LINGUISTIC CHARACTERISTICS OF FIRST-YEAR UNIVERSITY WRITING: A CORPUS INVESTIGATION

by

Aleksandra (Sasha) Turtova

Thesis submitted to the Department of English Language and Literature at Mary Immaculate College for the degree of Doctor of Philosophy

Supervisors: Dr Brian Clancy and Dr Joan O'Sullivan

Submitted to Mary Immaculate College, March 2025

Abstract

This study examines the characteristics of first-year composition writing using a corpus-driven approach. The corpus analyzed, COMP 101, consists of 383 texts totaling 188,184 words. These texts include seven essay genres: descriptive, narrative, classification, process analysis, compare and contrast, cause and effect, and argumentative. Using a genre-based theoretical framework, the study explores the most common language features found in first-year writing in a typical university classroom of native and non-native English speakers and how these features are represented across the different genres. According to the corpus analysis, the most frequently used features in COMP 101 include first-person pronouns, second-person pronouns, conjunctions, and punctuation marks. The study compares these findings to academic writing represented by the British Corpus of Academic Written English (BAWE). The findings reveal how first-year students engage with readers and establish identities in their texts using first and second-person pronouns. Also, the study notes a tendency to avoid using past participles and an overall conversational tone, demonstrated by a preference for contractions, question marks, and exclamation marks. The implications of these findings for the teaching of writing in this context are then discussed.

Acknowledgments

I am forever grateful to my supervisors, Dr. Brian Clancy and Dr. Joan O'Sullivan, without whom this research would not be where it is today. They have continuously challenged and encouraged me throughout this journey, dedicating time and effort to reviewing numerous drafts—some far from perfect—and providing invaluable feedback. It is through their guidance and encouragement that I was able to navigate the complexities of the corpus and bring the pieces together.

I am forever grateful to Dr. Kathaleen Reid-Martinez, who has been instrumental in my academic journey by mentoring, challenging, and believing in me. Without her guidance, I would not be where I am today.

I am deeply grateful to my sons for their love and understanding as they patiently supported me in my studies throughout the years. A heartfelt thank you! I will always be grateful to my parents for their unwavering support and belief in my ability to succeed. Also, I want to express my sincere gratitude to my sister for her encouragement and support throughout this time.

I am incredibly grateful to my work supervisors, Dr. Evan Culp, Dr. Andrew Lang, and Dr. Boyd, for their understanding, encouragement, and support throughout this journey.

Finally, I would like to echo the words of the Apostle Paul: "My grace is sufficient for you, for my strength is made perfect in weakness" (2 Corinthians 12:9). I am infinitely grateful to the Lord Jesus Christ for His mercy and grace, which have sustained me through the many challenges of this journey.

Declaration of Originality

I declare that the following thesis has not been submitted as an exercise for a degree at this or any other university. It is entirely my work, except when other sources are used. These sources are indicated in the text as in-text citations and listed in the references of this document.

Signed:

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.0 Introduction

This study is an integral part of my career as an English teacher. I have taught English as a foreign language in Bulgaria and English composition classes in the United States. When comparing the challenges of teaching English as a foreign language with teaching English composition, I find that both involve similar levels of complexity in terms of teaching and learning strategies. In both cases, students are dealing with new rules, vocabulary, and communication systems and are either newly exposed or struggling to adapt. As some researchers observe, no one is native to academic English, and speaking English as a first language does not always mean being able to write in academic English (Viana and O'Boyle 2022). The university classroom comprises both native and non-native speakers of English with one common purpose: to gain certain skills and professionalism and eventually enter a career.

My research focuses on first-year university students and is motivated by the interest to examine the most common textual features that characterize this type of writing and how they differ from upper-level writing or writing in later stages at university. Knowing the writing competencies of students is important for gaining insight into their strengths and weaknesses, developing suitable strategies, and targeting the necessary skills they should acquire. In this regard, corpus linguistics is a logical choice because it provides the methodological framework for conducting such research. This chapter serves as an introduction to provide background information about the struggles universities face in addressing the needs of first-year students, the rationale for this study, and the research questions.

1.1 Background and context

Every year, over 15 million young people enroll in university education in the United States, but less than 60 percent of first-year students graduate within six years, which is two years longer than the standard four-year timeline, and despite this extended timeframe, the dropout rate has not decreased (Kirp 2019). This study focuses on first-year students who face challenges because they often lack the necessary skills to complete academic assignments. Consequently, many of them are unable to continue beyond their first year of studies. According to the American Talent Initiative (Strikwerda 2019), out of 3,000 four-year colleges in the United States, fewer than 300 manage to graduate at least 70 percent of their students within six years, which leads to a national average of 59 percent of students graduating within six years. If community colleges are considered separately, which is the largest sector in the U.S. education system, then this percentage drops even further to just 14 percent of students who are able to complete their degree within six years of starting. Therefore, increasing the number of students who successfully graduate is one of the most significant challenges faced by colleges and universities in the United States.

To help students graduate, college and university teachers must meet the needs of a diverse population of students from diverse ethnic, linguistic, and educational backgrounds (Hyland 2006). The increase in the refugee population and international migration has contributed to the changes in the classroom's homogeneous context, where both native and non-native English students face challenges with academic language. Biber (2006) discusses that the academic challenge is for “all students—whether native speakers of English or non-native speakers—need to adjust to a wide range of tasks in the university accomplished through language” (p.1). Hyland (2006) acknowledges the same need for training all students in academic English and confirms

the changing educational context and the growing “awareness that students, including native English speakers, have to take on new roles and engage with knowledge in new ways when they enter university” (p.2). The roles that students acquire at university require the ability to engage critically with academic texts and communicate effectively in the university environment.

This study will focus on the texts submitted in the Composition 101 (COMP 101) module at a university in the United States. The COMP 101 classroom consists of a diverse mix of international and domestic students, reflecting the varied population typically seen in entry-level composition classes, as noted by Hyland (2006). There are no special sections for international students; all students are assigned based on their schedules and majors. International students are defined as non-U.S. citizens or Green Card holders and must meet specific English proficiency requirements through tests like TOEFL, IELTS, ACT, or SAT. The class includes students from various countries around the globe. Chapter 3 provides detailed information on the participants' demographics and the English proficiency requirements. Regarding majors, students frequently change their majors, and only a few tend to keep their chosen majors after the first year of their studies. Thus, the class accommodates both native and non-native students from a wide range of disciplines, from Science to Humanities.

1.2 Rationale for the study

In the words of Hyland (2005a), “effective teaching and learning crucially depend on understanding how language works and using this understanding to help students communicate appropriately and successfully in their communities” (Hyland 2005a, p.3734). This implies that by examining students’ discourse in school genres, researchers are able to discover the language features that characterize student writing, compare these findings with what is considered

successful practice, and gain knowledge and insight on how to scaffold the different genre assignments to incorporate the language features that mark successful student writing. For example, Aull (2020) examines the differences between the different levels of writing at university, observing that the writing of first-year students differs from that of upper-level students in the degree of confidence and scope about the topics. First-year students avoided assertiveness and complexity, while upper-level students felt more comfortable expressing confidence and greater depth of ideas and details.

Other research has also explored the features of academic writing contexts of both first-language and second-language classrooms, focusing on various textual features, such as lexical bundles (Biber *et al.* 2004; Wood 2015; Durrant 2017), metadiscoursal language (Crismore *et al.* 1993; Thompson 2001; Hyland 2005a), civility (Aull 2020), compression (Biber and Gray 2010), variation (Hyland 2002a; Charles 2007), and informal features (Chang and Swales 1999; Hyland and Jiang 2017). Such research studies have been based on large corpora comprised of millions of words, such as the Michigan Corpus of Upper-level Student Papers (MICUSP), consisting of 2.6 million words (Lee and Deakin 2016; Lee *et al.* 2019) or the Longman Spoken and Written Corpus, consisting of 5 million words (Biber and Gray 2010). On the other hand, some researchers have chosen relatively small corpora, such as 322,750 words, as demonstrated in Hinkel's (2003) research on successful patterns of academic writing.

When discussing corpus size, it is important to consider the practical steps of corpus design and the role of length in the overall design. O'Keeffe, McCarthy, and Carter (2007) discuss the steps in building corpora, highlighting the representativeness and design rationale. The size of the corpora depends on the type of the corpora, whether spoken or written, and what it intends to represent. For example, a spoken corpus is considered large if it is over a million words, while a

large written corpus must carry more than five million words to qualify beyond the small norm. The suitability of the corpus depends on the design rather than the size, which motivates the corpus design rationale. Small specialized corpora "can lead to insights that cannot as easily be gained by looking at large general corpora" (O'Keeffe 2007, p.198). In the context of this study, analyzing a small, specialized corpus of student writings can provide unique insights into their writing competencies and discourse patterns, offering valuable guidance for tailored instruction and curriculum development.

Specialized corpora enable a targeted collection of data. According to O'Keeffe et al. (2007), it is important to use specialized corpora to collect carefully targeted data that accurately represents the target domain. Such data collection is more effective than capturing all aspects of a language. Corpora such as COMP 101 are comprised of students' essays and coursework, which allows linguists to explore the use of grammar, vocabulary, or common errors in student writing. Technical vocabulary and structures are likely to exhibit consistent patterns and distribution, even with a relatively small amount of data. For example, in universities and schools, the texts produced by the students correspond to various genres and demonstrate the students' writing competencies through norms and discourse (Miller 1984). Regarding the data analysis, Aull (2015; 2020) indicates that discourse patterns that communicate the writers' attitudes to the readers are implicit and require computer-aided insights to discover the repeated use of lexical and grammatical choices and effects. These repeated patterns are often studied in the context of genre, as Bawarshi (2010) notes, and can reveal the relations between social contexts and the actions of individuals within activity systems such as a college course.

This study aims to analyze the writing styles in COMP 101 and uncover what student writing can reveal about how students handle academic writing requirements. By using computer-aided tools

to recognize implicit discourse patterns, educators can focus their teaching on textual organization and engagement with the reader in order to help students gain a better grasp of both the large picture and the sentence details. With the same purpose, the study also delves into the genre contexts, such as descriptive, narrative, classification, process analysis, compare and contrast, cause and effect, and argumentative writing to examine the repeated textual features in the texts.

1.3 Aims and research questions

This study aims to analyze the vthe key linguistic features of texts written by first-year students as part of their COMP 101 course. The main research question focuses on identifying these characteristics in the corpus COMP 101 and compares them to upper levels of writing represented in the corpus of British Academic Written English (BAWE):

(1) What are the key linguistic features that characterize first-year composition writing, and how do they differ from upper levels or later stages of writing at university?

A sub-question that aids the investigation of the language features in the main research question is:

(2) How are these key linguistic features distributed in the descriptive, narrative, classification, process analysis, compare-and-contrast cause-and-effect, and argumentative texts?

The purpose of the sub-question is to address any of the repeated COMP 101 textual characteristics and their distribution in the narrative, descriptive, classification, process analysis, comparison and contrast, cause and effect, and argumentative writing.

1.4 Overview of the research structure

This research has nine chapters. Chapter 2 provides a literature review, starting with traditions in teaching academic English, then discussing the field of English for Specific Purposes, English in global academia, as well as English as a Lingua Franca, the levels of academic writing proficiency, and the features in English academic writing as reflected in corpus linguistic research. The literature review aims to provide a comprehensive background of the topic and offer the reader detailed information about academic writing. Chapter 3 presents the theoretical framework of the study, which follows a genre-based model and discusses the distinct elements surrounding the genre concept, such as register and style. The chapter also examines genre in a conventionalized setting, considering the roles of common discourse conventions, discourse communities, and text producers. Chapter 4 outlines corpus linguistics as the research methodology and the corpus design, participants, and corpus tools such as frequency lists, concordance lines, collocation, and keywords.

Chapters 5 through 8 provide an analysis discussion based on the frequency findings. Chapter 5 focuses on the first-person personal pronouns in COMP 101, how they differ from BAWE, and their distribution across the sub-corpora, such as the narrative, descriptive, classification, process analysis, comparison and contrast, cause and effect, and argumentative writing. Chapter 6 discusses the second-person personal pronouns in COMP 101 and their usage reflected in BAWE, finishing with the sub-corpora. Chapter 7 offers a detailed analysis of the most frequently used conjunctions and how they impact sentence structure and clausal patterning in COMP 101, the differences with the BAWE texts, and the distribution and use of these patterns across the sub-corpora. Chapter 8 explores the most frequently used punctuation in COMP 101 and BAWE, as well as the most frequently used punctuation in the sub-corpora. Chapter 9 serves

as the concluding chapter, summarizing the findings, presenting the limitations, and providing suggestions for the classroom and future research.

Chapter 2

Literature Review

2.0 Introduction

At first glance, academic writing appears to be a straightforward concept that should be easy to define. However, upon closer examination, as this chapter aims to accomplish, unveiling the different varieties, traditions, and levels of competence becomes challenging. Hyland (2009) discusses this challenge in its larger framework of academic discourse, referring to it as "unpacking the black box of academic discourse" (p.5). According to him, the complexity of this task is based on multiple factors, such as the diverse teaching practices, the emergence of English as the lingua franca of scholarship, and the growing diversity of incoming students. All three factors have one common denominator—English as the standard medium to facilitate both teaching and learning. Swales (1997) refers to this power of English, comparing it to a Tyrannosaurus Rex, "a powerful carnivore gobbling up the other denizens of the academic linguistic grazing grounds" (p.375). Having such a prominent language status, English impacts writers from different backgrounds and levels of competencies in countries that accept it as an official language and those that do not, and it is the job of academia to provide appropriate standards, guidelines, and support for students and researchers.

This chapter attempts to "unpack the black box" of academic writing by examining the teaching traditions, the global status of English as the language of academia alongside its impact on the curriculum, and the research demonstrating the textual features of academic English. In doing so, the chapter contextualizes this study and discusses the most frequent features in academic writing reflected in literature.

2.1 Traditions in teaching academic writing in English

Tracing the historical traditions of teaching academic writing in English can be challenging due to various approaches used over the years that continue to be utilized today despite their period of occurrence. Each approach has its focus and roots in different academic schools. In his work, Paltridge (2004) provides a detailed overview of the primary methods used in teaching academic writing, beginning with controlled composition in the 1940s through the 1960s. In this approach, language is perceived as a set of fixed patterns that writers can use to create new sentences without focusing on the patterns beyond the grammatical structures. Thus, it takes a prescriptive stance, focusing on the proper usage of grammatical rules and limiting the flexibility in writing styles.

In the mid-1960s, the so-called current-traditional rhetoric approach emerged, particularly successful in US universities (Paltridge 2004; Tribble 2009). Connors (1989) documents the rhetorical studies in the US, starting with the first courses in composition studies at Harvard University, initially designed to provide necessary writing skills to all students. Current-traditional rhetoric emphasizes the role of the text at the discourse level, moving beyond the level of sentence structure and showing students the rhetorical functions such as descriptions, narrations, classification, or argumentation. These rhetorical functions are still used in US universities (Tribble 2009), including the university where the texts of this study were collected.

In the 1970s, teachers became concerned that academic writing was not allowing enough opportunities for students to express themselves, which gave rise to the process approach. The process approach allows students to find the proper form of the text based on their message rather than letting the form control the message. The major criticism of the approach was the over-emphasis on expression rather than the academic requirements and learners' needs, which

led to the so-called needs analysis approach (Braine 2001). The need analysis considered the needs of the students, their learning gaps, and their personal views, which demonstrated both the subjective and objective assessment of writing. The approach also considered teachers' academic requirements when looking at the writing content, which gave rise to content-based instruction (Paltridge 2004) that helped students immerse in the educational environment but sometimes compromised the role of grammar.

In the 1990s, the genre approach to academic writing was popularized by Swales (1990) and adopted by other scholars such as Hyland (2003) and Flowerdew (2011), who proposed a mixture of an integrated approach between process and genre, focusing on the text procedure rather than the outcome. Tribble (2015) discusses the main characteristics of the genre approach as the conceptualization of disciplinary discourse communities, textual analysis of multiple text samples, understanding the realization of the text structure, and the use of corpus linguistics to analyze the lexico-grammar underlying the writers' choices.

This brief overview of writing instruction in English demonstrates that approaches have moved from strict, rule-based methods to more adaptable, holistic approaches, highlighting the dual nature of writing as both a form of expression and analysis. According to Paltridge (2004), effective academic writing instruction should not be reduced to the use of one approach only, but it needs to consider or incorporate diverse methods in targeting the individual needs of students. Such a comprehensive view on teaching academic writing emphasizes the importance of adjusting teaching strategies to suit a diverse academic population and unique individual learning requirements.

Building on this foundation, it is essential to explore the field of English for Academic Purposes (EAP), which focuses on equipping learners with the necessary skills to effectively study and

conduct research in academic environments. Hyland (2006) defines EAP as “teaching English to assist learners’ study or research in that language” (p.1). The field involves various teaching practices and traditions, often customized to the specific academic disciplines and cultural contexts of the learners. By offering targeted instruction that addresses both language proficiency and disciplinary conventions, EAP aims to empower students to navigate the demands of academic writing and research.

2.2 Traditions in English for Academic Purposes

EAP studies are typically associated with non-native English speakers and aim to provide strategies for students unfamiliar with using English in academic contexts. However, these studies are not exclusive to non-native speakers and can help all students struggling with specialized vocabulary, technical writing, and various academic style guides to adapt to the academic demands of the university. EAP was first coined by Tim Johns in 1974 and later popularized in a collection of papers by Cowie and Heaton in 1977 (Jordan 2002). The goal of EAP is to aid students with varying proficiency levels in universities by providing them with strategies for effective engagement with academic texts and meeting the demands of the educational context. In all this, it is important to recognize the changing demographics of the contemporary university classroom, which embraces both native and non-native English students who require guidance in creating academic texts. Thus, exploring the main traditions in EAP studies is important for the contextualization of this study and helps to unpack another layer of the complexity surrounding academic writing.

There is a limited amount of research discussing the chronology of EAP traditions, and in an effort to identify these traditions, this study found the research by Tribble (2009) to be very

helpful in establishing a timeline. The study also uses the work of other researchers (Halliday and Hasan 1985; Connors 1989; Swales 1990; Hyland 2004; Paltridge 2004; Biber 2006; Flowerdew 2016) to supplement the practices related to the timeline. Tribble (2009) identifies three main EAP traditions, which split between the practices of the UK and the US, acknowledging the challenge in reviewing these traditions by having "to deal with fragmented and sometimes contradictory accounts of what is meant by EAP" (2009, p.400). In the UK, Tribble (2009) notes that EAP has been closely associated with the practices related to English for Specific Purposes (ESP) and thus is founded in the work of Halliday and Hasan (1985) in the register and genre studies. In this regard, the EAP practices are based on both the communicative context and the linguistic behavior as part of the context as the baseline for addressing the needs of the learners. Like Tribble (2009), Flowerdew (2016) notes that EAP considers writing a social action in a specific situational context, and teachers help students socialize in their discourse communities.

In the discussion of EAP and ESP, it is important to mention that ESP evolved over 20 years from the 1960s to the 1980s, making the relationship between context, language, and teaching methods clear (Tribble 2009; Dou *et al.* 2023). Since EAP is rooted in ESP, it seeks to understand the learner's context, how it influences the learner's needs, and what kind of teaching approach is needed to meet those needs. This specific focus of EAP has allowed teachers in the field of corpus linguistics, text analysis, and related research projects to make efforts to find innovative approaches to addressing the needs of the learners. Tribble (2009) names the design of British Academic Written English (BAWE) and British Academic Spoken English (BASE) as part of these efforts to find suitable solutions for educational practices in academic writing. Being driven

by the same motivation, linguists like Swales (1990), Hyland (2004), and Biber (2006) have published a series of studies and contributed to the field of ESP and EAP.

Within this dynamic context, Tribble (2009; 2015) identifies the first tradition in EAP that primarily develops in UK Social/Genre and is rooted in the work of M. A. K. Halliday (1985) in Register Analysis and further explored by John Swales (1990), and Ken Hyland (2004). The key characteristics of this approach include the conceptualization of the disciplinary discourse communities, textual analysis in a social context, move analysis and its role in the textual structure, corpus linguistics, and scaffolded methodology (Tribble 2009; 2015). The scope of the Social/Genre approach goes beyond the needs of the language learners and addresses the specific areas in text development that are beneficial to all students involved in academic writing.

The second tradition in EAP is Intellectual/Rhetorical, and it evolved from the teaching practices in US Composition Studies (Tribble 2009; 2015). To understand the context of the EAP programs in the US, it is important to contrast it with the UK context, whose primary focus was on the English language spreading to other educational cultures and the rapid growth in the number of second-language students coming to UK universities. The EAP in the US, however, developed as a response to the expansion of higher education in the US during the 1960s and 1970s, placing its main goal on strengthening the competencies of students who had been previously excluded from university studies (Tribble 2009). This tradition is also listed by Connors (1989) regarding the emergence and development of composition courses with the same focus on addressing the needs of students coming from lower socioeconomic backgrounds who lack the necessary academic writing skills. Like Tribble (2009; 2015), Paltridge (2004) discusses the grammar, rhetorical approach, and process approach but addresses them as separately developed approaches motivated by the needs of the students. It is important to mention that both

Tribble and Paltridge emphasize the importance of integrated approaches rather than only one that fits all needs.

The key characteristics of the Intellectual/Rhetorical tradition include the emphasis on formal textual organization and rhetorical modes such as narration, classification, argumentation, or causal analysis. It also emphasizes collaborative writing, the process approach, and grammar practice to develop correct patterns at the sentence level. This tradition continues to impact the curriculum across US universities and has been the main influence on the composition curriculum for this study. The works submitted in the COMP 101 collection—narrative, classification, and argumentation, to name a few—reflect this tradition.

The third main tradition in EAP became known as "writing in the disciplines" in the US and "academic literacies" in the UK (Lea and Street 1998; Tribble 2009). According to Tribble (2009; 2015), writing in the disciplines contributes more critical and disciplinary-specific alternatives to composition and genre approaches. It mainly developed in the US during the 1980s and 1990s, shifting the focus from the predominant rhetorical approach to the discipline-specific use of language. In the UK, the tradition kept its close connection to its US counterpart but is known as academic literacies and emerged from the need to address the challenges of the changing student population, similar to what happened in the US universities as educators began to respond to the needs of the students who previously would have been excluded from higher education.

The focus of EAP has been on teaching English to help students study and conduct research in English, which covers all areas of academic, communicative practices—undergraduate and postgraduate teaching, classroom interactions, research genres, student writing, and administrative procedure—as they are situated in their local contexts and the needs of the

particular students. Considering the diverse context of the college and university classroom in the United States (Hyland 2006), the challenges that teachers in higher education have in helping students graduate, and the universal struggles that students, both native and non-native, face in using academic English seem to bring the role of EAP studies to the forefront in helping both students and teachers bridge the language gap in methods on the side of teachers and knowledge and skills on the side of students. In this regard, Durrant (2017) observes that the most significant challenge for learners, teachers, and researchers in the field of EAP is "the fact that, in many respects, academic texts have turned out to be highly heterogeneous" (p.165). Teachers notice these variations in the discourse and linguistic conventions in the textbooks, student assignments, and genres across disciplines as they try to help incoming students deal with the challenges of using academic discourse.

Looking back on the EAP evolution, it can be noted that it initially started as a curriculum that catered to the needs of English learners in universities, but it soon recognized the importance of addressing the needs of first-language speakers lacking essential higher education skills. Beyond the scope of undergraduate and graduate education, researchers and scholars publish articles around the globe, using English as a common medium. As Flowerdew (2016) observes English is "indisputably the international language of academic research" (p.6), which suggests that it is considered a common communication medium outside and inside English-speaking communities. This spread and internationalization of English around the globe is the topic of the next section.

2.3 English in global academia

This study aligns with Flowerdew's (2016) observation, noted earlier, that English is the predominant language used in academic research and writing worldwide. Despite debates over its impact on local and native languages (Cabral-Cardoso 2021) and arguments for the decolonization of academic writing (Canagarajah 2024), this study aligns with Swales' (1997) practical perspective on the global influence of English. This perspective is reflected by Flowerdew (2016) on the international status of English in academic writing and Hyland's (2009) view of the global language impact of English in the academic discourse and writing patterns. This section discusses the global use of English as a lingua franca and its role in academic writing.

2.3.1 English as a lingua franca

David Crystal (1997) explains the factors contributing to a language being considered global. He emphasizes that the number of people who speak the language is not the only factor; instead, he focuses on the type of people who use the language. He cites Latin as an example of a language that became international during the Roman Empire, not because of the native population's size but due to political and territorial expansion that allowed people of different languages to use it. Even though political or military expansion may be a factor in the potential of a language to become global, they are not the only ones. A language becomes a lingua franca or common language when people of diverse language groups are met by the need to communicate with each other for different purposes, including trade, education, or diplomatic relations (Crystal 1997). This need to drive communication and create a space for multiethnic and multinational dialogue in business, education, technology, or social levels is being facilitated by English.

Similar to the factors discussed above, Mauranen (2010; 2019) supports the concept of English as a Lingua Franca (ELF), emphasizing the importance of mobility and communication among speakers of different native languages and its role in facilitating cross-cultural interactions in an increasingly interconnected world. This powerful language interaction is described as "a somewhat paradoxical situation" by Seidlhofer (2005, p.339), where the predominant interaction happens outside of the native speakers. In this regard, Firth (1996) provides a helpful definition of English used as a lingua franca, "a contact language between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen foreign language of communication" (p.240). As a contact language, thousands of non-native English speakers use English to enter universities or enroll in online academic programs where education is conducted in English. After completing their education, many of them contribute to academic journals, the scientific community, or the university classroom—often in contexts where English remains the main communication medium. This reinforces its role in bringing educators, scholars, and students from diverse language backgrounds together. Thus, English impacts the spoken and written exchange of ideas beyond given geographical areas, reaching across cultures and languages across the world.

Regarding the areas of ELF and EAP, the COMP 101 classroom comprises a mix of native and non-native English language speakers. For non-native English speakers, English becomes the language of contact or ELF when communicating with one another. On the other hand, native English speakers who take COMP 101 must enroll in the class based on their low-performance high-school grades in English and must develop the same academic writing proficiency as their non-native counterparts. In a way, both groups benefit from the EAP-driven curriculum and thus represent EAP students. The next subsection discusses the global impact of English in academic

writing, which may seem too broad of a context to this study, given the fact that this study is conducted in the context of a US university, but the discussion of the global impact of English is necessary because it stresses the importance of academic standards ensuring a proper exchange of ideas within the larger academic community.

2.3.2 Impact of English on global academic writing

Viana and O'Boyle (2022) point out the forces contributing to the global academic landscape, such as the growing number of journal publications in English, courses in English beyond the Anglophone-speaking countries, and the implementation of the English curriculum on a global scale. Among these dynamic forces, academic publications in English seem to impact academic writing and its users significantly. Narvaez-Berthelemot and Russell (2001) find English as the main publication language in a UNESCO social sciences database. Another study by Testa (2009) discusses the four main criteria for evaluating the Thomson Reuters selection process: timeliness of publication, compliance with international editorial conventions, peer review, and the provision of English-language bibliographic information. This growing prevalence of English in academic settings understandably is one of the reasons for the increase in courses offered in English in countries where English is not an official language.

The increase in English-language courses can be attributed to the Bologna Agreement (Murphy and Dyrenfurth 2006), which aimed to standardize higher education across European institutions and promote mobility among students, researchers, and educators. As a result of the changes in the curriculum following the Agreement, there has been a growth in courses taught in English as a medium of instruction (EMI). These courses have gained popularity in South America and Southeast Asia (Rose *et al.* 2020; Viana and O'Boyle 2022). The rising number of EMI courses

globally has sparked discussions about developing a comprehensive international education model in English, incorporating a multi-source information and evaluation system to effectively meet the future trends in English education (Yang *et al.* 2021). The idea of a global English curriculum, the popularity of EMI courses, and the need to publish in prominent academic journals emphasize a common goal: the ability to express one's thoughts and knowledge in academic English, which is understandable to others.

Given the wide range of influence on academic literacy and performance, English should be an equal concern for both first and second-language learners. As Viana and O'Boyle (2022) observe, "no one is born writing academic English" (p.11), and speaking English as a first language does not always mean being able to write in academic English. One of the primary challenges in academic writing lies in its specialized context, which differs between various university disciplines in terms of terminology and style (Biber 2006; Biber and Gray 2010). Apart from the specialized content, the academic setting demands that students possess knowledge and critical thinking skills to effectively integrate new information with their existing understanding. The use of academic English enables all students, researchers, and educators to learn, collaborate, and share knowledge, emphasizing the need for an effective English curriculum that serves the needs of diverse learners.

Flowerdew (2019) notes the idea of standardization in the English academic curriculum for universities to facilitate common standards across educational systems that support academic mobility. However, creating common standards may be challenging due to the diverse backgrounds and writing abilities of stakeholders in the academic space. Corpus linguistic research is crucial in investigating the different levels of academic writing, identifying their characteristics, and providing guidelines for best practices or further research. Research clearly

divides student writers from expert writers (Aull *et al.* 2017; Whong and Godfrey 2020) as the main stakeholders in academic writing. Student writers are mainly categorized into three groups: first-year university students, upper-level university students, and expert writers (Aull and Lancaster 2014; Lancaster 2014; Aull 2015; Aull *et al.* 2017) (Sections 2.4.2 discusses the three levels of academic writing). Each group faces distinct challenges and requires appropriate support to fulfill their academic tasks.

