551-560). IGI Global.

Eriksson, A. (2012). Pedagogical perspectives on bundles: Teaching bundles to doctoral students of biochemistry. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora,* (pp. 195-211). Brno: Masaryk University Press.

Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

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

Fischer-Starcke, B. (2012). Corpus Analysis of Business English. In C. Chappelle (Ed.), *The encyclopedia of applied linguistics*. New Jersey: Wiley-Blackwell.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T.A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11-33). Amsterdam: J. Benjamins.

Flowerdew, L. (2012). Needs Analysis and Curriculum Development in ESP. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 325-346). New Jersey: Wiley-Blackwell.

Fuster-Marquez, M., & Pennock-Speck, B. (2015). Target frames in British hotel websites. *International Journal of English Studies, 15*(1), 51-69.

Geva, E. (1983). Facilitating Reading Comprehension through Flowcharting. *Reading Research Quarterly, 18*(4), 384-405.

Gick, M. (1989). Two functions of diagrams in problem solving by analogy. In H. Mandl & J.R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 215-231). Amsterdam: North-Holland.

Gray, B., Biber, D., & Geluso, J. (2015, July). Methods of characterizing discontinuous lexical frames: Quantitative measurements of predictability and variability. Paper presented at Corpus Linguistics, Lancaster University, UK.

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*, (pp. 212-226). New York, NY: Routledge.

Gries, S. (2013). 50-something years of work on collocations: What is or should be next… *International Journal of Corpus Linguistics, 18*(1), 137-166.

Hall, R., Dansereau, D., & Skaggs, L. (1992). Knowledge Maps and the Presentation of Related Information Domains. *The Journal of Experimental Education, 61*(1), 5-18.

Harley, H. (2006). *English words: A linguistic introduction*. New York, NY: John Wiley & Sons

Hinkel, E. (1997). Indirectness in L1 and L2 academic writing. *Journal of Pragmatics, 27*(3), 361-386.

Hinkel E. (2003). Simplicity Without Elegance: Features of Sentences in L1 and L2 Academic Texts. *TESOL Quarterly, 37*(2), 275-301.

Hinkel, E. (2005). Handbook of Research in Second Language Teaching and Learning. New York, NY: Routledge.

Ho, D. (2019). Notepad++[software]. Available at: https://notepad.software

Hu, G., & Cao, F. (2011). Hedging and boosting in abstracts of applied linguistics articles: A comparative study of English- and Chinese-medium journals. *Journal of Pragmatics, 43*(11), 2795-2809.

Hunston, S. (2010). How can a corpus be used to explore patterns? In M. McCarthy & A. O'Keeffe (Eds.*), The Routledge handbook of corpus linguistics* (152-166). London: Routledge.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies, 7*, 173-192.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-63.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Hyland, K., & Tse, P. (2005). Hooking the reader: A corpus study of evaluative *that* in abstracts. *English for Specific Purposes, 24*, 123–139.

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society, 32*(2), 257-279.

Jaworska, S. & Themistocleous, C. (2018). Public discourses on multilingualism in the UK: triangulating a corpus study with a sociolinguistic attitude survey. *Language in Society, 47*(1), 57-88.

Jones, M., & Haywood, S. (2014). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic Sequences,* (pp. 269-300). Philadelphia, PA: John Benjamins.

Juilland, A., & Chang-Rodriguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton de Gruyter.

Karras, J. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL, 28*(2), 166-186.

Kessler, J. (2017). Scattertext: A Browser-Based Tool for Visualizing how Corpora Differ. *ArXiv.org,* ArXiv.org, Apr 20, 2017.

Kettemann, B., & Marko, G. (2002). *Teaching and learning by doing corpus analysis:* Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July 2000. Amsterdam: Rodopi.

Koester, A. 2010. Building small specialised corpora. In M. McCarthy & A. O'Keeffe (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). London: Routledge.

Maswana, Kanamaru, & Tajino. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand, 2*(C), 1-11.

Lee, D. 2010. What corpora are available? In M. McCarthy and A. O'Keeffe (Eds.), *The Routledge handbook of corpus linguistics* (107-121). London: Routledge.

Lee, D., & Chen, S. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*(4), 281-296.

Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*, 56-75.

Levie, W., & Lentz, H. (1982). Effects of text illustrations: A review of research. *ECTJ, 30*(4), 195–232.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*(2), 85-102.

Liu, D. (2011). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes 31*(1), 25-35.

Llurda, E. (2005). Non-native language teachers: Perceptions, challenges, and contributions to the profession (Educational linguistics vol. 5). New York: Springer.

Mandl, H., & Levin, J.R. (1989). Knowledge acquisition from text and pictures. Amsterdam: North-Holland.

May 2018 First Destination Survey Outcome Report. (n.d.) Retrieved from https://www.cco.purdue.edu/Alumni/AlumniStartingSalaries

Mayer, R.E., Bove. W., Bryman, A., Mars, R. and Tapangco, L. (1996). When less is more: meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology 88*, 64–73.

Mayer, R.E. and Gallini, J.K. 1990: When is an illustration worth ten thousand words? *Journal of Educational Psychology 82*, 715–26.

Mento A., Martinelli P., & Jones R. 1999. Mind mapping in executive education: applications and outcomes. *The Journal of Management Development, 18*(4), 390–416.

Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics, 20(*1), 30-53.

Mondal, H., Mondal, S., & Das, D. (2016). Learning style preference for basic medical science: A key to instructional design. *International Journal of Clinical and Experimental Physiology, 3*(3), 122-126.

Nwogu, K.N. 1997. The medical research paper: Structure and functions. *English for Specific Purposes*, *16*(2), 119-38.

Oakey, D. (2002). Formulaic language in English academic writing. In R. Reppen, S. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 111-129). Amsterdam: John Benjamins.

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics* (1st ed., Routledge Handbooks in Applied Linguistics). London: Routledge.

Öztürk, Y., & Köse, G. (2016). Turkish and native English academic writers' use of lexical bundles. *Journal of Language and Linguistic Studies, 12*(1), 149-166.

Pan, F., Reppen, R., & Biber, B. (2016). Comparing patterns of L1 versus L2 English academic professionals: lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes 21*, 60-71.

Paranyushkin, D. 2011. Visualization of text's polysingularity using network analysis. *Prototype Letters 2*(3), p. 256-278.

Parkinson, J., & Musgrave, J. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, 14, p. 48-59.

Parodi, G. (2009). Written genres in university studies: Evidence from an academic corpus of Spanish in four disciplines. In C. Bazerman, A. Bonini, and D. Figueiredo (Eds.), *Genre in a changing world* (pp. 483-501). Fort Collins: Parlor Press.

Parodi, G. (2010). Academic and professional discourse genres in Spanish. Amsterdam: John Benjamins.

Peacock, M. (2015). Stance adverbials in research writing. *Ibérica, 29*(Apr), 35-48.

Peeck, J. (1989). Trends in the delayed use of information from an illustrated text. In H. Mandl & J.R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 23-277). Amsterdam: North-Holland.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes, 14*, 84-94.

Purdue Office of International Students and Scholars. (n.d.). *International Students and Scholars Enrollment & Statistical Report, Fall 2017*. Retrieved from https://www.iss.purdue.edu/Resources/Docs/Reports/ISS_StatisticalReportFall17.pdf

Rayson, P. (2009) Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Rayson, P., & Mariani, J. (2009). Visualising corpus linguistics. CL2009 Proceedings of the Corpus Linguistics Conference, Liverpool, UK.

Richards, J. (2001). *Curriculum development in language teaching* (Cambridge language education). Cambridge, UK ; New York: Cambridge University Press.

Rozycki, W., & Johnson, N. (2013). Non-canonical grammar in Best Paper award winners in engineering. *English for Specific Purposes, 32*(3), 157-169.

Salomon, G. 1989: Learning from texts and pictures: reflections on a metalevel. In Mandl, H. and Levin, J.R., editors, *Knowledge acquisition from text and pictures.* Amsterdam: North-Holland, 73–82.

Samraj, B. (2002). Disciplinary variation in abstracts: The case of Wildlife Behavior and Conservation Biology. In J. Flowerdew (Ed.), *Academic Discourse* (Applied linguistics and language study). New York: Longman.

Scott, M. (2018). WordSmith Tools version 7, Stroud: Lexical Analysis Software.

Shahriari, H. (2017). Comparing lexical bundles across the introduction, method and results sections of the research article. *Corpora 12*(1), 1-22.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writings: The ESL research and its implications. *TESOL Quarterly 27*(4), 657-677.

Simpson, R., & Mendis, D. (2003). A Corpus-Based Study of Idioms in Academic Speech. *TESOL Quarterly, 37*(3), 419-41.

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

Singh, J., Zerr, S., & Siersdorfer, S. (2017). Structure-Aware Visualization of Text Corpora. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval,* 107-116.

Skelton, J. 1994. Analysis of the structure of original research papers: An aid to writing original papers for publication. *British Journal of General Practice, 44*, 455-59.

Smith, M., Ceni A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., Dunne, C., (2010). NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016, from the Social Media Research Foundation.

Soruc, A., & Tekin, B. (2017). Vocabulary learning through data-driven learning in an English as a second language setting. *Educational Sciences-Theory & Practice, 17*(6), 1811-1832.

Sripicharn, P. 2010. How can we prepare learners for using language corpora? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics,* (pp. 371-384). New York, NY: Routledge.

Staples, S., Egbert, J., Biber, D., & McClair, A. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*, 214-225.

Swaab, R.I., Postmes, T., Neijens, P., Kiers, M.H., & Dumai, A.C.M. (2002). Multiparty negotiation support: The role of visualization's influence on the development of shared mental models. *Journal of Management Information Systems, 19*(1), 129-150.

Swales, J.M. 1990. Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

Swales, J., & Feak, C. (2004). *Academic writing for graduate students: Essential tasks and skills* Ann Arbor: University of Michigan Press.

Tanenbaum, W., Brand, J. (2008). Using AutoMap for Social and Textual Network Analysis. Army Research Lab Aberdeen Proving Ground MD.

Tyler, R. W., Gagné, R. M., & Scriven, M. (1967). *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Chicago, IL: Rand McNally

Valente, T., Coronges, K., Lakon, C., & Costenbader, E. (2008). How Correlated Are Network Centrality Measures? *Connections, 28*:1, 16-26.

Warchal, K. (2010). Moulding Interpersonal Relations through Conditional Clauses: Consensus-Building Strategies in Written Academic Discourse. *Journal of English for Academic Purposes, 9*(2), 140-150.

Wei N.X. (2007) Phraseological characteristics of Chinese learners spoken English: evidence of lexical chunks from COLSEC. Modern Foreign Languages 30: 280-91

Wei, Y., & Lei, L. (2011). Lexical Bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal: A Journal of Language Teaching and Research, 42*(2), 155-166.

Willis, D. (2003). Rules, patterns and words: Grammar and lexis in English language teaching. Cambridge University Press, Cambridge.

Winn, W. (1989). The design and use of instructional graphics. In H. Mandl & J.R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 125-144). Amsterdam: North-Holland.

Wood, D., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1-13.

Wray, A. **(**2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics, 32*, 231-254.

Yli-Jokipii, H. E., & Jorgensen, P. (2004). Academic journalese for the Internet: A study of native English-speaking editors' changes to texts written by Danish and Finnish professionals. *Journal of English for Academic Purposes, 3*(4), 341-359.

APPENDIX A. CORPUS INFORMATION

Journal Name	Branch	Issues Used	Type/Token
Nature Nanotechnology	Electrical Engineering	13:9-12, 14:1	6,663 / 104,125
Automatica	Electrical Engineering	96-100	6,216 / 121,969
IEEE Transactions on Pattern Analysis and Machine Intelligence	EE: Image Processing & Computer Vision	40:10-12, 41:1-2	6,076 / 143,060
International Journal of Computer Vision	EE: Image Processing & Computer Vision	126: 8-12	7,256 / 166,889
Total EE:			**16,448 / 536,043**
Nature Materials	Mechanical Engineering	17:9-13, 18:1	5,312 / 85,255
Applied Energy	Mechanical Engineering	230-234	5,664 / 120,373
Lab on Chip	ME: Biomedical Electromechanical	2018:22-24, 2019:1,2	5,494 / 98,307
Journal of Microelectromechanical Systems	ME: Biomedical Electromechanical	17: 2-6	4,532 / 85,026
Total ME:			**12,134 / 388,961**
Journal of Operations Management	Industrial Engineering	60-64	7,931 / 202,650
International Journal of Machine Tools and Manufacture	Industrial Engineering	133-137	5,074 / 120,713
Journal of Biomedical Informatics	IE: Biomedical Informatics	84-88	5,882 / 111,938
Applied Clinical Informatics Journal	IE: Biomedical Informatics	9:1-4, 10:1	4,725 / 84,385
Total IE:			**14,629 / 519,686**
International Journal of Impact Engineering	Aerospace Engineering	122-126	4,751 / 112,437
International Journal of Robust and Nonlinear Control	Aerospace Engineering	28:16-18, 29:1-2	5,890/107,596
Journal of Guidance, Control and Dynamic	AE: Aerodynamics	41:9-12, 42:1	6,030 / 144,978
Journal of Spacecraft and Rockets	AE: Aerodynamics	55:2-6	5,164 / 115,729
Total AE:			**13,330 / 480,740**
Total Corpus			**35,670 /1,925,430**

APPENDIX B. ARTICLES IN CORPUS

The articles in the corpus are listed by journal, with each articles' title and authors. For the sake of space, when there were more than three authors, only the first three authors are listed.

Article Name	First Three Authors
Nature Nanotechnology, Vols 13:9-12, 14:1	
Colloidal nanoelectronic state machines based on 2D materials for aerosolizable electronics	Volodymyr B. Koman, Pingwei Liu, Daichi Kozawa
Tailoring sample-wide pseudo-magnetic fields on a graphene–black phosphorus heterostructure	Yanpeng Liu, J. N. B. Rodrigues, Yong Zheng Luo
Plasmonic meta-electrodes allow intracellular recordings at network level on high-density CMOS-multi-electrode arrays	Michele Dipalo, Giovanni Melle, Laura Lovato
Enhanced water splitting under modal strong coupling conditions	Xu Shi, Kosei Ueno, Tomoya Oshikiri
Pre-adsorption of antibodies enables targeting of nanocarriers despite a biomolecular corona	Manuel Tonigold, Johanna Simon, Diego Estupiñán
Direct observation of noble metal nanoparticles transforming to thermally stable single atoms	Shengjie Wei, Ang Li, Jin-Cheng Liu, Zhi Li
Dissipative adaptation in driven self-assembly leading to self-dividing fibrils	Esra te Brinke, Joost Groen, Andreas Herrmann
Physical activation of innate immunity by spiky particles	Ji Wang, Hui-Jiuan Chen, Tian Hang
Gold nanoparticle biodissolution by a freshwater macrophyte and its associated microbiome	Astrid Avellan, Marie Simonin, Eric McGivney
Nucleic acid hybridization on an electrically reconfigurable network of gold-coated magnetic nanoparticles enables microRNA detection in blood	Roya Tavallaie, Joshua McCarroll, Marion Le Grand
Highly conductive, stretchable and biocompatible Ag–Au core–sheath nanowire composite for wearable and implantable bioelectronics	Suji Choi, Sang Ihn Han, Dongjun Jung
Directional lasing in resonant semiconductor nanoantenna arrays	Son Tung Ha, Yuan Hsing Fu, Naresh Kumar Emani
Nano-imaging of intersubband transitions in van der Waals quantum wells	Peter Schmidt, Fabien Vialla, Simone Latini
Electrical half-wave rectification at ferroelectric domain walls	Jakob Schaab, Sandra H. Skjærvø, Stephan Krohns
Neutrophil membrane-coated nanoparticles inhibit synovial inflammation and alleviate joint damage in inflammatory arthritis	Qiangzhe Zhang, Diana Dehaini, Yue Zhang
Fast current-driven domain walls and small skyrmions in a compensated ferrimagnet	Lucas Caretta, Maxwell Mann, Felix Büttner
High-efficiency colloidal quantum dot infrared light-emitting diodes via engineering at the supra-nanocrystalline level	Santanu Pradhan, Francesco Di Stasio, Yu Bi
Automatica, Vols. 96-100	
A distributed economic MPC framework for cooperative control under conflicting objectives	Philipp N. Köhler, Matthias A. Müller Frank Allgöwer
Maximum likelihood identification of stable linear dynamical systems	Jack Umenberger, Johan Wågberg Ian, R. Manchester
Stability analysis of networked linear control systems with direct-feedthrough terms	Stefan Heijmans, Romain Postoyan, Dragan Nešić
General linear forward and backward Stochastic difference equations with applications	Juanjuan Xu, Huanshui Zhang, Lihua Xie
Optimal scheduling of multiple sensors over shared channels with packet transmission constrain	Shuang Wu, Xiaoqiang Ren, Subhrakanti Dey
On input design for regularized LTI system identification: Power-constrained input	Biquiang Mu, Tianshi Chen
Symmetry reduction for dynamic programming	John Maidens, Axel Barrau, Silvère Bonnabel