Since the primary focus of this study is analyzing the frequency features found in the writing of first-year students and comparing them to the frequencies observed in upper-level texts, it is important to describe the spectrum of academic writing proficiency based on existing literature to contextualize the group under investigation. The following section briefly discusses the different levels within academic writing, focusing on first-year university writing, which is the specific group represented in the COMP 101 corpus, and how it differs from the upper level of writing.

2.4 Levels of academic writing proficiency

The progression of writing proficiency in academic writing spans from first-year university composition to upper-level student papers, and eventually to published scholarly works. Research by authors such as Aguila (2014), Aull (2015), and Staples and Reppen (2016) highlights the initial challenges first-year students face as they transition from structured, formulaic writing to more sophisticated and analytical. This foundational stage is marked by a struggle to meet higher academic expectations and to develop a nuanced understanding of disciplinary conventions. As students progress from first-year university writing to upper-level writing, they must master genre-specific features and the effective use of stance markers to

convey authority and engagement in their arguments (Aull 2019). Whicker (2022) discusses the significance of effective instruction in transferring skills to upper-level writing tasks. Beyond the university setting, expert writers demonstrate the highest level of proficiency, adhering to rigorous publication standards and contributing original knowledge to the academic community. Hyland (2001b; 2002a; 2002b), Biber et al. (2004), and Harwood (2005) examine the features of academic writing in journal articles and textbooks, and their studies illustrate the evolving demands of academic writing and the continuous development of writing skills necessary for academic success and professional scholarly communication.

2.4.1 First-year university students

According to Aull (2015), the common approach to preparing incoming students in the U.S. higher education system for academic assignments is to provide required first-year writing courses. The curriculum in these courses is designed to offer first-year students the skills and knowledge to bridge the gap between secondary and post-secondary education. Typically, the assignments in these writing courses consist of academic essays that aim to enhance students' critical thinking, genre awareness, and general academic writing abilities. Some universities have the means and resources to offer writing courses exclusively for non-native speakers of English and help them deal with academic vocabulary and grammar, but not all colleges and universities can afford such training. As Hyland (2006) notes, the reality is that first-year writing courses often train mixed groups of students—native and non-native—and help both groups learn how to structure their writing.

The primary emphasis of first-year writing courses is on the overall rhetorical structure and organization of texts rather than on the linguistic features at the micro level (Aull 2015). In this line of thought, writing in the context of first-year composition courses has focused on individual texts rather than language-level patterns across them, thus placing rhetoric-composition outside of the scope of linguistic studies, resulting in the fact that despite the regular presence of genres in the first year of university writing, very little is known about the micro-level linguistic features that describe those genres. Wardle (2009) also observes the challenges in these entry-level writing courses and how the composition assignments in the first year of university carry the heavy task of preparing students to write in a variety of fields even though the majority of the composition instructors are trained in literature or rhetoric, and many of the same instructors think of the genres in English studies as genres-in-general.

While rhetoric-composition appears central to first-year composition courses, it cannot encompass and provide answers to the diverse student population's broad spectrum of needs, but it needs the knowledge and methods of applied linguistics in EAP. Looking at the historical tradition in rhetoric-composition and EAP, Aull (2015) notes that the two fields approached academic writing from different perspectives: first-year composition courses were developed primarily under the guidance of rhetoric-composition to serve the needs of native English-speaking students, while the EAP studies were formed to serve non-native English-speaking students. Thus, rhetoric-composition and EAP developed separately and kept their separate disciplinary traditions. Motivated by the efficacy of genre awareness, both fields focus on the theoretical and pedagogical concepts of genre. In comparing the connection of these two fields to academic writing, Aull observes that EAP focuses on “linguistic features and methods like corpus linguistic analysis,” while rhetorical-composition studies see academic writing through

"the rhetorical achievements of whole texts in specific contexts" (2015, p.19). Thus, both fields intersect in their aim to help students gain writing competencies, but they differ in their approaches. This leaves room for EAP studies and corpus linguistics to investigate the linguistic features that characterize texts written by both native and non-native English speakers, providing a basis for teaching strategies. It is reasonable to conclude that first-year writing courses, such as COMP 101 related to this study, would benefit from incorporating corpus studies and genre analysis from EAP.

2.4.2 Upper-level student writers

Upper-level students, as explored by Lancaster (2014), Aull (2019), and Whicker (2022), are the students who typically find themselves in their third, fourth, or later years of university studies. They are already aware of their discipline-specific writing and better understand the academic conventions and writing strategies than first-year students. Research related to this group of writers mainly focuses on the stance and engagement markers (Lancaster 2014; Aull 2019) and the transfer of writing knowledge from the first year to the upper level of writing (Whicker 2022). This is an important stage of the academic writing journey as students develop more skills and grow discipline-specific vocabulary.

Lancaster's (2014) study notes that high-performing essays frequently adopt a stance that demonstrates commitment, critical distance, and efforts to engage readers. On the other hand, low-performing essays carry less commitment and critical distance. Also, engagement patterns, such as sentence-initial connectors, are more varied than low-performing essays. For example, high-performers demonstrate more connectors, such as "however," "but," and "nevertheless" than the low-performers. The verb choices in the high-performing essays indicate objectively

grounded expressions, such as “suggests” and “indicates,” rather than subjectively grounded, like “I feel” or “in my opinion.” Lastly, the use of personal pronouns by high performers shows strategic use, contributing to the dialogue, whereas their counterpart uses them to reduce the objectivity of the argument.

Aull’s (2019) study adds another aspect to the growing competence of upper-level students by examining the stance markers, indicating the writer’s position and engagement with the reader. She divides the markers into epistemic cues (hedges and boosters) and textual cues (contrastive connectors and code glosses). Effective use of hedges and boosters at this stage shows a more balanced academic stance, which leaves room for alternatives. Section 2.5.3 discusses further Aull’s research regarding academic writing. Finally, Whicker (2022) presents a study that shows that upper-level students can recall and appreciate knowledge about different types, styles, or formats of writing. This ability to distinguish between writing types demonstrates accumulated experience with text and an understanding of the various writing selections. Another important characteristic at this level is the ability of students to successfully retrieve the previously learned skills and transfer them to the next level, which emphasizes the importance of good teaching strategies and effective guidelines.

As upper-level students progress through their journey at university, they continue to exhibit a more refined understanding of discipline-specific writing and academic conventions. Both Aull (2019) and Lancaster (2014) indicate the importance of knowing a balanced use of stance markers and the ability to create a dialogic space with the readers. Whicker (2022) suggests that further research is necessary to examine upper-level writing to create better guidelines and teaching strategies. Within this context, this study focuses on texts created by first-year students in COMP 101 and compares them to texts of upper-level students as comprised in BAWE. The

goal is to identify the frequency features in first-year university writing, the possible patterns that they create, and how these features and patterns differ from upper-level writing. In a way, this study does not start with predetermined ideas about the most frequent textual features but seeks to investigate, identify, and then compare them to the ones used in the upper level of writing.

2.4.3 Expert writers

Based on Syrewizc's (2022) study of the current research in academic writing, expert writers are characterized by high levels of both declarative and procedural knowledge, strong motivation and interest, as well as effective self-regulating strategies. These writers engage in complex and highly recursive writing processes that require continuous planning and drafting. Many aspects of this writing work are related to social and professional relationships with peers and building networks. Another insight into expert writing comes from Aull's (2017) research on the epistemic stance in published articles, which reveals that expert writing contains significantly fewer generalization markers than student writing. This suggests that expert writers prefer to make more precise and narrowly defined claims. Additionally, expert writers strategically use indefinite pronouns and amplifiers to emphasize wide applicability or shared ideas, ensuring their statements are grounded and context-specific. These findings show that expert writers use generality in specific and limited ways, reflecting their understanding of academic conventions.

Regarding sentence structure, Biber and Gray (2010) use a corpus of research articles and university textbooks to demonstrate that academic writing is concise, with non-clausal modifiers embedded in noun phrases as the predominant structural units. Another characteristic of sentence structure in academic writing is its lack of explicitness, with structures being predominantly nominal and passive, often not revealing the agent or the occurrence of an activity. In science

research articles, Biber and Gray (2010) note that appositive nouns are some of the common structures, adding detail with no explicit grammatical markers to show explicit relationships, such as the examples in the underlined structures "depending on whether enrollment (first cohort visit) occurred within 6 months of the first physician diagnosis of systemic sclerosis (incident case) or whether diagnosis had preceded the first visit by >6 months (prevalent case)" [the underlining appears in the original text] (p.14). Such cases suggest that academic writing relies on readers' background knowledge, ability to infer meaning, and careful reading to fully grasp the nuances and connections between ideas.

Another interesting area in expert writing is the author's identity (Hyland 2002b; Harwood 2005) and reader engagement (Hyland 2001a). Hyland explores how author pronouns are used in published articles, finding their predominant use in humanities as authors "make a personal standing in their texts to establish a credible scholarly identity, and to underline what they have to say" (2002b, p.351). Hyland also explores the use of metadiscourse, which is discussed in detail in Section 2.5.2, and its role in engaging and guiding the reader. Harwood (2005) investigates the use of the author's pronouns in self-promotion, creating research space, organizing discourse, reporting findings, and disputing claims. Such research shows the strategic use of "I" and "we" in academic writing as it serves to organize the text, present the findings, and promote the authors' contributions, indicating that authorial identities matter even in highly objective texts as research articles.

Overall, based on the research (Hyland 2001a; Hyland 2002b; Harwood 2005; Biber and Gray 2010; Syrewicz 2022), expert writing can be distinguished from other types by the extensive knowledge, motivation, self-regulation, and good understanding of conventions that authors have. Additionally, it is concise and non-explicit, relying on readers' inferences and background

knowledge. While objectivity is highly regarded in academic writing, at an expert level, authors strategically use personal pronouns to emphasize certain points and promote their findings. With its distinguished characteristics, this type of writing differs from upper-level and first-year writing, demonstrating professional writing competencies. Since the study looks at first-year writing as the lower end of the proficiency writing spectrum, it is important to list expert writers as the representatives of the higher end. Understanding the differences indicated in the textual patterns can help gain a better perspective of the learning curve that exists between first-year students and expert writers and contribute to research on the linguistic features in academic writing. The following section provides an overview of this research and the various features discussed in academic literature.

2.5 Features in English academic writing

English academic writing has been the focus of long-time research and analysis among linguists in first and second language studies. Biber (2006) notes that "all students—whether native speakers of English or non-native speakers—need to adjust to a wide range of tasks in the university accomplished through language" (p.1). He further argues that most universities do not do a lot to prepare students for "the linguistic demands of the academic prose" (Biber 2006, p.1). Mauranen et al. (2010) argue that English has replaced Latin as the modern academic Lingua Franca. While style manuals have focused on providing rules and guidelines for academic writing, researchers have looked into the validity of the prescriptive rules in texts produced by students (Lee *et al.* 2019) tracing the linguistic features in academic writing: (1) the use of lexical bundles; (2) the use of metadiscourse (Hyland 2005a); (3) the use of civility, cohesion, and compression (Biber and Gray 2010; Aull 2020): (Biber and Gray 2010); (4) the patterns of

variation across academic disciplines (K. Hyland 1999; Hyland 2002b); and (5) the use of informal language (Chang and Swales 1999; Hyland and Jiang 2017). All these features of interest in academic writing have been investigated and analyzed through corpus linguistics. By analyzing large corpora of academic texts, researchers can identify patterns and features characteristic of academic writing. This approach allows for a more objective and comprehensive understanding of language use in academic contexts, thus informing better teaching practices and resources.

2.5.1 Lexical bundles in academic writing

One of the important areas of research in academic writing is the role of lexical bundles in texts. These multi-word expressions are extremely common and serve crucial functions by acting as pragmatic leading units or "heads" for larger phrases and clauses (Biber and Barbieri 2007, p.270). Unlike formulaic language, which comprises multi-word expressions with simple meanings or functions that are mentally stored and retrieved as single units, such as "good morning, look up, on the other hand, at top speed, etc." (Wood 2015, p.3) lexical bundles are not idiomatic in meaning and do not stand out to the reader. Instead, they provide structural support within academic texts, helping to frame discourse and signal new information effectively. Biber and Barbieri (2007) emphasize that lexical bundles do not represent complete structural units but are integral to the coherence and flow of academic writing. This usage differs from the more self-contained nature of formulaic language, which often serves specific social or communicative functions. Thus, while both formulaic language and lexical bundles play important roles in language use, lexical bundles are particularly vital in academic contexts for their ability to organize the text, creating a functional framework.

Lexical bundles are described as an important building block in academic discourse (Biber and Barbieri 2007) and are used to study development stages and levels of proficiency in academic writing between novice learners and expert writers (Biber *et al.* 2004; Cortes 2004; Biber and Barbieri 2007; Chen and Baker 2010; Pérez-Llantada 2014) and even their different structural usage between native and non-native English writers. Non-native writers tend to form lexical bundles with verbs and clause fragments (e.g., “we assume that the”), while native writers prefer noun and prepositional phrases (e.g., “the size of the”) (Pérez-Llantada 2014, p.64). According to Biber and Barbieri's (2007) research, lexical bundles in academic writing are used to create a functional framework of three major categories: stance expressions, discourse organizers, and referential expressions. In this framework, stance bundles indicate certainty, discourse organizers show relationships between prior and upcoming discourse, and referential bundles define specific parts of entities.

While bundles are important in academic writing, they are not static and vary with conditions and contexts. This characteristic leads to variations in their usage across disciplines. For example, disciplines like engineering demonstrate the highest use of bundles over time, while in other fields like sociology, the use of bundles has decreased by more than fifty percent over fifty years (Hyland and Jiang 2018). These differences can be attributed to the technical and formulaic nature of engineering, where precise and standardized expressions are crucial. In sociology, the decline in bundle usage may indicate a shift towards a more diverse and less fixed language. In both fields, these discrepancies may suggest a trend towards a more nuanced and adaptable discourse that corresponds to evolving theoretical frameworks and research practices.

Durrant (2017) observes the disciplinary variations and heterogeneous ranges in texts among disciplines and examines the use of lexical bundles in different disciplines without assuming the

discipline categories at the outset of the analysis but letting these categories emerge from the initial research across all texts in the corpus. Another important insight in this study is that lexical bundles serve to bridge linguistic units and connect two or more phrases or clauses (e.g., "if you look at," "that's one of the," "it's important to") and possess three distinct characteristics that draw researchers' interest: "they can be identified automatically; they play definable functional roles; and they are highly sensitive to differences between text types" (Durrant 2017, p.166). These linguistic qualities that lexical bundles carry make them useful in examining text differences and describing them in functional terms. The results of the study led to the differentiation of four disciplines: humanities and social sciences, science and technology, life sciences, and commerce. The instances of the stance-oriented bundles characterize the texts in humanities and social sciences more than the ones in science and technology writing.

Discussing the role of lexical bundles in academic writing, it is important to note Swales' (2019) research related to the formulaic sequences, emphasizing multi-word units, N-grams, and lexical bundles. He questions the usefulness of published material on short multi-word units, pointing out that most researchers use four-word bundles because they are easy to manage in size. However, these bundles often represent the most common sequence units in English (e.g. "as a result" or "in other words") (2019, p.77), and Swales concludes that bundles purely based on frequency data (e.g., "and of the" or "which is a") have little use for English language teachers and learners. Swales states that such studies "are indeed valuable and stand in sharp contrast to those who provide frequency data without any consideration of possible or potential pedagogical usage" (2019, p.77). Despite his skepticism towards multi-word units based solely on frequency data, he acknowledges the significant pedagogical value in research that categorizes phrases based on whether they represent chunks, cohesive meaning, or function and whether they are

worth teaching. This study does not focus on the use of lexical bundles themselves but rather on the most frequently used textual features in the context of first-year composition writing. Nevertheless, discussing lexical bundles in this section is important because they represent a key research topic in academic writing, highlighting differences in their usage across disciplines and emphasizing the need to connect this research with teaching practices.

2.5.2 Metadiscourse in academic writing

Similar to the functional role of lexical bundles in the text, *metadiscourse*—first introduced by Harris (1959) and later developed by Vande Kopple (1985) and Crismore (1989; 1993)—also plays an essential role. One of the most prominent names in the metadiscourse research in academic writing is Hyland (2004; 2005a; 2010; 2017). According to Hyland (2005a), metadiscourse is widely used in language interaction for two main reasons: (1) to understand the relationship between language and its context and (2) to apply this knowledge in literacy education. In discussing the meaning of metadiscourse, it is essential to distinguish the term from *metalanguage* and *metapragmatics* since literature indicates that the terms are often confused. The concept of metalanguage relates to "people's knowledge about language and representations of language, so it is the terms used by teachers, learners, and analysts to make statements about an object language" (Hyland 2017, p.17). Metapragmatics relates to "speakers' judgments of appropriateness of communicative behavior, both their own and that of others" (Hyland 2017, p.17). Hyland (2005a) describes the term metadiscourse as a language representing the attempts of the writer or speaker to guide the reader or listener in understanding the text or the speech. The concept includes a wide range of discourse features, such as hedges, connectives, and words signaling attitude to engage and guide the reader through the text.

Metadiscourse is a dynamic view of language that is based on the idea that communication reflects the personalities and assumptions of the communicators. It provides a framework for understanding communication in the text as a form of social engagement. When metadiscourse features are removed from texts, the texts become less personal, less interesting, and less easy to follow. By using interactive language through attitude markers, boosters, engagement markers, or hedges, the writer helps the reader establish a connection with the text. The effective use of metadiscourse is considered a feature of successful writing (Hyland 2005a). Crismore et al. (1993) explain metadiscourse as linguistic elements present in texts, whether they are written or spoken, that do not contribute to the core propositional content but are designed to assist the listener or reader.

Thompson (2001) argues that interaction involves two types of resources: interactive and interactional. Interactive resources guide the reader through the text, while interactional resources engage the reader with the text. Hyland (2005a) talks about these two categories of metadiscourse in similar terms: *interactive dimension* and *interactional dimension*. The first category, the *interactive dimension*, is about how the writer shapes the text to meet the needs of the readers and communicate interpretations and purposes. Table 2.1 shows examples of the interactive dimension, such as transitional markers, frame markers, and code glosses.

Table 2.1: Examples of interactive metadiscourse

Transitional markers	**Frame markers**	**Code glosses**
and, furthermore, moreover, by the way, similarly, likewise, equally, in the same way, in contrast, however, but, therefore, in conclusion	*First, then, ½, a/b, at the same time, next, to summarize, I argue here, my purpose is, there are several reasons why, let us return*	*this is called, in other words, that is, this can be defined as, for example, etc. (e.g.)*

The resources of this dimension address the organization of the discourse regarding the reader. Hyland (2005a) groups them into five broad sub-categories: transitional markers (signaling additive, causative, and contrastive relationships), frame markers (indicating elements of schematic text structure such as item sequence, labeling, and predicting arguments), endophoric markers (referencing other parts of the text to support arguments by referring to earlier material), evidential (representing different sources and establishing authorial command in academic genres), and code glosses (providing additional information by rephrasing or explaining what was said previously to help the reader understand the writer's intent).

Within the second category, the *interactional dimension*, the writer conducts the interaction by engaging with the text, and the metadiscourse in this dimension is evaluative and engaging, anticipating objections from the imagined audience. Based on Hyland (2005a), the rhetorical features in this dimension demonstrate the writer's personality displayed in the text and also alert the readers to the writer's perspective on the content. Table 2.2 provides a few examples of these rhetorical features.

Table 2.2: Examples of interactional metadiscourse

Hedges	Boosters	Attitude markers	Self-mention	Engagement markers
possible, might, perhaps	*clearly, obviously, demonstrate*	*agree, prefer* (attitude verbs), *unfortunately, hopefully* (adverbs), *appropriate, logical* (adjectives)	*I, me, mine, we, our, ours*	*you, your, we* (pronouns that include the reader), *by the way, you may notice* (interjections), *see, note, consider, should, must, have* (directives and modal verbs that show obligation)

The features include hedges (indicating the writer's acknowledgment of alternative perspectives and opinions, allowing information to be presented as opinions rather than facts), boosters

(helping the writer emphasize certainty and narrow the range of possible opinions), attitude markers (demonstrating the writer's attitude towards propositions and also expressing surprise, agreement, importance, obligation, or frustration), self-mention (showing the degree of the writer's presence in the text, often using first-person pronouns or possessive adjectives), and engagement markers (addressing readers by focusing their attention on the message or including them as participants in the text).

Hyland (2005a) makes explicitly clear that the rhetorical features representing each dimension are "meaningful only in the context where they occur and help explain why discourses are structured in a particular way among a particular group of users" (2005a, p.890). As part of their contextual meaning, he finds it important to openly criticize research that uses the lists of discourse markers as prompts for generating data without considering the context of the discourse markers: "This is a lazy approach as the list is just a starting point, the first fix on high-frequency items that commonly function as metadiscourse in a particular register" (2005a, p.125). His advice to researchers is to focus on reading the concordance lines rather than just the frequency data since the rhetorical purposes determine the use of the discourse forms.

Similar to the use of lexical bundles in tracing proficiency, researchers use metadiscourse and corpus linguistics to investigate the development stages in academic writing and discipline variations (Lee and Deakin 2016; Farahani and Sabetifard 2017; Ramoroka 2017; Ho and Li 2018). Two of the studies (Lee and Deakin 2016; Ho and Li 2018) examine the use of metadiscourse language in argumentative essays. Lee and Deakin (2016) compare argumentative texts at different levels of university writing to trace the extent to which successful (essays with letter grade A) and less-successful essays (essays with letter grade B) differ in the use of stance and engagement features, discovering that successful essays carry greater instances of hedging

devices (e.g. *apparently, broadly, largely, typically, unlikely*) than the less-successful essays. Using the metadiscourse features, Ho and Li (2018) focused on argumentative essays in the first year of university and discovered the same trend as Lee and Deakin (2016) that low-graded essays differed significantly in the use of the metadiscourse features from the high-rated essays. The high-rated essays also showed that the features were deployed in different sentence positions, enhancing the persuasiveness of the arguments and demonstrating better skill and knowledge in using the features. Ramoroka (2017) also uses the metadiscourse features to examine variations between texts submitted in education majors and those submitted in media studies. The findings reveal that although students in both disciplines engage with interactional features, those in media studies exhibit a higher frequency of self-mention compared to their counterparts in primary education.

The purpose of this study is to investigate the frequency features, underlining the most common textual patterns in first-year composition writing while also following Hyland's (2005a) advice to consider not only the frequency features but also the contextual usage revealed through the concordance lines. This approach can effectively reveal how writers employ various rhetorical features and engage the audience when conveying their attitudes toward the text or focusing on specific perspectives. The writer's ability to use interactive and interactional rhetorical features can demonstrate effective writing and may also highlight areas where such writing is missing. In this context, examining the frequency features in COMP 101 and comparing them with those in BAWE can be an effective way to identify the successful writing strategies revealed in both groups and areas that require further improvement.

2.5.3 Civility, cohesion, and compression in academic writing

The next feature of interest in academic writing research relates to the discourse choices that characterize successful writing (Aull 2020). Using corpus linguistics, Aull (2020) examines the discourse patterns within and across genres and levels, reinforcing sociocognitive practices and discursive identities. Based on this research, stance words (see Table 2.3) "explicitly frame written ideas, guiding readers to perceive propositional information in particular ways" (2020, p.30). This concept is quite similar to Hyland's description of metadiscourse and how it "offers a framework for understanding communication as social engagement" (2005a, p.215).

Table 2.3: Epistemic Stance Words

Stance Word	Examples Uses	Rhetorical Effect
Hedges	*May be argued, most likely, to a certain extent, to a certain level, it is not clear/not indisputable that, it may be true/may prove true, this might suggest*	Opens dialogic space, leaves room for alternatives downplays certainty and directness
Boosters	*Certainly, a case of, certainly a consideration, certainly a factor, has clear import, there is a clear connection between, clearly agree, clearly appear*	Closes dialogic space, leaves little to no room for alternatives, emphasizes certainty
Generality words	*Never been the case, has always been, people today, anybody can see that, none of the examples, no one knows, nothing can be done, all cases show, never/always, society, the world, Americans, human beings, people*	Shows wide applicability, extrapolates across contexts, time, and/or groups

The metadiscourse analysis in Aull's research results in three distinctive qualities--civility, cohesion, and compression—that scarcely appear in first-year composition writing but show mostly in upper-level writing regardless of discipline or assignment. These three qualities demonstrate how patterned academic discourse functions through interpersonal, intrapersonal,

and cognitive writing domains, giving the text the ability to be diplomatic and, at times, even inaccessible.

Civility or "diplomatic evidentially" as the first quality is mostly demonstrated in upper-level writing. The author explains that because this quality requires both open-mindedness and well-informed conviction, it marks later development stages in writing. Civility is achieved through discourse patterns that include balanced hedging and boosting, concessive and countering clauses, and grammatical structures that support critique or ideas, not people. Civility signals the opening and closing of dialogic space. In opening the dialogic space, the discourse features "acknowledge possible concerns, limitations, and disagreements, and, less often, close the dialogic space to endorse substantiated ideas" (2020, p.6). Based on the same research, this balance is not typically found in first-year argumentative writing. In first-year argumentative writing, students often close the dialogic space, whereas, in advanced writing, the tone of civility is a common characteristic regardless of the assignment genre.

Cohesion, the second rhetorical marker, is demonstrated in explicit coherence across parts of a text through "text consecutiveness and reformation words, often in predictable moves, which showcase writers' reasoning and help create a shared understanding between readers and writers" (Aull 2020). Again, the study indicates a wide range of cohesive ties in upper-level writing, while first-level writing does not indicate the same attentiveness to the readers' needs, as it includes a similar number of cohesive devices.

Compression, the third rhetorical quality, is demonstrated using dense, phrasal detail. Biber et al. (2010) describe academic writing as compressed rather than elaborate, which also makes academic writing "one of the most distinctive registers in English. In its grammatical characteristics, it is dramatically different from all spoken registers and most other written

registers. It does occasionally use spoken features (like first person pronouns), but the basic grammatical structure of discourse is nominal/phrasal rather than clausal" (Biber and Gray 2010, p.18). Such characteristic of academic writing seems to be mostly related to expert writers, where information is packed by using the nominal and phrasal structures, and the first-person pronouns are occasionally used to express authorial power and state positions. In contrast, first-year composition writers may not have the same level of experience to effectively convey large amounts of information through synthesis and may rely more on clausal structures. This study will use the frequency items in COMP 101 and corpus linguistics tools to demonstrate how novice writers typically engage with information and express their positions.

2.5.4 Variation patterns in academic writing

Hyland (1999; 2002b) examines the use of reporting verbs and writer pronouns in academic writing and concludes the existence of considerable differences between empirical sciences and humanities. In a similar study, he (Hyland 2001b) also focuses on self-mention in academic writing as a powerful rhetorical strategy for emphasizing a writer's contribution, with a particular focus on the use of self-citation and exclusive first-person pronouns. Some of the examples related to the use of the first person pronouns show that in philosophy, applied linguistics, and sociology, 75% of all authors' pronouns are *I*, *me*, and *my*, while empirical sciences mostly use plural pronouns, such as *we*, *us*, and *our* (Hyland 2002b). The variation is related to the type of knowledge featured in the disciplines. The disciplines in humanities allow more freedom to authors' perspectives and interpretations, but sciences rely on evidence and facts with less personal connection in expressing positions.

Charles (2007) argues for the use of corpus linguistics in studying variations in patterns across different academic disciplines. She compares a corpus in politics/international relations with one in materials science and notes that the noun + *that* pattern varies based on the cultures, epistemologies, and knowledge-building practices in different disciplines. She identifies five main noun groups: idea (wishes and thoughts), argument (written or spoken), evidence (signs or evidence), possibility (likely or unlikely), and other (fact, case, concern, and sense). In politics and international relations, it is more common to observe noun groups related to ideas and arguments. Charles (2007) also discusses the pedagogical implications of her research, providing guidelines for graduate students when writing academic papers in their specific disciplines.

Carter-Thomas and Rowley-Jolivet (2008) analyze the variations of *if-conditionals* across three genres of medical discourse: research articles, conference presentations, and editorials. Alongside the corpus study, the authors describe how conditionals are featured in textbooks: initial conditional with *if*-clauses appear to be the norm, while conditionals inside the main clause receive almost no attention. On the other hand, in the medical discourse in all three genres, conditionals include initial (59%), medial (8%), and final (33%). Carter-Thomas and Rowley-Jolivet demonstrate the discrepancies between information about conditionals in textbooks, focused predominately on initial conditionals, and real-life use of conditionals across three different genres, promoting the design of instructional materials based on real data in corpus-based studies.

Since this study examines not only the overall frequencies in COMP 101 but also the subcorpora comprised of various essays—such as descriptive, narrative, classification, process analysis, compare and contrast, cause and effect, and argumentative—it allows for identifying variations across different texts. These findings can offer deeper insights into first-year composition writing

and offer evidence for patterns associated with specific genres, aiming to revisit or design more effective teaching strategies.

2.5.5 Informal language in academic writing

The use of informal language in academic writing has been the focus of several studies (Fairclough 2001; Foster 2005), suggesting that academic English has shifted towards a more informal style (Hyland and Jiang 2017). To understand the concept of informality, it is crucial to define its meaning and explore how it is expressed in academic writing. One of the most comprehensive studies on this topic is by Hyland and Jiang (2017), which analyzes a 2.2-million-word corpus of articles from prominent academic journals across four disciplines (applied linguistics, biology, engineering, and sociology). The authors provide a detailed examination of informal features by contrasting them with formal ones. They define informality by referencing Heylighen and Dewaele (1999), who characterize formal style as detached, accurate, rigid, and heavy. In contrast, the informal style is described as more flexible, direct, implicit, and engaged, but it tends to be less informative. According to Hyland and Jiang (2017), informality can be a slippery concept and difficult to define with clarity, making the most practical way to describe it through contrast with formality. Formal writing aims to prevent ambiguity and misunderstandings by following clear standards, while informal writing avoids strict conventions to maintain a relaxed and friendly tone.