Exponential convergence under distributed averaging integral frequency control	Erik Weitenberg, Claudio De Persis, Nim Monshizadeh
Stability of Kalman filtering with a random measurement equation: Application to sensor scheduling with intermittent observations	Damián Edgard Marellia, Tianju Suic, Eduardo Rath
Detectability and observer design for switched differential–algebraic equations	Aneel Tanwan, Stephan Trenn
Ensuring privacy with constrained additive noise by minimizing Fisher information	Farhed Farokhi, Henrik Sandberg
Analysis and synthesis for a class of stochastic switching systems against delayed mode switching: A framework of integrating mode weights	Lixian Zhang, Zepeng Ning, Yang Shi
Subspace identification with moment matching	Masaki Inoue
Free finite horizon LQR: A bilevel perspective and its application to model predictive control	Alberto De Marchi, Matthias Gerdts
Finite time stability of sets for hybrid dynamical systems	Yuchun Li, Ricardo G. Sanfelice
A new class of pursuer strategies for the discrete-time lion and man problem	Marco Casini, Andrea Garulli
Complexity and convergence certification of a block principal pivoting method for box-constrained quadratic programs	Gionata Ciminia, Alberto Bemporad
IEEE Transactions on Pattern Analysis and Machine Intelligence, Vols. 40:10-12, 41:1-2	
Bayesian Helmholtz Stereopsis with Integrability Prior	Nadejda Roubtsova, Jean-Yves Guillemaut
Learning from Narrated Instruction Videos	Jean-Baptiste Alayrac , Piotr Bojanowski, Nishant Agrawal
Colour Constancy Beyond the Classical Receptive Field	Arash Akbarinia, C. Alejandro Parraga
Bilinear Factor Matrix Norm Minimization for Robust PCA: Algorithms and Applications	Fanhua Shang , James Cheng, Yuanyuan Liu
Robust Light Field Depth Estimation Using Occlusion-Noise Aware Data Costs	Williem, In Kyu Park , Kyoung Mu Lee Williem, In Kyu Park , Kyoung Mu Lee
Partition Level Constrained Clustering	Hongfu Liu , Zhiqiang Tao, Yun Fu
Error-Correcting Factorization	Miguel Ángel Bautista Martin, Oriol Pujol, Fernando De la Torre
Dynamic Video Deblurring Using a Locally Adaptive Blur Model	Tae Hyun Kim, Seungjun Nah, Kyoung Mu Lee
Characterization of Color Images with Multiscale Monogenic Maxima	Raphaël Soulard, Philippe Carré
One-Pass Learning with Incremental and Decremental Features	Chenping Hou, Zhi-Hua Zhou
Recovering Joint and Individual Components in Facial Data	Christos Sagonas, Evangelos Ververas, Yannis Panagakis
Visual and Semantic Knowledge Transfer for Large Scale Semi-Supervised Object Detection	Yuxing Tang, Josiah Wang, Xiaofang Wang
Learning Kinematic Structure Correspondences Using Multi-Order Similarities	Shuhui Jiang, Ming Shao, Chengcheng Jia
Learning Consensus Representation for Weak Style Classification	Hyung Jin Chang, Tobias Fischer, Maxime Petit
Kernel Clustering: Density Biases and Solutions	Dmitrii Marin, Meng Tang, Ismail Ben Ayed
Face Alignment in Full Pose Range: A 3D Total Solution	Xiangyu Zhu, Xiaoming Liu, Zhen Lei
International Journal of Computer Vision, Vols 126:8-12	Boris Cigale, Damjan Zazula
Multi-label Learning with Missing Labels Using Mixed Dependency Graphs	Baoyuan Wu, Fan Jia. Wei Liu
Robust Detection and Affine Rectification of Planar Homogeneous Texture for Scene Understanding	Shahzor Ahmad, Loong- Fah Cheong
Adaptive Correlation Filters with Long-Term and Short-Term Memory for Object Tracking	Chao Ma, Jia- Bin Huang, Xiaokang Yang
Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes	Hassan Abu, Alhaija Siva, Karthik Mustikovela
What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation?	Nikolaus Mayer, Eddy Ilg, Philipp Fischer
Configurable 3D Scene Synthesis and 2D Image Rendering with Per-pixel Ground Truth Using Stochastic Grammars	Chenfanfu Jiang, Siyuan Qi, Yixin Zhu
Sim4CV: A Photo-Realistic Simulator for Computer Vision Applications	Matthias Müller, Vincent Casser, Jean Lahoud

Subspace Learning by Induced Sparsity	Yingzhen Yang, Jiashi Feng, Nebojsa Jojic Jianchao Yang
RED-Net: A Recurrent Encoder–Decoder Network for Video-Based Face Alignment	Xi Peng, Rogerio S. Feris, Xiaoyu Wang
Elastic Alignment of Triangular Surface Meshes	Zsolt Sánta, Zoltan Kato
Artistic Style Transfer for Videos and Spherical Images	Manuel Ruder, Alexey Dosovitskiy, Thomas Brox
Describing Upper-Body Motions Based on Labanotation for Learning-from-Observation Robots	Katsushi Ikeuchi, Zhaoyuan Ma, Zengqiang Yan
EMVS: Event-Based Multi-View Stereo—3D Reconstruction with an Event Camera in Real-Time	Henri Rebec, Guillermo Gallego, Elias Mueggler
Combining Shape from Shading and Stereo: A Joint Variational Method for Estimating Depth, Illumination and Albedo	Daniel Maurer, Yong Chul Ju, Michael Breuß
Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs	Oscar Koller, Sepehr Zargaran, Hermann Ney
Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis	Bernhard Egger, Sandro Schönborn, Andreas Schneider
Depth-Based Hand Pose Estimation: Methods, Data, and Challenges	James Steven, Supančič Grégory Rogez, Yi Yang
Nature Materials, Vols. 126: 8-12	
Therapeutic luminal coating of the intestine	Yuhan Lee, Tara E. Deelman, Keyue Chen
Imaging orbital-selective quasiparticles in the Hund's metal state of FeSe	A. Kostin, P. O. Sprau, A. Kreisel
Decoupling the role of stress and corrosion in the intergranular cracking of noble-metal alloys	N. Badwe, X. Chen, D. K. Schreiber
Robust microscale superlubricity in graphite/hexagonal boron nitride layered heterojunctions	Yiming Song, Davide Mandelli, Oded Hod
Fluid-enhanced surface diffusion controls intraparticle phase transformations	Yiyang Li, Hungru Chen, Kipil Lim
Ultrafast water harvesting and transport in hierarchical microchannels	Huawei Chen, Tong Ran, Yang Gan
On-chip valley topological materials for elastic wave manipulation	Mou Yan, Jiuyang Lu, Feng Li
Polarity governs atomic interaction through two-dimensional materials	Wei Kong, Huashan Li, Kuan Qin
Chemical nature of ferroelastic twin domains in CH3NH3PbI3 perovskite	Yongtao Liu, Liam Collins, Roger Proksch
High-mobility band-like charge transport in a semiconducting two-dimensional metal–organic framework	Renhao Dong, Peng Han, Himani Arora
Golden single-atomic-site platinum electrocatalysts	Paul N. Duchesne, Z. Y. Li, Christopher P. Deming
A rhombohedral ferroelectric phase in epitaxially strained Hf0.5Zr0.5O2 thin films	A. Fang, K. Kroenlein, D. Riccardi
Long spin coherence length and bulk-like spin–orbit torque in ferrimagnetic multilayers	Jiawei Yu, Do Bang, Rahul Mishra
Magneto-ionic control of magnetism using a solid-state proton pump	Aik Jun Tan, Mantao Huang, Can Onur Avci
Highly mechanosensitive ion channels from graphene-embedded crown ethers	A. Fang, K. Kroenlien, D. Riccardi
Real-time insight into the doping mechanism of redox-active organic radical polymers	Shaoyang Wang, Fei Li, Alexandra D. Easley
An efficient nanosieve	Anastasiya Bavykina & Jorge Gascon
Applied Energy, Vols. 230-234	
Sensitivity analysis for optimization of renewable-energy-based air-circulation-type temperature-control system	Haksung Leea, Akihito Ozakib
Numerical analysis of experimental studies of methane hydrate dissociation induced by depressurization in a sandy porous medium	Zhenyuan Yina, George Moridiscda, Zheng Ron
Impact of bio-alcohol fuels combustion on particulate matter morphology from efficient gasoline direct injection engines	C. Hergueta, A. Tsolakis, J. M. Herrerosa
Optimal decarbonization pathways for urban residential building energy services	Benjamin D. Leibowicz, Christopher M. Lanham, Max T. Brozynski
Techno-economic assessment of a renewable bio-jet-fuel production using power-to-gas	Konstantin M. Zecha, Sebastian Dietricha, Matthias Reichmuth

Analysis and forecasting of the carbon price using multi—resolution singular value decomposition and extreme learning machine optimized by adaptive whale optimization algorithm	Wei Sun, Chongchong Zhang
Optimal placement of distributed energy storage systems in distribution networks using artificial bee colony algorithm	Choton K. Das, Octavian Bass, Ganesh Kothapalli
Network-constrained unit commitment under significant wind penetration: A multistage robust approach with non-fixed recourse	Noemi G. Cobosa, José M. Arroyo, Natalia Alguacil
Optimal active and reactive power allocation in distribution networks using a novel heuristic approach	A. Bayat, A. Bagheri
Hybrid thermomagnetic oscillator for cooling and direct waste heat conversion to electricity	K. Deepak, V. B. Varma, G. Prasanna
A room-temperature activated graphite felt as the cost-effective, highly active and stable electrode for vanadium redox flow batteries	H. R. Jiang, W. Shyy, Y. X. Ren
Oil price volatility and economic growth: Evidence from advanced economies using more than a century's data	Reneévan Eyden, Mamothoana Difeto, Rangan Gupta
Trading power instead of energy in day-ahead electricity markets	Rens Philipsena, Germán Morales-España, Mathijs Weerdt
Intelligent simultaneous fault diagnosis for solid oxide fuel cell system based on deep learning	Zehan Zhang, Shuanghong Li, Yawen Xiao
Evaluation and transient control of an advanced multi-cylinder engine based on partially premixed combustion	Lianhao Yin, Gabriel Turesson, Per Tunestål
Parked electric car's cabin heat management using photovoltaic powered ventilation system	M. Kolhe, S. K. Adhikari, T. Muneer
Possible design with equity and responsibility in China's renewable portfolio standards	Bing Wang, Yi- Ming Wei, Xiao-Chen Yuan
Lab on a Chip, Vols. 2018:22-24, 2019:1,2	
Microfluidic diamagnetic water-in-water droplets: a biocompatible cell encapsulation and manipulation platform	Maryam Navi, Niki Abbasi, Morteza Jeyhani
Measurement and mitigation of free convection in microfluidic gradient generators	Yang Gu, Varun Hegdem Kyle J. M. Bishop
Development of a biomimetic liver tumor-on-a-chip model based on decellularized liver matrix for toxicity testing	Siming Lu, Fabio Cuzzucoli, Jing Jiang
Label-free isolation of rare tumor cells from untreated whole blood by interfacial viscoelastic microfluidics	Fei Tian, Lili Cai, Jianqiao Chang
Microfluidic-based solid phase extraction of cell free DNA	Camila D. M. Campos, Sachindra S. T. Gamage
Extraction of electrokinetically separated analytes with on-demand encapsulation	Xander F. van Kooten, Moran Bercovici, Govind V. Kaigala
Integration of sample preparation and analysis into an optofluidic chip for multi-target disease detection	Gopikrishnan G. Meena, Aadhar Jain, Joshua W. Parks
Manipulation of a floating liquid marble using dielectrophoresis	Chin Hong Ooi, Jing Jin, Kamalalayam Rajan Sreejith
A radial microfluidic platform for higher throughput chemotaxis studies with individual gradient control	Jiandong Wu, Aditya Kumar-Kanojia, Sabine Hombach-Klonisch
Disposable silicon-glass microfluidic devices: precise, robust and cheap	ZhenBang Qi, Lining Xu, Yi Xu
Urine-based liquid biopsy: non-invasive and sensitive AR-V7 detection in urinary EVs from patients with prostate cancer	Hyun-Kyung Woo, Juhee Park, Ja Yoon Ku
Sample pre-concentration on a digital microfluidic platform for rapid AMR detection in urine	Sumit Kalsi, Martha Valiadi, Carrie Turner
Digital nanoliter to milliliter flow rate sensor with in vivo demonstration for continuous sweat rate measurement	Jessica Francis, Isaac Stamper, Jason Heikenfeld
Mechanical-activated digital microfluidics with gradient surface wettability	Lin Qi, Ye Niu, Cody Ruck
Acoustophoretic focusing effects on particle synthesis and clogging in microreactors	Zhengya Dong, David Fernandez Rivas, Simon Kuhn
Dynamic control of capillary flow in porous media by electroosmotic pumping	Tally Rosenfeld, Moran Bercovici
A scalable filtration method for high throughput screening based on cell deformability	Navjot Kaur Gill, Chau Ly, Kendra D. Nyberg
Journal of Microelectromechanical Systems, Vols. 17:2-6	

The Multi-Mode Resonance in AlN Lamb Wave Resonators	Jie Zou, Chih-Ming Lin, Anming Gao
Design and Development of Piezoelectric Composite-Based Micropump	S. Revathi, R. Padmanabhan
A Proactive Plastic Deformation Method for Fine-Tuning of Metal-Based MEMS Devices After Fabrication	Yong-Hoon Yoon, Chang-Hoon Han, Jae-Shin Lee
High Performance, Continuously Tunable Microwave Filters Using MEMS Devices With Very Large, Controlled, Out-of-Plane Actuation	Jackson Chang, Michael J. Holyoak, George K. Kannell
A Novel Topology for Process Variation-Tolerant Piezoelectric Micromachined Ultrasonic Transducers	Alexandre Robichaud, Dominic Deslandes, Paul-Vahé Cicek
A MEMS Nonlinear Dynamic Approach for Neural Computing	Fadi M. Alsaleem, Mohammad H. H. Hasan, Mehari K. Tesfay
Solving FSR Versus Offset-Drift Trade-Offs With Three-Axis Time-Switched FM MEMS Accelerometer	Cristiano Rocco Marra, Alessandro Tocchio, Francesco Rizzini
Direct Detection of Anchor Damping in MEMS Tuning Fork Resonators	Janna Rodriguez, Saurabh Arun Chandorkar, Grant M. Glaze
Inverse Eigenvalue Sensing in Coupled Micro/Nano System	Guowei Tao, Hemin Zhang, Honglong Chang
Optimization of a Collapsed Mode CMUT Receiver for Maximum Off-Resonance Sensitivity	Mansoor Khan, Talha M. Khan, Akif Sinan Taşdelen
Electrical Stimulation, Recording and Impedance-Based Real-Time Position Detection of Cultured Neurons Using Thin-Film-Transistor Array	Faruk Azam Shaik, Grant Alexander Cathcart, Satoshi Ihida
Monitoring of Water Transportation in Plant Stem With Microneedle Sap Flow Sensor	Sangwoong Baek, Eunyong Jeon, Kyoung Sub Park
3-D Printed Biocompatible Micro-Bellows Membranes	Khalil Moussi, Jurgen Kosel
A Closed-Loop Mode-Localized Accelerometer	Jing Yang, Jiming Zhong, Honglong Chang
Design of a Dual Quantization Electromechanical Sigma–Delta Modulator MEMS Vibratory Wheel Gyroscope	Bin Sheng, Fang Chen, Chao Qian
Investigation of Multimodal Electret-Based MEMS Energy Harvester With Impact-Induced Nonlinearity	Kai Tao, Lihua Tang, Jin Wu
High Shock-Resistant Design for Wafer-Level-Packaged Three-Axis Accelerometer With Ring-Shaped Beam	Atsushi Kazama, Takanori Aono, Ryoji Okada
Journal of Operations Management, Vols. 60-64	
Does the meaningful use of electronic health records improve patient outcomes?	Deepa Wani, Manoj Malhotra
Designing crowdsourced delivery systems: The effect of driver disclosure and ethnic similarity	Ha Ta, Terry L. Esper, Adriana Rossiter Hofer
Judgmental selection of forecasting models	Fotios Petropoulos, Nikolaos Kourentzes, Konstantinos Nikolopoulos
Supplier non-retention post disruption: What role does anger play?	Mikaella Polyviou, M. Johnny Rungtusanatham, Rebecca W. Reczek
Valuing supply-chain responsiveness under demand jumps	Işık Biçer, Verena Hagspiel, Suzanne de Treville
An empirical examination of surgeon experience, surgeon rating, and costs in perioperative services	Sriram Venkataraman, Lawrence D. Fredendall, Kevin M. Taaffe
The decision to recall: A behavioral investigation in the medical device industry	George P. Ball, Rachna Shah, Karen Donohue
Inventory agility upon demand shocks: Empirical evidence from the financial crisis	Maximiliano Udenio, Kai Hoberg, Jan C. Fransoo
Purchasing managers' willingness to pay for attributes that constitute sustainability	Philipp Goebel, Carsten Reuter, Richard Pibernik
Effectiveness of bonus and penalty incentive contracts in supply chain exchanges: Does national culture matter?	Yun Shin Lee, Dina Ribbink, Stephanie Eckerd
Avoiding epistemological silos and empirical elephants in OM: How to combine empirical and simulation methods?	Aravind Chandrasekaran, Kevin Linderman, Fabian J. Sting
Asset supply networks in humanitarian operations: A combined empirical-simulation approach	Jon M. Stauffer, Alfonso J. Pedraza Martinez, Lu (Lucy) Yan