The research conducted by Hyland and Jiang (2017) notes an increased rate of informal features over the last fifty years and acknowledges the growing interest in evidence of greater interactivity and expression of personal identity among applied linguists. One possible reason for

this increase might be the greater frequency of informal discourse in research. Hyland and Jiang (2017) use a list of informal features originally developed by Chang and Swales (1999). Some of these informal features include first-person pronouns to refer to the authors (*I* and *we*), unattended anaphoric pronouns (*this*, *these*, *that*, *those*, *it*), split infinitives (*to sharply admonish*), sentence-initial conjunctions (*And*), sentence-final preposition (…*that he can think about*), listing expressions used when ending list (*and so on*, *etc*., and *so forth*), second person pronouns to refer to the reader (*you* and *your*), contractions (*won't*), and exclamations at the end of sentences. The study concludes that the increase in informal features is incremental, with only a two percent rise over the last fifty years when estimated across all published words.

Another interesting finding in Hyland and Jiang's (2017) study is that not all disciplines show the same increase in informal language. While some fields, such as biology (with a 24.8% increase) and engineering (with a 9% increase), have shown a rise in informal features, others, like applied linguistics (with a 10.3% decrease) and sociology (with a 3% decrease), have experienced a steady decline in this aspect. These differences may suggest that each field has its unique communication expectations and norms. The increase in informal language in biology and engineering may indicate a shift towards more accessible and engaging communication with readers, while fields like applied linguistics prioritize precision and detail in expressing ideas.

While looking at the use of informal features in professional writing, it is interesting to see the trend of informal language in student writing at university. One such study by Lee, Bychkovska, and Maxwell (2019) explores the same informal features proposed by Chang and Swales (1999), but this time, in the context of the argumentative genre in the university setting. The authors discover that senior students who have English as a second language frequently use the anaphoric pronoun *it* and the second person pronouns *you* and *your,* while freshman students

with English as their first language use a wider variety of informal language. This suggests that both groups make use of informal language to some extent and may require instruction in using the correct stylistic features in academic writing.

Continuing the exploration of informal features, the study of Liardet et al. (2019) focuses on the use of informal features in undergraduate writing and places the study of formality as a central concept in EAP. The informal features they investigate include the use of informal vocabulary (*plenty of*, *lots of*), abbreviations (*Aussie*), contractions (*didn't*, *can't*, *haven't*), and personal pronouns (*I*, *you*, *we*, *our*, *my*). Liardet et al. (2019) use a corpus comprising 140 student essays that are analyzed and rated for formality by EAP instructors. The results of the study conclude that in only 20 of the essays, the instructors show consistent ratings. Out of these 20 texts, 12 were rated as informal, and 8 were rated as formal. Interestingly, only fourteen percent of the total essays are rated consistently, which speaks to the challenge of creating a clear dividing line between formal and informal language. The researchers also agree with the difficulty in finding the difference between formal and informal language in academic writing by emphasizing formality as "the lack of a clear definition of formality and the range of instructional guidelines mostly focusing on discrete features to avoid can be problematic and limiting for students and EAP instructors alike" (Liardet *et al.* 2019, p.12). These findings highlight the difficulty of teaching and evaluating formality in academic writing, suggesting that the line between formal and informal language is becoming less distinct. This indicates a broader trend towards more flexible and varied language use in academic contexts. In this context, the use of informal language emphasizes the need for guidelines and further research to understand the shift towards informality and the necessity to update current guidelines. This study seeks to address the necessity for further research by analyzing the frequency patterns in COMP 101 and identifying

any informal elements used by first-year composition writers. It also aims to explore any differences observed in the writing of upper-level students.

2.6 Conclusion

The purpose of this chapter was to unpack the "black box" of academic writing and discuss different teaching traditions in the field of English for Academic Purposes (EAP), the role of English as the global language of academia, as well as the major stakeholders within the academic writing field. Also, the chapter demonstrates that effective strategies provided by EAP are beneficial not only for language learners but also for native language speakers in improving their academic writing. The last section of the chapter explores various features of academic writing, such as lexical bundles, metadiscourse, civility, variation patterns, and informal language. These features demonstrate language in context and highlight the role of corpus linguistics in uncovering salient patterns and providing information for educators to update manuals and guidelines effectively. As John Swales states, "the training of people to process and produce academic and research English remains a major international endeavor, whether in contexts where English is a first language, a second language, or a foreign language" (1990, p.1). This study attempts to contribute to this endeavor by investigating the most frequent textual features in first-year university writing through the tools of corpus linguistics and comparing these features to those of the upper level. This literature review provides the broader context of academic writing, while the next chapter will discuss the theoretical framework of this study, contextualizing it within existing theories and providing a point of reference.

Chapter 3

Theoretical Framework

3.0 Introduction

The primary objective of this chapter is to establish the theoretical framework for analyzing the most frequent linguistic features in first-year composition writing. The theoretical framework seeks to ground this study within the established theories and provide an appropriate lens for interpreting the research findings and their relation to the existing theories. In determining the theoretical lens, it is essential to consider that COMP 101 is a collection of written work students submit during their first year of university composition writing. Since the study relies on corpus linguistics as methodology, it focuses on text analysis to investigate the most common linguistic features demonstrated in first-year university writing. In such analysis, Stubbs (1996) emphasizes the importance of defining the textual data, the setting of its production, and its reception. A central aspect of this text production relates to the common principles encompassing the various disciplines and regulating the text for all students, creating an initial lens in the textual features. As with other corpus-based studies (Hyland 2002a; Biber and Gray 2010; Aull 2015), this research will rely on genre and its related concepts to define the texts and investigate the corpus findings that relate them to specific types of language use. This explores the differences between genre and register, defines how these terms will be used in the study, discusses relevant scholarly research, categorizes genres, and proposes a theoretical framework model for this study.

3.1 Register, genre, and their related concepts

Before delving into the definitions of register and genre, it is essential to briefly list their origins and discuss the philosophical schools of thought supporting them. In language description and analysis, *mentalism* and *functionalism* are two primary schools of thought (Nunan 2008, p.18).

Mentalism regards language as a psychological phenomenon embedded in the human brain, focusing on the abstract rules governing well-formulated sentences (Pinker 2010). On the other hand, functionalism views language as a social construct with meaning as its focal point (Halliday and Matthiessen 2014). The most comprehensive approach to functionalism is systemic-functional linguistics (SFL), which gives rise to the concepts of genre and register. Both register and genre have played important roles in text classification to enable meaningful interpretation of corpus findings and either have focused on specific genres of texts (Steen 1999; Gere *et al.* 2013; Hyland 2015b) or different types of registers (Zacharof and Charalambidou 2018; Fang *et al.* 2020) instead of language in general, thus, leading researchers to make observations about patterns within particular groups of texts. However, genre and register have often been used interchangeably, mainly due to the overlap in their meaning (Lee 2001). Because of their terminological differences and overlapping, this section discusses the meaning behind genre and register and their distinctive characteristics. Additionally, it differentiates these terms from other related terms—text types and style—that may also contribute to the terminological quagmire.

3.1.1 Genre, register and style

In an effort to define the terminology in this chapter, the first distinction to be made is between the terms register, genre and style. The term style is included in this discussion because genre is sometimes mentioned alongside style (Swales 1990; Lee 2001; Biber and Gray 2010), necessitating an exploration of the boundaries and possible overlaps between these somewhat interconnected terms. The discussion of the three terms is important to the theoretical framework of this study, which is based on a corpus of various essays aiming to analyze the most frequently

used textual features. A secondary goal of the study relates to the examination of the textual features across the different essay genres, such as descriptive, narrative, classification, process-analysis, compare and contrast, cause and effect, and argumentative (see also Chapter 4). This section provides a discussion of register, genre, and style in an effort to define the use of genre and how it differs from the other two terms.

Since the most comprehensive treatment of genre and register is offered in SFL, it is appropriate to start their examination from that perspective. Halliday and Hasan (1985) and Martin (1985) discuss both terms as two essential semiotic systems in understanding and investigating texts. They see register as defined by J.R. Firth in the categories of *field* (the institutional focus of text), *mode* (the medium through which text is realized), and *tenor* (social distance between speaker and addressee). Contrary to register, Martin (1985) views genre as a more abstract concept encompassing literary and non-literary forms addressing how language accomplishes things. The connection between genre and register, in Martin's (1985) words, is the genre's responsibility "to constrain the possible combinations of field, mode and tenor variables used by a given culture" (p.250) and its ability to represent "at an abstract level the verbal strategies used to accomplish social purposes of many kinds" (p.251). Within this description, Martin considers that genre goes above and beyond the register and facilitates the language processes and communication, constraining the field, mode, and tenor, or the register.

In critical discourse analysis, Kress (2012) uses register in similar ways to the SFL terms but does not support the all-encompassing nature of genre with the argument that such a multi-faceted treatment is "uncontrollable" (p.31). His interpretation of genre covers only one aspect of the textual organization, "the part which has to do with the structuring effect on the text of sets of complex social relations between consumers and producers of texts" (2012, p.33). Even though

he disagrees with the SFL description of the genre, his stance does not seem to present a solid argument against genre boundaries and territory. Lee (2001), in his discussion of genre and register, evaluates the disagreement as insubstantial since, in both descriptions, the genre is situated within the broader context of situational and social structure.

Biber and Conrad (2009) discuss register, genre, and style, defining register as "an analysis of linguistic characteristics common in a text variety with analysis of the situation of use of the variety" (p.2). In this work, register is viewed as a sample of text excerpts, focused on the linguistic characteristics that include any lexico-grammatical features. The distribution of linguistic characteristics is focused on the frequent and pervasive features in the texts from the variety. The interpretation of linguistic features is based on their communicative functions in the register. On the other hand, Biber and Conrad (2009) see genre as being focused on complete texts with linguistic characteristics, including specialized expressions, rhetorical organization, and formatting. The distribution of these linguistic characteristics includes features that usually occur once in a particular place in a text. The interpretation of these features seeks to understand their conventional association with genre, which is often the format and not the function. They view style as similar to register in its focus on the linguistic features but different in its purpose; style views the linguistic features as preferred, while register views the linguistic features as functional. In a later study, however, Goulart, Biber, and Reppen (2022) assume a more interchangeable definition of register and genre, which views texts as both register and genre, since "register distinctions are fluid with fuzzy boundaries and extensive internal variation, and then developing empirical methods to describe the nature of that continuous space of variation" (Goulart *et al.* 2022, p.3). Within this more fluid view of register and genre, they (2022)

acknowledge that the two terms may not be distinctly separate elements as thought previously but interconnected.

Another perspective of the terms register, genre, and style is offered by Swales (1990), who sees register as a well-established and central linguistic concept that defines the larger context, while the genre is part of different discourse communities and their specific communicative purpose. These discourse communities are formed around common goals, specialized language, and expertise, such as business, linguistics, or engineering. In Swales (1990), the term style is part of the genre concept where the genre's rationale determines the style's choice, and genre exemplars carry similarities in structure, style, and content. Thus, style is considered a visible characteristic in genres, which operate alongside structure and content and serve the rationale recognized by the given discourse community.

The work of Bhatia (1993) is closely related to the view that Swales (1990) holds of genre analysis. Bhatia's (1993) initial interest is focused on legal discourse and the differences between professional genres, which ties his work to genre theory in discipline-specific discourse communities. Bhatia (2004) defines genre as the use of language in a conventionalized setting with specific communicative purposes, which leads to stable structural forms with constraints on the lexico-grammatical features. Even though the linguistic features in different genres function under the institutionalized goals' constraints, they are not static but dynamic, and in this sense Bhatia's (2004) view resonates with the SFL view of the genre. Bhatia (2012) explains the emergence of genre analysis as a theory with an emphasis placed on English for Specific Purposes (ESP) as it relates to producing and communicating meaning in professional contexts, which later motivates critical genre analysis as a step further to define professional practices in academic settings. He distinguishes critical genre analysis from critical discourse analysis,

relating the former to what is explicitly said in genres by professional writers to understand professional practices and the latter as cultural critique, which "focuses on social relations, including race and gender" (Bhatia 2004, p.23). Similarly to Swales (1990), Bhatia (2012) distinguishes register from genre based on Halliday's terminology - field, mode, and tenor of discourse. In his view, disciplinary discourse as a concept interacts both with register and genre. According to Bhatia (2012), disciplines are identified in terms of their content and thus by the field of discourse rather than by all three variables: field, mode, and tenor. On the other hand, the genre remains in its distinct category and relates to discipline and register.

Another prominent linguist who follows Swales (1990) and Bhatia (1993; 2004) is Hyland (2002a; 2003; 2004; 2005a; 2015b), who agrees with the Hallidayan terms of genre and register and focuses his work on the particular linguistic features demonstrated in genre as a goal-oriented social process that is part of disciplinary discourse communities. It is a social process because genres are created by members of a culture who interact with each other to achieve particular goals or purposes. These goals can be anything from sharing information to persuading an audience to take a specific action, which denotes that genres meet particular purposes and are shaped by their social context. This means community members recognize that genre through similarities in texts that correlate with specific purposes, and readers expect similarities in texts with the same purpose.

To conclude the discussion on genre, register, and style, it is worth noting that most scholars agree on the definition of register as expressed in SFL terms as an instantiation of the meta-functions field, tenor, and mode tied to different societal situations. The scholarly differences towards the SFL understanding of register are expressed by Kress (2012), in the field of critical discourse analysis. In this particular view, the territory of genre is defined as only one part of the

register, the part that textually structures the impact of complex social relationships between text consumers and producers. This supposedly different terminological difference is not considered to be substantive by Lee's (2001) assessment of register and genre in the corpora text categorization schemes with the argument that both Martin (1985) and Kress (2012) place genre as an instrument of navigating social structure. On the other hand, style is treated in Swales (1990) as part of the genre concept and inhabits a visible aesthetic characteristic of the genre, while in Biber and Conrad (2009), the style is similar to register since it relates to the lexico-grammatical features writers choose in creating texts at language, not text level. Even though the style is recognized as belonging to different camps—genre or register—in both perspectives, it refers to the writer's choice of words.

3.1.2 Genre and text types

Genre is related not only to register and style but also to text types. The confusion between the two terms results from the different criteria surrounding them. Biber (1988) distinguishes between genre and text types by associating genre with external criteria, such as the purpose of use, while associating text types with internal textual characteristics, such as lexical and grammatical features. According to Biber (1988), the external criteria include the intended audience, purpose, and conventionally or culturally recognized groupings of text, and text types are "groupings of texts that are similar in their linguistic form, irrespective of genre" (Biber 1988, p.170). On the other hand, the internal characteristics focus on the lexico-grammatical features of the text, and based on Biber (1988), the relationship between genre and text types seems quite fluid. Thus, he observes that a genre of academic exposition may exhibit features of narration, which might be considered more typical to fiction than academic writing, so the genre

might be academic exposition, but the text type is academic narrative. In discussing the distinctions between genre and text types, he leaves the conversation at the criteria level—external versus internal—without defining the exact recommendations for identifying text types.

In an attempt to clarify text types, Paltridge (1996) refers to them as rhetorical patternings or types, such as narration, description, or argumentation. He observes that more than one genre can carry the same text type; for example, police reports and advertisements may share description as a text type. In addition, a single genre may have more than one text type, such as the genre of formal letters, which may combine exposition and problem-solution. He maintains that these discourse or rhetorical patterns are based on internal text criteria but also notes that despite the structural differences, they represent complementary perspectives of the texts.

Contrary to Paltridge (1996), Stubbs (1996) uses genre and text types interchangeably to suggest that they carry the same meanings and cover the same ground when discussing the text and its analysis, a use that Lee (2001) defines as common with most other linguists. One of the main reasons for the interchangeable use between genre and text types is most probably not having established parameters for internal criteria, which leads to variations in the use of the two terms. In this regard, Lee (2001) makes a helpful observation that there are no explicit recommendations to identify the internal criteria that "cut across traditionally recognizable genres" (p.39) and concludes that all corpora currently use only external standards to classify texts, making genres a useful text classification.

Evaluating the distinctions between genre and text type, it is fair to say that linguists agree that genres are based on external criteria. In contrast, text types are based on internal criteria, but not having a clear definition of the internal criteria makes it problematic to define text types. A practical way to look at the differences between genres and text types is to classify corpora based

on external criteria, such as genre, and focus on the internal features, such as lexical and grammatical characteristics. While Paltridge's (1996) study contributes to a deeper understanding of the genre's structural patterns and complex nature, providing practical distinctions between the two terms, Stubbs (1996) demonstrates a synonymous use and views them as encompassing the same content. Understanding the historical background between genre and text type enables this study to establish a well-defined theoretical framework and basis for the subsequent analysis of findings.

3.1.3 Genre and its related terms in this study

This study, which explores the prevalent linguistic features in first-year composition texts compiled in COMP 101, classifies the texts by genre (Section 3.3 discusses the categorization of genres and their hierarchical structure). Genre is a more appropriate choice than register because it looks at the text as a member of a category, while register looks at the text as language. In making this distinction, it is useful to use Ferguson's (1994) distinction between the two terms: register is a "communicative situation that recurs regularly in a society", and genre is a "message type that recurs regularly in the community" (p.21). To illustrate this relationship between the two terms, it might be useful to mention the existence of a formal register in which official documents, such as affidavits, wills, and courtroom debates, represent some of the genres.

Another confirmation of the genre usefulness in the text categorization as applicable to this study is Lee's (2001) evaluation, proposing genre as the level of text categorization that is "theoretically and pedagogically most useful and most practical to work with" (2001, p.37). This differentiation between register and genre highlights the register's connection to societal communication situations and the genre's association with recurring message types in a

community, which stresses the genre applicability in text analysis. The text analysis of this study investigates first-year composition writing, represented by the essays collected in COMP 101.

Based on the SFL school of thought, the terms register and genre overlap, and Lee (2001) explains this overlap as two different ways of looking at the same object. In this sense, register refers to text as language, which represents a variety according to a specific use, which might be seen as "somewhat static and uncritical" behavior (Lee 2001, p.46). In contrast, genre is "more dynamic" (Lee 2001, p.46) based on its internal structure or its various categories established through cultural consensus. It is subject to change over time as generic conventions defining genres are contested and undergo changes.

The connection between genre and register is practically demonstrated by Lee's (2001) theoretical definition that genres represent specific instances of registers. To unpack this connection further, Lee uses Eggins and Martin's (1997) conceptual framework, which views the linguistic features within the text as encoding the contextual dimensions, both of "its immediate context of production (i.e., register) and of its generic identity (i.e., genre)" (p.237). This implies that language is not uniform across situations but varies based on the context in which it is used. Thus, language use is regarded as a tool that is selected and adopted based on particular situations, and the linguistic features express these various contextual situations. The linguistic features are contained within the genre and serve specific purposes. Thus, the architecture and substance of genres allow a practical lens for examining the linguistic features in COMP 101 (Chapters 5 through 8 discuss the most frequently used linguistic features in COMP 101). By using genre lenses, it is possible to examine these linguistic features in the context of first-year composition writing and gain an understanding of the writers, as well as compare them to upper levels of university writing.

Within the discussion of the linguistic features displayed in the genre, it is unavoidable to notice the various styles demonstrated by the writers. Despite the differences of opinions between Biber and Conrad (2009) and Swales (1990) concerning the affiliation of style, the former associating it with register and the latter with the genre characteristics, it is reasonable to conclude that it is related to the aesthetic choices of the writers. Another way to look at these choices is through Lee's (2001) definition of style as the internal properties of the text that reflect the individual writer's word choices. Writers vary their use of language between formal, informal, colloquial, intimate, or humorous styles. The current study will not focus on the individual writers' styles but on the most common linguistic features that COMP 101 demonstrates in the wider context of academic writing. Nevertheless, acknowledging the significance of style is essential when delving into the textual attributes that define COMP 101's writers since style may play a crucial role in unveiling any distinctive expressions and offering insights into the writers' personalities within the context of first-year university writing. The next section aims to discuss the importance of the conventionalized setting of academic writing shaped by the common conventions, academic disciplines, and text producers.

3.2 Genre in the conventionalized academic setting

The setting plays a fundamental role in creating meaning and text by expressing its importance in social experiences as a driving force in language learning (Halliday and Hasan 1985). By recognizing the significance of the context, writers learn to adapt their language choices when creating content to meet specific purposes within their community or disciplines. Genre within this framework depends on the user's experience within the given context, which suggests that as the user gains more experience, their interaction with genre becomes more proficient.

Consequently, the text producers play an important role in creating content for specific purposes within the community, in this case, academic disciplines. This aligns with Eggins and Martin's (1997) view of genre as "staged, goal-oriented social processes" (p.243) that members of the community or academic disciplines use in specific contexts to achieve different purposes. In this regard, Bhatia (2004) stresses the importance of conventions in shaping genre within situational contexts, emphasizing that "genre essentially refers to language use in a conventionalized communicative setting" to address communicative goals within structural categories "by imposing constraints on the use of lexico-grammatical as well as discourse resources" (p.27). In academic writing, text producers learn to adjust their language according to common conventions and specific goals within their disciplines. This reflects an organic interaction between conventions, academic disciplines, and text producers, all of which contribute to shaping genre. The following sections will explore each of these three factors—common conventions, academic disciplines, and text producers—in more detail.

3.2.1 The role of the common conventions

Upon entering the university context, all students must engage with complex ideas and theories, which requires not only critical-thinking capacity but also following specific academic conventions. Every academic discipline follows specific writing standards, which govern the way scholars communicate with one another in written form. Even though disciplinary discourses demonstrate nuanced use of language features and format variations, there are several general conventions or principles that are commonly practiced across disciplines and can be identified. In her study, Bennett (2009) discusses the regulatory criteria of academic writing as outlined in academic manuals to demonstrate the expected standards that provide guidelines and

ensure best practices in students' performance across educational institutions. The study finds that despite the disciplinary differences, manuals agree on common conventions regarding academic writing, such as general principles, text structure, grammatical issues, and bibliographic references.

The general principles correspond to clarity, conciseness, and objectivity (Macmillan and Weyes 2007; Bennett 2009). Clarity requires that writers in academic settings avoid unnecessary language and ambiguity to help readers focus on the main ideas and follow the arguments. Concise wording targets efficiency and economy of expression to promote logic and understanding. Objectivity focuses on evidence-based arguments without personal bias and subjective interpretations, avoiding emotional language and preferences. These general principles are relevant to the curriculum of first-year composition writing, which also emphasizes clarity, conciseness, and objectivity. At this stage, students begin to wrestle with a new level of language complexity when faced with academic literature and writing. They strive to understand new terminology and condensed texts (see Chapter 2.5.3 for details on compression) and produce clear, concise, and objective texts.

The text structure is a critical aspect of academic writing that refers to the overall organization of a written piece, including the introduction, body paragraphs, and conclusion (Halliday and Hasan 1976; Northedge 2005; Bennett 2009). The introduction sets the stage for the rest of the paper, highlighting the main arguments. The body paragraphs present the main arguments and supporting evidence logically and organized, using topic sentences, supporting details, cohesion, and unity. The conclusion summarizes the main arguments presented in the body paragraphs. Effective text structure ensures that the reader can understand and engage with the writer's ideas and arguments, achieved through careful planning and attention to detail. The text organization

that navigates the reader is a particularly challenging area for first-year composition students, as the analysis in Chapter 5 reveals, which stresses the importance of text structure as a general principle in academic writing.

Effective academic writing requires attention to grammatical correctness that generally covers sentence structure and parallelism (Barrass 2002; Bennett 2009). Sentence structure addresses engagement with readers and recommends that writers avoid text monotony and use varied sentence lengths and structure. The varied sentence structure implies short, simple, and complex compound sentences. Parallelism refers to using grammatically similar structures within a sentence that convey a sense of balance and symmetry. Another aspect of grammatical standards is proper punctuation, avoiding sentence fragments and run-on sentences, and following subject-verb agreement and pronoun usage rules. For example, *The Little Brown Handbook*—one of the popular course materials on academic writing in U.S. universities—recommends that students primarily utilize the third-person perspective and avoid the first and second-person pronouns (Fowler and Aaron 2016) to obtain objectivity in their texts.

An essential aspect of academic writing includes referencing and citations since they allow writers to acknowledge and credit the sources used in the texts (Bennett 2009). The specific style guides within disciplines ensure consistency across texts by establishing rules for in-text citations, footnotes, and bibliographic references to help readers identify the sources in the text and verify the credibility of the information presented. By following the required protocols for referencing and citations, writers can demonstrate their understanding of the academic conventions of their discipline and show respect for the work of experts in their field. This helps to build credibility and authority for the writer and strengthens their argument. This is one of the

challenging areas for first-year composition writers in learning the appropriate referencing styles and integrating research into their texts as Chapter 8 reveals.

Bennett's (2009) study provides evidence supporting shared conventions that regulate the organization and structure of genres in the institutionalized academic setting such as the university context. While the discipline-specific communities have their expectations and guidelines within the given subject areas, these generic academic conventions, such as *The Little Brown Handbook* (Fowler and Aaron 2016) or *About Writing* (Jeffrey 2016), provide common academic guidelines for structuring texts. These writing guidelines cover a wide range of topics, including concise writing, establishing logical connections, maintaining consistent verb tenses, conducting thorough research, proper citation practices, and ensuring objectivity in writing.

These generic guidelines might be useful to all university students, but particularly for those who are just starting out on their academic journey and may not be familiar yet with the specific regulations of their discipline. Different disciplines have different writing requirements and expectations, which can be challenging for students as they learn new vocabulary and writing styles within their majors. These academic disciplines represent communities where students learn to engage in professional dialogues and develop experiences related to their chosen disciplines. Even though first-year students are still adjusting to their chosen fields of study or still looking for their specialized area of study, these disciplines play an important role in text production. The next section discusses the importance of academic disciplines as they relate to genre.

3.2.2 The role of academic disciplines

The academic disciplines provide the socio-cultural setting of genres and, as such, play a crucial role in the analysis of genre. The topic of academic disciplines has gained a prominent place in several studies, such as Swales (Swales 1990), Bhatia (1993; 2004), and Hyland (2001b; 2002a; 2004; 2015b). Swales (1990) develops the idea of discourse communities, which can relate to academic disciplines. He separates the term discourse communities from speech communities to emphasize the critical role of shared literary activities between members within the shared cultural setting. According to Swales (1990), discourse communities consist of broadly agreed public goals, intercommunication among their members, one or more genres related to the discoursal aims, specific lexis, members with suitable degrees of the subject matter, discoursal expertise, silent relations, and horizon expectations. These principles focus on the significance of the community's expectations regarding genres and the constraints imposed on text producers within these communities, with the primary goal of helping students flourish in their various educational environments. Both spoken and written communication play a role in the initiation of novice members into the standards upheld by the community (Swales 1990). Within composition modules, this type of communication is upheld through a blend of in-class guidance and assigned textbook readings aimed at helping students comprehend customary practices and writing conventions.

Bhatia (1993; 2004) expands on Swales' (1990) framework and discusses the role of the linguistic features related to academic disciplines, shaping professional genres, and focusing on academic and professional expectations. In his view, disciplines are "primarily understood in terms of the specific knowledge, methodologies and shared practices of their community members, especially their ways of thinking, constructing and consuming knowledge" (Bhatia

2004, p.35). An important part of Bhatia's view as part of the disciplinary variation is that "genres cut across disciplines" (Bhatia 2004, p.35). An example of genre-transcending multiple fields is the textbook genre, which shares lexico-grammatical features across disciplinary boundaries. These linguistic features might be difficult to observe in first-year university writing, such as COMP 101, but may be seen in later stages, such as the texts in BAWE. For example, in Section 7.4, the texts in BAWE reveal specialized language features related to medical care, such as "Because you are still having difficulty swallowing after your strokes, we would like to get a swallowing assessment to make sure that you are able to get the right intake of food and drink to try and prevent this from happening again" or information technology as in "Basically when you type the URL address into your web browser you sent a request using HTTP. In this case, a browser is the HTTP client and a web page server is the HTTP." The texts in COMP 101 do not demonstrate this level of engagement in given academic disciplines. In this regard, considering the role of disciplinary variations as a factor influencing genres provides insight into the differences between first-year composition students and upper-level students.

According to Hyland (2015b), genre represents one of the most powerful tools in understanding and teaching language. He uses corpus linguistics to examine the relationships between community expectations and the individual writer and demonstrates the options of authorial identity and engagement across disciplines (K. Hyland 1999; Hyland 2001b; Hyland 2015b). Hyland (2015b) shows that authors in soft knowledge disciplines use the first person singular and plural forms (*I* and *we*) three times more frequently than in science disciplines. The choices of authorial identity are not random but motivated by the degree of personal engagement with the reader. In soft sciences, the writers demonstrate claims through personal convictions, while in science, they emphasize their contribution to the field. This relates to Bhatia's (2004) view that

the expectations of the communities constrain genres and show contrasting features between disciplines.

Within these processes, novice writers like first-year composition students do not immediately show a command of the academic language and the discipline-specific vocabulary but gradually acquire the patterns that encode "disciplinary preferences for opinions, arguments, styles, attitudes to knowledge, and the relationship between individuals and between individuals and ideas" (Hyland 2015b, p.32) Through genre, researchers can understand how writers understand their immediate setting and the more considerable constraints of the broader academic community that impact the interaction. During the first year of composition writing, the students are introduced to the specificity of academic writing and get familiarized with the common academic expectations that cut across disciplines and serve as boundaries.