Structural anatomy and evolution of supply chain alliance networks: A multi-method approach	Hyunwoo Park, Marcus A. Bellamy, Rahul C. Basole
Supplier dependence and R&D intensity: The moderating role of network centrality and interconnectedness	Dong-Young Kim, Pengcheng Zhu
On doing relevant and rigorous experiments: Review and recommendations	Sirio Lonati, Bernardo F. Quiroga, Christian Zehnder
Hedging weather risk and coordinating supply chains	Xavier Brusset, Jean-Louis Bertrand
Addressing endogeneity in operations management research: Recent developments, common problems, and directions for future research	Guanyi Lu, Xin (David) Ding, David Xiaosong Peng
International Journal of Machine Tools and Manufacture, Vols. 133-137	
Thermal, mechanical and chemical material removal mechanism of carbon fiber reinforced polymers in electrical discharge machining	Xiaoming Yue, Xiaodong Yang, Jing Tian
On the generation of chatter marks in peripheral milling: A spectral interpretation	N. Grossi, A. Scippa, L. Sallese, F. Montevecchi
A theoretical and experimental study of spindle imbalance induced forced vibration and its effect on surface generation in diamond turning	Shaojian Zhang, Jinjie Yu, Suet To
Forming characteristics of tube free-bending with small bending radii based on a new spherical connection	Xunzhong Guo, Hao Xiong, Heng Li
A physics-based single crystal plasticity model for crystal orientation and length scale dependence of machining response	Eralp Demir, Canset Mercan
An investigation of resolved shear stress on activation of slip systems during ultraprecision rotary cutting of local anisotropic Ti-6Al-4V alloy: Models and experiments	Zejia Zhao, Suet To
Heat flux in metal cutting: Experiment, model, and comparative analysis	V. Kryzhanivskyy, V. Bushlya, O. Gutnichenko
A novel ductile machining model of single-crystal silicon for freeform surfaces with large azimuthal height variation by ultra-precision fly cutting	Zhanwen Sun, Suet To, Shaojian Zhang
Predictive modeling of chatter stability considering force-induced deformation effect in milling thin-walled parts	Yuwen Sun, Shanglei Jiang
Hardness, grainsize and porosity formation prediction on the Laser Metal Deposition of AISI 304 stainless steel	Jon Iñaki Arrizubieta, Aitzol Lamikiz, Magdalena Cortina
Real-time feedrate scheduling for five-axis machining by simultaneously planning linear and angular trajectories	Jie Huang, Yaoan Lu, Li-Min Zhu
Microstructure evolution in Ti64 subjected to laser-assisted ultrasonic nanocrystal surface modification	Jun Liu, Sergey Suslov, Zhencheng Ren
Adaptive machining for curved contour on deformed large skin based on on-machine measurement and isometric mapping	Qingzhen Bi, Nuodi Huang, Shaokun Zhang
Cutting forces in fast-/slow tool servo diamond turning of micro-structured surfaces	Zhiwei Zhu, Suet To, Wu-Le Zhu
Diamond cutting of micro-structure array on brittle material assisted by multi-ion implantation	Jinshi Wang, Xiaodong Zhang, Fengzhou Fang
On the role of powder flow behavior in fluid thermodynamics and laser processability of Ni-based composites by selective laser melting	Dongdong Gu, Mujian Xia, Donghua Dai
Modelling of elliptical dimples generated by five-axis milling for surface texturing	Miguel Arizmendi, Amaia Jiménez, Wilmer E. Cumbicus
Medical concept normalization in social media posts with recurrent neural networks	Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko
Chemical-induced disease relation extraction with dependency information and prior knowledge	Huiwei Zhou, Shixian Ning, Yunlong YangLin
From narrative descriptions to MedDRA: automagically encoding adverse drug reactions	Carlo Combi, Margherita Zorzi, Gabriele Pozzani
A data-driven method to detect adverse drug events from prescription data	Chen Zhan, Elizabeth Roughead, Lin Liu, Nicole Pratt, Jiuyong Li
When to re-order laboratory tests? Learning laboratory test shelf-life	Gal Levy-Fix, Sharon Lipsky Gorman, Jorge L. Sepulveda
An unsupervised machine learning method for discovering patient clusters based on genetic signatures	Christian Lopez, Scott Tucker, Tarik Salameh
Transferability of artificial neural networks for clinical document classification across hospitals: A case study on abnormality detection from radiology reports	Hamed Hassanzadeh, Anthony Nguyen, Sarvnaz Karimi

Trie-based rule processing for clinical NLP: A use-case study of n-trie, making the ConText algorithm more efficient and scalable	Jianlin Shi, John F. Hurdle
POPCORN: A web service for individual PrognOsisprediction based on multi-center clinical dataCollabORatioN without patient-level data sharing	Yu Tian, Yong Shang, Dan-Yang Tong
Semantic relation extraction aware of N-gram features from unstructured biomedical text	Zheng Wang, Shuo Xu, Lijun Zhu
A multi-level usability evaluation of mobile health applications: A case study	Hwayoung Cho, Po-Yin Yen, Dawn Dowding
Predicting the risk of acute care readmissions among rehabilitation inpatients: A machine learning approach	Yajiong Xue, Huigang Liang, John Norbury
Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power	Alex Graudenzi, Davide Maspero, Marzia Di Filippo
relSCAN – A system for extracting chemical-induced disease relation from biomedical literature	Stanley Chika Onye, Arif Akkeleş, Nazife Dimililer
An evaluation method of risk grades for prostate cancer using similarity measure of cubic hesitant fuzzy sets	Jing Fu, Jun Ye, Wenhua Cui
Integrated Bioinformatics Analysis for Identificating the Therapeutic Targets of Aspirin in Small Cell Lung Cancer	Liuyun Gong, Dan Zhang, Yiping Dong
A method for harmonization of clinical abbreviation and acronym sense inventories	Lisa V. Grossman, Elliot G. Mitchell, George Hripcsak
Applied Clinical Informatics Journal, Vols. 9:1-4, 10:1	
Measuring Electronic Health Record Use in Primary Care: A Scoping Review	Huang, Michael Z., Gibson, Candace J., Terry, Amanda L
Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record	Erik Joukes, Ameen Abu-Hanna, Ronald Cornet
Exploring Data Quality Management within Clinical Trials	Lauren Houston, Yasmine Probst, Ping Yu
Optimizing the User Experience: Identifying Opportunities to Improve Use of an Inpatient Portal	Daniel Walker, Terri Menser, Po-Yin Yen
Development and Validation of a Natural Language Processing Tool to Identify Patients Treated for Pneumonia across VA Emergency Departments	B.E. Jones, B.R. South., Shao Y. Lu,
Validation and Refinement of a Pain Information Model from EHR Flowsheet Data	Bonnie L. Westra, Steven G. Johnson, Samira Ali
Using EHR Data to Detect Prescribing Errors in Rapidly Discontinued Medication Orders	Jonathan D. Burlison, Robert B. McDaniel, Donald K. Baker
How Do Experienced Physicians Access and Evaluate Laboratory Test Results for the Chronic Patient? A Qualitative Analysis	Torbjørn Torsvik, Børge Lilliebo, Morten Hertzum
Leveraging Patient-Reported Outcomes Using Data Visualization	Lisa Grossman, Steven K. Feiner, Elliot Michetll
Adoption of Electronic Dental Records: Examining the Influence of Practice Characteristics on Adoption in One State	Zain Chauhan, Mohammad Samarah, Kim M. Unertl
Using Clinical Data Standards to Measure Quality: A New Approach	John D. D'Amore, Chun Li, Laura McCrary
SNOMED CT Concept Hierarchies for Sharing Definitions of Clinical Conditions Using Electronic Health Record Data	Duwane L. Willett, Vaishnavi Kannan, Ling Chu
Integrating Multimodal Radiation Therapy Data into i2b2	Eric Zapletal, Jean-Emmanuel Bibault, Philippe Giraud
Integrated Electronic Discharge Summaries—Experience of a Tertiary Pediatric Institution	Daryl Cheng, Merav Katz, Mike South
Simple Workflow Changes Enable Effective Patient Identity Matching in Poison Control	Mollie R. Cummins, Pallavi Ranade-Kharkar, Cody Johansen
Automatic Detection of Front-Line Clinician Hospital Shifts: A Novel Use of Electronic Health Record Timestamp Data	Adam C. Dziorny, Evan W. Orenstein, Robert B. Lindell
Communicating with Vulnerable Patient Populations: A Randomized Intervention to Teach Inpatients to Use the Electronic Patient Portal	Jacob Stein, Jared Klein, Thomas Payne
International Journal of Impact Engineering, Vols. 122-126	
Hydrocode simulations of a hypervelocity impact experiment over a range of velocities	Tamra Heberling, Guillermo Terrones, Wayne Weseloh

High-velocity impact behaviour of aluminium honeycomb sandwich panels with different structural configurations	Guangyong Sun, Dongdong Chen, Hongxu Wang
Mechanism of velocity enhancement of asymmetrically two lines initiated warhead	Yuan Li, Shihui Xiong, Xiaogang Li
X-FEM Analysis of dynamic crack growth under transient loading in thick shells	Thomas Elguedj, Yannick Jan, Alain Combescure
Analytical model of hypervelocity penetration into rock	Jie Li, Mingyang Wang, Yihao Cheng
Skin-stringer interface failure investigation of stringer-stiffened curved composite panels under hail ice impact	Zhenhua Song, Jacqueline Le, Daniel Whisler
Debris dispersion analysis for the determination of impact conditions via traceback technology	Jong-Tak Kim, Sung-Choong Woo, Jin-Young Kim
Experimental evaluation of the impact behavior of partially melted ice particles	Miguel Alvarez, Richard E. Kreeger, Jose Palacios
Influence of lateral dimensions, obliquity, and target thickness toward the efficiency of unconfined ceramic tiles for the defeat of rod penetrators	T. Behner, A. Heine
Experimental and numerical studies on the drop impact resistance of prestressed concrete plates	M.A. Iqbal, V. Kumar, A.K. Mittal
Experimental, numerical, and theoretical studies of the response of short cylindrical stainless steel tubes under lateral air blast loading	Shiquiang Li, Boli u, Dora Karagiozova, Zhifang Liu
Measurement of shock pressure and shock-wave attenuation near a blast hole in rock	Li Yuan Chi, Zong-Xian Zhang, Arne Aalberg
Evolutionary characteristics of thermal radiation induced by 2A12 aluminum plate under hypervelocity impact loading	Han Yafei, Tang Enling, He Liping
A scaled framework for strain rate sensitive structures subjected to high rate impact loading	Hamed Sadeghi, Keith Davey, Rooholamin Darvizeh
Protection effectiveness of perforated plates made of high strength steel	Wojciech Burian, Paweł Żochowski, Michał Gmitrzuk
Comparison of shielding performance of Al/Mg impedance-graded-material-enhanced and aluminum Whipple shields	Pinliang Zhang, Zizheng Gong, Dongbo Tian
Ballistic performance and energy absorption characteristics of thin nickel-based alloy plates at elevated temperatures	Jiao Liu, Bailin Zheng, Kai Zhang
International Journal of Robust and Nonlinear Control, Vols. 28:16-18, 29:1-2	
On the complexity and dynamical properties of mixed logical dynamical systems via an automaton-based realization of discrete-time hybrid automaton	Mohammad Hejri, Alessandro Giua, Hossein Mokhtari
Global finite-time output stabilization of nonlinear systems with unknown measurement sensitivity	Jin-Xi Zhang, Guang-Hong Yang
Computing multiple Lyapunov-like functions for inner estimates of domains of attraction of switched hybrid systems	Xiuliang Zheng, Zhikun She, Junjie Lu
Adaptive estimation and output feedback FTC for nonlinear systems with unknown nonlinearities and faults	Sheng-Juan Huang, Da-Qing Zhang, Liang-Dong Guo
Reliable finite-time H∞ control for singular Markovian jump system with actuator saturation via sliding-mode approach	Yuechao Ma, Yangfan Liu, Na Liu
Gain-scheduled continuous-time control using polytope-bounded inexact scheduling parameters	Arash Sadeghzadeh
Integrated fault estimation and fault-tolerant control for uncertain time-varying delay nonlinear Markovian jump systems	Fucang Qi, Yuechao Ma
Adaptive sliding-mode control for spacecraft relative position tracking with maneuvering target	Kai Zhang, Guangren Duan, Mingda Ma
Robust output-feedback control of 3D directional drilling systems	O.A. Villarreal, Magaña F.H.A. Monsieurs, E. Detournay
Tracking control for systems with slope-restricted hysteresis nonlinearities	Arnab Dey, Sourav Patra, Siddhartha Sen
Input/output stability of a damped string equation coupled with ordinary differential system	Matthieu Barreau, Frédéric Gouaisbaut, Alexandre Seuret
Distributed adaptive model-based event-triggered predictive control for consensus of multiagent systems	Xiuxia Yin, Dong Yue, Songlin Hu
Semi-global containment control for linear systems in the presence of actuator position and rate saturation	Zhiyun Zhao, Hongbo Shi

Stability analysis of a class of uncertain switched time-delay systems with sliding modes	Mohammad Hasan, H. Kani Mohammad, Javad Yazdanpanah
Neuroadaptive quantized PID sliding-mode control for heterogeneous vehicular platoon with unknown actuator deadzone	Xianggui Guo, Jianliang Wang, Fang Liao
Distributed leader-follower consensus for a class of semilinear second-order multiagent systems using time scale theory	Serhii Babenko, Michael Defoort, Mohamed Djemai
Dynamic event-triggered control for linear time-invariant systems with c2 gain performance	Dan Liu, Guang-Hong Yang
Journal of Guidance, Control and Dynamic, Vols. 41:9-12, 42:1	
Probabilistic Trajectory Optimization Under Uncertain Path Constraints for Close Proximity Operations	Christopher Jewison, David W. Miller
Square-Root Consider Filters with Hyperbolic Householder Reflections	James S. McCabe, Kyle J. DeMars
Study of Correction Maneuver for Lunar Flyby Transfers in the Real Ephemeris	Yi Qi, Anton de Ruiter
Closed-Loop Linear Covariance Analysis for Hosted Payloads	Randall S. Christensen, David K. Geller
Analytical Conditions for Bounded Mean Inter-Satellite Distances in the J2 Problem	Tao Nie, Pini Gurfil, Shijie Zhang
Robust Orbit Determination with Flash Lidar Around Small Bodies	Ann B. Dietrich, Jay W. McMahon
Finite-Time Input-to-State Stability Guidance Law	Guilin Li, Ming Xin, Changxin Miao
Linear Parameter-Varying Control for Quadrotors in Case of Complete Actuator Loss	Johannes Stephan, Lorenz Schmitt, Walter Fichter
Dynamics and Control of In-Flight Wing Tip Docking	John R. Cooper, Paul M. Rothhaar
Weapon–Target Assignment Algorithm for Simultaneous and Sequenced Arrival	Kyle Volle, Jonathan Rogers
Robust Constant-Amplitude Input Shapers with Selectable Duration	Christopher Adams, William Singhose
Min-Max Differential Dynamic Programming: Continuous and Discrete Time Formulations	Wei Sun, Yunpeng Pan, Jaein Lim
Semisynchronizing Strategy for Capturing a High-Speed Tumbling Target	Chuan Ma, Caisheng Wei, Jianping Yuan
Fuel-Optimal Rocket Landing with Aerodynamic Controls	Xinfu Liu
Bounded Relative Orbits in the Zonal Problem via High-Order Poincaré Maps	Yanchao He, Roberto Armellin, Ming Xu
Trajectory Tracking Near Small Bodies Using Only Attitude Control	Xiangyu Li, Rakesh R. Warier, Amit K. Sanyal
Sliding-Mode Impact Time Guidance Law Design for Various Target Motions	Qinglei Hu, Tuo Han, Ming Xin
Journal of Spacecraft and Rockets, Vols. 55:2-6	
Optimization of a Mach-6 Quiet Wind-Tunnel Nozzle	Matthew T. Lakebrink, Kevin G. Bowcutt, Troy Winfree
Force and Moment Measurements on a Free-Flying Capsule in a Shock Tunnel	S. J. Laurence, C. S. Butler, J. Martinez Schramm
Constellation Phasing with Differential Drag on Planet Labs Satellites	Cyrus Foster, James Mason, Vivek Vittaldev
Trajectories for Flyby Sample Return at Icy Moons	Drew Ryan Jones
Planetary Probe Entry Atmosphere Estimation Using Synthetic Air Data System	Christopher D. Karlgaard, Mark Schoenenberger
Conceptual Design of an Air-Breathing Electric Thruster for CubeSat Applications	Stephen W. Jackson, Robert Marshall
Surface Chemical Effects on Hypersonic Nonequilibrium Aeroheating in Dissociated Carbon–Oxygen Mixture	Xiaofeng Yang, Yewei Gui, Wei Tang
Computational Analysis of Subsonic Jets from Rectangular Nozzles with and Without Bevel	M. Sandhya, P. S. Tide
Simplified Numerical Approach for the Prediction of Aerodynamic Forces on Grid Fins	Erdem Dikbas, Özgür Uğraş Baran, Cuneyt Sert
Particle Simulation of Plasma Drag Force Generation in the Magnetic Plasma Deorbit	Rei Kawashima, Junhwi Bak, Shinji Matsuzawa
Feasibility Assessment of Aerocapture for Future Titan Orbiter Missions	Ye Lu, Sarag J. Saikia
Superposition Method for Force Estimations on Bodies in Supersonic and Hypersonic Flows	Ansgar Marwege, Sebastian Willems, Ali Gülhan
Robust Attitude Control Using a Double-Gimbal Variable-Speed Control Moment Gyroscope	Takahiro Sasaki, Takashi Shimomura, Hanspeter Schaub

Modular, Fast Model for Design and Optimization of Hypersonic Vehicle Propulsion Systems	Alessandro Mogavero, Richard E. Brown
Hypersonic Boundary-Layer Duplication Methodology Downstream of the Stagnation Point	Alan Viladegut, Jean-Etienne Durand, Fabio Pinna
Liquid Crystal Device with Reflective Microstructure for Attitude Control	Toshihiro Chujo, Hirokazu Ishida, Osamu Mori
Effect of Rocket Engine Layouts on Jet Flowfield Inside a Launch Pad	Seiji Tsutsumi, Wataru Sarae, Hiroyuki Ueda

APPENDIX C. ADAPTATIONS TO WMATRIX SEMANTIC CATEOGRIES

Passive Verb P-frame Verb Semantic Categories

Final Category	WMatrix Category	Example Verbs
Cause & Effect/Connect	Cause & Effect/Connection	link, determine, produce, relate
Change	Change	modify, develop, rework, transform
	Clothes	tailor
Change Quantity	Quantities: Many/Much	add, increase
	Size: Big	grow, expand, magnify
	Quantities: Little	reduce, minimize, diminish
Compare	Comparing: Similar/Different	compare, tend
	Comparing: Similar	converge, replicate
	Comparing: Different	vary, distinguish
Describe	Generally, Kinds, Groups, Examples	illustrate, exemplify, characterize
Desire	Wanted	aim, desire, schedule target
Evaluate	Evaluation: Good/Bad	
Find	Knowledgeable	identify, now, recognize
	Open, Finding; Showing	reveal, demonstrate, detect, indicate
	Sensory: Sight	see
Get	Getting and Giving	exchange
	Getting and Possession	obtain, achieve, take, capture, keep
Improve	Evaluation: Good	
Investigate	Education	study, test
	investigate, Examine, Test, Search	
Leave out	Avoiding	avoid, omit, neglect
	Time: Ending	cancel, disconnect
	Inattentive	ignore, disregard
	Entirety, Maximum	limit
Make	Architecture	Building
	Objects Generally	clamp, trigger, drill
	General Actions/Making	construct, create, perform, conduct, make, enforce, project
	Arts & Crafts	design, depict
	Science and Technology	engineer
	Inclusion	integrate, include, contain, embed
	Substances	solidify, deposit, drop
Measure	Measurement: General	calibrate, measure
	Measurement: Distance	extend

	Measurement: Size	scale
	Measurement: Area	stretch
Move	Vehicles and Transport on Land	drive, transport, track
	Location and Direction	locate, orient, transpose
	Putting, Pulling, Pushing, Transporting	remove, carry, attach, eject, download, extract
	Moving, Coming and Going	run, scatter, disperse, follow, maneuver
Record	Sensory: Sight	observe
	Linguistic Actions, States, and Processes	represent, embody
	Paper Documents and Writing	write, record, document, note
Speech acts	Likely	confirm, guarantee, assure
	Speech Acts	recommend, report, summarize, propose
	Speech: Communicative	talk, say, argue,
Think	Expected	anticipate, expect, envisage
	Thought, belief	consider, assume, suppose, formulate
	Deciding	decide, estimate, conclude
	Understanding	realize, understand, infer
Use	Work and Employment	employ, work
	Using	use, utilize

Noun + *Preposition* P-frames Noun Semantic Categories

Final Category	WMatrix Categories	Example Nouns
Research	Investigate, examine, test	research, assessment, analysis
	Knowledge	database, identification, information
	Language, speech, grammar	terminology, vocabulary, abbreviation
	Mental Objects	system, method, target, pattern
	Paper documents and writing	paper, flowchart, graph, text
	Generally Kinds, Groups	case, category, illustration, instance
Evaluation	Likelihood	potential, viability, prospect
	Quantity	increase, reduction, peak, scarcity
	Cause & Effect /Connection	result, impact, effect
	Comparing	norm, variance, comparison
	Evaluation	quality, accuracy, error, validity
	Helping/Hindering	recovery, benefit, drawback, obstacle
Engineering	Anatomy and physiology	eye, pore, tissue, optic
	Business	vendor, office, retailer
	Electricity	anode, radar, voltage
	Medical	hospital, physician, diagnosis, injection
	Abbreviations	CSA, RSE, ATP
	Objects generally	bundle, pipelines, pendulum
	Science & technology	engineer, topography, wavelength
	Substances & materials	chemical atom, granite, substrate
	Technical	flowfield, accelerometer, ferrofluid
	Flying & aircraft	trajectory, asteroid, Jacobean
	Light	light, beam, laser
	Vehicles and transport	bumper, cart, transport
	Location and Direction	orientation, position, midst
Math	Mathematics	ratio, computation, equation, sum
	Measurement	radius, diameter, gauge
	Shape	sphere, spiral, geometry
Objects	Objects Generally	core, buckle, frame, machine, rock
	Living Creatures	mice, plant, shell, tail
	No Constraint	discharge, release, scope
	Sports	goal, striker
Places	Places	area, environment, region, site, vicinity
	Geography	midway, channel, earth, inlet, stream
Not a Word	Not a word	miscellaneous abbreviations

APPENDIX D. LESSON PLANS

Original Lesson Plan

Day 1

Minute	Schedule
0	Pretest
10	Intro Sheet: educational goals, and the corpus
15	Phrase 1: *due to the;* check comfort with parts of speech *Questions 1 individually, report back to class*
25	Explain the four sentence locations of *due to the* *Question 2, individually then together*
35	Introduce concepts of strengthening and hedging *Questions 3,4 in groups*
45	Phrase 2: *there be no.* Introduce transitions *Questions 1 individually*
55	Introduce verb phrases: citing authority vs. own work; strengthening vs. hedging verbs *Questions 2-3 in groups*
65	Strengthening and hedging adjectives *Question 4 in pairs*
	Assign the final questions on both sheets and a feedback form for homework

Day 2

Minute	**Schedule**
0	Collect homework
5	Phrase 3: *it be * to:* introduce dummy it, author stance, adjectives of possibility & importance *Question 1 in groups, 2 individually then pairs*
20	Introduce 3 rhetorical uses *Questions 3, 4 in pairs*
35	Phrase 4: *it be * that*: explain new clause group, review dummy it, introduce qualifying phrases *Question 1 in pairs*
45	Introduce the phrases' three rhetorical uses *Questions 2, 3 individually, then in groups*
55	Identify the passive "knower" *Question 4 in pairs*
65	Assign final Qs as homework, do posttest, collect feedback

Revised Lesson Plan

Day 1

M**inute**	**Schedule**
0	Pretest
10	Intro Sheet: educational goals, and the corpus
15	Phrase 1: *due to the;* check comfort with parts of speech *Questions 1 individually, report back to class*
25	Explain the four sentence locations of *due to the* *Question 2, individually then together*

35	Introduce concepts of strengthening and hedging *Questions 3,4 in groups* Feedback/Questions?
45	Question 5
55	Phrase 3: *it be * to:* introduce dummy it, author stance, adjectives of possibility & importance *Questions 1 & 2 in pairs, as a class*

Day 2

Minute	**Schedule**
0	Return to Phrase 3, Introduce 3 rhetorical uses *Questions 3, 4 in pairs, as a class* *Question 5: share with class*
25	Phrase 4: *it be * that*: explain new clause group, review dummy it, introduce qualifying phrases *Question 1 as a group*
35	Introduce the phrases' three rhetorical uses *Questions 2, 3 individually, then in groups*
45	Identify the passive "knower" *Questions 4, 5 in pairs, review as class*
55	*Start Question 6 as a class*
60	Do posttest, distribute and collect feedback form

APPENDIX E. CLASSROOM MATERIALS

The contents of Appendix E are the nine documents given to the students and used in the classroom, as follows:

I. Introductory Sheet

II. Worksheet for *due to the*

III. Summary for *due to the*

IV. Worksheet for *there be no*

V. Summary for *there be no*

VI. Worksheet for *it is * to*

VII. Summary for *it is * to*

VIII. Worksheet for *it be * that*

IX. Summary for *it be * that*

X. PowerPoint Slides

I. Introductory Sheet

What are these phrases, where is the data from, and how can they improve my writing?

Much of academic writing is composed of phrases that authors use very frequently to build arguments, explain their work and report results. When you practice using these structures, you will also become aware of some of the underlying patterns in academic writing and how you can re-write and change around your sentences to express yourself more effectively.

The information on the worksheets and summaries have been drawn from a **corpus** of published articles in the fields of aerospace, electrical, industrial and mechanical engineering. The corpus has 1.9 million words from 262 articles. The corpus has been analyzed to pull out the most common fixed and variable phrases. Verb tense information was removed, so capital letters indicate a variety of tenses. For example, "BE" originally appeared as *is, are, was, were,* and *been*.

A **fixed phrase** was a string of unchanging words: for example, "due to the," or "there BE no." A **variable phrase** is when one word in the string changes, but the meaning of the whole phrase is similar. For example, in "BE shown/believed/considered to be," the verb changes, but the function of the phrases is similar. Variable phrases are helpful for writers, as they show you the many nuanced ways you can say one thing.

For each phrase, we will discuss their context and their uses. Each phrase has a different **context**. In the illustrations, you will find the phrases in the middle, surrounded by the words that commonly appeared before and after the phrases. The larger the dot, the more frequently this word appeared. These words have been color coded – generally, transitions are bright red, verbs are dark red, green is adverbs and light blue is adjectives, though some phrases have special categories with other colors. We will discuss the **uses** of the phrase – what you as a writer can do with them. This information will help you to use the phrases **fluently**, as you will identify how & where published writers use these phrases and you will practice using them in similar patterns.

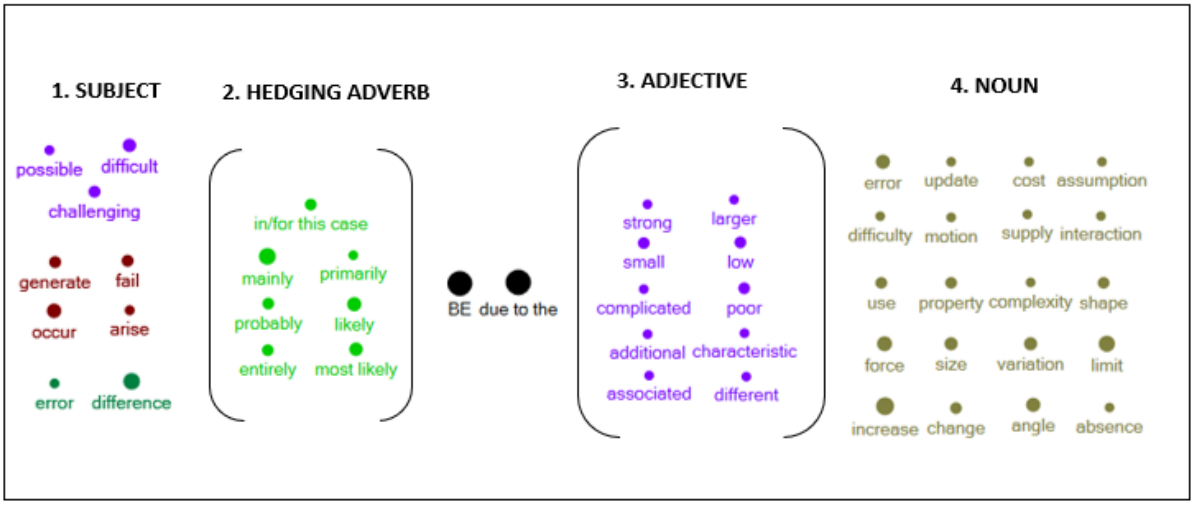
The sections titled **corpus examples** contain **real examples** from the corpus. This gives you examples of their natural use and will help you in learning to use them fluently. On the left of each example, you can see what engineering discipline in came from: AE is aeronautical engineering, EE is electrical engineering, IE is industrial engineering, ME is mechanical engineering.

Journals in the Corpus and their Word Counts

Journal Name	Branch	Issues	Word Count
Nature Nanotechnology	Electrical Engineering	13:9-12, 14:1	104,125
Automatica	Electrical Engineering	96-100	121,969
IEEE Transactions on Pattern Analysis and Machine Intelligence	EE: image Processing & Computer Vision	40:10-12, 41:1-2	143,060
International Journal of Computer Vision	EE: image Processing & Computer Vision	126: 8-12	166,889
Total EE:			**536,043**
Nature Materials	Mechanical Engineering	17:9-13, 18:1	85,255
Applied Energy	Mechanical Engineering	230-234	120,373
Lab on Chip	ME: Biomedical Electromechanical	2018:22-24, 2019:1,2	98,307
Journal of Microelectromechanical Systems	ME: biomedical electromechanical	17: 2-6	85,026
Total ME:			**388,961**
Journal of Operations Management	Industrial Engineering	60-64	202,650
International Journal of Machine Tools and Manufacture	Industrial Engineering	133-137	120,713
Journal of Biomedical Informatics	IE: biomedical informatics	84-88	111,938
Applied Clinical Informatics Journal	IE: biomedical informatics	9:1-4, 10:1	84,385
Total IE:			**519,686**
International Journal of Impact Engineering	Aerospace Engineering	122-126	112,437
International Journal of Robust and Nonlinear Control	Aerospace Engineering	28:16-18, 29:1-2	107,596
Journal of Guidance, Control and Dynamic	AE: Aerodynamics	41:9-12, 42:1	144,978
Journal of Spacecraft and Rockets	AE: Aerodynamics	55:2-6	115,729
Total AE:			**480,740**
Total Corpus			**1,925,430**

II. Worksheet for *due to the*

Fixed Phrase 1: "due to the"



Exercises

1. See corpus examples 49, 2 and 116 for noun, adjective and verb, respectively. Read through 10-15 more lines, underlining the subject. Is the subject usually a noun, adjective or verb?