3.2.3 The role of the text producers

Text producers are another important factor in understanding the situational context based on the writers' interaction and experiences with conventional practices when expressing their individual positions (Hyland 2015b). The ability to comprehend and adhere to the fundamental principles of academic writing is crucial for students' success in their academic programs, regardless of their native or non-native status (Biber 2006). Academic writing covers a vast ground of subjects and topics, such as undergraduate students (Aull 2015; Lee and Deakin 2016; Staples and Reppen 2016; Durrant *et al.* 2019), postgraduate students and expert writers (Aull and Lancaster 2014; Gil and Caro 2019), and disciplinary discourses (Hyland 2004; Cheung and Lau 2020). In this array of writing levels and fields, expert writers, undergraduate, graduate, or postgraduate student writers represent various categories engaged in academic writing within different

disciplines. Discipline discourses demonstrate language features such as specific reporting verbs and authorial stances that vary between disciplines in humanities, science, and business (Hyland 2001b; Harwood 2005) and seem characteristic of expert and upper-level writing. In contrast, undergraduate and first-year writing can hardly be characterized as disciplinary-specific, especially at the start of academic writing. Thus, many writers fall into a large group still developing their academic writing skills.

As text producers, writers play an essential part in text development and demonstrate their perception of the community's expectations through the lexico-grammatical choices in the texts. These rhetorical choices might demonstrate an objective and formal voice by framing positions using the third person and avoiding conversational language. On the other hand, these choices may reveal a personal voice indicated by the frequent use of the first person or addressing the reader with the use of the second person and maintaining a casual tone rather than serious and academic, underlined by objectivity. In this regard, genres help researchers understand texts as the product of the text producers as well as the common conventions and discourse communities. In COMP 101, the text producers are the first-year composition writers who are new to the writing conventions. During the first year of their university writing, these students are inducted into the broad academic community and begin the journey of their discipline-specific practices. COMP 101 captures this first year of writing and aims to trace the most common linguistic patterns that the students exhibit in the essay assignments. Thus, the essential investigative role that genres play in the texts emphasizes the need for their classification, which is the purpose of the next section.

3.3 Categorization of genres

One main challenge with genre categories is the existence of many different levels of generality One of the best examples of that generality is the genre of academic writing, which encompasses a vast area of texts and represents a very high-level genre category within which many sub-genres vary (Lee 2001). To categorize genres, Paltridge (1995) underscores the importance of prototypicality as a model to categorize items and concepts to represent a hierarchy of their relations and its usefulness to genre taxonomy. Similarly, Steen (1999) uses prototype theory to categorize genres, creating three categories: basic level, subordinate, and superordinate. The basic level represents the informational level at which the concepts are commonly known and understood, while subordinate and superordinate concepts are less differentiated from their corresponding alternatives. The prototype example Steen (1999) provides is the word *fruit* for the superordinate level, *apple* for the basic level, and *Golden Delicious* for the subordinate level. The hierarchical level begins with the superordinate, the broadest category. Then it is followed by the basic level, often considered the middle level, and concludes with the subordinate level representing the fine distinctions of the preceding categories. Lee (2001) illustrates the hierarchical structure of genre with the example of literature. At the superordinate level, he places literature as the most comprehensive category, followed by the basic level, represented by forms such as novels, poems, and dramas. Lastly, the subordinate level is illustrated by western, romance, and adventure. According to Steen (1999) and Lee (2001), the members of the basic-level category carry the maximal distinctive features, which make them most easily recognized.

This study uses academic writing as the superordinate level, or the broadest category of genre (Lee 2001; Hempel and Degand 2008; Kostrova and Kulinich 2015). The classification of academic writing as a genre may seem overly broad, given genre theory's definition of genre as

"a class of communicative events, the members of which share some set of communicative purposes" (Swales, 1990, p. 58). The diversity of academic texts—including essays, reports, reviews, and beyond—might suggest alternative categorizations, such as a register shaped by field, mode, and tenor (Halliday & Hasan, 1985) or a collection of distinct genres. To address this concern, this study frames academic writing as a superordinate genre, aligned with Lee's (2001) concept of academic prose, encompassing texts produced in academic settings that share the communicative purpose of creating and sharing knowledge. Despite their variety, these texts are unified by shared conventions, including clarity, conciseness, text structure, and referencing (Bennett, 2009), which support social goals across disciplines. Drawing on Martin's (1985) functional approach, this classification operates at an abstract level, capturing verbal strategies that achieve knowledge-building goals within academic discourse communities (Swales, 1990). Lee's (2001) hierarchical framework also accommodates an internal variation using the basic-level genres like the academic essay, subdivided into narration or classification, specifying distinct communicative purposes while still containing the superordinate genre's conventions.

This genre approach enables the analysis of the register, discipline, and writer's identity (Bhatia, 2004; Hyland, 2000) displayed by the first-year composition writers in their texts, as the model for the theoretical framework shows in Figure 3.1. To clarify the boundaries of this definition, academic writing, in this case, excludes administrative documents (e.g., syllabi) or informal communication (e.g., discussion posts). While a register-based approach emphasizes linguistic features, it overlooks the rhetorical and social functions central to genre theory (Hyland, 2004), making it less suited to analyzing the communicative purposes and disciplinary contexts of student writing. A limitation of this superordinate genre framing is the heterogeneity of academic writing, as disciplinary differences (e.g., humanities vs. sciences) may challenge its unity.

Additionally, the genre approach may underexplore linguistic nuances compared to a register perspective. However, Lee's (2001) hierarchical structure and shared conventions ensure a coherent framework for analyzing the linguistic features of first-year student essays, comparing them with upper-level texts, and highlighting unified communicative strategies across academic contexts.

The basic level, or the middle, is represented in this study as the essay. Nikolaev et al. (2021) describe the academic essay as the most common basic genre level used in university settings, and it is also the genre that most appropriately defines the text in COMP 101. The essay, the primary genre in COMP 101, encompasses seven subtypes or subgenres: descriptive, narrative, classification, process analysis, comparison and contrast, cause and effect, and argumentative. The composition course views the essay as a focused piece of nonfiction writing that explores or presents the writer's perspective on a topic, organized around a central idea or thesis. The essays may vary in length and structure, ranging from a single, well-constructed paragraph to multiple paragraphs. Regardless of length, an essay must include a clear thesis, supporting evidence or reasoning, and a conclusion. The seven subtypes or subgenres feature as the main seven assignments in COMP 101 that are spread throughout a fifteen-week semester. Through these assignments, students acquire the necessary skills to analyze and respond to academic tasks by recognizing and using descriptions, narratives, classification, process analysis, comparison and contrast, cause and effect, and argumentation. Also, the students learn how to write focused papers, clearly organized, correctly phrased, and referenced. Even though COMP 101 does not focus specifically on research papers, it does encourage students to use outside sources and introduces students to the university library database and free web citation management tools.

The subordinate level is represented by different types of essays, such as descriptive, narrative, classification, process, compare-and-contrast, cause-and-effect, and argumentative. Categorizing these various essay patterns is important, and it is necessary to refer to the earlier Section 3.1.2 in this chapter, which shows the varied perspectives of linguists related to such essay categories. As discussed previously, Paltridge (1996) refers to them as text types that show the internal linguistic features of essays, while Stubbs (1996) and Lee (2001) view text types and genres as the same conceptual territory, in other words, as the same concept. In this regard, this study aligns with Stubbs (1996) and Lee (2001), viewing the descriptive, narrative, and other subordinate categories as genres. For clarity, the study defines each of the subtypes or subgenres below, based on the textbook used in COMP 101, and it also includes the tasks or the assignment instructions. All tasks allow students to choose a topic that interests them, encouraging personal engagement and the use of prior knowledge.

The descriptive essay is the only assignment that requires a single paragraph. It emphasizes a dominant impression, conveying a sense of a person, place, or object through sensory details that engage imagination, association, and symbolism (Rys 2019). It requires a central idea, supporting details, logical organization, and a conclusion.

The assignment instructions include the following: Write a one-paragraph descriptive essay describing a place, thing, or person.

Instructions:

- Include a clear topic sentence that creates a single dominant impression.
- Make sure all sentences relate to the topic sentence.
- Arrange the details in a logical (such as spatial) order.

- Use nouns, verbs, and modifiers that are specific and appeal to the senses (sight, hearing, touch, taste, smell).
- Include a concluding sentence that "wraps things up."
- The paragraph needs to include at least seven sentences. There is no minimum word requirement, but adequately developing a paragraph means adding good, specific details.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.
- The final draft should contain a minimum of grammar, usage, and punctuation errors.

The narrative essay requires five paragraphs focused on a meaningful event and people (Rys 2019). It requires a fluid and clear organization, developed through the introductory, body paragraphs, and conclusion. The narrative may include one to multiple characters, dialogue, perspective, and a setting.

The assignment instructions include the following: Write a narrative essay in five paragraphs.

Instructions:

- Give the essay an interesting title.
- Build an appealing introduction and a strong thesis.
- In each body paragraph, build clear topic, supporting, and concluding sentences.
- Tell the story in a logical (such as chronological) order.

- Provide all necessary background information. Describe important scenes and people in enough detail using specific nouns, verbs, and modifiers. Be sure to include examples, names, numbers, dates, and appeals to the senses of sight, sound, smell, taste, and touch.
- Limit the time the story covers -- one day is best. The story may take place over a period but must not include a period longer than one year.
- Leave out extraneous or irrelevant details.
- Include dialogue where appropriate and punctuate it correctly.
- Use transitions, repetition, or parallelism to achieve coherence.
- Each paragraph should include at least eight sentences in total. There is no minimum word requirement, but adequately developing a paragraph means adding good, specific details.
- Include a concluding sentence that "wraps things up."
- The story you tell should be true, not fictional.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

The classification essay requires four paragraphs that focus on an organization of complex information, showing how members of the group are related and differentiated. The classification items might be people, places, things, or concepts (Rys 2019). The text requires a focused introduction with a clear thesis, body paragraphs, and conclusion. The use of outside sources is encouraged.

The assignment instructions include the following: Write a four-paragraph classification essay.

Instructions:

- The title captures interest and the main point of the essay
- The introduction captures the reader's attention and clearly introduces the thesis sentence.
- Each paragraph has a clear topic sentence that creates a single dominant Impression and supporting sentences.
- Categories and details are organized in a logical order.
- Categories are introduced by name and explained thoroughly.
- The essay makes the classification scheme clear.
- Nouns, verbs, and modifiers are specific and original.
- Uses clear and effective phrasing.
- Contains a clear concluding paragraph that wraps things up.
- The essay contains four paragraphs.
- Essay is in first or third person (the essay does not use the second person).
- The use of outside sources must follow MLA citations.
- The text consistently uses correct sentence structure and maintains agreement in grammar throughout.
- The text uses proper punctuation, capitalization, and word usage.
- The text uses correct spelling & heading
- Text is double-spaced & paragraphs are indented

- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

The process-analysis essay requires four paragraphs focusing on explaining or directing how something happens, works, is made, or is done. This might be natural, like a phenomenon that occurs in nature, performative, or a historical process (Rys 2019). The text requires a focused introduction with a clear thesis, body paragraphs, and conclusion. The use of outside sources is encouraged.

The assignment instructions include the following: Write a four-paragraph process analysis essay.

Instructions:

- Include an interesting title and a clear thesis statement that expresses the point of the essay.
- Each paragraph has a clear topic sentence that creates a single dominant impression and supporting sentences.
- Paragraphs use transitions to create coherence within and between paragraphs.
- Details are organized in logical order.
- Examples and explanations of the process are provided.
- Uses specific nouns, verbs, and modifiers.
- Use clear and effective phrasing (no wordiness).
- Include a clear concluding paragraph that wraps things up.

- Write four paragraphs.
- The use of outside sources must follow MLA citations.
- Use correct sentence structure (no sentence fragments, run-ons, comma splices) without errors in agreement.
- Use correct capitalization and punctuation.
- Use correct word usage.
- Use the correct spelling.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

The compare-and-contrast essay requires five paragraphs that focus on two or more things, phenomena, or concepts side by side that show either comparison, focusing on the similarities, or contrast, focusing on the differences (Rys 2019). The text requires a focused introduction with a clear thesis, body paragraphs, and conclusion. The use of outside sources is encouraged.

The assignment instructions include the following: Write a five-paragraph comparison-contrast essay.

Instructions:

- The title captures interest and the main point of the essay.
- The introduction captures readers' interest. It includes interesting background information and a clear and effective thesis that makes an interesting point about the comparison.
- The text has a clear thesis statement expressing the point of the essay.

- Topic sentences clearly relate to the thesis, include vivid details such as names and appeals to the senses, and make points in the same order for both subjects using block or alternating pattern, and use transitions to create coherence within and between paragraphs.
- The essay shows clear and consistent categories for comparison/contrast.
- The essay uses a proper comparison/contrast structure (point-by-point or block method).
- Conclusion repeats the main idea without repeating the thesis word-for-word and echoes the main points, leaving a strong impression.
- Nouns, verbs, and modifiers are specific and original.
- The essay is in first or third person (the essay does not use the second person).
- The use of outside sources must follow MLA citations.
- Word choice is interesting and appropriate.
- Sentence structure is varied
- Use correct sentence structure (no sentence fragments, run-ons, comma splices) without errors in agreement.
- Use correct capitalization and punctuation.
- Use correct word usage.
- Use the correct spelling.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

The cause-and-effect essay consists of five paragraphs that may examine either the effects of a specific event, action, or phenomenon, or the causes that led to that event, action, or phenomenon, or both (Rys 2019). The text requires a focused introduction with a clear thesis, body paragraphs, and conclusion. The use of outside sources is encouraged.

The assignment instructions include the following: Write a five-paragraph cause-and-effect essay.

Instructions:

- The essay title captures interest and the main point of the essay.
- The introduction captures readers' interest. It includes interesting background information and provides a clear and effective thesis, which makes an interesting point about causes or effects.
- The topic sentences in each paragraph clearly relate to the thesis.
- Each paragraph follows the correct development structure of the topic sentence and supporting details. Paragraphs use transitions to create coherence within and between paragraphs.
- Causes/effects are appropriately developed and relate to the thesis of the essay.
- Each paragraph develops a single cause or effect by providing details and a clear explanation to support the analysis.
- The conclusion repeats the main idea without repeating the thesis word-for-word and echoes the main points, leaving a strong impression.
- Nouns, verbs, and modifiers are specific and original.

- The essay is in the first or third person (the essay does not use the second person).
- Word choice is appropriate.
- Use of outside sources must follow MLA citation.
- Sentence structure is varied.
- Use correct sentence structure (no sentence fragments, run-ons, comma splices) without errors in agreement.
- Use correct capitalization and punctuation.
- Use correct word usage.
- Use the correct spelling.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

The argumentative essay consists of five paragraphs that focus on a debatable issue or a problem through a reasonable and measured stance, rooted in reliable evidence, as well as acknowledging the opposition in respectful terms and developing a rebuttal (Rys 2019). The text requires a focused introduction with a clear thesis, body paragraphs, and conclusion. The use of outside sources is encouraged.

The assignment instructions include the following: Write an argumentative essay supporting your opinion on a topic of your choice. In five paragraphs, introduce the topic, present your argument while also acknowledging the counterclaim, and provide reasons and evidence for both.

The essay must include all of the following:

- Title captures interest and the main point of the essay
- The introduction captures readers' interest. It includes interesting background information and a clear and debatable thesis that makes an assertion and indicates the main points.
- The topic sentences clearly relate to the thesis. Each paragraph follows the correct development structure of the topic sentence and supporting details. Paragraphs use transitions to create coherence within and between paragraphs.
- The body paragraphs include vivid details such as names and appeal to the senses, making points logically.
- The essay contains an opposition argument; it concedes/refutes/counteracts the opposition argument.
- Evidence supports all claims, and quotations are introduced.
- Conclusion repeats the main idea without repeating the thesis word-for-word and echoes main points, leaving a strong impression.
- Nouns, verbs, and modifiers are specific.
- The essay is in the first or third person (the essay does not use the second person).
- Word choice is appropriate.
- Sentence structure is varied.
- Quotations are properly punctuated.
- Use of outside sources must follow MLA citation.
- Paraphrases and quotations are accurate.

- Correct in-text citations & correct works cited.
- Use correct sentence structure (no sentence fragments, run-ons, comma splices) without errors in agreement.
- Use correct capitalization and punctuation.
- Use correct word usage.
- Use the correct spelling.
- The final draft must follow the MLA format, including the heading format illustrated in the MLA sample.

Each of these assignments covers a two-week period during which students discuss a specific subset or subgenre in class. The class meets for 55 minutes three days a week. During the class time, students explore the unique characteristics of each subtype or subgenre through samples provided from the textbook, discussion groups focused on tasks contributing to the assignment (e.g., classifying information on a given topic, comparing or contrasting, etc.), constructing individual outlines and rough drafts that lead to final drafts of the assignment. Additionally, students have designated writing time both in class and at the writing center, with support from the instructor. At the end of the two-week period, students are required to submit their completed assignments. The chronological order of the assignments is as follows: Week 1 (Overview & Writing Process), Week 2 (Description), Week 3 (Description), Week 4 (Narration), Week 5 (Narration), Week 6 (Classification), Week 7 (Classification), Week 8 (Process Analysis), Week 9 (Process Analysis), Week 10 (Compare and Contrast), Week 11 (Compare and Contrast), Week 12 (Cause and Effect), Week 13 (Cause and Effect), Week 14 (Argumentation), Week 15 (Argumentation).

This study focuses on COMP 101, which consists of essays written by first-year students and also represents one of the most common writing tasks that students encounter during this stage. The study evaluates the linguistic features in COMP 101 at the basic level of the essay genre, using BAWE as a comparative reference to evaluate the differences in linguistic features between the two corpora. Given that the BAWE corpus encompasses a wide range of genres, such as essays, case studies, and summaries (Lee 2001; Nesi 2011), it can be viewed from a superordinate level of genre as encompassing the genre of academic writing, representing the broadest or most comprehensive category of genre.

This categorization helps the study address the first research question of the most frequent linguistic features characterizing first-year composition writing and also examine how these features differ when viewed from upper-level academic writing. The internal categorization relates to COMP 101 and views the descriptive, narrative, classification, process analysis, compare and contrast, cause and effect, and argumentative texts as essays representing the subordinate genre. The subordinate level of categorization helps the study address the sub-question that seeks to identify the most frequent linguistic features distributed in the various essays. The next section will propose a model for the theoretical framework for this study based on genre.

3.4. A model for the theoretical framework

Writing practices have covered a wide range of topics, such as syntactic complexity (Ortega 2015; Staples and Reppen 2016), formulaic patterns (Biber and Barbieri 2007; Gil and Caro 2019), and linguistic features (Ramoroka 2017; Kim and Nam 2019; Crossley 2020). This study specifically focuses on the linguistic features or the internal characteristics of the texts submitted

in first-year university writing, analyzing them through genre-based theory. The following two subsections provide the rationale and the model behind the theoretical framework used by the study.

3.4.1 Rationale for the genre-based model

The genre-based theory offers valuable insights into how writers understand their setting and the broader constraints imposed by their immediate community and the larger cultural context. It also pinpoints the attitudes displayed by text producers regarding their topics, viewpoints, arguments, or relationships with others. According to Hyland (2015b), genre theory has played one of the most profound roles in how researchers and educators understand and teach language. Through genre theory, researchers can learn about the conventions characterizing particular genres and discourse communities and how such conventions originate and develop.

Bhatia (2004) proposes a practical view of a genre framework showing the influence of both register and discipline on genre. Hyland (1998; 2001b; 2005a; 2015b) further explores the various genre settings demonstrated through the role of the writers and their linguistic choices. These language choices coin different writers' identities in advanced academic writing (Harwood 2005) and student academic writing (Tang and John 1999; Aull 2019). Analyzing the writers' identities helps uncover the patterns of self-awareness writers manifest across different levels of writing (Hyland 2002b) establishing proximity (relationship between self and community) and building positions (relationship between speaker and content) (Hyland 2015b). It is through genre theory that researchers are able to focus on the texts as products not only of their conventionalized setting but also the text producers and their backgrounds. Thus, the genre is

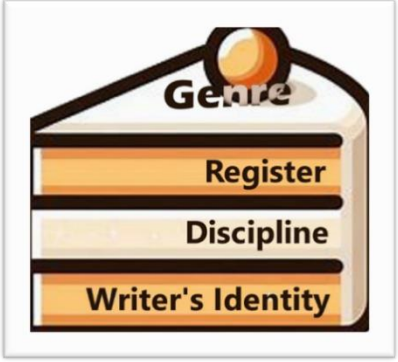
able to reveal the diverse disciplinary discourses, writing conventions, and writers' language choices.

At the superordinate or broadest level of genre, the texts in COMP 101 can be categorized as academic writing (Lee 2001; Hempel and Degand 2008; Kostrova and Kulinich 2015). The basic level of text categorization in COMP 101 is the essay genre, and the internal categorization at the subordinate level divides the texts into seven genres: descriptive, narrative, classification, process-analysis, compare-and-contrast, cause-and-effect, and argumentative essays. Even though the descriptive and narrative genres may be categorized as personal writing, they still have to adhere to the general academic conventions, as discussed in Section 3.2.1, and fall within the broadest genre category of academic writing. This genre categorization allows the study to concentrate on a given genre across the spectrum and investigate the most common linguistic features characterizing first-year university writing. Thus, the first comparison provides insights into the linguistic features revealed in first-year academic writing in COMP 101 with the BAWE corpus, a more extensive collection of texts representing academic writing in multiple levels of university writing, and as such, provides a comparative base with upper-level writing. The second investigation drills down into the textual features within the subordinate levels of genre and aims to uncover any patterns across their spectrum.

3.4.2 Model of the theoretical framework

A summary of the relationship between the framework elements constituting genre: register, discipline, and writer's identity is visualized in Figure 3.1. This visualization is based on Bhatia's (2004) and Hyland's (2015b) view of genre, academic disciplines, and writer's identity.

Figure 3.1: Elements of the theoretical framework



It is challenging to separate these elements as they are interconnected in their role to impact genre. One way to visually represent their involvement in genre might be to depict them figuratively, using a layered cake analogy, where multiple layers come together to form one cohesive item, like a layered cake. The first layer that genre reveals is the register (see Section 3.1.1 for more details). Looking at academic writing as the superordinate or the broadest genre category, the register used in that genre is typically formal. The genre-based analysis can identify the degree of formality writers use in COMP 101 and how it differs from BAWE, as well as the formality across the subordinate level demonstrated in the various essays comprising COMP 101. Formal academic writing is governed by conventions (see Section 3.2.1 for more details) that require objectivity, standard grammar and syntax, and an impersonal tone. Looking at formality can also reveal areas where writers do not comply with it, as some studies indicate (Hyland and Jiang 2017) (see Section 2.5.5 for more details). The opposite of the formal register is the informal, which is demonstrated by the use of first and second-person pronouns, contractions, and exclamation marks (Chang and Swales 1999; Liardet *et al.* 2019). It is worthwhile exploring the language features demonstrating formality or informality among first-year composition writers and how they differ from the upper level. A secondary objective is to see these language features across the different essay genres in COMP 101 and identify any variations in formality.

The second layer of the cake in Figure 3 shows the discipline, which relates to the academic disciplines. The concept of academic disciplines, as discussed in Section 3.2.2, is well described and discussed by Swales (1990), Hyland (1999; 2001b; 2015b), and Bhatia (2004). The linguistic features related to the academic disciplines are typically found in expert-level writing since expert writers are members of the disciplines and communicate within the specialized conventions (see Section 2.5.4 for more details). Contrary to expert writers, students are not as proficient in their chosen disciplinary writing yet, and as they progress in their fields, they begin to demonstrate the specialized features of various disciplines. In their first year of composition, students might not be expected to demonstrate a deep understanding of academic disciplines because they are either new to their chosen field or have not yet chosen one. In Chapter 4, section 4.1.2, the study provides the declared majors of the participants in the study when they entered the university. These majors, however, do not always stay consistent, and students often change their majors in the second year of their studies. Thus, the features related to academic disciplines become observable as students progress in their academic journey, like at upper levels, where they become more engaged with their chosen fields. Using the genre-based theoretical framework, the degree of language features related to academic disciplines between COMP 101 and BAWE can be revealed. The study also explores whether disciplinary features emerge across different essay genres in COMP 101 as students progress through the semester.

Finally, the last layer of the cake in Figure 3 shows the writer's identity, providing information about the text producers (see Section 3.2.3), who play a significant role in shaping genres. This study specifically focuses on first-year composition writing and how it differs from upper-level writing based on the most frequently used linguistic features by the two different groups of text producers. These two levels demonstrate different proficiency levels in academic writing, which

relates to the writer's identity. The most frequently used language features can help reveal these identities and any unique language choices found in the texts.

3.5. Conclusion

The goal of this chapter was to provide a clear theoretical framework that enables the investigation of the most common linguistic features in first-year composition writing. To establish this framework, the chapter demonstrated the relationships and distinctions among register, genre, and style and discussed their related terms and relationships with the implication to this study. Furthermore, special attention was placed on genre in the conventionalized academic setting influenced by the common conventions, academic disciplines, and the text producers. The chapter also provided a categorization of genres that helps in understanding the levels of the analysis: it primarily focuses on the linguistic features revealed in COMP 101 and how they might differ from BAWE. Secondary to this primary focus, the study seeks to investigate the linguistic features across the essay genres in COMP 101. In developing the framework, the chapter provided a rationale and a model, drawing on Bhatia's (2004) and Hyland's (2015b) research. The model for the theoretical framework focuses on the register, academic disciplines, and the writer's identity in comprising genre. This model provides a practical and effective lens for investigating first-year composition writers through corpus linguistics to uncover the most frequently used patterns and determine any differences with the upper level of writing. Gaining information about the writers is important for the instructors building and delivering the curriculum, having greater insight about the students through the lenses of the research. Also, through this type of framework and analysis, the study is able to investigate the linguistic features in COMP 101 through a corpus-driven methodology that

examines the patterns based on the corpus investigation. The next chapter delves into the methodology of this study, namely corpus linguistics, discussing the corpus and tools used by the study.

Chapter 4

Methodology

4.0 Introduction

This study examines the most common language features found in a corpus of first-year academic writing and compares them to those found at other levels at university. The study also looks at how these features are distributed across different genre types, such as descriptive, narrative, classification, process analysis, compare-and-contrast, cause-and-effect, and argumentative essays. The research methodology used is corpus linguistics, which involves analyzing a collection of texts stored electronically using digital tools like frequency lists, concordances, collocations, and keywords. By taking a corpus-driven approach, the study aims to investigate linguistic patterns, including words and grammatical structures, some of which may be underrepresented in research or writing manuals. This aspect of corpus linguistics is crucial in academic writing research, especially in the field of EAP (see Chapter 2 for details), as it helps uncover language patterns across different levels of writing and provide data, informing curriculum, teaching, and best practice guidelines. Since this study focuses on researching the frequency features in first-year academic writing, corpus linguistics is a suitable choice of methodology to explore these patterns. The chapter is organized into five sections: Section 4.1 outlines the corpus size, data, and participants; Section 4.2 examines corpus representativeness; Section 4.3 describes the corpus tools used in the methodology; and Section 4.4 concludes the chapter.

4.1 Corpus size, data, and participants

The study uses a corpus to investigate the language features of first-year composition writing. A *corpus*, or *corpora* in plural, refers to a specialized form of linguistic data. It comprises "a collection of written texts or transcripts of spoken language that can be searched by a computer

using specialized software" (Brezina 2018, p.6). The corpus used in this research, COMP 101, includes all written assignments submitted during three consecutive semesters and completed by three different cohorts of students - Fall 2019, Spring 2020, and Fall 2020. The researcher chose this specific timeline for data collection to maintain consistency across the dataset and to handle the data compilation and analysis effectively. Consistency was ensured by collecting data in the same course throughout the three consecutive semesters, during which the researcher also served as the course instructor, which added another layer of uniformity and reduced potential variability that might be caused by different instructors. The three consecutive semesters also contributed to consistency in the curriculum, which remained unchanged, and the three cohorts of students submitted identical assignments throughout the three semesters. Following the end of each semester, the researcher collected and uploaded the data to Sketch Engine, a corpus manager (see Section 4.1.1 for more details).

The texts in COMP 101 are assignments for an entry-level university course called *Composition 101*, which is mandatory for first-year students. The texts submitted in this corpus focus on entry-level academic writing competencies, representing a specific language domain in academic writing. Therefore, COMP 101 represents a specialized corpus. Specialized corpora help researchers gain insights about specific areas or types of language that are not easily gained through the analysis of large corpora. According to O'Keeffe et al. (2007) and Jones and Waller (2015), specialized corpora is a collection of texts representing language use within a specific domain or discipline, thus focusing on a particular aspect of the language.

The use of corpus linguistics in this research allows the representation of language use in a collection of electronically stored texts. Digital tools such as frequency lists, concordances, collocations, and keywords are utilized to describe and examine the data, supporting the study's

main research question, which focuses on examining the language features of entry-level composition writing.