2. Write two sentences about your research where you use this phrase to explain a cause. Now change the place of the *due to the* phrase in each sentence. What differences the changes in order and grammar make?

3. How can you use words from category 2 and 3 to *hedge* or *strengthen* a claim? What other words can you use?

4. Re-write your sentences from (4) and add hedging and strengthening adverbs where appropriate, and exchange sentences with a colleague. What elements can you identify as hedging or strengthening?

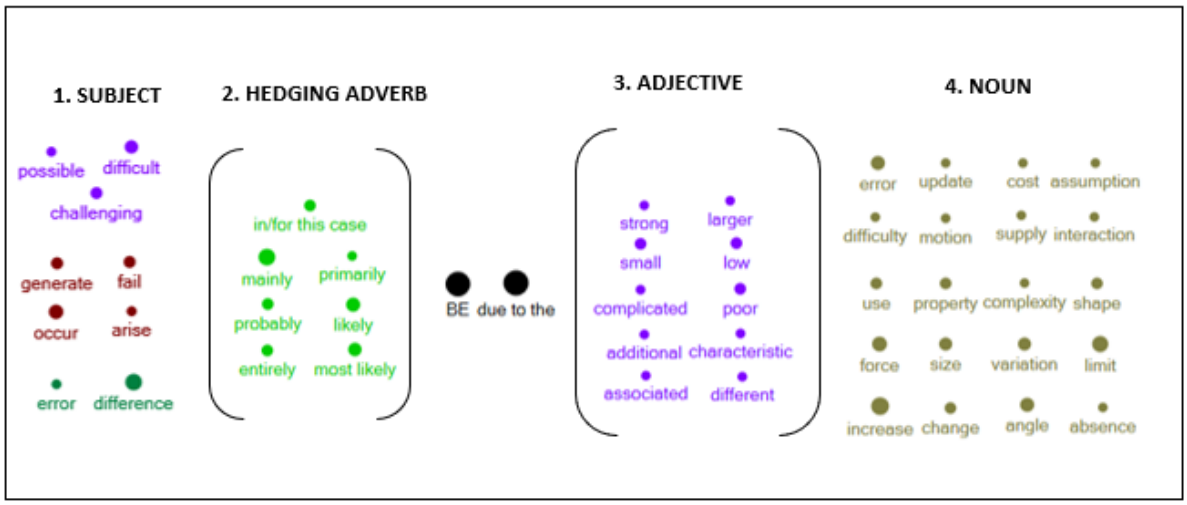
5. Write four new sentences with the phrase *due to the*. In your four sentences, use at least two kinds of subjects (verb, noun or adjective), two hedging adverbs, and all four sentence placements.

Corpus Examples

EX. #		SOURCE
2	reacting case, a direct comparison is difficult due to the absence of averaged data, but	AE4.14
3	to be beneficial for reducing PM emissions due to the absence of aromatics, which are	ME2.3
15	section III. The small discrepancies may be due to the actual design parameter variations during	ME4.16
20	generator is usually difficult to scale up due to the additional requirement of flow infusion	ME3.9
35	The velocity and jerk violations are due to the approximation for linear and angular	IE2.11
49	can be seen, the impact time errors due to the autopilot lag are within an	AE3.17
57	correspond to Eqs. (2), (3), and (6) respectively. Due to the binary constraint on Z, it	EE4.1
68	partition is demanding in an electron system due to the challenges associated with atomic-scale	ME1.7
88	This result could be due to the higher temperature observed in the	EE3.5
93	on a single silicon chip, remains challenging due to the complexity of elastic multimode coupling	ME1.7
99	they are inefficient for large point sets, due to the computational cost. Triangular surface	EE4.10
110	well as short ones. This is likely due to the contextual anomaly detection approach,	IE3.5
116	by Lowe et al. [29] is, however, limited due to the cost and also the time-	IE3.14
133	and (d) show the varying entry angles due to the deformation along the axial depth	IE2.9
139	to avoid any kind of flow blockage due to the development of the boundary layer.	AE4.15
147	systems, which simplifies controller design. Due to the differences between the estimated and	AE2.9
159	frequency, as shown in Table 1. However, due to the difficulty of purchasing commercial silicon	ME3.15
181	exhibits ferromagnetic response and rises due to the effect of the magnetic field	ME2.10
215	in Refs. [12, 13], that effect is not simply due to the extra mass or its capability	AE1.9
231	application of numerical techniques is easier due to the fact that differential equations are	AE2.1
254	of the aspect ratio used. This is due to the fact that the difference between	ME4.2
310	that these problems are entirely due to the resulting electricity programs not giving	EE4.3
334	both assays was found to be similar, due to the higher noise from the protein	ME3.7
362	increase in aspect ratio may be either due to the increase in length or decrease	ME4.2
380	"A" is set to 0.201, see Eq. (1). Besides, due to the interaction between the laser beam	IE2.10
412	, and the results are presented in Fig. 18. Due to the lack of experimental data, solutions	AE4.9
413	data has recently become a research focus. Due to the lack of external validation, prediction	IE3.9
432	task. The last layer's size is due to the large amount of sign-labels	EE4.15
459	a cutting tool is never infinitely sharp due to the limitations in manufacturing. As the	IE2.5
476	39s trained on the same video frames. Due to the low variation in the training	EE4.11
512	There are inconsistencies as well, possibly Due to the manner in which IT or EHR was	IE1.1
516	of the immune system is not surprising due to the nature of the Immunochip assay,	IE3.6
577	that is transferred into a tool body. Due to the practical impossibility of measuring flux	IE2.7
597	apoptosis and catabolic status, probably due to the variation of pro-arthritogenic profiles of	EE1.15
618	the typical cases for this strategy. However, due to the relatively small amount of electricity	ME2.17
619	hard to accurately detect within the image due to the relatively large movement associated with	ME3.8
639	MEMS deflection is higher than the deflection due to the same level of voltage applied	ME4.6
657	pixel colour changes considerably more quickly due to the short duration of the extracellular	EE1.3
669	leakage: Fuel leakage is likely to occur due to the small size and high diffusivity	ME2.14
678	especially for sandwich panels due to the sophistication of various geometric	AE1.2
690	the beginning of the face model adaptation due to the strong influence of illumination on	EE4.16
703	errors, namely mistakes that are entirely due to the supplier's poor judgements and	IE1.4
749	end, recognizing that the decision involves risk. Due to the uncertainty of the verification process,	IE1.10
750	directly to the nonlinear vehicular system (1) due to the uncertainty of mechanical drag and	AE2.15
788	to guarantee input/output stability. This is due to the well-known small-delay phenomenon,	AE2.11

III. Summary for *due to the*

"due to the" Summary



Pieces of the Sentence:

(1) The subject of the phrase: an adjective, noun or verb that is caused. For example, "The *errors* are due to the autopilot lag," (N) "A comparison is *difficult* due to the absence of averaged data," (ADJ) and "The mistakes *were detected* due to the verification process," (V). Nouns & adjectives are most common.

(2) (optional) a hedging or strengthening adverb

(BE) due to the

(3) (optional) adjective, usually strengthening the claim

(4) Noun: whatever caused the adjective, noun or verb. The noun may be based on an adjective (absence, difficulty) or a verb (variation, interaction, assumption).

Place in Sentence

The options for the place of *due to the* slowly increase the emphasis on *why* it happened, instead of *what* happened:

1. after the subject	"The powder aggregates quickly due to the cohesive force."
2. before the subject	"Due to the cohesive force, the powder aggregates quickly."
3. linked to subject with a form of "BE"	"The quick aggregation of powder is due to the cohesive force."
4. summarized with "THIS" and linked with "BE"	"The powder aggregates quickly. This is due to its cohesive force."

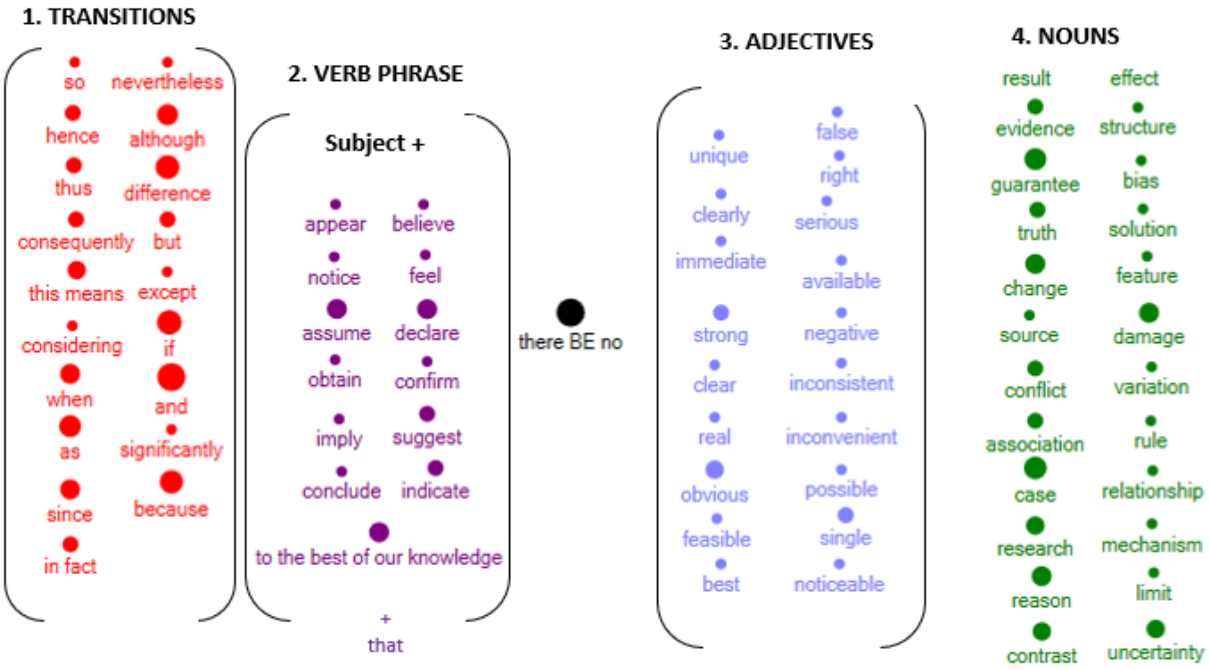
Hedging & Strengthening

The adverb and adjective can be used to hedge or strengthen a claim. The most frequent adverb hedges for a claim of causation are in the diagram above. Less frequent hedges were *might be, may be, could be, is thought to be, largely, possibly*. The infrequent adverbs that were used to strengthen were *entirely,* and *solely*.

You can also add an adjective that emphasizes an aspect of the noun (*small* size, *low* error rate); these adjectives primarily strengthen the claim you are making.

IV. Worksheet for *there BE no*

Fixed Phrase 2: "there BE no"



Exercises

1. Look at the transitions in Category (1) and in the Corpus Examples. Which transitions do you use most often? How can you group the transitions by their meaning?

2. The verbs in Category (2) appear with a subject before them. The full verb phrases can be found in the Corpus Examples 175-224. Who is the subject of the verb in these sentences?

3. Looking at the Category (2) list and the corpus examples, which verbs most indicate opinion or possibility? Which indicate fact? Arrange them from most certain to least certain.

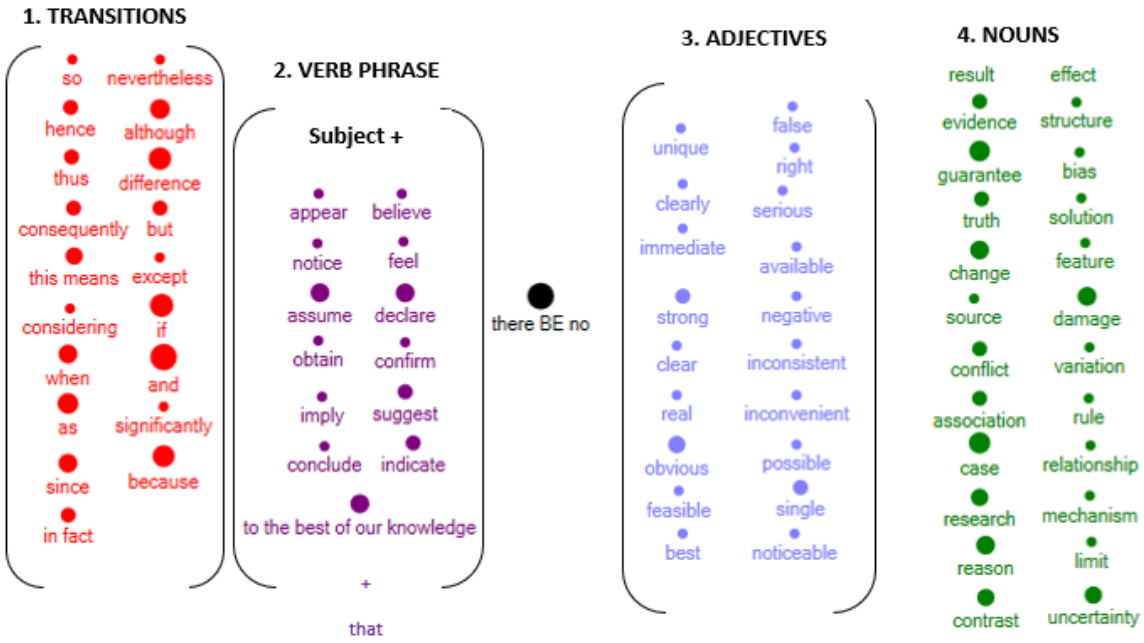
4. Find the adjectives before the nouns in the Corpus Examples. How do they strengthen or hedge the author's argument?

5. Write four sentences with "there BE no" in different contexts to discuss (hypothetical) results of your current research. Practice using new transitions, verb phrases, and hedging or strengthening adjectives.

Corpus Examples

#			TXT
4	the hybrid information. As mentioned above,	there are no studies on the expression of	IE3.15
15	efficiency of the scheduled generators, and	there is no uncertainty to be taken into	ME2.13
24	not present in the UGF simulations, and	there is no way to get the proper	AE4.9
34	0s z—' Sw t—1 is insufficient, as	there is no guarantee that this will reduce	EE2.2
37	equals ~0.006 (see Methods section). As	there is no rigorous estimation of the values	ME1.4
42	space. This is illustrated in Fig. 3b:	there is no need to associate an event	EE4.13
45	rows are ignored from the calculation because	there is no negative effect of inserting extra	EE4.12
49	regional IT adoption. However, because	there is no reason to believe that referral	IE4.10
52	difficult to make use of this because	there was no easy way to connect what	IE4.8
54	ohmic losses would increase. In this case,	there was no dependency between bias and Q	ME4.8
60	cultivated under cyanide-forming condition,	there is no understanding of the mechanisms of	EE1.9
65	constraint is not taken into consideration,	there is no limitation in the feasible region	AE3.3
69	of laboratory test orders. Since currently,	there are no other proposed methods to	IE3.5
72	in the primary care setting. To date,	there are no sources which comprehensively	IE4.1
84	decide to recall a product. In fact,	there are no behavioral studies, to our knowledge,	IE1.7
93	can increase (and often does), and hence	there is no guarantee that the last rounded	EE3.2
98	major changes in AR vari-ants.33,34 However,	there are no published reports comparing AR-V7	ME3.11
100	further validated the top K sequences. However,	there is no ground truth that can be	IE3.4
109	(e.g., shoulder) considering its location. If	there is no body part node between two	EE3.13
115	porosity that extends ~400 nm. Importantly,	there was no hint of porosity along the	ME1.3
126	To the best of the authors' knowledge,	there is no current literature contribution referring	ME2.8
127	Unfortunately, to our knowledge,	there are no alternate models for anchor damping	ME4.8
130	systems. To the best of our knowledge,	there is no relevant results on containment control	AE2.13
137	feasible regions after lunar flyby are narrow.	There is no obvious difference among the three	AE3.3
138	of the quantum Hall effect5. Nevertheless,	there is no practical means yet discovered to	EE1.2
143	union of follicular and non-follicular ones.	There is no general rule to separate the	EE3.17
145	confirming that from a continuum perspective	there is no limitation on although in reality	AE1.14
157	, each particle is viewed individually. In short,	there is no interaction between the molecules; it	AE4.6
165	patients, which are shown in Table 9. Since	there are no other state of the art	IE3.15
175	depicted in Fig. 9. It was apparent that	there were no small-sized powder particles	IE2.16
176	impact energy of 68.1 ± 1.3 J. It appears that	there was no clear trend for the effects	AE1.2
179	the bending process, it is assumed that	there is no shear deformation between the	IE2.4
180	the second is endothermic. Assuming that	there are no catalytic reactions, simulations with	AE4.7
	Additionally, marginal residual plots confirm that	there are no residual patterns. Table 2 provides a	
195	Moreover, although CSSAG ensures that	there are no inconsistent labels in its binary	EE4.1
196	simplicity, it is assumed in Fig. 3 that	there is no tool parallel axis offset and	IE2.17
199	Earth to LPO, it is found that	there is no abnormal region in the portion	AE3.3
202	Finally, our research findings imply that	there is no known reason to believe that	IE1.4
208	used for this task. We note that	there is no agreed definition of a disease	IE3.1
212	ending up with 1. It turned out that	there was no change in sensitivity and specificity	EE3.17
216	from the first image. This shows that	there is no shift invariance unless the shift	EE4.11
221	the case study hospital. They stated that	there was no reason to think that surgeons	IE1.6
223	and the spray area, which suggests that	there is no change in the state of	AE1.16
224	change. While many economists argue that that	there was no single world crude oil price	ME2.12
229	the others belong to relation type. Thus,	there are no n-gram features needed to	IE3.10
232	is met. With algorithms of this type,	there is no guaranteed convergence rate, and so	AE3.10
234	is limited to a single camera view,	there is no possible way to augment the	EE4.7
235	with higher scales, i.e., larger wavelets.	There is no need to increase their size	EE3.17
244	since statistical models can be applicable where	there are no strong edges. To be specific,	EE3.8
253	factors like vehicle physics into account. While	there is no dedicated measurement for	EE4.7