This study uses a corpus-driven approach to investigate the data. Tognini-Bonelli (2001) defines a corpus-driven approach in terms of "the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" (p.84). Thus, the corpus-driven approach documents linguistic constructs, such as words and grammatical constructions, some of which may not be recognized by current linguistic theories. The corpus findings follow the empirical data and are analyzed based on researchers' intuition and observation, which verifies and assists the investigation of the research questions. Thus, corpus linguistics combines the quantitative aspect of empirically supported findings and qualitative analysis based on the researchers' intuition regarding the targeted texts (Bonelli-Tognini 2001; Sinclair and Carter 2004; Biber and Reppen 2015). The following three subsections focus on the corpus data (4.1.1), the corpus participants (4.1.2), and conclude with a discussion on the issue of corpus size (4.1.3).

4.1.1 Corpus size and data

The data in COMP 101 is collected over three consecutive semesters: Fall 2019, Spring 2020, and Fall 2020 and is displayed in Table 4.1.

Table 4.1: COMP 101 Corpus matrix

Semester	**Fall 2019**	**Spring 2020**	**Fall 2020**	**TOTAL**
No of students	27	10	25	62
No of texts	168	58	157	383
No of distinct tasks	7	7	7	7
No of words	82327	28831	77670	188,828

The total corpus word count is 188,828. The study involved participants between the ages of 18 and 25, with 41 percent being male and 59 percent female. In Fall 2019, 27 out of 50 students agreed to participate. In Spring 2020, 10 out of 23 students participated, and in Fall 2020, 25 out of 40 students consented to participate. Based on these rates, the participation rate for Fall 2019 was 54 percent, falling within the medium range. In Spring 2020, the participation rate dropped to approximately 43 percent, below the medium range. However, Fall 2020 saw an increase in participation, with a rate of 63 percent, categorizing it as above the medium range. These fluctuations indicate varying levels of student engagement across the terms, with Fall 2019 having medium participation, Spring 2020 below medium, and Fall 2020 above medium.

The data in COMP 101 is collected from documents saved as Word or PDF files and submitted as assignments during English composition writing. COMP 101 is compiled using Sketch Engine, a corpus management tool that can analyze texts of billions of words and instantly identify typical, rare, or unusual patterns (Kilgarriff *et al.* 2014).

The course of Composition 101 covers seven types of writing genres: narrative, descriptive, classification, process analysis, compare and contrast, cause and effect, and argumentative writing, which are divided into individual sub-corpora. Table 4.2 shows the words and their corresponding percentages for each sub-corpus to the overall number of words.

Table 4.2: Words & percentages in the genre corpora

Type	Words	Percentage
Narrative	37342	20%
Argumentative	33326	18%
Cause & Effect	30117	16%
Process Analysis	27102	14%
Classification	24960	13%

Compare & Contrast	22171	12%
Descriptive	13166	7%

The distribution of text types in the Composition I curriculum is as follows: narrative text accounts for 20 percent, argumentative text for 18 percent, cause and effect for 16 percent, compare and contrast for 14 percent, process analysis for 14 percent, classification for 13 percent, and descriptive text for 7 percent. In Composition I, narrative, argumentative, cause and effect, and comparison and contrast essays are expected to be five paragraphs long, while process analysis and classification essays are expected to be three paragraphs long, and descriptive essays are expected to be a single paragraph. The varying paragraph requirements contribute to the differences in the sizes of the different text type categories.

4.1.2 Participants

The texts were written by a group of first-year university students of mixed nationalities who agreed to take part in the study. The students in Composition 101 come from international and domestic backgrounds. Table 4.3 displays the nationalities of the students in COMP 101, demonstrating the diverse makeup of the class.

Table 4.3: Students' Nationalities

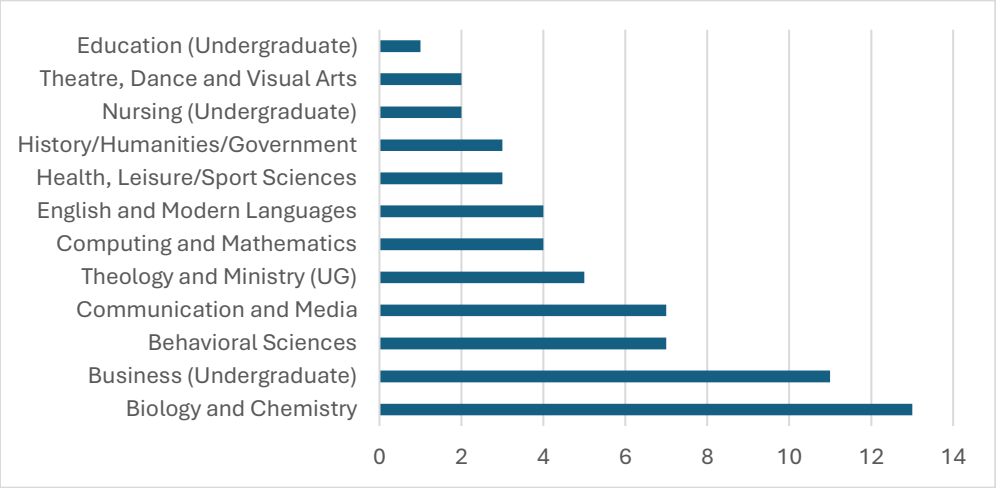
Nationality	Fall 2019	Spring 2020	Fall 2020	Total
Brazil	1		1	2
China	3	1		4
Colombia	1			1
El Salvador	1			1
France	1			1
Honduras	1			1
Israel	1	1		2

Mongolia	1			1
Myanmar	1		1	2
South Africa	1			1
South Korea	1	1		2
Tanzania			1	1
United States	14	7	22	43
Total	**27**	**10**	**25**	**62**

The university does not have a separate section for the course Composition 101, offered only for international students; instead, all students are grouped into Composition 101 based on their semester schedule within their respective majors. International students, who are not citizens of the United States or Green Card holders, must provide a test score in TOEFL (61 internet-based, 173 computer-based, or 500 paper-based), IELTS (6.0 or higher), ACT (20 or higher), or SAT (1020 or higher) as part of the admissions process. The students involved in the study come from various countries, including the United States, Brazil, Colombia, El Salvador, Honduras, China, South Korea, Mongolia, France, Israel, and South Africa.

Even though students declare specific majors upon their entry to the University, they often change their majors in their second or even third year. Therefore, it would be misleading to assume that the majors indicated upon admission are representative throughout the entire duration of their studies at the University. Figure 4.1 illustrates the majors that students indicated upon admission to the University.

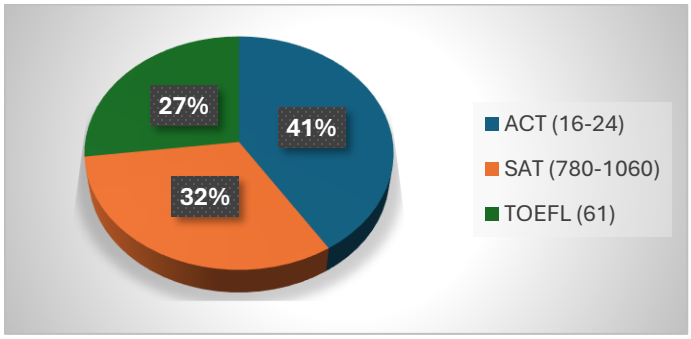
Figure 4.1: Students' Majors



The figure shows the distribution of students across different majors, with the smallest number in Theatre, Dance, and Visual Arts and the largest in Biology and Chemistry.

Figure 4.2 shows the ranges and percentages of the students' proficiencies scores submitted to the University and represented by ACT and SAT for native students and TOEFL for international students. Students with lower scores than 30 on the ACT and 1130 on the SAT are required to take COMP 101. For international students, the required TOEFL score is 61 Internet-based and 500 paper-based. An alternative to TOEFL is an IELTS score of 6.0 or higher.

Figure 4.2: Standard testing scores: ACT, SAT, and TOEFL



According to the ETS TOEFL website (*The TOEFL Family of Assessments* 2020), scores lower than 72 correspond with a B1 proficiency level in the Common European Framework of Reference (CEFR) (*The CEFR Levels* 2020). The English Profile (Profile 2011) defines B1 as independent users of English who "can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst traveling in an area where the language is spoken. Can produce simple connected text on topics that are familiar or of personal interest. Can describe experiences and events, dreams, hopes, and ambitions and briefly give reasons and expectations for opinions and plans" (Profile 2011). The ACT scores of the students participating in COMP 101 are between 16-24, which indicates that "students who achieve this score on ACT English Test have a 50 percent likelihood of achieving a B or better in a first-year English Composition course at a typical college" (*ACT College & Career Readiness Standards - CCRS-English Standards.pdf* 2020). The SAT scores that range between 780 and 1060 indicate poor to average skills.

4.1.3 The issue of corpus size

It is important to consider the effectiveness of using a small corpus in the investigation of the linguistic features. The overall number of words in COMP 101 is 188,828, which is not in the range of mega-corpora like the British National Corpus (BNC), totaling 96,134,547 words. According to O'Keeffe et al. (2007), a spoken corpus comprising over a million words is considered large, while a written corpus containing less than five million words is rather small. The significant difference between the numbers representing each type is due to the required time in collecting data—spoken corpora take longer for compilation than written ones. When considering the size of corpora, it is important to note Sinclair and Carter's perspective (2004).

They emphasize the importance of large corpora in providing reliable search results for various inquiries. From a lexicographic point of view, they admit the possible use of small corpora with the reservation of its limitations in representing varieties of lexis and phraseology. Even though small corpora do not appear suitable for lexis and phraseology, Carter and McCarthy claim that they can enable the study of pronouns, prepositions, and auxiliary and model verbs based on their high frequencies in texts (1995). Also, according to Flowerdew (2004), small corpora focusing on a specific question for research or teaching purposes may contain up to 250,000 words.

Even though small corpora lack the richness of lexis and phraseology, they provide better opportunities for researchers to study the original context of language features. Smaller corpora specializing in a specific language variation provide insights into the patterns of language use in particular settings (Koester 2010; Nelson 2010; Rippen 2010; Vaughan and Clancy 2013). Since this study seeks to investigate the most frequently used linguistic features in first-year composition writing, it is suitable to use a small, specialized corpus. Flowerdew (2004, p.21) lists several criteria that qualify a specialized corpus: (1) specific purpose for compilation; (2) contextualization (e.g., specific setting, participants, and communicative purpose; (3) genre (Table 4.2 shows the genre in COMP 101); (4) type of text or discourse (see Table 4.2 for the text types); (5) subject matter; and (6) variety of English. The purpose of COMP 101 is to investigate the most frequent language features in the entry-level writing course at university, and the corpus is created to serve the purpose of this study. Table 4.1 identifies the number of texts that comprise COMP 101, and Section 4.1.2 describes the participants—first-year university students—who wrote the essays as part of the assignments in Composition 101. The collection of these texts allows the researcher to control the text variety, the target group of writers, and the

context of language production. Thus, COMP 101 is a collection of texts that reflect a specific group of writers and text variety known as the specialized corpus (Friginal and Hardy 2014). The next important issue to consider, which is as important as the corpus size, is the corpus representativeness.

4.2 Corpus Representativeness

Since a corpus is a sample, it needs to be representative of a given aspect or aspects of language, and sampling is challenging because researchers need to determine the number of samples that will be representative of the language, their size, and what type the samples should be (Nelson 2010). In terms of representativeness, Sinclair (2005) emphasizes that structural criteria selection is crucial for a trustworthy corpus, as balance and representativeness rely on these choices. Common structural criteria include the mode of text (speech, writing, electronic), text type, language varieties, text locations, and dates. When compiling a small, specialized corpus, which is in the case of COMP 101, Koester (2010) notes that the most important consideration is representativeness. According to Leech (1991), a corpus is representative when "the findings based on its contents can be generalized to a larger hypothetical corpus" (p.27). Adding to the same concept is Biber's (1993) definition of representativeness, which is "the extent to which a sample includes the full range of variability in a population" (p.243), a definition used by Tognini-Bonelli (2001), Friginal and Hardy (2014), and Collins (2019) in their discussion of this issue.

Representativeness relates to two perspectives: situational and linguistic. Genre or register "are situationally defined text categories," and text types are "linguistically defined text categories" (Biber 1993, p.380). When considering the situational or contextual variety, a corpus should

include a full range of text that characterizes the language variety while the linguistic variability relates to the range of language characteristics within that language variety, and very often, this is what the researchers are looking to identify through the analysis (Biber 1993; Friginal and Hardy 2014; Collins 2019). Thus, situational criteria need to come before linguistics. When investigating a specific genre, it is quite straightforward to establish situational representativeness, where all samples represent the types of genre under research (Koester 2010).

According to Biber (1993), linguistic representativeness depends on the number of words per text sample and the number of samples per genre. Based on statistical analysis, he finds that the occurrence of the most linguistic features is relatively stable across 1,000-word samples. For the linguistic features to adequately represent a genre, they need to be stable with ten text samples per genre. This is encouraging for researchers of specialized corpora, as it suggests that reliable results can be obtained from small corpora when these conditions are met (Biber, 19993). Koester's (2010) advice to researchers is to ensure that any sub-corpus within the corpus (for example, a particular genre or sub-genre) is represented by at least 1,000 words (even if these are spread across different texts or conversations) and that every sub-corpus contains at least five, if possible ten, different samples. In this regard, COMP 101, as Table 4.2 shows (listed in Section 4.1.1), comprises more than ten samples per text type for each genre, e.g., 60 text samples of descriptive writing that have 13,166 words in total and 55 narrative essays of 37,342 words.

Next to the issue of representativeness is balance. McEnery et al. (2012) note that balance and representativeness "are ideals which corpus builders strive for but rarely, if ever, attain. In truth, the measures of balance and representativeness are matters of degree" (p.10). Discussing balance, Nelson (2010) challenges corpus builders who attempt to balance a corpus to give a representative view of the language and define the area of the language they want to represent.

Biber et al. (1999) refer to the area of language a corpus strives to represent as a register, and it is balanced if it covers a sufficient portion of the language variation. In building a specialized corpus, Vaughan and Clancy (2013) note that balance relies on the range of texts that are typical for what the corpus is meant to represent, and the issues of representativeness are addressed "by ensuring that the samples collected are typical of the speech domain represented in the corpus" (p.5). Since COMP 101 is a specialized corpus, Vaughan and Clancy's (2013) guidelines help address the balance issue. Specifically, Vaughan and Clancy's (2013) guidelines emphasize that a specialized corpus is balanced when it includes the full range of text types that characterize the linguistic practices of the target domain. For COMP 101, a specialized corpus of first-year university writing, balance is achieved by incorporating all seven essay subtypes taught in Composition 101: description, narration, classification, process analysis, comparison and contrast, cause and effect, and argumentation. These subtypes encompass the complete set of writing tasks assigned in the course, ensuring that the corpus reflects the diversity of linguistic features and textual structure expected in first-year composition. By aligning the content of the corpus with the essay types emphasized in the curriculum, COMP 101 ensures both representativeness and balance in relation to the domain of first-year composition.

Based on these guidelines, it is reasonable to conclude in this section that COMP 101 is a balanced corpus that sufficiently represents the specific linguistic context under study, which is first-year university writing. The next section turns to the methodology used in corpus linguistics and explores the digital tools used for analyzing the dataset comprising the corpus, such as frequency lists, keywords, concordance lines, and collocations.

4.3 Methodology

The current study analyzes the language features in first-year academic writing through the use of corpus linguistics, aiming to facilitate "empirical investigations of language variation and use, resulting in research findings that have much greater generalizability and validity than would otherwise be feasible" (Biber and Gray 2015, p.182). Corpus linguistics involves studying language by analyzing authentic electronic texts and relies on empirical data to identify patterns and structures in language. McEnery and Hardie (2012) distinguish corpus linguistics from other topics in linguistics as not directly connected to the study of any particular aspect of language, but rather "an area which focuses upon a set of procedures, or methods, for studying language" (p.2). The digital tools used in corpus linguistics to analyze data include frequency lists, concordances, collocations, and keywords.

This study follows a corpus-driven approach, identifying patterns emerging from the data instead of starting with a specific hypothesis. The main question is to examine the high-frequency features in first-year composition writing and identify the patterns they create in the text, as well as how these patterns compare to upper-level writing. In corpus-based studies, researchers use data from a corpus to investigate existing theories, while corpus-driven studies aim to identify patterns in raw corpus data to investigate emerging structures (Bonelli-Tognini 2001; McIntyre and Walker 2019). The next subsection discusses the frequency lists and their use in the context of this study.

4.3.1 Frequency list

Frequency lists are a common tool used in corpus linguistics to identify patterns in language usage among a group of people engaged in a specific activity. According to Hyland (2006), these

lists reveal the regularities and exceptions in language use, while high-frequency items represent common choices in academic writing. Leech (2011) also emphasizes the importance of frequency in corpus research, highlighting three main uses of frequencies: raw frequency, normalized frequency, and ordinal frequency. These measures provide valuable information for comparing and analyzing different corpora. O'Keeffe et al. (2007) and Rayson (2015) note that frequency lists are often the starting point for analyzing words in a corpus. Since frequency lists are often the first step in analyzing the words in a corpus, the COMP 101 frequency list serves as the starting point for the research findings. Table 3.2 displays the first twenty-five most frequent words in the corpus generated through Sketch Engine, a corpus management tool already mentioned in Section 4.1.1.

Table 4.4: COMP 101 ordinal frequency list

Number	Item	Freq
1	the	10644
2	,	9998
3	.	9947
4	to	6165
5	and	5626
6	of	4510
7	a	4097
8	in	3275
9	is	3224
10	I	2459
11	that	2433
12	it	2195
13	for	1851
14	are	1775
15	they	1397
16	with	1394
17	my	1393
18	“	1367

19	you	1361
20	was	1357
21	on	1290
22	be	1283
23	not	1282
24	have	1222
25	as	1152

Considering the requirements in academic writing for an impersonal tone (Bennett 2009), it is very interesting to notice a high number of pronouns typical for a conversational register rather than academic writing (Biber and Conrad 2009), such as *I* (2459) in 10th position, *you* (1361) in the 19th, and *my* (1393) in the 17th. The results of the COMP 101 ordinal frequencies reveal evidence only for the corpus under investigation, and without a reference point, it is difficult to interpret the data. Based on Gablasova, Brezina, and McEnery's (2017) research, a comparative design with a larger corpus, or evaluating corpus evidence with a reference point in mind makes the findings in the generated frequency list more informative. Leech (1998) recommends that corpora used in the comparative design should differ regarding the variable under research but be similar in other respects. Gablasova et al. (2017) discuss that this type of comparative comparison highlights "the ability to attribute the observed difference in corpus frequencies with a reasonable degree of certainty to the predictor variable and measure the strength (size) of its impact" (p.136). To avoid skewing the results that will follow from the comparative design, it is also important to consider the comparability of the reference corpora.

Since COMP 101 represents first-year composition writing at university and thus entry-level academic writing (see Section 4.2 for a complete discussion on representativeness), it needs to be compared to corpora in academic English, which cover a wide range and level of university writing, e.g., different academic disciplines and years of study. The British Academic Written

English (BAWE) corpus comprises academic texts of university students' writing in the UK, covering Arts, Humanities, Social Sciences, Life Sciences, and Physical Sciences. The participating students vary from first-year undergraduates to those in their fourth year of study. The corpus has 2858 texts that range between 500 and 5000 words. The assignments were collected between 2005 and 2007 from the universities of Reading, Warwick, and Oxford Brookes. The complete corpus totals 6,968,089 words. The corpus developers are Hilary Nesi and Sheena Gardner, Paul Thompson, and Paul Wickens (Hyland 2008; Nesi and Gardner 2018). BAWE provides a larger database of texts produced by undergraduate and graduate writers, and thus it is similar to COMP 101 in the target language but differs in the range, volume, and quality of writing; therefore, it serves as a good reference corpus.

An important factor to consider when comparing frequency lists is their normalization. Raw frequencies simply show how often an item appears in a corpus without considering the overall size of the corpus. This makes it difficult to compare raw frequencies between corpora of different sizes. To make accurate comparisons, we need to normalize the frequencies. This means dividing the raw frequency by the total number of words in the corpus and then multiplying by a standard unit, usually a million words (e.g., (Number of Occurrences/Total Number of Words in the Corpus) x Normalization Unit) (McEnery and Hardie 2012).

The frequency lists generated In COMP 101 provide Information for the first-year or the entry-level students in a mixed classroom. Since COMP 101 is 188,828 and BAWE is 6,968,089 words, the frequency counts must be normalized to make the two corpora comparable. Table 4.5 compares the normalized frequency counts in COMP 101 and BAWE. The basis for the normalization is 1,000,000 to reflect the larger corpora size of BAWE. For example, *the* in COMP 101 is normalized by dividing the number of occurrences 10644 by the total number of

words in COMP 101 and multiplying the result by 1,000,000. Once the formula is applied to normalize the result, then the number of normalized occurrences is 56369 per 1,000,000 words. In BAWE, then, *the* is again the first word on the list, but with a slightly higher number of occurrences—70646 per million words.

Table 4.5: Frequency words in COMP 101 and BAWE

COMP 101				BAWE		
Number	**Item**	**Freq**	**N Freq**	**Item**	**R Freq**	**N Freq**
1	the	10644	56369	the	492270	70646
2	,	9998	52948	,	391643	56205
3	.	9947	52678	.	313580	45002
4	to	6165	32649	of	271079	38903
5	and	5626	29794	and	208693	29950
6	of	4510	23884	to	191604	27497
7	a	4097	21697	in	153326	22004
8	in	3275	17344	a	136398	19575
9	is	3224	17074	is	111307	15974
10	I	2459	13022	)	91843	13181
11	that	2433	12885	(	90538	12993
12	it	2195	11624	that	79337	11386
13	for	1851	9803	‘	72584	10417
14	are	1775	9400	as	68072	9769
15	they	1397	7398	for	59564	8548
16	with	1394	7382	be	58120	8341
17	my	1393	7377	this	54393	7806
18	“	1367	7239	it	51248	7355
19	you	1361	7208	“	47283	6786
20	was	1357	7186	:	47060	6754
21	on	1290	6832	are	42739	6134
22	be	1283	6795	with	42310	6072
23	not	1282	6789	on	40642	5833
24	have	1222	6471	by	40564	5821
25	as	1152	6101	was	36855	5289

The normalized results show similar results in the ranking of the first 25 frequent words in the two corpora and six differences in COMP 101. The words that are equally ranked in both corpora

are colored in light grey (▁), and the ones that are unique to each of the corpora are in dark grey (▇). Most of the frequencies in COMP 101 are closely ranked to the ones in BAWE, and only five of the instances are equally ranked: *the* in position 1 in both lists, the comma punctuation sign (,) in position 2, the period sign (.) in position 3, *and* in position 5, and *is* in position 9. The words that show only in COMP 101 include *I* in position 10, *they* in place 15, *me* in 17, *you* in 19, *not* in 23, and *have* in 24. In BAWE, the differences with COMP 101 are open parenthesis in position 10 and closing parenthesis in 11, the apostrophe in 13, *this* in 17, the colon (:) in 20, and *by* in 24. The most striking variances between the two corpora seem to be the use of the personal pronouns *I*, *you*, and *my* in COMP 101, which are typical for spoken English but not academic (Biber and Conrad 2009). On the other hand, in BAWE, the high ranking of the parenthesis, which is not displayed in COMP 101, might imply a frequent use of in-text citations that characterize academic writing (Sun and Wang 2019). The high use of the apostrophe (‘) could show expressions of possessive form that often are used with the author’s name to show authorship of ideas, theories, views, or philosophies. The colon is usually used to introduce lists, sub-titles, sub-divisions, references, and quoted speech (Lester 2018). The high frequency of *this* in BAWE might relate to what Biber *et al*. (1999) discuss regarding its usage as a determiner and a pronoun in academic writing, helping writers make immediate textual references (p.349). The result of the comparison between the frequency words in BAWE and COMP 101 reveals that despite the similar results between the ranking of words, COMP 101 has higher frequencies of personal pronouns that are not typical for academic writing and do not appear in BAWE.

The comparison of COMP 101 and BAWE corpora shows similar frequency rankings for most top 25 words, with COMP 101 featuring personal pronouns (e.g., *I, you, my*) and BAWE displaying academic conventions like parentheses and colons. However, using BAWE, a UK-

based corpus, with the US-based COMP 101 could be questioned due to contextual differences in educational settings. To validate BAWE's use, the Cambridge Academic English (CAE) corpus was included as a supplementary dataset. This section discusses the CAE validation and its implications before examining keywords in corpus linguistics.

The CAE corpus includes 3.16 million words of academic texts, such as essays, journals, and lectures, from US and UK undergraduate and postgraduate students (*Cambridge English Corpus available for academic use | Cambridge Language Sciences* 2019; Sketch Engine 2020) . The mix of US and UK texts in CAE can provide a good test of whether BAWE's frequency patterns remain consistent across different academic contexts. Appendix A (Table A.1) presents the normalized frequencies of the top 25 items in COMP 101, BAWE, and CAE, with equally ranked words in light grey (▁) and unique ones in dark grey (▇). BAWE and CAE share ten equally ranked items (e.g., *the, comma, in, a*), with others varying slightly (e.g., *of* in position 4 in BAWE, 3 in CAE). The colon (:), ranked 20 in BAWE, is absent in CAE and COMP 101, suggesting a UK-specific use for lists or quotations. COMP 101's unique items (*I, my, you, they, have*) confirm the personal style of the COMP 101 students, consistent with the BAWE comparison.

The CAE results show BAWE's frequency rankings align with a broader US-UK academic corpus, supporting its role as the primary corpus for comparison. COMP 101's differences, validated by CAE, reflect the first-year writing practices. While the UK context of BAWE may influence features like the colon, CAE confirms that core academic words are consistent across contexts. This comparison justifies the use of BAWE as a comparison corpus and confirms that the observed differences in the frequency lists are a result of student level rather than national context.

The next tool used in corpus linguistics relates to identifying the keywords, and the following subsection discusses the usage of the keywords in corpus studies.

4.3.2 Keywords

Rayson (2015) describes keywords as a metric that "provides complementary information to word frequency alone and gives an indication of the aboutness of a text, or what items are worthy of further investigation" (p.41). Similar to this description, Culpeper and Demmen (2015) note that "a keyword has a quantitative basis: it is a term for a word that is statistically characteristic of a text or set of texts" (p.90) and the keyness of words is calculated based on a cross-tabulation and a chi-square or log-likelihood significance test. The statistical nature of keywords reveals linguistic features in corpora that are not based on subjective judgments of researchers but on statistical formulas or quantitative tools. Text analysis tools like Wordsmith Tools and Sketch Engine determine keyness by comparing frequency lists in one corpus with another, indicating statistically more frequent occurrences using a log-likelihood formula (Hyland 2015a). Keywords are significant because they carry a frequency that is unusually high or unusually low in comparison with some norm, and they show a comparison between two sets of frequency lists: one in the corpus being analyzed and the other in a bigger corpus. The role of the bigger corpus is to provide the background data that serves as the comparison reference (O'Keeffe and Carter 2007; Rayson 2015). Hyland (2015a) sees keyness as a helpful tool for comparing the most frequent keywords between different disciplines. For example, Scott and Tribble's (2006) research shows that the most frequent words in humanities are *of, the, in, early, war, theory, as, century,* and *between*, but in medicine, these words are *clinical, patients, treatment, disease, of, study,* and *diagnosis*. Hyland (2015a) suggests that keyness can provide points of similarities and differences between disciplines and references for further research.

The focus of this study is on first-year composition writing, and at this level, students often transition between disciplines and may not have a specialized vocabulary for each field, which makes identifying lexical differences at a disciplinary level a challenge. However, the keywords may provide insight into the statistical significance of these items. Table 4.6 uses BAWE as the reference corpus to generate a list of the keywords in COMP 101. This list focuses on the first 25 keywords, providing their score, frequency in the focus corpus, and frequency in the reference corpus. The "Score" column represents the keyness score, calculated by Sketch Engine, using "simple math" formula, which identifies statistically significant keywords by comparing their relative frequencies in COMP 101 and BAWE. A higher score indicates a word is more characteristic of COMP 101. BAWE is the larger corpus, also focused on academic English, (see Section 4.3.1 for more details), and it serves as a reference corpus. According to Culpeper (2009), "the choice of reference corpus will affect whether you acquire keyword results that are all relevant to the particular aspect of the text(s) you are researching. The closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus" (p.35). Since COMP 101 and BAWE focus on academic writing, the results displayed in Table 4.6 show some of the specific language features in COMP 101.

Table 4.6: Keywords in COMP 101 with reference corpus BAWE

N	Term	Score	Frequency (focus)	Frequency (reference)
1	I	11465.3	2457	0
2	them	2240.66	480	0
3	God	1092.84	234	0
4	me	668.07	538	23
5	him	626.24	134	0
6	Christian	467.6	100	0
7	Jesus	448.93	96	0

8	United	378.94	81	0
9	One	378.94	81	0
10	States	360.28	77	0
11	Christmas	341.62	73	0
12	America	332.28	71	0
13	To	308.95	66	0
14	Chinese	308.95	66	0
15	American	304.29	65	0
16	African	294.96	63	0
17	What	290.29	62	0
18	English	271.63	58	0
19	France	248.3	53	0
20	mom	206.61	60	3
21	How	201.64	43	0
22	Americans	196.97	42	0
23	China	192.3	41	0
24	Internet	182.97	39	0
25	basketball	179.16	52	3

The first keyword in Table 4.6 is the first person pronoun *I*, which is also one of the most frequent words in COMP 101 and is also examined further through the concordance tool in Section 4.3.3 and collocation in Section 4.3.4, but in the case of the keywords, *I* is recognized as a word of statistical significance with a score of 11465, which means that the writers in COMP 101 deliberately chose it in structuring their text, which falls within the type of style keywords. Scott et al. (2006) differentiate between three types of keywords: proper nouns, aboutness words (lexical words such as nouns, verbs, adjectives, or adverbs), and frequency grammatical words that indicate style. Commenting on the second type or the aboutness of words, Baker (2008) adds that, generally, they are most interesting to analyze. Most of the keywords in Table 4.6 are of "aboutness" type, for example, *Christian*, *American*, and *Chinese* in COMP point out items that the writers wrote about in their texts.

Another list that the study generated in Sketch Engine used COMP 101 as a reference corpus to identify the keywords in BAWE. In this case, the list does not show any personal pronouns like

in COMP 101, but mostly aboutness words related to specific disciplines, such as *organization* at position 4, *variable* at position 7, *output* at position 8, *graph* at position 9, and *ibid* at position 18.

Table 4.7: Keywords in BAWE with reference corpus COMP 101

N	Term	Score	Frequency (focus)	Frequency (reference)
1	London	590.473	4914	0
2	according	291.178	2419	0
3	appendix	285.3	2370	0
4	organization	268.266	2228	0
5	species	258.19	2144	0
6	labour	252.192	2094	0
7	variable	247.873	2058	0
8	output	241.156	2002	0
9	graph	238.996	1984	0
10	Cambridge	216.324	1795	0
11	terms	209.727	1740	0
12	including	193.052	1601	0
13	tourism	190.653	1581	0
14	manager	179.737	1490	0
15	analyze	173.739	1440	0
16	input	165.582	1372	0
17	revolution	162.223	1344	0
18	ibid	160.904	1333	0
19	electron	155.866	1291	0
20	increased	151.187	1252	0
21	EU	143.27	1186	0
22	sales	140.391	1162	0
23	goods	127.196	1052	0
24	increasing	124.916	1033	0
25	recognize	124.077	1026	0

In comparing the keywords in COMP 101, displayed in Table 4.6, and BAWE, displayed in Table 4.7, it is evident that the focus of the first-year students is more on the general topics that are related to everyday life. In BAWE, the upper-level students use *appendix*, *variable*, *output*, *graph*, *input*, and *ibid*, which show a more specialized use of language suggesting technical aspects of writing and research. This comparison shows the overall differences between these

two groups of writers that support the findings of the frequency list in Section 4.3.1. The frequency list offers a more comprehensive understanding of the frequency features that characterize COMP 101. Since this study aims to understand the overall language patterns rather than specific themes or topics, the frequency list helps identify these high-frequency patterns, which makes it an effective tool for analyzing the language features during the first year of composition writing. The next subsection discusses the meaning of concordance lines, their use in corpus linguistics, and their role in this study.

4.3.3 Concordance lines

While frequency lists show the focus of given texts, concordance lines allow searches by a single word, suffixes, multiple phrases, or part-of-speech tags (O'Keeffe and Carter 2007; Hyland 2015a; Rayson 2015). Concordance lines show the repeated co-occurrences of words in texts, thus showing the preferred meanings indicated by individuals in a given community (Hyland 2015a). Concordance analysis allows the exploration of frequency functions in the context of their immediate texts, which provides information about the use of the discourse forms (Hyland 2005a; Baker 2008). It lists all the occurrences of a particular search term in a corpus as it is presented in its context, providing the words to the left and right. Through the concordance lines, researchers can search for language patterns by looking for their reoccurrences in the text and then use the patterns to identify the textual characteristics. Corpus tools like Sketch Engine and WordSmith refer to this visual representation of results from a concordance search as a “key word in context” (KWIC). It is important to note that KWIC is a different concept from the keywords discussed in Section 4.3.2 of this chapter.

To examine the key words in context, this study uses the guidelines recommended by Sinclair (2003) and Baker (2008), starting with scanning the concordance lines and searching for similarities in language use. This can be done by looking at the words and phrases on the left and right-hand sides of the randomized lines with the KWIC and identifying any interesting patterns in the text. Another way to examine the lines is by sorting them alphabetically and checking for possible repetitions through the remaining portion of concordance lines. To illustrate the work with the concordance lines, the study examines the context of the personal pronoun "I," listed as a high-frequency word in COMP 101, which does not appear in BAWE (see Table 4.5 in Section 4.3.1). The results of the first twenty-five randomized lines with *I* as a KWIC are displayed in Figure 4.3.

Figure 4.3: The first twenty-five randomized lines with *I* as a KWIC

	Left context	KWIC	Right context
1.	s may not be known for decades to come. </s><s>	I	feel the evidence discovered in this research demonstra
2.	s involved. </s><s> The argument is life and death,	I	do not believe in abortion because no one has the right
3.	t depending on the trimester, a fetus is shown but	I	believe that once a mother becomes pregnant, there is
4.	icent is just misunderstood can be seen in the text	I	just read by a little girl from a friend; the little girl wrote,
5.	things happen to. </s><s> " To explain this briefly,	I	had a couple of my online friends watch Maleficent. </s
6.	igh school. </s><s> What I do differently, now that	I	'm in college, is add color to my notes. </s><s> I try to c
7.	o color- coding specific titles, the information that	I	have a hard time remembering, and the topics that the p
8.	ge, it has been a challenge, but I try to do the best	I	can. </s><s> The high school experience is very differe
9.	cific way of doing something doesn't work for me,	I	try finding another way to be more beneficial. </s><s> T
10.	s. </s><s> I have three pairs of Nike leggings that	I	work out in almost every day. </s><s> While they are gr
11.	many challenges along the way. </s><s> The book	I	Know Why The Caged Bird Sings is an autobiography of
12.	Jasmines offers quality above popularity. </s><s>	I	will be comparing the best and the worst of both Shops
13.	he up with fun games to play with my cuisines and	I	would find a place to play that game at. </s><s> My gra
14.	problem with sharing my faith, until then. </s><s>	I	suddenly felt out of place and different than the rest of
15.	ive essay, but I do not believe it is perfect. </s><s>	I	then realized the time, and I started doing a WPA assign
16.	/s><s> My dad is a prime example of this attitude,	I	can usually hear him yelling at the top of his lungs at th
17.	merica as Commander in chief. </s><s> However,	I	want to urge you to make an informed decision. </s><s:
18.	wise to look at the running mates as well. </s><s>	I	want to remind you that the constitution starts with "We
19.	ve is an Electric drum; when I saw that in my mind,	I	thought this could be not worked out, some reason I do
20.	e you? </s><s> " "Oh you're a 5 arent you? </s><s>	I	can tell" statements like these whether you know it or n
21.	ozier. </s><s> Ultimately, I think my roommate and	I	have done a great job at turning this tiny dorm room int
22.	relatives, and reliving family traditions. </s><s> As	I	enter my grandmother's house, the birds of the air sing
23.	e red lights help you concentrate. </s><s> My Dog	I	am the owner of a black and white, Blue Heeler mix dog
24.	ng can be found if one is looking hard enough, but	I	believe that Bierson was bias with the intentions only to
25.	th will change the world. </s><s> When I say faith,	I	am meaning has enough faith to take one step for what

Once the lines are generated and randomized, the next step is to examine the immediate right and left context of the first-person pronoun and look for similar patterns. Table 4.8 shows some of the repeated patterns in the right context. The first column indicates the number of the line that corresponds to the concordance in Figure 3.4, the second column shows the words in the context, and the last one provides the researcher's comments.

Table 4.8: *I* right context

N	Phrase	Comment
1	I feel the evidence	Followed by a verb that expresses modality
2	I do not believe	Followed by the present negative form of a verb that expresses modality
3	I believe	Followed by the present form of a verb that expresses modality
4	I just read	Past tense of an active verb
5	I had a couple of minutes	Followed by a chunk (had a couple of minutes)
7	I have a hard time remembering	Followed by a chunk (had a hard time)
12	I will be comparing	Followed by a modal verb to express a purpose
13	I would find a place to play	Followed by the past form of a modal verb to express a habitual action in the past
16	I can usually hear	Followed by a modal verb to express an ability
21	I have done a great job	Followed by an idiom
24	I believe that Bierson was biased	Followed by the present form of a verb to express modality

The patterns observed in using the pronoun *I* in the given phrases reflect personal voice in various modalities and functions in language. For instance, phrases like “I feel the evidence” and “I believe” are followed by verbs that express modality, indicating the speaker's stance or belief. This use of modality can reflect personal opinions or judgments. On the other hand, phrases such as “I just read” and “I had a couple of minutes” show past actions and experiences in specific time contexts in personal narratives. Additionally, expressions like “I can usually hear” highlight abilities, while “I have done a great job” shows idiomatic language of self-assessment or achievement. Overall, these patterns demonstrate a range of personal abilities or experiences regarding past events and achievements. Table 4.9 shows the patterns in the left context with *I* in the first column, indicating the numbered lines corresponding to Figure 3.4 and the researcher’s comments in the last column.

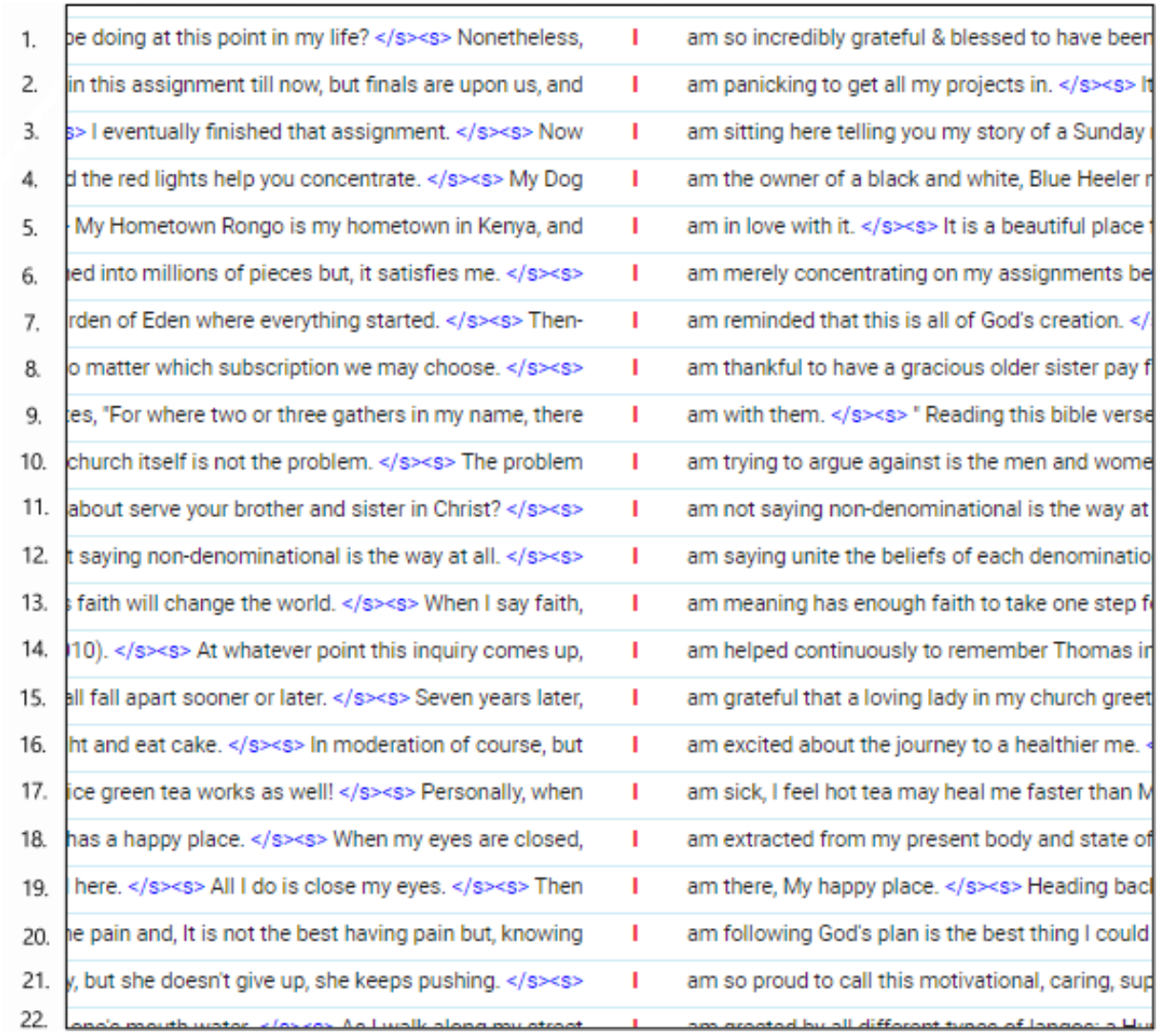
Table 4.9: *I* left context

N	Phrase	Comment
3	a fetus is shown but I	Preceded by a coordinating conjunction
5	To explain this briefly, I	Transitional phrase
6	now that I	That-clause
7	the information that I	That-clause
12	However, I	Transitional word
22	When I say faith, I	Dependent clause

The right context (see Table 4.8) shows a predominant co-occurrence of modal verbs with *I* that express present ability, inability, past ability, habitual actions in the past, and evaluative expressions that help the authors express possibilities based on certain conditions, provide context and explanation through dependent clauses and prepositional phrases, and build conversations through idiomatic expressions. In its left context (see Table 4.9), *I* co-occurs with coordinating conjunctions, transitional words, and dependent clauses. Based on the selected patterns, a possible hypothesis is that the writers use predominately present tense with modal verbs and dependent clauses or transitions when engaging with a given topic, using the first-person pronoun.

In the next step, the concordance lines are sorted alphabetically, a method viewed as more effective in discovering patterns (Baker 2008), but also might be seen as confirming patterns already revealed in the randomization. The subsequent step involves sorting the lines alphabetically to the right of KWIC and analyzing the patterns. Since there are over 2000 lines in the search results, Figure 4.4 only displays the results for the first significant pattern of occurrence in the right context, which is the verb *be* used with the first person.

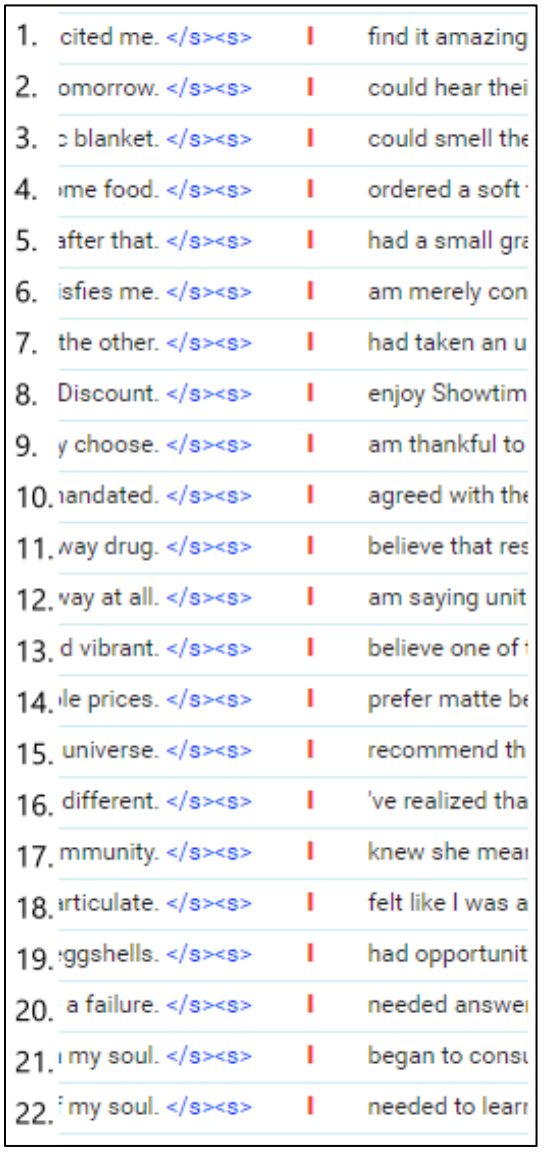
Figure 4.4: Alphabetized right context of *I* with predominant pattern



1.	be doing at this point in my life? </s><s> Nonetheless,	I	am so incredibly grateful & blessed to have been
2.	in this assignment till now, but finals are upon us, and	I	am panicking to get all my projects in. </s><s> It
3.	s> I eventually finished that assignment. </s><s> Now	I	am sitting here telling you my story of a Sunday
4.	d the red lights help you concentrate. </s><s> My Dog	I	am the owner of a black and white, Blue Heeler n
5.	My Hometown Rongo is my hometown in Kenya, and	I	am in love with it. </s><s> It is a beautiful place
6.	ed into millions of pieces but, it satisfies me. </s><s>	I	am merely concentrating on my assignments be
7.	rden of Eden where everything started. </s><s> Then-	I	am reminded that this is all of God's creation. </
8.	o matter which subscription we may choose. </s><s>	I	am thankful to have a gracious older sister pay f
9.	es, "For where two or three gathers in my name, there	I	am with them. </s><s> " Reading this bible verse
10.	church itself is not the problem. </s><s> The problem	I	am trying to argue against is the men and wome
11.	about serve your brother and sister in Christ? </s><s>	I	am not saying non-denominational is the way at
12.	t saying non-denominational is the way at all. </s><s>	I	am saying unite the beliefs of each denominatio
13.	faith will change the world. </s><s> When I say faith,	I	am meaning has enough faith to take one step f
14.	10). </s><s> At whatever point this inquiry comes up,	I	am helped continuously to remember Thomas in
15.	ll fall apart sooner or later. </s><s> Seven years later,	I	am grateful that a loving lady in my church greet
16.	ht and eat cake. </s><s> In moderation of course, but	I	am excited about the journey to a healthier me.
17.	ice green tea works as well! </s><s> Personally, when	I	am sick, I feel hot tea may heal me faster than M
18.	has a happy place. </s><s> When my eyes are closed,	I	am extracted from my present body and state of
19.	here. </s><s> All I do is close my eyes. </s><s> Then	I	am there, My happy place. </s><s> Heading bac
20.	e pain and, It is not the best having pain but, knowing	I	am following God's plan is the best thing I could
21.	, but she doesn't give up, she keeps pushing. </s><s>	I	am so proud to call this motivational, caring, sup
22.	one's mouth water. </s><s> As I walk along my street	I	am greeted by all different types of langos: a Hu

The first person is used in multiple contexts with the verb *be* to describe different personal experiences, highlighting the writers' personal focus as illustrated in Figure 4.3. Next, Figure 4.5 shows the alphabetized left context of *I* with its predominant patterns. A significant recurring pattern in the immediate left context of *I* is the use of a period and the first-person singular at the beginning of the sentence.

Figure 4.5: Alphabetized left context of I with predominant pattern



1.	cited me. </s><s>	I	find it amazing
2.	omorrow. </s><s>	I	could hear thei
3.	ɔ blanket. </s><s>	I	could smell the
4.	ıme food. </s><s>	I	ordered a soft
5.	after that. </s><s>	I	had a small gra
6.	isfies me. </s><s>	I	am merely con
7.	the other. </s><s>	I	had taken an u
8.	Discount. </s><s>	I	enjoy Showtim
9.	y choose. </s><s>	I	am thankful to
10.	ıandated. </s><s>	I	agreed with the
11.	way drug. </s><s>	I	believe that res
12.	vay at all. </s><s>	I	am saying unit
13.	d vibrant. </s><s>	I	believe one of
14.	ıle prices. </s><s>	I	prefer matte be
15.	universe. </s><s>	I	recommend th
16.	different. </s><s>	I	've realized tha
17.	mmunity. </s><s>	I	knew she meai
18.	ırticulate. </s><s>	I	felt like I was a
19.	ɡgshells. </s><s>	I	had opportunit
20.	a failure. </s><s>	I	needed answei
21.	ı my soul. </s><s>	I	began to consu
22.	my soul. </s><s>	I	needed to learı

Some of the most frequent words to the right of *I* in Figure 4.5 include *believe, can, could, feel, had, have, know, thought, want, will*, and *would*, which again indicates that writers tend to engage with the topic by using modality and predominantly present tense when showing personal opinions, impressions or experiences.

These examples illustrate the value of concordance lines in revealing language patterns within their immediate context. As such, they are an essential tool for this study, helping to identify the

underlying characteristics of first-year composition writing. For instance, the concordance lines are crucial for examining the roles of the first and second person in COMP 101 in Chapters 5 and 6, as well as the roles of conjunctions in Chapter 7. Also, the randomized concordance lines are essential in identifying the punctuation patterns discussed in Chapter 8. In addition to frequency lists, concordance lines are also used with collocations to examine the word pairs, their context, and any patterns related to them. The collocations and their use in corpus studies are discussed in the next subsection of this chapter.

4.3.4 Collocation

Concordances pose a challenge for researchers because they often contain a large number of lines requiring extensive analysis to identify word patterns. Although sorting functions can expedite the review of these patterns, the concordance lines may not always uncover the most crucial features of the words under study. This is where collocations come into play. They aid researchers by revealing the strength of the association between words that frequently occur together. Collocates are words that commonly appear near another word, and this relationship is statistically significant in some way. The phenomenon of certain words frequently appearing next to each other is known as collocation (Baker 2008). One method for determining collocations involves counting the number of times a word appears within a certain number of slots to the right or left of a search word.

The collocation functions are enabled through statistical tests to calculate saliency by “taking into account the frequency of words in a corpus and their relative number of occurrences both next to and away from each other” (Baker 2008). Several statistical formulae are used in corpus linguistics to identify statistically significant collocations: mutual information, t-test, Log Dice,

and log-likelihood (Baker 2008; Xiao 2015; D. Gablasova *et al.* 2017). Table 4.12 summarizes the values for the different tests, followed by a discussion of each test.

Table 4.10: Conventionally accepted score values for Log-likelihood, T-score, MI, and Log Dice

Statistical Test	Value
Log-likelihood	3.84 or above
T-score	2.576 or above
MI	3.0 or above
Log Dice	7.00 (roughly 100 times collocation; maximum: 14)

Mutual Information or MI is a statistical test that calculates all of the places where two words occur in a corpus and, through an algorithm, computes the expected probability of these two words appearing near each other based on their relative frequencies and the overall size of the corpus. Next, the algorithm compares the expected figure to the actual result, converting the difference between the expected and the actual into a number, which becomes the identifier of the collocation strength. Researchers generally accept that an MI score of 3.0 or above is regarded as evidence of significant collocation (Baker 2008; Xiao 2015). Thus, the higher the MI score, the stronger the collocation, and the closer the MI score is to zero, the more likely the two words appear by chance.

T-scores are another test for measuring collocations. The algorithm in t-score tests considers the amount of evidence available for a collocation or the size of the corpus in identifying the statistical significance. Xiao (2015) describes that the t-score is calculated based on "the difference between the observed and expected means, scaled by the variance, to determine the probability of a particular sample of that mean and variance with the assumption of the normal

distribution of the dataset" (p.109). Researchers generally accept a t-score of 2.576 or above as a measure of statistically significant collocation (Xiao 2015).

Log Dice is not only based on the co-occurrence of the node and the collocate but also on the size of the corpus, which makes it very useful in comparing scores between different corpora. In a way, it is similar to the MI-score because it emphasizes exclusive but not necessarily rare combinations. Gablasova *et al*. describe Log Dice as "a standardized measure operating on a scale with a fixed maximum value of 14, which makes Log Dice directly comparable across different corpora and somewhat preferable to the MI-Score" (2017, p.164). Negative values and 0 show that there is no statistical significance of the collocates, while the value of 7 means roughly 100 times collocation (Rychlý 2008). Finally, the log-likelihood or LL is noted by Xiao (2015) to be one of the most complex since its score is calculated on the basis of a contingency table that adds every cell to the logarithm of that cell and applies it to multiple combinations of table cells, multiplying the final result by two. The results based on this test are regarded as consistently better in generating collocations based on including common and rare lexical items. When choosing statistical tests to calculate collocations, Baker (2008) recommends that researchers focus on the type of words they want to investigate because different statistical tests tend to favor different types of words. He recommends MI tests for lexical and low-frequency words, rank by frequency for high-frequency words, log-likelihood for grammatical words, or a mixture of tests in examining low-frequency words.

In this study, the collocation tool provides additional information about the patterns or the words that, for example, co-occur frequently near *I* and *you* as node words. The study uses the collocate function in Sketch Engine with a three-word span on the left of the node word (*I/you*) or expressed as -3-to-three-word span to the right of the node word (I/you) or +3. The statistical test

in Sketch Engine allows algorithms for Log-likelihood, T-score, MI, and Log Dice, which are also used to measure the strength of the collocations of *I* and *you*. Table 4.13 displays the results for the first 25 collocates used with *I* and their test scores in Log-likelihood, T-score, MI, and Log Dice. The collocation tables display results similar to the observations based on the concordance lines where *I* is used frequently with modal verbs, the present and past forms of the verb *be*—the cells of the collocates are highlighted in .

Table 4.11: Log-likelihood, T-Score, MI, and MI3 with *I* as KWIC

Collocate	Collocate	T-score	MI	Log Dice	Log Likelihood
.	929	27	3	11	2662
,	854	25	3	11	2272
to	402	17	3	11	810
the	378	14	2	10	451
was	336	17	4	11	1552
that	324	16	4	11	1088
and	307	14	2	10	514
my	234	14	4	11	936
have	166	12	4	11	549
a	154	9	2	10	168
had	151	12	5	11	789
I	151	10	2	10	274
in	132	9	2	10	171
n’t	129	11	5	10	617
not	125	10	3	10	338
am	110	10	6	10	960
would	110	10	4	10	462
could	108	10	5	10	587
“	94	8	3	10	187
up	92	9	4	10	372
did	90	9	5	10	471
me	87	9	4	10	315
do	86	8	4	10	281
know	67	8	5	10	338
felt	59	8	6	10	434

In Table 4.13, Log-likelihood and T-score display high values of statistical significance, while MI and Log Dice, which measure collocation strength rather than statistical significance, show

lower values for some collocates like the grammatical items *the* (2), *and* (2), *a* (2), *I* (2), and *in* (2) as not significant collocations, which confirms Baker's recommendation that MI is usually used with lexical items. The collocations that confirm the concordance lines' findings and also show statistically significant values are *have, had, am, would*, and *could*. One of the surprising findings through the collocation tool is the high occurrence and statistical values of *be* in the past tense—*was*, and its occurrence is 336 or 226 times more than the present form *am*. The statistical scores display the highest values in the Log-likelihood column for the punctuation marks based on their high frequency as collocates: the period occurs 929 times with *I* while the comma 854. The concordance lines also show that periods often precede *I* (Figure 4.5), which places the first-person pronoun at the beginning of the sentence without a transitional word or phrase. Thus, using collocations together with concordance lines enhances the ability of the researcher to see the strength of associations between words that frequently occur together and indicate a notable relationship. Chapters 5 and 6 use collocations to examine the role of the first and second person in COMP 101 and also compare that use to upper-level writing.

4.4 Conclusion

Corpus linguistics enables the analysis of language features in large volumes of texts using frequencies, keywords, concordance lines, and collocations. The applied corpus linguistics methodology to COMP 101 reveals that among the unique frequencies in the corpus, the first-and second person pronouns seem to be very unusual in the context of academic writing and also compared to BAWE. These features indicate a high interaction between the authors and readers, which is more suitable for informal conversational settings than formal academic ones. The study examines the patterns associated with the first-person pronoun in Chapter 5 and the use of the

second-person pronoun in Chapter 6. The other two frequency features analyzed by the study include the use of conjunctions, discussed in Chapter 7, and punctuation marks, covered in Chapter 8. Together, these four chapters analyze the high-frequency features in COMP 101 and their distribution across different genres (for instance, whether narrative writing exhibits more conversational elements than argumentative writing). Thus, the analysis explores how engagement with the reader is achieved and the role of the writer through the use of first- and second-person pronouns, as well as sentence structure and punctuation, in the final two chapters of the analysis.

Chapter 5

First-person personal pronouns

5.0 Introduction

The initial review of the texts discussed in the previous chapter showed that students often use the first-person pronoun in their texts to express their feelings and experiences. This chapter aims to analyze the use of first-person personal pronouns in COMP 101. In examining the most frequent features in COMP 101, a corpus comprised of entry-level composition texts, this chapter focuses on using *I* and *we* as devices that express the authors' identities in academic texts. The frequency list identified both the singular and plural first-person pronouns as items of frequent usage (See Table 5.2), suggesting that students vary the use between the exclusive *I* and the inclusive *we*. This chapter focuses on the analysis of both pronouns to provide a holistic overview of the use of first-person personal pronouns in the context of first-year composition writers and to compare the use with upper-level writing, such as BAWE. The findings of this study seek to contribute to the research focused on authorial identities and specifically address these features in first-year composition writing.

5.1 Analytical Framework

A broad look at the academic literature indicates that the personal pronoun *I* is used not only in first-year composition writing but throughout academic writing as the means to create space and presence in the text, organize the discourse, outline procedures, report findings, dispute the research of others, or promote one's work. The first-person pronouns have engaged researchers' attention in undergraduate, graduate, and professional academic wiring (Tang and John 1999; Hyland 2001a; Hyland 2001b; Ivanic and Camps 2001; Harwood 2005; Hyland 2005a; Ramoroka 2017; Taylor and Goodall 2019).

Tang and John (1999) define the use of the first-person pronouns as the most observable manifestation of a writer's presence in the text and classify six different identities (i.e.., representative, guide, architect, recounter, opinion-holder, and organizer) that manifest "degree of power wielded by the authorial presence through a particular instance of the use of the first person pronoun" (p.26), expanding on the four-identity-roles concept (autobiographical self, discoursal self, self as author, possibilities for self-hood in the socio-cultural and institutional context), previously developed by Ivanic (1998), the latter based on Halliday's (2014) taxonomy of verbs (material, relational, and mental).