V. Summary for *there BE no*



Pieces of the Sentence

(1) transitions to indicate the connection

(2) a verb phrase indicating who gives us the information. This is almost always followed by *that*. An example of a verb phrase: "*our findings imply* that there is no reason to believe"

there BE no

(3) adjectives that describe the noun

(4) Noun

Transitions

Transition words connected this phrase to other sentences in the following ways:

mark importance	*significantly, importantly, in fact*
link	*so, hence, thus, consequently, this means*
contrast	*nevertheless, although, except, in contrast*
summarize	*in short, in general, currently*
introduce a reason	*as, since, because, if, when, considering*

Verb Phrases Before the Phrase

Examples of full verb phrases from the corpus include: *as mentioned above, the above observations reveal, the results show, we conclude, we assumed, control experiments confirmed, the process ensures, our findings imply, we find, it can be seen, we believe, our results suggest, they state, it indicates, many economists argue,*

These verb phrases tell the audience on what authority you make the claim: either someone you are citing, or the work done in your paper.

Verbs that Strengthen/Hedge

Ranked from most certain (strengthening) to least certain (hedging):

- declare, confirm, obtain, conclude, show, ensure
- find, see, state, appear, notice, mention
- believe, feel
- imply, indicate, suggest

If you are going to disagree with or question a finding, use a hedging verb prepares your reader for your counterargument.

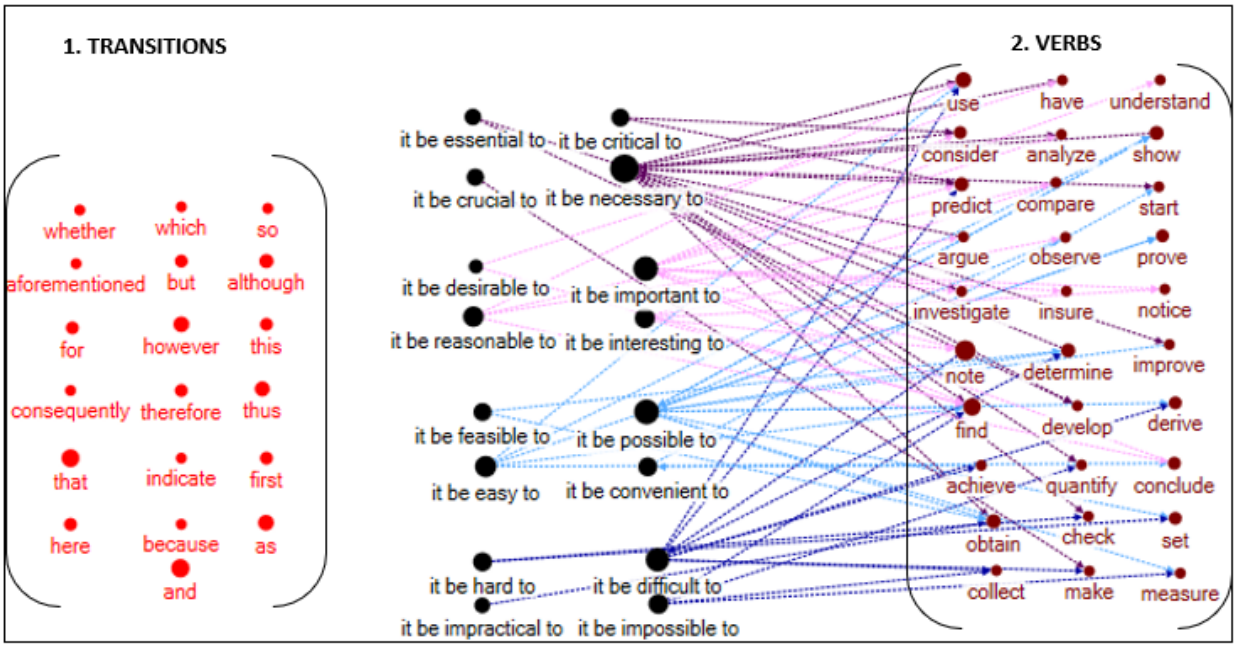
Adjectives that Strengthen/Hedge

Adjectives are generally used here to hedge claims. For example, saying “there is no perfect method” is a smaller claim than “there is no method.” And “there is no significant difference,” is a smaller claim than “there is no difference.”

However, they can occasionally be used to strengthen a claim by using “possible,” (“there is no possible entrance,” is a stronger claim than “there is no entrance.”)

VI. Worksheet for *it BE * to*

Variable Phrase 1: "it BE * to"



Exercises

1. Adjectives in this phrase not included above that occurred 2+ times: *advantageous, applicable, appropriate, beneficial, challenging, common, helpful, imperative, inappropriate, indispensable, infeasible, instructive, meaningful, meaningless, noteworthy, optimal, plausible, proven, prudent, rare, remarkable, robust, safe, sensible, simple, straightforward, suboptimal, tedious, trivial, unavoidable, unlikely, unnecessary, unreasonable, unwise, useful, vital, worthwhile*

    a) Which of these adjectives can be used as synonyms for the adjectives in the four groups?
    b) Which are used for separate purposes?

2. Rewrite five of the sentences from the Corpus Examples with different adjectives. How does the new adjective change the author stance?

3. *It BE * to* has three rhetorical uses. Identify a few sentences in the Corpus Examples where you think the author was employing each of those uses.

4. Write one sentence for each of the three rhetorical uses of *it BE * to*.

5. Write six sentences on your research using some of the less-common adjectives in the "It is X to" format. Can you employ each of the three uses above?

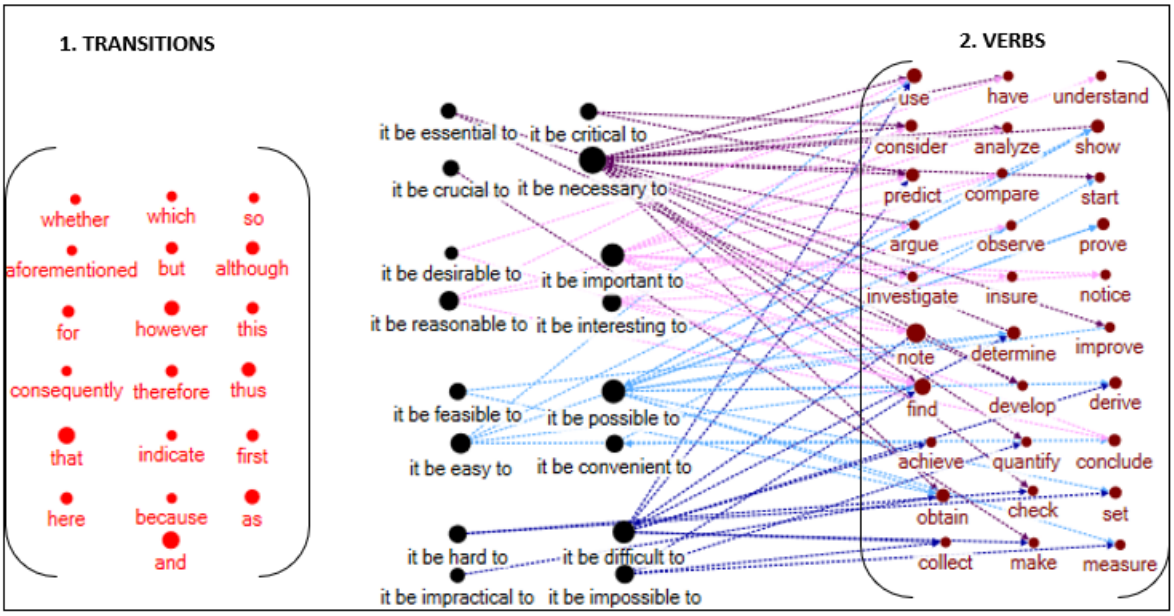
Corpus Examples

#			TXT
10	increase by 500 times after impact is engaged.	It is adequate to get an intuitive insight	ME4.16
11	first-stage regression is sufficiently high and	it is advantageous to R-square because the	IE1.17
12	rate, and so as a practical matter,	it is advisable to use an early-termination	AE3.10
26	scribing each individual engine module in detail,	it is appropriate to summarize the general gas	AE4.14
33	Before stating the problem formulation,	it is beneficial to note that J is	EE2.11
40	want to explore the question of whether	it is better to use the CNN's	EE4.15
41	information about the illuminant. In general,	it is challenging to automatically tune p for	EE3.3
51	in the same way colour constancy does,	it is common to impose several assumptions	EE3.3
60	surface of section crossing. To achieve this,	it is convenient to change the independent variabl	AE3.15
67	in dealing with more complex tasks where	it is critical to quickly review relevant information	IE1.1
68	on particular aspects of the problem. However,	it is crucial to make such assumptions explicit (	EE4.17
78	and key features for their future design.	It is desirable to have some degree of	ME1.16
114	for the generalization error of this problem,	it is difficult to employ traditional strategies since	EE3.10
126	several important closed systems from which	it is difficult to extract information or interface	EE1.1
154	for linear combination of two PSD matrices,	it is easy to conclude that L is	EE4.1
169	relate to organizational environments; thus,	it is easy to see why they might	IE1.15
178	the process in milling thin-walled parts,	it is essential to develop accurate chatter stability	IE2.9
180	understanding of their characteristics. As such,	it is essential to establish an effective and	ME2.6
185	, with such a small number of iterations,	it is feasible to determine an optimal S1	AE3.17
189	learning simulations have also indicated that	it is feasible to routinely collect labeled data	IE3.7
196	within the allowable range [-0.02,0.02]. However,	it is hard to achieve this case by	AE2.4
197	what is considered a good eye pattern,	it was hard to set standards for a	IE3.11
202	calculated one. different. Thus, we believe that	it is helpful to predict the plastic deformation	ME4.3
206	, their performance varies in different contexts.	It is imperative to understand which model	IE3.12
208	To ensure data integrity in clinical research,	it is imperative to introduce a "gold standard	IE4.3
212	the plume that it generates, is air.	It is important to note that this assumption	AE4.14
223	car penetration is increasing exponentially.	It is important to optimize the power flow	ME2.16
227	this mechanism to be consequential. Finally,	it is important to emphasize that the comments	IE1.15
230	but also on pressure [4,5]. Going even further,	it is important to remark the contribution of	AE4.15
233	on the final CID relation obtained. However,	it is important to notice that when compared	IE3.14
242	t-triggered consensus control protocols for MASs,	it is important to investigate the methods for	AE2.12
260	institutions (Henrich et al., 2005). As such,	it is important to understand the influence of	IE1.10
266	the work of Guo et al.9,10 Therefore,	it is important to consider the effect of	AE2.15
272	sequel, from the technical point of view,	it is important to make the following assumption.	AE2.6
279	research organizations and kept in-house.49 As	it is impossible to collect data about how	IE4.3
284	ever, in the spacecraft attitude control problem,	it is impossible to design a GS controller	AE4.13
286	model (see Fig. 2). At the same time,	it is impossible to detect occlusions without strong	EE4.16
291	Due to the complexity of damping mechanism,	it is impractical to obtain the damping ratio	IE2.9
292	geometric parameters on the impact properties.	It is inappropriate to compare the results of	AE1.2
294	the mishaps of YF-22 fighter aircraft. Thus,	it is indispensable to take into consideration both	AE2.13
296	data comes like a stream and thus	it is infeasible to keep the whole data	EE3.10
297	to contribute to the averaged carrier scattering.	It is instructive to compare the optical and	ME1.10
301	improved the perforation resistance effectively.	It is interesting to note that the effects	AE1.2
309	When the regularization part is skipped,	it is interesting to compare our method with	EE2.6
310	and Heemels (2017), and the references therein.	It is interesting to observe that none of	EE2.3
311	increasing dimensionality of the problem and	it is intractable to determine the explicit control	AE2.1
316	cutting edge trajectory of the ideal case.	It is likely to think of surface profile	IE2.2
321	was more accurate overall. This indicates that	it is meaningful to optimize the parameter of	ME2.6
323	current data in the training set but	it is meaningless to use historical data to	IE3.12

332	time, in the language of functional analysis [33],	it is necessary to minimize the distance between	IE2.7
341	fabricated parts generates. Consequently,	it is necessary to consider the combined effects	IE2.16
369	the family of fixed points, and thus	it is necessary to assess its accuracy. To	AE3.15
375	of five DGs installation. capacitor (1580 kVAr).	It is noteworthy to mention that the results	ME2.9
381	operations that include crowdsourced processes.	It is plausible to infer that ethnicity biases	IE1.2
383	for the panel of countries. In addition,	it is possible to conclude which countries in	ME2.12
394	-consumer sales data. In all three cases,	it was possible to identify the most significant	IE1.16
395	partial differential equation. For some cases,	it is possible to find a solution to	EE2.11
417	different flight conditions. In the literature,	it is possible to find a number of	AE4.9
456	a lower Q factor. In this way,	it is possible to obtain directional lasing at	EE1.12
465	Durbin and Koopman (2012, §4.5). Once again,	it is prudent to use square-root implementations	EE2.2
466	free of charge through its online Dataport.	It is rare to have such disaggregated, temporally	ME2.4
468	and thrust force deriving from moving blade.	It was reasonable to conclude that the size	IE2.16
476	of response to a mega disaster. Therefore,	it is reasonable to expect that IHOs with	IE1.12
485	Despite the subsampling process for STERGM,	it is remarkable to note that the two	IE1.13
505	the lower-resolution shape models are formed,	it is simple to match the high-resolution	AE3.6
506	[30], if A[— and A[-+1] are well conditioned,	it is stable to use the Woodbury formula.	EE3.10
520	level rather than pairwise comparisons. Besides,	it is tedious to build a pairwise constraint	EE3.6
530	in meteoroid entry trajectory simulation codes,	it is typical to approximate the meteoroid	AE4.12
531	index of the next mode b. Thus,	it is unavoidable to utilize the probability	EE2.12
532	also be used for semi-supervised classification,	it is unfair to compare the clustering results	EE3.6
533	less than the required L/D). Although	it is unlikely to be the case for	AE4.11
534	simplistic linear model will be sufficient and	it is unnecessary to use more complex models.	IE3.12
536	metallic framework lattice. As an aside, observe	it is unreasonable to expect an algorithm to	EE4.2
548	above to varying degree. In this light,	it is useful to consider strictly decreasing energy	EE2.8
550	in all publications. Clinical Relevance Statement:	It is vital to ensure the scientific rigor	IE4.3
559	based image retrieval results in Figs. 12 and 13.	It is worthwhile to note that parent classes	EE4.1

VII. Summary for *it BE *to*

"it BE * to" Summary



Order

(1) Transition

it BE (adjective) to

(2) Verb

The Dummy "It"