The authorial presence and textual features are further examined by Hyland (Hyland 2001a; 2001b; 2002a; 2002b; 2005a), leading to the conceptual model of metadiscourse, consisting of interactive and interactional dimensions which is described in detail in Chapter 2 (see Section 2.5.2). A considerable number of researchers based their analysis on Tang and John (1999) and Hyland (2001b); for example, Harwood (2005) evaluates the use of both first-person personal pronouns--*I* and *we*--in research articles to show the self-promotional role of the first-person pronouns in journal articles, Aull (2014) investigates the role of stance in early and advanced writing, Ramoroka (2017) examines the metadiscourse features across university disciplines, and Taylor and Goodall (2019) explore the rhetorical function of *I* in business writing. Undisputedly, this considerable number of research cases speaks of the high popularity of *I* in academic writing and its multi-functionality in creating writers' identities and stances. However, the pertinent question for the scope of the present research is not about the use of *I* and *we* in academic discourse at large but in the context of first-year composition writing and how that use relates to the roles already described in previous research.

Tang and John (1999) offer a helpful framework of reference that distinguishes the different roles of *I* and *we* in academic writing. According to this framework, the first-person pronouns show different degrees of authorial power, which can be explained through a continuum of six roles: Representative, guide, architect, and recounter of the research process, opinion-holder, and the originator (Tang and John 1999, pp.27-29). The framework includes personal pronouns in the first person, *I* and *we,* and contrasts their roles and usage based on verbs in academic texts. Table 5.1 identifies the six roles by placing them on a line representing a continuum of first-person pronouns. The continuum starts with the least powerful authorial presence and moves towards the most powerful authorial presence.

Table 5.1: First-person pronouns taxonomy based on Tang and John (1999)

Least powerful authorial presence **Most powerful authorial presence**

No First Person	Representative *We*	Guide *We* with mental perception verbs (e.g., *see, note, observe, feel, taste, smell, hear*)	Architect *I* (e.g., *In this essay, I will discuss*)	Recounter of research *I/we (‘doing’ verbs—work, read, collect*)	Opinion-holder *I/we* with mental cognitive verbs (*think, believe, suppose, expect, consider, know, understand*)	Originator *I* that claims authority

The columns under the continuum show the representative roles as they move from the least powerful to the most powerful. As already mentioned, the framework includes both the singular

and the plural personal pronouns. For example, the representative role is realized through the plural form *we*, referring to everyone or a small group of people without providing the reader with a lot of information about the author, thus reducing the author's role to a non-entity and marking the least powerful degree of the first-person pronoun. Next, the role of the guide is to navigate the reader through the text and point out significant parts for consideration. Tang and John's (1999) study interprets the role of the guide through Halliday's (2014) clause classification of the process (e.g., *perceive, sense, see, notice, glimpse, hear, overhear, feel, taste, smell*). The guide is often realized through the plural form of the first-person pronoun—*we*. The role of an architect of the text is usually recognized via the singular form through which the writer carries the responsibility of organizing the text (e.g., *In this essay, I will discuss*). Since the guide and the architect appear similar, it is essential to differentiate between them. While the architect outlines the material, the guide only seems to walk the reader through the existing text. The fourth role is the recounter of the research process or describing the steps involved in the research process, and might be is used with *we* or *I* or both. Often the verbs that signal this role are related to the material process described by Halliday (2014) (e.g., *work, read, interview, collect*). The last two roles inhabit the highest degrees of authorial power: the opinion-holder and the originator. The opinion-holder is the one who shares opinions or attitudes, varying from agreement to disagreement or expressing mere interest in a topic. This role might be realized with both singular and plural forms of the pronoun and signaled by verbs related to the mental process of cognition based on Halliday (2014) (e.g*., think, believe, suppose, expect, consider, know, understand, realize, appreciate, imagine, dream, pretend*). The originator is the last role in the taxonomy and the most powerful one to express authorial presence among the six. It claims

authority and ownership of the text, identifying the writer or the self, using only the singular form of the first person as the one who can originate new ideas.

In addition to Tang et al. (1999), the study uses Hyland's (2001a; 2002b; 2005b) work on the author's identity in academic writing realized with the first-person pronouns. While Tang and John (1999) explore the role of the first-person pronoun in student writing, Hyland (2001a; 2001b; 2002a; 2002b) examines the impact of authorial presence in journal articles and compares it to student writing; thus, he provides a valuable perspective to this study. Hyland (2001a; 2002b) refers to Tang and John's (1999) taxonomy as a helpful framework for distinguishing the roles of the first-person pronouns and adds to the concept by examining professional academic writing. In a similar study, Harwood (2005) mainly focuses on the role of the inclusive *we* by contrasting it to *I*, thus adding another dimension to exploring the first-person roles in academic writing. A supplementary verb categorization incorporated by the study is Biber et al. (1999), which distinguishes the major verb classes and their semantic categories. Through the research discussed in this section, the study examines the patterns of the first-person pronouns in COMP 101. The following section outlines the procedures and tools used in this study to investigate the usage of these pronouns.

5.2 Procedure and tools

The COMP 101 frequency list ranks both *I* and *we* in the first twenty-five top frequencies. Table 5.2 shows the filtered frequencies in COMP 101 and BAWE, which excludes the punctuation marks (see Table 4.5 for comparison) with a primary focus on demonstrating the first-person pronouns ranking within the most frequently used words, displaying *I* at position 8 and *we* at 24. Table 5.2 shows the top twenty-five frequencies in COMP 101 and BAWE with their respective

normalized frequencies per 1,000,000. Contrary to COMP 101, the first-person pronouns do not feature within the 25 most frequent items in BAWE , indicating that their role is not as prominent as in entry-level composition writing.

Table 5.2: Top 25 most Frequent items in COMP 101 and the BAWE (normalized per million words)

N	COMP 101	N Freq	BAWE	N Feq
1	the	56562	the	70646
2	to	32760	of	38903
3	and	29896	and	29950
4	of	23966	to	27497
5	a	21771	in	22004
6	in	17403	a	19575
7	is	17132	is	15974
8	I	13067	that	11386
9	that	12929	as	9769
10	it	11664	for	8548
11	for	9836	be	8341
12	are	9432	this	7806
13	they	7424	it	7355
14	with	7408	are	6134
15	my	7402	with	6072
16	you	7232	on	5833
17	was	7211	by	5821
18	on	6855	was	5289
19	be	6818	not	4805
20	not	6812	from	4514
21	have	6494	an	4195
22	as	6122	which	4186
23	this	6047	's	4147
24	we	5819	have	3677
25	their	5208	can	3642

The frequency of words in Table 5.2 shows significant differences in the use of first-person pronouns between COMP 101 and BAWE. This variation likely reflects differences in the writing tasks assigned and the instructional contexts shaping students' linguistic choices. In

COMP 101, four of the seven tasks—classification, comparison and contrast, cause and effect, and argumentative essays—specifically require students to use either first- or third-person pronouns, while descriptive and narrative tasks allow flexibility in pronoun choice (see Section 3.3). The process analysis task, which permits a directional approach addressing the reader, does not mandate first-person pronouns but may still include them depending on students' choices (e.g., describing personal processes using *I*). In contrast, BAWE texts include a wide range of university-written texts that may involve research essays, case studies, or reports that prioritize objectivity with the use of the third-person or impersonal constructions. The higher frequency of first-person pronouns in COMP 101, driven by task instructions, raises questions about how the first-person pronouns function in students' writing. While COMP 101 task instructions specify the use of first- or third-person pronouns for certain assignments, they do not prescribe specific roles for first-person usage, such as acting as a representative, guide, architect, opinion holder, or recounter. The following sections analyze these roles of first-person pronouns in COMP 101, comparing their distribution and usage patterns with those in BAWE to explore the writers' linguistic choices.

To analyze further patterning of *I* and *we* in COMP 101, the study uses the collocation and concordance tools in the SketchEngine interface (Kilgarriff *et al.* 2014) and examines the verbs used with *I, me, my* and also *we, our, us*.

To identify the collocates, the study used the LogDice scores, which facilitate comparisons across various corpora and datasets. In theory, the highest LogDice score is 14, which occurs when one item always co-occurs with the other, while a moderately low score is 7. This moderately low score helps filter a large range of verbs and examine their different types. Based on these values, the study uses 7 as the cutoff score. Once the dataset is searched for *I* and *we* in

SketchEngine, the collocation tool is applied with -3 and +3 words from the node, identifying the collocates cooccurring with the first-person pronouns. The verbs are then filtered based on their LogDice score, leading to varied results between the singular and plural first-person pronouns.

The functional categories are based on Tang and John (1999) taxonomy, which identifies six roles of authorial presence (see Table 5.1): the representative, guide, architect, recounter, opinion-holder, and originator. In COMP 101, only four roles were observed: **representative**, **guide**, **architect**, and **opinion holder**, and each of these categories is described and illustrated in examples 1-4 below:

1. **Representative:** "However, considering that we are a nation that the pilgrims started." (collective identity).

This category uses *we* to represent a collective identity with minimal power. In this example, *we* identifies with a national community without asserting individual authority. This role is described to minimize the writer's individual presence, positioning it as the least powerful role.

2. **Guide:** "We see this in the previously discussed Supreme Court case of Jacobson v Massachusetts." (navigation).

This category uses *we* with perception verbs like *see* or *note* to guide the reader through the text by highlighting certain items or directing attention to specific points. This role positions the writer as navigating the content with the reader and conveying a shared perspective.

3. **Architect:** "I will explain the cooking process in three stages." (organization).

This category uses *I* to organize the text's structure. Unlike the guide role, which uses *we* to navigate the content collaboratively, the architect uses *I* with modal and activity verbs, such as *will*, *explain*, or *tell* to outline the essay's framework for the reader.

4. **Opinion-holder:** "I believe as long as both parties have clear intentions and share similar moral values and beliefs, either style of dating will be beneficial." (opinion expression).

The opinion-holder category is characterized by the use of *I* or *we* with cognitive or mental verbs, such as *believe*, *think*, or *feel,* to assert the writer's personal stance. This role is among the most powerful roles, as it directly conveys the writer's perspective on the topic.

The analysis process was as follows: first, the Sketch Engine concordance tool identified first-person pronouns and their collocating verbs. Each pronoun–verb pair was then contextually analyzed and classified according to Tang and John's roles. Accuracy and reliability of the classification were established by confirming the identified roles through consultation with supervisors Dr. Brian Clancy and Joan O'Sullivan. Tables 5.3 and 5.4 show all the collocate verbs cooccurring with the first-person pronouns that realize specific roles in COMP 101. The labels in the tables below show (1) numbered lines, (2) the alphabetized collocate verbs, (3) the corresponding LogDice with the first person pronoun, (4) the verb categories based on Biber (1999), and (5) the roles expressed in the immediate contexts based on Tang and John (1999). The personal pronouns in Tables 5.3 and 5.4—*I, me*, *my* for singular and *we*, *our*, *us* for plural—were combined into singular and plural categories to support the identification of the authorial roles based on Tang and John (1999). However, verbs were kept in their original forms, such as *am* (present, singular) and *were* (past, plural), to gain a better understanding how first-year composition writers use different tenses and forms to express time, perspective, or actions in their texts These verb forms support different patterns that help distinguish roles like the representative, architect, guide, or opinion holder.

Table 5.3: Singular first-person pronouns (*I, my, me*) associated with verbs and their LogDice score in COMP 101

N	Verb	LogDice	Verb Category	Roles
1	am	10.45268	Primary	Architect (I, 2)
2	believe	8.93051	Mental	Opinion (I, 16)
3	feel	9.31487	Mental	Opinion (I, 4)
4	going	9.03825	Auxiliary	Architect (I,10)
5	see	8.76241	Mental	Guide (we, 7)
6	think	9.32536	Mental	Opinion (I, 18)
7	will	8.7839	Modal	Architect (I, 8)
8	would	10.24269	Modal	Architect (I, 2)
9	started	8.23543	Activity	Architect (my, 1)

In Table 5.3, the focus is on the use of the singular first-person pronoun and identifies the collocates with a score between 14 and 7. The main verb category is related to mental activities and is associated with either expressing opinions or providing guidance. The architect role is most used with modal verbs.

In Table 5.4, the study lists the collocate verbs ranging between 14 and 7 with the plural forms of the first-person pronoun in COMP 101, demonstrating a preference for the plural form of the first-person pronoun to express the different roles. The prevalent category in the plural dataset is activity, indicating a focus on actions and processes in the plural context. The most frequently used role is the representative, which ranges between mental, primary, modal, and activity verbs.

Table 5.4: Plural first-person pronouns (*we, our, us*) associated with verbs and their LogDice score in COMP 101

N	Verb	LogDice	Verb Category	Roles
1	allow	9.13705	Mental	Representative (us, 4), Guide (us, 2)
2	are	10.07481	Primary	Representative (we, our, 27)
3	be	9.04091	Primary	Guide (our, 4)
4	can	9.96955	Modal	Representative (we, our, us, 21)
5	changed	8.40009	Activity	Representative (our, 4)
6	could	9.08373	Modal	Representative (we, 1)
7	get	9.70586	Activity	Representative (we, our, 6)
8	have/has	10.47143	Primary	Representative (we, us, 22), Guide (we, 4)

9	help	9.76798	Activity	Representative (us, 11)
10	know	10.16931	Mental	Representative (we, 18)
11	let	8.3323	Activity	Guide (us, 3)
12	live	8.37064	Existence	Representative (we, 4)
13	make	8.72595	Activity	Representative (we, us 7)
14	may	9.30063	Modal	Guide (our, 2)
15	need	10.0033	Mental	Representative (we, 5)
16	see	9.63565	Mental	Guide (we, 7), Representative (we, 2)
17	should	8.78298	Modal	Opinion (we)
18	take	8.67357	Activity	Representative (we, 6)
19	think	8.97064	Mental	Representative (we, 5)
20	were	10.45236	Primary	Representative (we, 2)
21	would	10.39065	Modal	Representative (we, 4)

The variety of verb choices related to the plural form of the first-person pronoun in COMP 101 is twice as much as the verb choices with the singular form. Section 5.3 discusses the underlying patterns based on the collocates related to the singular and plural forms of the first-person pronoun.

In addition to analyzing the roles and choices of pronouns and verbs in COMP 101 for first-year composition writers, the study also looks at the same word combinations in BAWE. Table 5.5 summarizes the results for the verb combinations used with singular first-person pronouns in BAWE. The columns in Table 5.5 have the same structure as in Table 5.4. The verbs are filtered based on the LogDice score, which ranges between 14 and 7, resulting in a varying number of verbs.

Table 5.5: Singular first-person pronouns (*I, my, me*) associated with verbs and their LogDice score in BAWE

N	Verb	LogDice	Verb Categories	Roles
1	am	10.43996	Primary	Architect (I, 7) Opinion (I, 50)
2	allow	8.31462	Mental	Architect (me, 25)
3	allowed	7.51556	Mental	Recounter (me, 13)
4	argue	8.21143	Communication	Architect (I, 85)
5	believe	9.85257	Mental	Opinion (I, 91)
6	can	8.13798	Modal	Architect (I, 2)
7	decided	8.32378	Mental	Architect (I, 3) Recounter (I, 6)
8	enable	7.9603	Activity	Architect (me, 14)

9	enabled	7.36428	Activity	Recounter (me, 8)
10	feel	9.94839	Mental	Opinion (I, 50)
11	find	8.07643	Activity	Opinion (I, 10)
12	found	8.86578	Activity	Recounter (I, 5)
13	have	10.13588	Primary	Opinion (I, 1)
14	help	8.30968	Activity	Architect (me, 33)
15	know	8.27208	Mental	Opinion (I, 7)
16	leads	8.14136	Activity	Architect (me, 19)
17	led	7.38529	Activity	Recounter (me, 13)
18	like	8.1735	Mental	Architect (I, 9)
19	look	8.18712	Mental	Architect (I, 53)
20	’m	8.12206	Primary	Architect (I, 3)
21	made	8.00919	Activity	Recounter (I, me, 49)
22	say	8.10468	Communication	Opinion (I, 5)
23	seemed	7.36903	Mental	Recounter (me, 9)
24	seems	8.08034	Mental	Opinion (me, 30)
25	shall	9.18117	Modal	Architect (I, 70) Opinion (I, 1)
26	think	10.01623	Mental	Opinion (I, 51)
27	thought	8.53602	Mental	Recounter (me, 6)
28	understand	7.2747	Mental	Architect (me, 13)
29	use	7.97525	Activity	Architect (I, 24)
30	will	10.13753	Modal	Architect (I, my, 91) Opinion (I, 1)
31	would	9.67709	Modal	Architect (I, 28) Opinion (I, 2)

Similar to COMP 101, the mental verb category is most frequently used with singular first-person pronouns in BAWE. The architect role is predominant, as it is in COMP 101. .

Table 5.6 shows the verb choices with the plural first-person pronoun, following the same labeling structure as in the previous three tables. The first impression, looking at Tables 5.5 and 5.6, is that BAWE writers have a balanced use between the singular and plural first-person pronouns. The collocate verbs used with the plural dataset are with only 6 more verbs than the singular.

Table 5.6: Plural first-person pronouns (*we, our, us*) associated with verbs and their LogDice score in BAWE

N	Verb	LogDice	Verb Categories	Roles
1	allow	9.08613	Mental	Guide (us, 117) Recounter (us, 24)
2	are	9.53384	Primary	Representative (we, 18)
3	assume	8.117	Mental	Opinion (we, 29)
4	brings	7.37612	Activity	Guide (us, 20)
5	can	10.75429	Modal	Opinion (we, 3) Guide (we, 18)
6	consider	8.74325	Mental	Opinion (we, 1) Guide (we, 6)
7	could	8.63038	Modal	Opinion (we, 9) Guide (we, 11)
8	enable	7.97625	Activity	Guide (us, 50)
9	enabled	7.31387	Activity	Recounter (us, 19)
10	find	9.0125	Activity	Guide (we, 8) Representative (we, 5)
11	get	8.43997	Activity	Guide (we, 10)
12	give	8.53801	Activity	Guide (us, 124)
13	have	10.38498	Primary	Guide (we, 3)
14	help	8.53805	Activity	Guide (us, 30)
15	helped	7.08254	Activity	Recounter (us, 5)
16	leads	8.24798	Activity	Guide (us, 60)
17	let	9.81018	Activity	Guide (us, 201)
18	look	8.45906	Mental	Guide (we, 17)
19	make	8.01896	Activity	Guide (we, us, 39)
20	may	8.0857	Modal	Representative (we, 6)
21	must	9.19817	Modal	Guide (we, 11) Opinion (we, 16)
22	need	9.33748	Mental	Opinion (we, 20)
23	presents	7.53701	Activity	Guide (us, 23)
24	provided	7.63138	Activity	Guide (us, 58)
25	reminds	8.17982	Mental	Guide (us, 20)
26	say	8.24699	Communication	Opinion (we, 26)
27	see	10.53836	Mental	Guide (we, 24)
28	shall	8.15669	Modal	Guide (we, 51) Opinion (we, 18)
29	should	8.97646	Modal	Guide (we, 4)
30	shows	7.98636	Communication	Guide (us, 33)
31	take	8.32835	Activity	Guide (we, 7)
32	tell	9.82785	Communication	Guide (us, 209)
33	think	8.1243	Mental	Opinion (we, 9) Guide (us, 26)
34	understand	8.3031	Mental	Representative (we, 12)
35	use	8.75883	Activity	Guide (we, 10)
36	will	9.14867	Modal	Guide (we, 31)
37	would	8.76844	Modal	Opinion (we, 10)

The use of the plural form differs in that there is a preference for the role of the guide over the representative, which was more common in COMP 101. The activity verb category is the most prevalent, but it is primarily used to guide readers through the text, showing that the author's

choices carry more weight. In the next section, the analysis will focus on each role in COMP 101 and BAWE.

5.3 First-person pronoun roles in COMP 101 and BAWE

The results of the demonstrated roles related to the first-person in COMP 101 and BAWE will be discussed in the following section. During the process of identifying these roles, the researcher consulted with Dr. Brian Clancy and Dr. Joan O'Sullivan, who are both supervisors in the study. The roles discussed below have been finalized based on the feedback received from them.

5.3.1 Representative role

Tang and John (1999) define the representative role as the least demanding and decisive role in academic writing. This role is viewed as predominately realized by *we*, not *I*, thus seeking to include and identify with the reader in the narrative. Generally, the role is considered non-threatening since it does not seek to promote one's views and challenge an alternative side. Instead, it assumes a favorable stand towards the reader (Harwood 2005). As Table 5.7 illustrates, in COMP 101, this role accounts for 63 percent of all occurring roles in the corpus, making it one of the most frequently used by writers. The significant use of the representative role may also be linked to the tasks assigned to the students. As discussed in Section 3.3, COMP 101 focuses on various essay genres that may encourage research but do not necessarily require students to write a research paper. The representative role enables students to engage in discussions about everyday topics within the essay genres described by the tasks, helping them connect with the larger society.

Table 5.7: Representative roles in COMP 101 and BAWE

Roles	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Representative (*we, our, us*) Accounts for 63% of all roles in COMP 101 Accounts for 6% of all roles in BAWE	allow	4	are	18
	are	27	can	11
	can	21	find	14
	changed	4	get	2
	could	1	have (main verb)	15
	get	6	know	36
	have	22	make	8
	know	18	may	6
	live	4	see	1
	make	7	take	1
	may	9	understand	12
	need	12	use	5
	see	2	will	5
	take	6		
	think	5		
	were	2		

Even though the role takes a good mixture of primary, modal, and lexical verbs (*are, can, could, get, have, know, live, make, may, need, see, take, think*, and *were*), the writers seem to gravitate towards the primary verbs *are, have*, modal *can*, and mental verbs *know* and *need* (see Table 5.4). In its representative role (see examples in 5.1), *we* demonstrates audience involvement (Harwood 2005) and everyday experiences (Vassileva 1998) that seek to bring the reader into the text and build rapport:

5.1 COMP 101

> **Descriptive Subcorpus:** ***As we all know****, texting and driving is never a good idea.*
> **Narrative Subcorpus:** *We **are divided** as a nation, and the bad side of humanity is seen in the U.S.*
> **Narrative Subcorpus:** *Many times, **we are** all focused on the destination and forget that the journey is important.*
> **Process-analysis Subcorpus:** *If **we are** hungry, we will seek out food.*
> **Cause-and-effect Subcorpus:** *However, if we analyze, **we can see** that all actions of society are influenced by advertising, driving all people's actions in the market.*
> **Process-analysis Subcorpus:** ***We all have** that one special day out of the year that was made for us.*

Classification Subcorpus: *Now, **we all know** about the different friend types, but do you know about the types of energy vibes we get when we are together with our true friends?*

In this early stage of writing, COMP 101 writers seem to recognize the importance of including the readers and identifying them by using the inclusive *we*, a feature that Hyland (Hyland 2001a) calls evidence of reader engagement. Since *we* functions as an engagement marker, the writers predominately use this first-person plural pronoun in the present tense (examples 5.1), which is very similar to its use in earlier research (Tang and John 1999; Harwood 2005). As an exception to the typical use of the present tense, the writers refer to the past in only three instances, and all of them relate to the global pandemic brought by Covid-19. These past experiences are evident in (5.2), via which the writers recall their personal experiences with the shared event.

5.2 COMP 101

Descriptive Subcorpus: ***We didn't know** this was how 2020 would go.*
Narrative Subcorpus: ***We thought** it would be a few weeks at home, and then things would get back to normal.*
Narrative Subcorpus: ***We were quarantined** for months, having to stay in our homes.*

Instead of using the first-person singular, the writers reference past experiences using the plural form, highlighting the commonality of the events (Harwood, 2005) and strengthening the relationship between the writer and reader (Hyland 2001a).

To further understand the use of *we* in COMP 101, the study discusses how it compares to BAWE. Table 5.7 shows that the representative role accounts for only 6 percent of all the roles in BAWE, which may suggest that their tasks could naturaly motivate them to use more potent roles, such as the guide, architect, and opinion holder than the representative. As a reminder, the texts in BAWE are composed of writers from upper level of academic writing and more proficient skills in academic writing.

Like the use of the representative role in COMP 101, in BAWE, the role continues to develop a commonality and engagement with the reader, as example (5.3) indicates.

5.3 BAWE

> ***We are constantly aware*** *of the city's power.*
> *It highlighted the fact that* ***we are creatures of habit*** *and will see things the way we expect them to be, even though we are vulnerable to trickery such as the distorted room.*
> *Still,* ***we can trace*** *a binding element between all users across time and social position.*
> *Humans create problems, at the same time* ***we will find*** *solutions.*

The difference seems apparent in the contexts of use where BAWE writers demonstrate a shift towards discipline-specific content rather than the shared experiences and events as in COMP 101. In BAWE, the writers continue to develop engagement with the reader but move to topics that require more information (example 5.4) than general common sense and engage the reader as a group member through the inclusive *we*. For example, in COMP 101, in example 5.2, the statement, *"*We didn't know this was how 2020 would go," reflects a general observation, which does not seem to require specialized knowledge. This aligns with the COMP 101 focus on personal experiences or common-knowledge topics, as students select subjects of personal interest. In contrast, the examples in BAWE, such as, "In India we find a similar situation to that which we found in the US. Imports and GDP are positively correlated, in fact marginally more so with a correlation coefficient of: 0.979607." and "We know every subgroup of an abelian group is normal..," involve specialized background knowledge and technical vocabulary, which may suggest discipline specific writing.

5.4 BAWE

> *In India **we find a similar situation** to that which we found in the US. Imports and GDP are positively correlated, in fact marginally more so with a correlation coefficient of: 0.979607.*
> ***We know every subgroup** of an abelian group is normal and each generates a cyclic subgroup.*

Even though the initial results of the pronoun-verb examination reveal the representative role is used similarly by the writers in COMP 101 and BAWE, closer scrutiny of the texts shows that the representative role in BAWE is discipline-specific in the use of *we*, which might be connected to what Hyland (2001) discusses as "a clear signal of membership" (558) of the writers who engage with various professional topics. In contrast, COMP 101, being at the beginning stage of the writing journey, features writers who do not address discipline-specific topics of their chosen majors and instead focus on everyday experiences and events because the writing tasks allow them freedom to choose their own topics and do not require research in discipline-specific areas. However, it is important to note that, despite their novice skills, COMP 101 writers show an interest in engaging with the reader, as evidenced by the high percentage of representative role usage across their texts.

5.3.2 Guide role

The role of the guide is a degree higher in authorial power than the representative. Tang et al. (1999) discuss its function as the writer guiding the reader's journey through unfamiliar territory, emphasizing or pointing out obvious or subtle points that will help the reader reach the conclusions the writer shares. Since the writer is the guide followed by the reader, the role is usually realized through the plural form *we* and verbs that denote the mental process of

perception, such as *see, note*, and *observe*. Another way to look at that role is through Hyland's (2002b) perspective, which interprets it as stating a purpose and signaling intentions. Similarly to Hyland is Vassileva's (1998) view on the role of *we* as the ability to clarify the text's methodology and procedures, thus providing a roadmap to the readers. Table 5.8 demonstrates that the guide is not as popular as the Representative in COMP 101, only 10 percent, versus BAWE where its use is 45 percent.

Table 5.8: Guide roles in COMP 101 and BAWE

Roles	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Guide (*we, our, us*)	be (*our, us*)	6	allow (*us*)	100
	can	1	brings (*us*)	20
	have	4	can	18
Accounts for 10% of all roles in COMP 101	let (*us*)	3	consider	6
	see	7	could	11
	will	3	enable (*us*)	50
Accounts for 45% of all the roles in BAWE			find	13
			get	10
			give (*us*)	100
			have	3
			help (*us*)	30
			look	17
			make	39
			must	11
			leads (*us*)	60
			let (only *us*)	100
			presents (*us*)	23
			provided (*us*)	58
			reminds (*us*)	20
			see	24
			shall	51
			should	4
			shows (*us*)	33
			take	7
			tell (*us*)	100
			think (*us*)	100
			use	10
			will	31

As entry-level COMP 101 writers indicate an emerging presence of the guide through the pronoun, *we,* and the verbs *can, have*, *see,* and *will*. Interestingly, the role does not occur in the

texts submitted at the beginning of the semester—descriptive and narrative—but in the later ones, such as classification, process-analysis, compare-and-contrast, and argumentative. The earlier texts often focus on personal writing, such as descriptive scenes or narrative, which rely on a natural order of description or the unfolding of a story. In contrast, analytical topics require a careful textual structure to effectively convey concepts. As the writers move from the personal essays to the analytical, they start to exhibit a specialized use of *we*, which navigates the reader through the various text components. The use of the guide by the first-year students is very interesting since the course instructions does not specifically focus on the role of the guide. This navigational aspect is evident in the use of the introductory phrases that further clarify the where and how in the text as example (5.5) indicates.

5.5 COMP 101

> **Classification Subcorpus:** ***First****,* ***we have*** *the crying case.* (Transitional signal)
> **Process-analysis Subcorpus:** ***Now that we have the terms defined and the variations specified****, it is important to clarify your price range.* (Discourse marker)
> **Compare-and-contrast Subcorpus:** ***In Shatter Island****,* ***we have*** *the same character from the beginning to the end.* (Prepositional phrase)
> **Compare-and-contrast Subcorpus:** ***Another difference*** *that* ***we see*** *in Ledger is the level of madness.* (Noun phrase)
> **Compare-and-contrast Subcorpus:** ***Now****,* ***we will see the difference*** *between the books and the movies.* (Discourse marker)

The introductory phrases are used as signals for the reader and point out specific components of the text that are important to the message. The collocate verbs vary between the limited choice of modal auxiliary—*can* and *will*, primary—*have*, and one lexical verb—*see*. The limited word choice in realizing the role might suggest the writer's struggle and insecurity carry an intentional navigation approach to the texts. Still, a few instances demonstrate that entry-level writers begin

to recognize the importance of creating a textual structure and providing a roadmap for the readers. The narrow lexical pool of choices might speak of the need for focused training in expanding the vocabulary expressions related to the guide.