Here, the *it* doesn't refer to anything specific; it is a holding structure for an adjective of opinion. This is useful in avoiding personal pronouns and in expressing *author stance.*

Uses

(1) Introduce a topic, where "It is important/interesting/etc. to do X" means, "I will do X in this paragraph"
(2) Draw attention to a claim or point (with an adjective of importance)
(3) Defend a methodological choice (with an adjective of importance or possibility)

Adjectives Types

Adjectives of **importance,** from most to least possible:

- essential, critical, crucial, imperative, indispensable, vital
- necessary, optimal
- important, worthwhile
- useful, advantageous, beneficial, helpful
- applicable, desirable, interesting, reasonable, meaningful, noteworthy
- appropriate, instructive, common
- meaningless, tedious, trivial, unnecessary, unwise

Adjectives of **possibility,** from most to least possible:

- easy, convenient, simple, straightforward
- feasible, possible, plausible
- hard, difficult, challenging
- suboptimal, impractical, inappropriate, infeasible, implausible, unlikely, unreasonable
- impossible

Other adjectives in this slot:

- proven, prudent, rare, remarkable, robust, safe, sensible, unavoidable

Bonus information

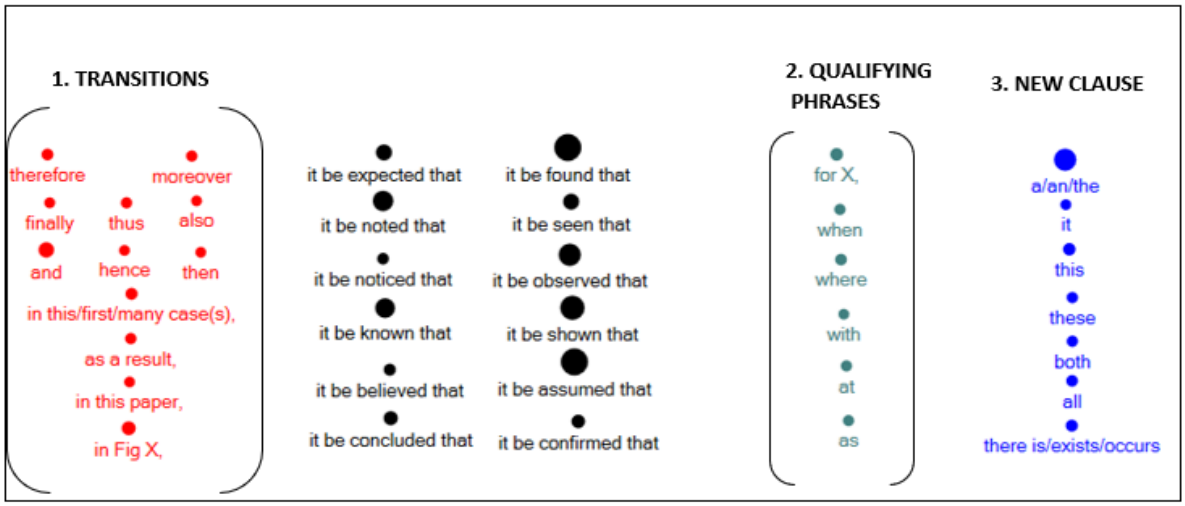
Place in Sentence & Prepositional Phrases

Of the 561 instances of this idiom in the corpus, 22% start a sentence (which gives additional emphasis), but 47% come after a prepositional phrase or transition with a comma. The transitions can be the words in Category (1) in the graph, or longer phrases. The prepositional phrases are not shown in the graph above, as they have a lot of variety. These phrases before the comma serve three main purposes:

- transition "however,: "thus," "going even further," "in light of this information"
- provide context: "in the literature," "in the simulation codes," "to minimize damage"
- hedge or limit the claim: "for some cases," "with this configuration," "if A is well conditioned,"

VIII. Worksheet for *it BE * that*

Variable Phrase 2: "it BE * that"



1. In the Corpus Examples below, you see many phrases that come after *it BE * that* that are highlighted in light blue. What purposes do these phrases serve?

2. Write two sentences with a qualifying phrase after the *it BE * that* phrase

3. Looking through the Corpus Examples, find example sentences for each of the three rhetorical uses.

4. In addition to the verbs above, less-common verbs in the variable slot were: *anticipated, claimed, considered, demonstrated, determined, discovered, discussed, emphasized, estimated, expected, explained, felt, guessed, highlighted, hypothesized, illustrated, inferred, known, postulated, proposed, realized, recognized, reported, revealed, said*
   Which verbs can you use for each of the rhetorical uses?

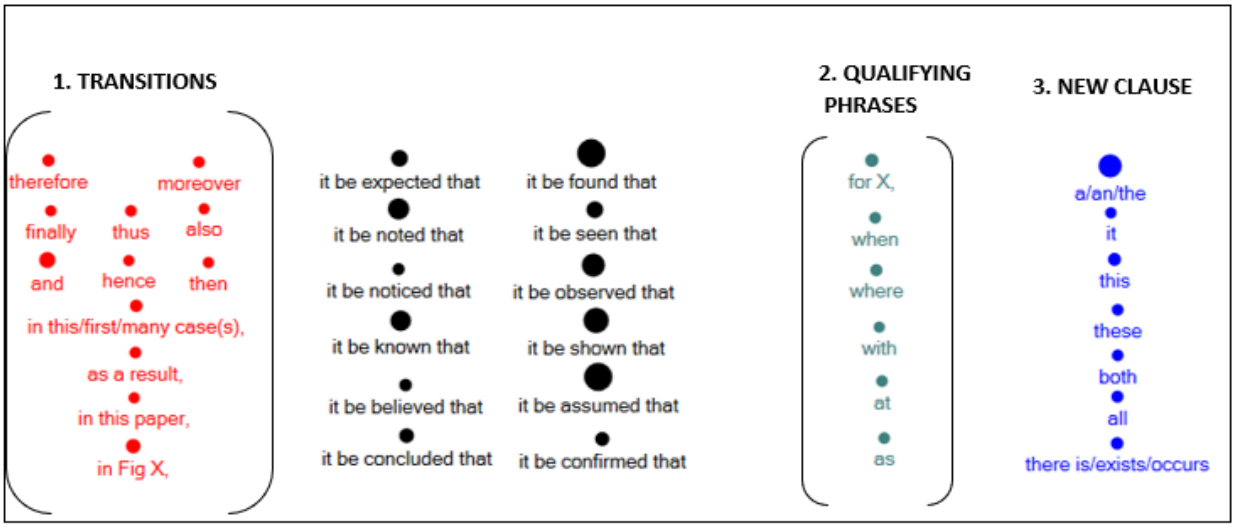
5. This phrase does not make it clear *who* knows. Read through the corpus examples and identify who does the knowing and how you as a reader can guess that.

Corpus Examples

#			TXT
3	to no information about the wind environment.	It is anticipated that the algorithm developed in	AE4.5
51	additive noise is considered in this paper.	It is assumed that anyone, including an adversary,	EE2.11
56	system by the input and output valves.	It is assumed that, when valve V1 is	AE2.1
69	wave with such a high amplitude. Thus	it is assumed that Eq. (10) holds when difference	AE1.5
81	and without the separation bubble. Moreover,	it is believed that other likely sources of	AE4.14
82	results are relatively big. Analyzing the reasons,	it is believed that when the central angle	AE1.3
85	method of de Caigny et al,16 where	it is claimed that a parameter-independent one	AE2.6
105	he CDP and classical metal-plasticity model.	It was concluded that for a dynamic numerical	AE1.10
148	theory is reviewed briefly in Section 2, where	it is confirmed that similitude is possible for	AE1.14
150	the interaction between SPH particles. Whereas,	it is considered that the minor axis of	AE1.7
156	shown in Fig. 6a and b, respectively.	It is demonstrated that among samples studied,	ME2.11
162	monetary deductions from a base payout if	it is determined that contract specifications are	IE1.10
164	of the tube used in the analysis.	It was discovered that for a constant value	AE1.11
165	compared with covalently coupled antibodies.	It is discussed that surface adsorption of proteins	EE1.5
166	size with unity aspect ratio be considered.	It was emphasized that the damage and strain	AE1.10
167	the work of Carrion and Arroyo. Nevertheless,	it is emphasized that the number of integers	EE2.11
196	See Appendix C.	It is expected that, by increasing w- w	AE3.9
198	behave in perforated armour constructions.	It was expected that higher mechanical properties,	AE1.15
207	In both the oxidation and reduction process,	it is expected that mass transport is dominated	ME1.14
211	second sphere is discussed. In Sec. V,	it was explained that the inflow conditions for	AE4.12
213	here. I think that would be helpful."	It was felt that this human interaction could	IE4.4
214	of target plates at 600 °C and 25 °C,	it is found that at 600 °C, the deformation	AE1.17
248	al simulation in our previous study [7]. Finally,	it was found that the failure criterion with	AE1.2
270	Fig. 2. As discussed in the later sections,	it was found that the optimum frequency	ME3.15
281	for feasibility of Problem 6. By the assumption,	it is guaranteed that the equality constraints do	EE2.13
282	fibers segments were flying from the CFRPs.	It was guessed that two possible reasons caused	IE2.1
284	) denotes Laplace Ei\S) transform of ei.	It is highlighted that, according to (12), Ei(s) (	AE2.15
285	vehicle and not just few operative points.	It is hoped that the model described here,	AE4.14
288	in mechanical stability is not well understood,	it is hypothesized that the Al2O3 coats	ME4.4
298	database of US Food and Drug Administration.	İt is known that the main limitations of	IE3.4
329	intensity-inverting data transformation. While	it is known [7] that NC is equivalent to	EE3.15
358	guidance response [37,38]. Specifically, in [37]	it was noted that the vehicle response during	AE4.5
380	of mass near the model extremes. Finally,	it is noted that, although the present experiments	ME2.12
393	In addition, from Figs. 5c and 5d,	It is noticeable that, with the proposed law,	AE3.7
399	analyzing the existing numerical models for LMD,	it is noticed that there is no complete	IE2.10
427	0 pm) at different positions in the microchannel.	It is observed that the length needed to	ME3.15
483	20 deg. As discussed in Sec. II.C,	it is preferred that the science orbit has	AE4.11
484	operations using a digital computer. Moreover,	it is projected that the power consumption of	ME4.6
486	major work of this paper. In Refs. [36,37],	it is proposed that general nonlinear dynamics can	AE3.14
490	those produced by the CFD analysis, once	it is realized that the CFD results are	AE4.14
492	to induce undesirable chatter phenomenon.	It is recognized that chatter greatly degrades the	IE2.9
493	external aerodynamics is not considered;	it is recommended that this be added in	AE4.14
495	the IMF affects the thrust characteristics [6].	It was reported that the IMF enhances the	AE4.10
500	To cast these problems into tractable ones,	it is required that the scheduling parameters be	AE2.6
505	the dimensional scaling one depicted in Fig. 7,	it is revealed that the error percentage of 18.53%	AE1.14
513	verified in this study. In 1DG case,	it is seen that by the lower size	ME2.9
565	orbit insertion at Titan. In this paper,	it is shown that different interplanetary arrival	AE4.11
576	o guarantee stability of the NCS. Moreover,	It was shown that for the standard SD,	EE2.3
591	blew away along the jetting direction. Thus,	it is thought that the jetting force serves	IE2.1
597	power factor on the distribution network.	It was verified that the proposed approach	ME2.9

IX. Summary for *it be * that*

Variable Phrase 2: "it BE * that"



Pieces of the Sentence

(1) Transition
(2) Qualifying Phrase that restricts the claim. e.g., "It is noted that, *with a higher-resolution instrument*, the differences may be more pronounced." This can also come before the phrase.

it BE (past participle verb) that

(3) Beginning of a new clause

The Dummy "It"

The *it* doesn't refer to anything specific. Instead, here it is a holding structure for a passive verb of knowing.

Rhetorical Uses

(1) to explain the author's expectations, assumptions, or thoughts without using personal pronouns
(2) to state general truths or background information
(3) to summarize and highlight certain research results

The "Knower" of the Phrase

Usually, the invisible subject is either the general scientific community, the authors, or someone who is cited in the sentence. The transition phrase or prepositional phrase before "it is X that," often indicates who the invisible subject is.

Position in Sentence

48% of the idioms were at the start of sentences and 36% followed commas. The phrase before *it BE * that* often gives information about who knows, and the phrase after *it BE * that* usually qualifies or contextualizes the claim.

Types of Verbs

Depending on the rhetorical use of the phrase, there are many different verbs you can use; all of them are different ways or types of knowing. Below, they have been categorized into families of similar words.

See	known, seen, found, observed, shown, recognized
Say	noted, noticed, said, explained, reported, emphasized, highlighted, claimed, discussed, reported, considered
Expect	anticipated, believed, expected, estimated, postulated, proposed, guessed, hypothesized
Confirm	realized, confirmed, concluded, demonstrated, determined, discovered, revealed
Other	inferred, illustrated, felt, assume

X. PowerPoint Slides

Pretest – 10 minutes

Understanding Common Academic Phrases

NOVEMBER 11 & 13, 2019

MARIA PRITCHETT - MCUPERY@PURDUE.EDU

Introduction

What We'll Do

- We will study three phrases commonly used in academic writing
- We will look for patterns in how they are used & the words around them
- We will practice writing sentences with those phrases

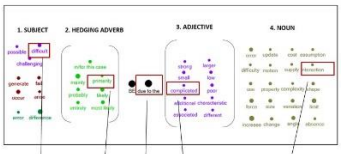
What Will I Learn?

- The **rhetorical uses** of the three phrases in academic writing
- What kind of words/information is used **around** the phrases
- How to **connect** these phrases with the other parts of the sentence

Where Does the Information Come From?

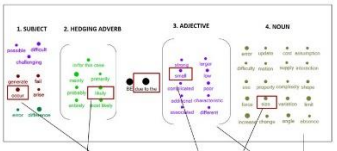
- All the information in the slides and handouts is based on my corpus (collection) of 262 published engineering articles, with 1.9 million words.
- This means that all the examples and information comes from real articles.

How to read the diagram



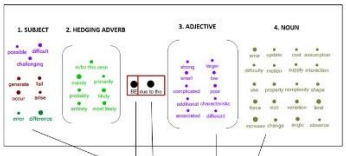
"The explanation is difficult, primarily **due to the** complicated interactions between A3 and C3."

How to read the diagram



"Fuel leakage likely occurs **due to the** small size and high diffusivity."

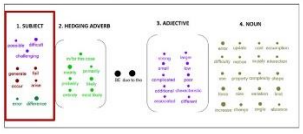
How to read the diagram



"The contribution of its harmonics is higher. This **is due to the** lower order of the pseudo moire effects."

Phrase 1: *due to the*

Parts of Speech	Definition	Example
Noun	thing, person, place, idea	inconsistencies, size
Adjective	a word that describes a noun	challenging, similar
Verb	an action	exhibits, rises



Noun, Adjective, or Verb?

In the following sentences (from your Corpus Examples), what things are the cause of the "due to the" phrase?

Is this subject an adjective, noun, or verb?

a direct comparison is difficult due to the absence of the average data
to be beneficial for reducing PM emissions due to the absence of aromatics, which are
section III. The small discrepancies may be due to the actual design parameter variations during
generator is usually difficult to scale up due to the additional requirement of flow infusion
The velocity and jerk violations are due to the approximation for linear and angular
can be seen, the impact time errors due to the autopilot lag are within an
it exhibits ferromagnetic response and rises due to the effect of the magnetic field
corresponds to Eq. (2), (3), and (6), respectively. Due to the binary constraint on Z, we see

Question 1

Phrase 1: *due to the*

Four primary options for the location:

Location of *due to the*	Example
1. after the subject	"The powder aggregates quickly **due to the** cohesive force."
2. before the subject	"**Due to the** cohesive force, the powder aggregates quickly."
3. linked to subject with a form of "BE"	"The quick aggregation of powder **is due to the** cohesive force."
4. summarized with "THIS" and linked with "BE"	"The powder aggregates quickly. **This is due to the** cohesive force."

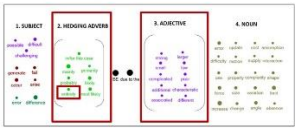
These options slowly increase the emphasis on *why* it happened, instead of *what* happened.

Question 2

Phrase 1: *due to the*

Strengthening language: language used to emphasize a point

Hedging language: language used to soften a point and avoid being overconfident



Strengthening & Hedging

Where do you see hedging adverbs or strengthening adjectives in these sentences?

"*Due to the* considerable complexities, theoretical developments in S-MJSs have advanced more slowly"

"*Due to the* critical significance of CIDs, resources such as the Comparative Toxicogenomic Database (CTD)1 are being manually curated."

"Generally, discharge phenomena cannot be directly observed *due to the* extremely bright plasma in the gap."

"They showed a larger deviation which might be *due to the* higher number of convolutions."

"the performance of different Lamb wave modes varies largely *due to the* dispersive characteristics."

"Unfortunately in most cases *due to the* complexity of the process, there can be significantly more than 3 unknowns."

Question 3 & 4

Phrase 1: *due to the* Practice

Question 5

Phrase 2: *it BE * to*

"Dummy it"

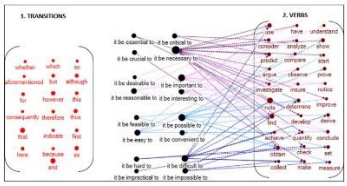
"Dummy it" describes sentences like the ones below, where there "it" doesn't refer to a specific noun:

"It is important to water your plants."

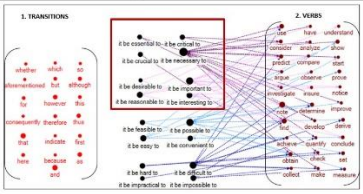
"It is easy to understand this principle."

In academic writing, *it BE *adjective* to"* is an easy way to express your opinion while sounding objective.

When an author expresses their opinion in their writing, this is called **author stance.**



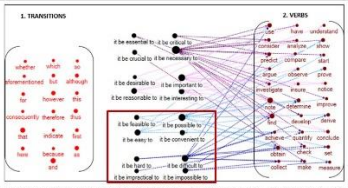
Adjectives of importance



"However, **it is crucial to** make such assumptions explicit , which much prior work does not. We do so below."

"**It is interesting to** notice that the IDT effectively excites the S modes."

Adjectives of possibility

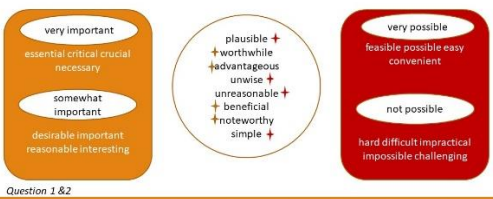


"Since there has been no rigorous standard of measuring a good eye pattern, **it was hard to** set standards."

"Thus, **it is convenient to** change the independent variable from t into t = f/Td."

Possibility or Importance Adjectives?

What category do these less-common adjectives belong to?



Question 1 &2

Phrase 2: *it BE * to*

Three most common rhetorical uses:

Use	Example
1. To introduce a topic, where "It is important/interesting/etc. to X" means "I will X"	**"It is important** to note that the cost functions presented here are designed separately." [followed by a discussion of why]
2. Draw attention to a claim or point	"Since this demands decarbonization, **it is imperative** to minimize the use of fossil fuels."
3. Defend a methodological choice.	"In the latent disturbances formulation of EM, **it is convenient to** work with the more general parametrization."

Phrase 2: *it BE * to*

3 Rhetorical Moves
1. To introduce a topic, where "It is *important* to X" means "I will X"
2. Draw attention to a claim or point
3. Defend a methodological choice.

1. "Although all the other gas dynamic models are well described, it is appropriate to give more detail regarding the balance equation solver"
2. "In conclusion, it is vital to highlight the scientific rigor of clinical trials."
1. "In this subsection, we will show it is important to consider the effect of nonzero initial spacing, velocity, and acceleration errors."
3. "It is straightforward to directly measure these eigenvectors by recording the amplitudes of both resonators."
2. "Despite the intuitive nature of such findings, it is interesting to note that even large LMs can be transported at such high speeds."
3. "Besides, it is tedious to build a matrix with only 100 instances."

Phrase 2: *it BE * to*

Three most common rhetorical uses:

Use	Example
1. To introduce a topic, where "It is important/interesting/etc. to X" means "I will X"	**"It is important** to note that the cost functions presented here are designed separately." [followed by a discussion of why]
2. Draw attention to a claim or point	"Since this demands decarbonization, **it is imperative** to minimize the use of fossil fuels."
3. Defend a methodological choice.	"In the latent disturbances formulation of EM, **it is convenient to** work with the more general parametrization."

Questions 3 & 4

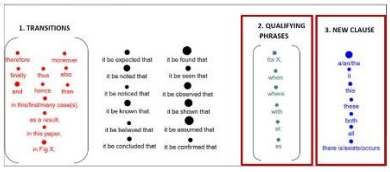
Phrase 2: *it BE * to*
Practice

Question 5

Phrase 3: *it BE * that*

Important Terms

- Dummy it (again)
- What do these verbs have in common?
- Qualifying phrases *restrict the claim that follows*



Qualifying Phrase & New Clause Practice

Where is the qualifying phrase in these sentences? Where does the new clause start?

1. It is assumed that if agents engage the same target, their effectiveness values combine as independent probabilities.
2. For HA1c, it was found that 22% of orders were made prior to the first shelf-life.
3. It is assumed that, when valve V1 is open, the liquid flows into the container at a fixed rate.
4. For any initial conditions in the zonal problem, it is observed that the resulting Td and AUd exhibit periodic-oscillation behaviors.
5. It is confirmed that the local indentation was more severe and longer-lasting.
6. It is expected that, by increasing the width, it becomes easier to preserve the privacy of the server.

Question 1 & 2

Phrase 3: *it be * that*

Three main rhetorical moves:

Use	Example
1. To explain the author's expectations, assumptions or thoughts impersonally.	"Hence, it was inferred that the variation of particles significantly influenced the packing state."
2. To state general truths or background information.	"It is assumed that the aerodynamic model produces outputs in some known coordinate frame."
3. To summarize and highlight specific results.	"It was observed that the car cabin temperature went up to 57 °C."

Rhetorical Move Practice

Uses
1. To explain the author's expectations & assumptions impersonally.
2. To state general truths or background information.
3. To summarize and highlight specific results.

1. 1. This new approach is developed for planetary entry, and it is anticipated that the method can be applied to other applications
2. 2. it is reported that the effective conversion rate is only half of the equilibrium rate.
3. 3. Different values of peak g-load were identified, and it was shown that peak g-load is not the limiting factor for aerocapture at Titan.
1. 4. Although it is postulated that transition would not occur, there is increased risk of transition
3. 5. The ROC curve is presented in Fig. 5C, and it is observed that the curve is very close to the upper left corner.
2. 6. It is known that the optimal solution to Equation (1) satisfies (see Zipkin, 2000).

Question 3

Verb Options

See	seen, found, observed, shown, recognized
Say	noted, noticed, said, explained, reported, emphasized, highlighted, claimed, discussed, reported, considered
Expect	anticipated, believed, expected, estimated, postulated, proposed, guessed, hypothesized
Confirm	realized, confirmed, concluded, demonstrated, determined, discovered, revealed
Other	inferred, illustrated, felt, assumed, known

Question 4

Phrase 3: *it be * that*

Who does the action behind these passive verbs?

"Also, from Fig. 12, *it is confirmed that* higher the voltage applied to the actuator, higher is the flow rate for all geometric parameters." (the authors/the data)

"*It is found that* EA of sandwich panels could be improved by increasing the face-sheet thickness." (the authors)

"*It is known* that the presence of PMFs will modify the band structure." (general scientific community)

"Importantly, several studies have shown similar discrepancies… In these studies, *it was concluded* that there must be an element of mechanical separation." (a cited source)

Question 5

Post Test (10 minutes)

Phrase 3: *it BE * that*
Practice

Question 6

Feedback

Thank you!

You can also email me at mcupery@purdue.edu with comments & feedback.

APPENDIX F. CLASS SYLLABUS

Course Schedule for Fall 2019, Session 1

Unit 1: Managing Author Responsibilities

Day 1	Introduction to Course and Course Website Writing Diagnostic and Discussion Working Groups
Day 2	People, Process, and Product: Who you are as a writer, why you write, where you write, when you write, how you write, and what you write.

Unit 2: Managing Audience expectations

Day 3	Rhetorical Situations and Genre;
Day 4	Academic Genres
Day 5	Moves and Functions in Academic Writing
Day 6	Moves and Functions in Academic Writing (continued)

Unit 3: Managing Language choices

Day 7	Sentence Structures
Day 8	Sentence Structures (continued)
Day 9	Common academic phrases (fixed structures)*
Day 10	Common academic phrases (variable structures)*

Unit 4: Managing Information and Ideas

Day 11	FoOD Principle: Focus, Organization, Development
Day 12	Paragraphs and Coherence

* On Day 9 and 10, Maria Pritchett will be guest teaching on common academic phrases. Her research is on teaching academic phrases, and if you accept, she will use data from the in-class pre-test and post-test for her dissertation. If you do not wish to participate, you do not need to take the pre- and post-test.

APPENDIX G. EVALUATION MATERIALS

The contents of Appendix G are the pre-, post-, and delayed posttests used in class, as well as the feedback form that the students filled out at the end of the second class:

I. Pretest

II. Posttest

III. Delayed Posttest

IV. Student Evaluation Form

I. Pretest

Pretest

Read the quote and then answer the question(s) about it. Don't worry about understanding the quote.

"Usually, in the laboratory conditions, it is difficult to meet the requirement of the standard regarding the shooting distance (30 m)."

1. What word can you substitute for *difficult* to mean not possible?

_____

2. What word can you substitute for *difficult* that means very possible?

_____

"Due to the minute contact zone for heat transfer, measurement of these parameters in metal cutting is very difficult."

3. Rewrite this sentence so that there is a form of the verb "is" before the highlighted clause. How does this change the emphasis of the sentence?

_____

_____

4. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

"Given these advantages, it is expected that the room-temperature activated graphite felt is a promising electrode."

5. Who *expects* something in this sentence?

_____

6. What is a more certain word to substitute for *expected*?

_____

"It is known that the considerable Van der Waals force… causes the generation of the cavity defects during powder paving."

7. What is a less certain word to substitute for *known*?

_____

8. Who *knows* something in this sentence?

_____

9. What word would you substitute for *known* if "Van der Waals forces causes the generation of the cavity defects" was an unproved hypothesis?

_____

"EHR technology can assist healthcare providers in dealing with more complex tasks where it is critical to quickly review relevant information."

10. What word can you substitute for *critical* that also means very important?

_____

11. What word can you substitute for *critical* that means not important?

_____

"The spacecraft can track a desired trajectory without thrust control. This is due to the fact that the error… is of the same order of magnitude as the amount of force."

12. Rewrite this sentence so that the highlighted clause follows its subject without interruption. How does this change the emphasis of the sentence?

_____

_____

13. What word could come before the highlighted clause to strengthen the author's claim?

_____

"His task might be challenging due to the differences in the use of medical terminology between health care professionals and social media texts."

14. What word could the author use in place of "might be" to hedge the claim?

_____

15. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

"It is highlighted that, according to (12), the results in (A) includes the effects of initial errors in (B),"

16. What is another word you can use instead of "highlighted" to report a result?

_____

"It is challenging to measure the pipe's width precisely, so the measurements are rounded to one decimal place."

17. Is the author using the highlighted phrase to:
a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

"Before describing each individual engine module in detail, it is appropriate to summarize the general gas dynamic models that are used to model the flow."

18. Is the author using the highlighted phrase to:
a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

II. Posttest

___ I was in class Sept 23 & 25

___ I was only in class today, Sep 25

Posttest

Read the quote and then answer the question(s) about it. Don't worry about understanding the quote.

"At this stage we can say that it is possible to reconstruct a color image based on a point-wise description of its contours."

1. What word can you substitute for *possible* that means more possible?

_____

2. What word can you substitute for *possible* to mean not possible?

_____

"The power handling ability becomes worse due to the extremely small sizes."

3. What word could the author add before the highlighted phrase to hedge the claim?

_____

4. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

"In many cases it was observed that the core of the projectile was completely damaged."

5. Who *observes* something in this sentence?

_____

6. What is another word you can use instead of "observed" to mean a result was observed?

_____

"More importantly, room-temperature activation methods cannot compete with high-temperature activation methods, primarily due to the poor performance of the fabricated electrodes."

7. What word could replace "primarily" to strengthen the author's claim?

_____

8. Rewrite this sentence so that the highlighted phrase with 'due to the' comes before what it caused. How does this change the emphasis of the sentence?

_____

"Eventually, after numerous investigations, including those completed by the National Aeronautics and Space Administration (NASA), it was determined that no significant issue existed."

9. Who *determines* something in this sentence?

_____

10. What word would you substitute for *determined* if "no significant issues existed" was a hypothesis rather than a result?

_____

"It is important to mention that the nonlinear function f is more general."

11. Is the author using the highlighted phrase to:

a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

"It is unreasonable to expect an algorithm to reliably compute the horizontal dominant vanishing point for this patch based on detecting straight lines."

12. Is the author using the highlighted phrase to:

a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

"It is claimed that, according to these figures, the local response of the full-scale model is also predicted with a reasonable accuracy"

1. What is a more certain word you could substitute for *claimed*?

_____

2. What is a less certain word you could substitute for *claimed?*

_____

"It is vital to take into consideration specific aspects of a mobile device when selecting evaluation methods."

3. What word can you substitute for *vital* that also means very important?

_____

4. What word can you substitute for *vital* that means not important?

_____

"Due to the large size data set and high feature dimensions at each layer, 200 samples are selected for each fault type."

5. Rewrite this sentence so that there is a form of the verb "is" before the highlighted clause. How does this change the emphasis of the sentence?

_____

6. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

___ I was in class Sept 23 & 25
___ I was only in class Sept 23
___ I was not in class Sept 23 or 25

III. Delayed Posttest

Delayed Posttest

Read the quote and then answer the question(s) about it. Don't worry about understanding the quote.

"In practice, it is sensible to assume that events can be grouped in time."

1. What word can you substitute for *sensible* to say that is a good idea?

_____

2. What word can you substitute for *sensible* to say that it is not a good idea?

_____

"It is found that the honeycomb core failed due to the buckling and folding of cell walls around the impact region."

3. What word could the author add before the highlighted phrase to hedge the claim?

_____

4. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

"It is claimed that, according to these figures, the local response of the full-scale model is also predicted with a reasonable accuracy"

5. Who *claims* something in this sentence?

_____

6. What is another word you can use instead of *claimed* to mean something was claimed?

_____

"Any observed difference in the time trends after the intervention is most likely due to the treatment."

7. What word could replace “mostly likely” to hedge the author’s claim?

_____

8. Rewrite this sentence so that the highlighted phrase comes before what it caused. How does this change the emphasis of the sentence?

_____

“In this study, it is confirmed that silicon can be also be shear after ion modification.”

9. Who *confirms* something in this sentence?

_____

10. What word would you substitute for *confirmed* if “silicon can also be shear after ion modification” was an unproved hypothesis?

_____

“Before stating the problem formulation, it is beneficial to note the properties of J.

11. Is the author using the highlighted phrase to:

a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

“It is critical to remember that the string equation considered here behaves like a communication channel with string dynamics.”

12. Is the author using the highlighted phrase to:

a) draw attention to a point
b) defend a methodological choice
c) Introduce the topic of a paragraph

“Based on this extended set of data, it is thought that the findings of Ref. [7] are further corroborated.”

13. What is a more certain word you could substitute for *thought*?

_____

14. What is a less certain word you could substitute for *thought*?

_____

"Our baseline uses 2D depth encodings. It is straightforward to contrast this with a "standard" 2D scanning-window template."

15. What word can you substitute for *straightforward* that also means very possible?

_____

16. What word can you substitute for *straightforward* that means less possible?

_____

"His task appears challenging due to the differences in the use of medical terminology between the two disciplines."

17. Rewrite this sentence so that there is a form of the verb "is" before the highlighted clause. How does this change the emphasis of the sentence?

_____

18. What word does the highlighted clause explain? Is it an adjective, noun or verb? Explain.

_____

IV. Student Evaluation Form

Feedback on Academic Phrases Module

1. **How useful were the pictures that showed the context of the phrases?**

Not Useful *Somewhat Useful* *Very Useful*

○ ○ ○ ○ ○

2. **How much of the information we covered was useful to you as a writer?**

Not Useful *Somewhat Useful* *Very Useful*

○ ○ ○ ○ ○

3. **How much of the information covered in class did you feel you understood?**

No information *Half the Information* *All the information*

○ ○ ○ ○ ○

4. **What was most helpful about the material?**

5. **What was most confusing or unhelpful about the material?**

6. **Other comments/questions?**