Shifting the focus to BAWE and referring to Table 5.8, it is immediately evident that 15 collocate verbs realize the guide. The wider lexical variety may indicate the different topics and perhaps required research as part of their tasks as in the following examples: “although not explicit in the chart, we can draw,” “to understand this method, we must first define” (examples in 5.6).

5.6 BAWE

> ***Although not explicit in the chart, we can draw*** *three possible conclusions from this.*
> ***We can also compare*** *the distances traveled by both samples in order to roughly determine the isoelectric point of each.*
> ***We shall then consider*** *the equivalence of Protagoras' views and, indeed, its relevance to modern ethical relativism.*
> ***This is why we consider*** *the interaction of political and economic dimensions among various factors that affect international cooperation and conflict.*
> ***To understand this method, we must first define*** *the stereographic projection, with which you may be familiar.*

In addition to the rise of vocabulary complexity and word choice, the writers seem to identify different disciplinary fields, a notion previously discussed by Hyland (2001a) as one of the rhetorical purposes of using first-person pronouns. Each of the examples in (5.6) relates to particular disciplines and specialized topics that require straightforward navigation. For example, the BAWE writers in example 5.6, write, “We can also compare the distances traveled by both samples in order to roughly determine the isoelectric point of each,” and “We shall then consider the equivalence of Protagoras’ views…” reflect discipline-specific concepts that require specialized background knowledge. In COMP 101, the issues relate to ordinary events,

experiences, or processes, as students are not required to engage in research papers but choose topics of their own interestOn the other hand, the BAWE writers show more specialized topics that might relate to particular tasks within their chosen majors. The collocate verbs go beyond *can* and *will* to include a more expressive array of modals—*could, must, shall, should*—and the lexical verbs dominate the primary—*consider, find, get, look*, *make, see, take*, and *use*. While Tang et al. (1999) indicate the mental verbs as the active force behind *we* to realize the role of the guide, in BAWE, writers successfully use activity verbs *find, get, make, take*, and *use* in creating a roadmap for the reader.

Based on the examples and discussions above, it is reasonable to conclude that in both COMP 101 and BAWE, the writers use modals and introductory phrases to direct the reader to specific actions. Interestingly, Hyland (2001a) makes similar observations about the function of directives realized through modals of obligation, complement to-clause, or introductory phrases—all driven by *we* to guide the reader metadiscoursally. Such signal phrases indicate an almost inseparable component to realizing the guide role in both COMP 101 and BAWE. Wang et al. (2021) refer to such introductory phrases as transitional signals that define specific aspects of the authors' ideas and focus on the inclusive use of *we*.

5.3.3 Architect role

The architect's role is realized through the exclusive *I* rather than inclusive *we*, which allows the writer to pinpoint their role as the creator of the textual structures. Since the role shapes the text, it exhibits similarities with the guide but differs significantly from it in its role to create a purpose and structure rather than stick to the mere roadmap directions (Tang and John 1999). Even though Hyland (2002a) does not refer to the role of *I* as an architect, he discusses in detail

the discoursal aims of the personal pronoun to enable the writers to signal their intentions and provide structure for their message. This role is popular among student writers because it does not assert high power and avoids making explicit claims. According to Table 5.9, the architect role accounts for 9 percent of all roles in COMP 101, while in BAWE it represents 25 percent of all roles.

Table 5.9: Architect roles in COMP 101 and BAWE

Roles	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Architect (*I, me, my*) Accounts for 9% of all the roles in COMP 101 Accounts for 25% of all the roles in BAWE	am	2	allow (*me*)	25
	going to	10	am	7
	started (*my*)	1	argue	85
	will	8	can	2
	would	2	decided	3
			enable (*me*)	14
			have	23
			help (*me*)	33
			leads (*me*)	19
			like	9
			look	53
			’m	3
			Shall	70
			understand	13
			use	24
			will	74
			will (*me*)	91
			would	28

In COMP 101, the role is realized through the collocate verbs *am, going to, will*, and *would*. As in the role of the guide, the signal phrases continue (example 5.7) to accompany the collocate verbs to contextualize the text purpose and structure—*in this essay, for now,* or *the descriptions that I will use*. The entry-level writers recognize the critical function of the signal phrases in providing additional information to frame the text.

5.7 COMP 101

> **Process-analysis subcorpus:** *Because of this fact,* ***I'm going to teach*** *you the process of building and how to decorate the best-looking tree.*
>
> **Process-analysis subcorpus:** ***I will explain*** *the cooking process in three stages.*
> **Classification subcorpus:** ***I'm going to tell*** *you the difference between each group.*
> **Compare-and-contrast subcorpus:** ***I'm going to tell*** *you the difference between each group.*
> **Argumentative subcorpus:** ***I will be arguing*** *why everyone should stay indoors while the covid- 10, or coronavirus, is sweeping across the nation.*
> **Argumentative subcorpus:** *The second thing* ***I would like to point out*** *is WIFI is not always reliable.*

Contrary to the 11 overall percentage of the architect in COMP 101, in BAWE, the architect accounts for 36 percent of all the roles, which makes it the most frequently used. BAWE writers use specific language to create purpose and structure such as the examples in (5.8): "In closing then I would argue," "For now, in this essay, I'm going to assume," or "The description that I will use."

5.8 BAWE

> ***Firstly, however, I am going to discuss*** *each of the styles and establish the clear difference between them before exploring how each approach affects the components of the different plays.*
> ***In closing then, I would argue that*** *the American movement cannot be argued to be anything other than a massive influence on German racial hygiene, but that the extent of this cannot be stretched to suggest that they ought to take responsibility for the way in which it was applied in.*
> ***I will look at the difference*** *between generic and profession-specific skills.*
> ***In this essay, I'm going to evaluate*** *the case for reform of Britain's law on industrial actions, especially in terms of the right of strike and the right of secondary action.*
> ***For now, in this essay, I'm going to assume*** *the benevolent God argument to be true, in order to understand Descartes's theories.*
> ***The description that I will use*** *is to have 'proof with an absence of doubt.'*
> ***In this essay, I would be dealing*** *with the Commission of European Communities (Commission) enforcement action under Art.*

The writers clearly articulate their explicit authors' presence and communicate various goals as they fit into the larger frame of the texts. Interestingly, in BAWE, the predominant main verb is *argue*, as Table 5.9 indicates (e.g., *argue* is used 85 times), which is not the case in COMP 101. The frequent use of the verb *argue* suggests a higher degree of writers' awareness in writing strategies when stating purposes.

5.9 BAWE

> ***I argue*** *this in three stages.*
> ***I would argue*** *that it is difficult to compare this finding with those for Western countries, as the measure used for husbands participation in household work is relative and so is dependent on wives perception of the level of participation of household work and how household work is defined.*
> ***However, I would argue*** *that such conclusions largely misinterpret what Goffman intended to tell us about the 'self', and ignore much of what he had to say about it.*
> ***I will argue that*** *fertility and mortality were important in influencing population change.*

The examples in (5.9) speak of the BAWE writers' critical engagement indicated by the strong choice of main verb, which is not the case in COMP 101 where the modal verbs are predominantly followed by *teach, explain*, or *tell*, and only once with *argue*, which also shows that the COMP 101 tasks are not heavily focused on building arguments or positions as the writing tasks in Section 3.3 indicate. The curriculum at this stage is concerned about the overall textual structure across the different tasks, which might be the reason first-year students do not show frequent expressions of arguing given points. The stronger choices of verbs demonstrated by the BAWE writers may speak of their specific task that might be part of their chosen disciplines since they represent an upper level and may come from already established majors.

Another difference in the use of the architect role in BAWE is the present perfect tense, which helps writers reflect on the texts in (5.10):

5.10 BAWE

> ***I have previously written*** *two or three functions that act on or use lists, but I've never written a function with two lists as its arguments before.*
> ***In conclusion, I have argued*** *that Heart of Darkness represents the modern journey to Foucault's argument that the objective authorial figure is dead.*
> ***I have also noted*** *from the images of the company's products that the designs are not the most aesthetically pleasing.*
> ***I have chosen*** *the Neuman's systems model to use to help explain the assessment process and to guide me when reflecting upon what affected this family.*

In COMP 101, the writers use the architect only regarding their intentions but not self-reflection of what has occurred in the text. In this sense, BAWE demonstrates in (5.10) a more intentional and critical approach to the text where writers move from creating purpose and structures to revisiting them and re-evaluating the text, reminding the reader of the organization and the goal. While the architect emerges in COMP 101 through modal verbs and signal phrases, it reaches a higher level of use in BAWE, expressed in the complexity of the topics, critical engagement, and reflective elements.

5.3.4 Recounter role

According to Tang et al. (1999), the recounter role conveys the steps and procedures in the research process, which implies documenting previously completed work to help the reader understand the text's implications and process. The role is described by both Harwood (2005) and Hyland (2002a): Hyland relates the exclusive use of *I* when students describe research procedures, noting its low degree of authorial power, while Harwood (2005) highlights the use of *we* by researchers to describe the general methodology. The latter may suggest that professional writers prefer *we* to *I* to associate themselves with the larger academic community. In contrast to researchers, student writers typically use *I*, which indicates that they seem to view the research

process from an exclusive view, not yet seeing the larger perspective of the professional community.

Table 5.10 shows the role of the recounter in BAWE, which accounts for 7 percent of all the roles in the corpus. As an entry-level writing corpus, COMP 101 does not include any occurrences with the role. The COMP 101 writing tasks provide students with a wide range of essay genres and basic textual organization to prepare them for the following semesters, but at this stage, students are not expected to adopt the role of a recounter.

Table 5.10: Recounter roles in COMP 101 and BAWE

Roles	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Recounter (*I, me, us*) It does not occur in COMP 101 Accounts for 7% of all the roles in BAWE			allowed (*us*) enabled (*me*) enabled (*us*) found had helped (*us*) led (*me*) made provided (*me*) seemed (*me*) thought (*me*)	24 8 19 6 4 5 13 49 12 9 6

The concordance lines in COMP 101 do not reveal occurrences of the first-person pronoun in documenting the research processes, which shows that research is not an essential area of entry-level writing. In contrast to COMP 101, the writers in BAWE demonstrate the recounter to the overall of 2 percent. A limited number of collocate verbs realize the role—*decided, founded, made, was*, and *had*—all relate to already completed steps. Rausova (2018) notes that the recounter is typically realized with verbs in the past or perfect tense, referring to completed actions, which distinguishes the role from the architect, which is realized through the present tense verbs that state purposes rather than documenting previously completed steps in processes.

In (5.11), the signal phrases continue to be an essential feature that supports *I* and the collocate verbs. Writers use the additional information to help readers see the particular elements of reflection over the steps, procedures, and methods involved in their studies.

5.11 BAWE

> *When I observed this aspect,* ***I decided*** *to test to what extent a reaction could be caused among customers by intruding their private sphere.*
> *However,* ***I did not find*** *any significant evidence for my null hypothesis of convergence in the sample, as the coefficient on initial GDP was negative but not significant.*
> ***I found*** *this search process extremely time consuming, but was able to quickly identify the key debates and primary authors on this subject.*
> *To do this,* ***I had to*** *initially spend time exploring different series by selecting them and displaying them on-screen.*
> *Using the 'data analysis' function of 'tools' in excel,* ***I was able*** *to compute the correlation coefficient of the two series, which worked out to be -0.9899 to four decimal places.*

The comparison between COMP 101 and BAWE shows that the exclusive use of *I* in recounting research marks mature university writing and is scarce even in larger corpora such as BAWE. Also, COMP 101 data indicates that the recounter is not a feature that characterizes entry-level writing but typically occurs in the later stages.

5.3.5 Opinion-holder role

The opinion-holder carries one of the highest degrees of authorial identity in the text and is often realized through *I* and occasionally *we,* along with collocate verbs that express mental processes. Tang et al. (1999) identify a limited use of this role in undergraduate student writing, while Hyland (2002a), categorizes the role as most frequent in professional versus student writing based on its high-risk function. Similarly to Hyland, Harwood (2005) discusses the first-person

pronouns, *I* and *we*, as means for researchers to move between inclusive *we* and exclusive *I* and represent either an academic community or express a personal stance. Hyland (2005b) defines the presence or absence of explicit author reference as a "conscious choice by writers to adopt a particular stance and disciplinary-situated authorial identity" (p.181).

Clearly, self-mention in academic writing is intentional and directly relates to writers' role and expertise in the academic community. Even though this role is more typical for established academic writers, it is not uncommon to see it realized among novice student writers like the ones in COMP 101. Table 5.11 identifies the percentages of this role usage in COMP 101 and BAWE—19 and 18 percent, respectively.

Table 5.11: Opinion-holder roles in COMP 101 and BAWE

Roles	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Opinion-holder (both I and we) Accounts for 19% of all the roles in COMP 101 Accounts for 18% of all the roles in BAWE	believe	16	assume (*we*)	25
	hope	1	can (*we*)	3
	feel	2	consider (*we*)	1
	know	2	could (*we*)	9
	should (*we*)	7	must (*we*)	16
	think	18	need (*we*)	20
	will (*we*)	1	say (*we*)	26
			should (*we*)	18
			think (*we*)	9
			would (*we*)	10
			am (*I*)	11
			believe (*I*)	91
			feel (*I*)	50
			find (*I*)	10
			have to (*I*)	1
			know (*I*)	7
			say (*I*)	5
			seems (*me*)	30
			shall (*I*)	1
			think (*I*)	51
			will (*I*)	1
			would (*I*)	2

In COMP 101, the two most frequent verbs realizing the opinion-holder are *believe* and *think*. The authorial stance in COMP 101 is almost exclusively realized with *I,* and collocate verbs in the present tense, as seen in example (5.12).

5.12 COMP 101

> **Process-analysis subcorpus:** *So now **I believe** that everyone should know how to make the perfect sandwich just in case they are in the time of need.*
> **Cause-and-effect subcorpus:** *Genetically modified foodstuffs have several disadvantages, which **I believe** are possibly avoidable through research and experiments.*
> **Argumentative subcorpus:** ***I strongly believe** that radium in drinking water should be of primary concern; this type of exposure can lead many people to get cancer, kidney damage, and cause birth defects.*
> **Argumentative subcorpus:** ***I think** that if people would educate themselves moreover certain vaccines than they will be able to make the correct decision.*
> **Argumentative subcorpus:** ***I think** that being there physically is the best option because I have experienced them both and in person gives you opportunities that online can never give you.*

In over half of the role occurrences, writers realize their stance, using *I think that* or *I believe that* about topics related to everyday processes, health issues, or education modes. Using the exclusive *I* versus *we* demonstrates the personal stance rather than the voice of the academic community (Harwood 2005), which is understandable given the fact that the writing tasks are not focused on research and discipline-specific writing. The only examples of *we* in COMP 101 as the opinion-holder role are in cases where the writers speak on behalf of the community and what their community should think or do.

5.13 COMP 101

Argumentative subcorpus: ***I think*** *if we inform ourselves about the vaccines than e a choice on which vaccines we need and which ones we really don't need and* ***we should be allowed*** *not to get the ones that we know won't benefit or harm us in any way.*
Argumentative subcorpus: *As God's creation,* ***we should*** *realize He is in existence and hence should simply evaluate ourselves better.*
Argumentative subcorpus: *One reason why* ***we should be involved*** *is the fact that God created government, ordained it, and He commanded us to be "salt" and "light" in everything whether it be government, business or education.*

All the scenarios involving the use of *we* as an opinion-holder in example (5.13) relate to current societal issues where the writers assume that the community or government should act or move in a certain way. None of the uses of *we* shows the writers identifying themselves as part of a professional community or discipline, such as the cases identified by Harwood (2005), which can be expected based on the different focus that COMP 101 has

In BAWE, on the other hand, the opinion-holder accounts for 36 percent of all the first-person pronoun roles. In this corpus, writers use both *I* and *we* to express an opinion; for example, *I* realizes the role 230 times and *we* 137. Table 5.11 indicates a certain tendency in verb choices when writers use *I—believe* (91 times), *feel* (50), and *think* (51)—and when they move to *we—assume* (25), *need* (20), *say* (26), and *should* (18).

Similar to COMP 101, the BAWE writers also deploy *think, believe*, and *feel* with *I*, but based on the contexts, they seem to contextualize their stands by prepositional, adverb, or infinitive phrases.

5.14 BAWE

While they are described as courtesans, ***I think*** *it is fair to say that they are courtesans of a much higher order, and that the Western image of Geisha is grossly misunderstood.*
We assume *that firms value current profits over future profits and therefore take into account a* ***discount factor () to obtain the present value of future profits****.*

Although I do not deny the importance of the logical consistency problem on its own, ***I do think*** *that there are more interesting insights to be gained by considering the problem of* ***Nietzsche's authorial credibility****.*
To allow a consistent approach to patients, I think *it is necessary for a health care professional to participate in regular reflection of their own behaviour and attitudes.*

In example (5.14), the exclusive *I* seem to articulate positions within specific disciplines, for example, "it has increased my clinical understanding" or "to allow a consistent approach to patients". The topics do not relate to common issues of concern but appeal to particular readers, which is not a feature of first-year composition writing but one that emerges with years of university experience. In addition to *I*, writers use *we*, *you* and *I*, assuming that the readers will accept the message, and as such, *we* expresses collective identity, becoming a communal pronoun, which indicates a set of mutual disciplinary understandings (Hyland 2001a; Harwood 2005). This communal use of *we* sets it apart from the exclusive *I*, demonstrating that the BAWE writers make statements on behalf of their academic community, a characteristic particular only to this role. This use of the opinion holder may also suggest that the BAWE tasks are more related to their chosen disciplines than independently chosen topics, as COMP 101 tasks allow.

The BAWE writers also demonstrate an easy move from the exclusive *I* to the inclusive *we* in example (5.15):

5.15 BAWE

We assume *that firms value current profits over future profits and therefore take into account a discount factor () to obtain the present value of future profits.*
We need *now though to amplify the output signal of the circuit as the strain gauges in the circuit will only be changing their resistance by a small amount, so the output voltage will therefore only change by a small amount too.*
We need *to see these and other global companies not only work on a more local basis but also see their service become more intimate and personal from the workforce side.*
We should *not forget that courts are not helpless against arbitrary pursuits of alleged bias and may put the party that raises such allegations at risk of an order for costs.*

> ***To apply*** *this to the London bombers,* ***we need*** *to assume that they were frustrated vicariously as a result of the perception of their in-group's frustration in the middle east. Furthermore* ***we can say*** *that the main cause of technological change lies elsewhere, perhaps due to commercial pressures from increasing demand.*

In (5.15), the writers assume that their readers will accept the stated positions and express collective identity, which seems to indicate mutual disciplinary understandings (Hyland 2001a; Harwood 2005). This communal use of *we* sets it apart from the exclusive *I*, marking writers' maturity in BAWE and their ability to make statements on behalf of their academic community, a characteristic particular only for this role. While the exclusive *I* leaves the individual mark of the writer in the field, the inclusive *we* demonstrate membership and identification with a specific group. Both identities unpack essential features that the writers carry and navigate the audience through the message.

Another interesting observation in using *we* is the absence of introductory phrases or concise introductory phrases. It might be speculated that the inclusive pronoun allows writers to use a more direct approach in expressing their stance since it carries a diplomatic tone by including the reader on the author's side (Harwood 2005). In contrast, the exclusive *I* is a high-risk-taking feature and requires more immediate support to gain the reader. COMP 101 realizes the opinion-holder only through the exclusive *I* and demonstrates writers' stands and perspectives on common issues. Even though the role has a limited occurrence, it still shows that writers are aware of their voices and are brave enough to speak as individuals.

Looking at all four roles, first-year students rely heavily on the representative role, which accounts for 63 percent of all roles in COMP 101. This appears to be the most convenient role at this stage, primarily because it is non-confrontational and stands with the reader at large, as well as it is more suitable for the writing tasks which focus on seven essay genres (e.g., descriptive,

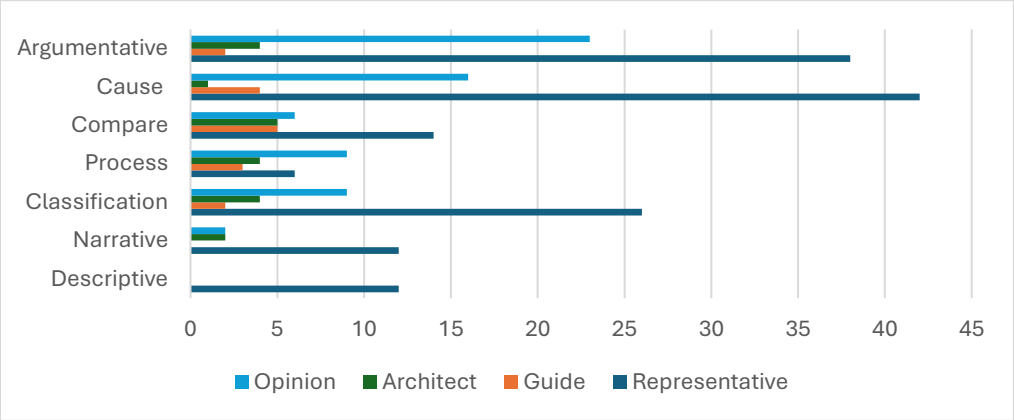
narrative, etc.). Contrary to this use, in BAWE, the representative role accounts for only 6 percent, suggesting that upper-level writers may have different tasks that require more specific engagement with the textual structure and sharing a stance on given topics. The next prominent role in COMP 101 is the guide, which makes up 10 percent of all roles, but in BAWE, this role is used 45 percent of the time, making it the most frequently utilized role among upper-level students. The guide's main function is to navigate the reader through the text by pointing out specific items and clarifying procedures. Similar to the guide, the architect is used 9 percent of the time in COMP 101 and 25 percent in BAWE. The main focus of the role is on the structure of texts and the objectives. Both the guide and the architect roles highlight emerging patterns among first-year students, indicating a need for continuous classroom instruction that helps students grasp the importance of both building their texts and guiding the reader through them.

The two last and most potent roles are the recounter and the opinion-holder. The recounter role is only used by upper-level students, accounting for just 7 percent of all roles in BAWE, but it is not used in COMP 101 since first-year students are not required at this stage to write research papers. Lastly, the opinion-holder role shows 19 percent in COMP 101 and 18 percent in BAWE. Although this role seems to rise among first-year students, it is generally used with generic or everyday topics, while upper-level writers apply it in a discipline-specific context, demonstrating growing expertise and skill. The everyday topics related to the first-year students are based on the free choices that the COMP 101 curriculum allows when students need to select a topic for the different essay genres or tasks.To encourage first-year students to use discipline-specific language, it may be beneficial for the curriculum to include research projects related to their chosen or intended majors. Such a strategy could help students focus on relevant topics and enhance their knowledge and writing voice.

5.4 Role distribution of *I* and *we* in the subcorpora

COMP 101 is divided into seven subcorpora based on the text type: descriptive, narrative, classification, process-analysis, compare-and-contrast, cause-and-effect, and argumentative. The COMP 101 role distribution is displayed in Figure 5.1. The texts are ordered in descending order—the latest texts on the top (e.g., argumentative) and moving down to the earliest at the bottom (e.g., descriptive). Since the recounter is used only in BAWE, the figure shows the four roles present in COMP 101.

Figure 5.1: Role distribution in the subcorpora



The data suggests several inferences about the use of different writing roles across the various genres. In descriptive and narrative texts, the representative role is most common, indicating that at the beginning of the semester, students tend to focus on basic descriptions and storytelling without asserting strong authorial control, which also reflects the writing tasks. In classification, the representative role is still frequently used, but writers begin to use the opinion holder, the guide, and the architect, suggesting that students engage more with complex organizational structures as they move to analytical writing. In process-analysis texts, the representative role

continues to decline, which may indicate that the process task requires a different approach in framing the topics. . It is interesting to observe that the guide role becomes significantly more present as the essay genre and tasks shift to analytical writing that demand clear and structured guidance. The role of the opinion holder continues to be used steadily, suggesting that writers are becoming more comfortable expressing opinions about processes. In compare-and-contrast texts, all four roles are well represented, suggesting a balanced use. The role of the guide peaks in this genre, which may indicate improvement in the organizational structure of the text as students gain more experience with the previous tasks and knowledge during the semester. Finally, in causal and argumentative writing, the opinion holder has its highest representation, despite the fact that the representative role continues to be predominant, which also demonstrates the nature of the persuasive tasks. Both the guide and architect roles are present, as these essays require clear explanation and organization.

Considering Tang et al. (1999) taxonomy of authorial power with its various degrees reflected in each role—assigning the representative to the least confident level, moving through the moderate degree of the guide and architect, and ultimately reaching the demanding and potent role of the opinion-holder—may unfold students engagement with the different essay genres or tasks, indicating higher authorial power in the genres near the end of the semester such as cause-and-effect and argumentative writing while lower authorial control goes in the earlier part, such as the descriptive and narrative tasks. As students progress through their writing journey, they gradually begin to deploy a larger number of roles since they are introduced to different essay genres and tasks. Hyland (2001a; 2002b) and Harwood (2005) discuss the assets of the intentional *I* and *we* usage in academic writing to create structure and stance and distinguish such use as the mark of mature or professional writing.

5.5 Conclusion

The analysis of the first-person personal pronouns in COMP 101 and BAWE finds that the exclusive *I* and the inclusive *we* realize important roles in university writing, clearly distinguishing the entry-level writers in COMP 101 from the upper-level writers included in BAWE. The striking difference between COMP 101 and BAWE is in the number of times the pronoun-verb collocates exhibit specific roles. In COMP 101, the authorial roles gravitate toward lower stakes and slowly move to the more assertive ones at the end of the semester, which also correlates with the type of tasks they have to produce during the semester. On the other hand, in BAWE, almost every collocate verb is used to realize a role in the texts. Also, considering the degrees of authorial power, writers in BAWE demonstrate significant attention in establishing roles, such as the architect, opinion-holder, and even the recounter of the research process, which may also indicate the different tasks these students have to produce in their setting. . The analysis of subcorpora provided insights into the writing trajectory of the COMP 101 students in their first semester of composition writing and the genres corresponding to the overarching structure of each subcorpus. The next chapter will focus on using the second person in COMP 101 and explore how this usage differs in upper-level writing, such as in BAWE.

Chapter 6

Second-person personal pronouns

6.0 Introduction

This chapter focuses on the use of the second-person pronoun—*you*—in COMP 101. The second-person pronoun is ranked as one of the top twenty-five most frequent words in COMP 101, indicating its popularity among these writers (Table 6.2). Since this study aims to examine the most frequent linguistic features in entry-level composition writing and understand the functions that the second-person pronoun serves in this context, this feature qualifies for the examination. Traditionally, academic writing is described as focused on objectivity and formality, which makes the second person not a typical feature of this writing type (Bennett 2009; Liardet et al. 2019). The second-person pronoun occurs in academic writing only occasionally and is discussed by Hyland (2001, 2005) as an engagement marker, helping writers acknowledge their readers and establish a connection with them. This chapter examines the uses of *you* in first-year composition writing based on the analytical framework outlined in the following section, and how those uses differ from upper-level writing as in BAWE.

6.1 Analytical Framework

Generally, the second-person pronoun is regarded as taboo by style manuals, emphasizing the importance of authors' distancing themselves from the readers, suggesting impartiality and precision (Hyland and Jiang 2017). Biber et al. (1999) characterize the second person as a feature of everyday conversations rather than academic writing and estimate its use to be 25 percent more common in conversations than in formal writing, while Liardet et al. 2019 characterize it as a marker of informality. Since formality is associated with objectivity, absence of ambiguity, and misinterpretation, it requires writers to acquire an impersonal voice and distance themselves from colloquial language related to a relaxed and approachable persona, such as the use of *you*

(Hyland and Jiang 2017; Liardet *et al.* 2019). Thus, across disciplines, academic conventions agree on the strict avoidance of the second person in academic texts, thus excluding it from formal writing and deeming its properties informal.

Several scholars (Whitley 1978; Bolinger 1979; Kitagawa and Lehrer 1990; O'Connor 1994; Berry 2009; Stirling and Manderson 2011) distinguish between two types of *you*: generic and specific. In academic writing, even though very seldomly used, the second-person pronoun occurs mainly in its generic type (Wales 1996; Carter and McCarthy 2006). This generic occurrence of *you* has become one of the distinguished markers of informality in academic writing (Hyland and Jiang 2017; Liardet *et al.* 2019). While Chang and Swales (1999), Hyland and Jiang (2017), and Liardet et al. (2019) discuss the occurrence of the second person, they do not distinguish between the different functions that the second-person pronoun realizes and note the need for further research.

Whitley (1978) points out the challenge in defining *you* beyond its deictic use established by the speaker in a given situation, addressing its complex make-up in person and number, with no distinguishing characteristics between singular and plural when creating an impersonal meaning that is interchangeable with *one*. Such impersonal use of *you*, according to Whitley (1978), tends to occur in sentences expressing viewpoint, obligation, possibility, procedure, and narration. In this type of usage, *you* is usually paired with expressions such as *never*, *always*, or modal verbs. Bolinger (1979) confirms the interpretational challenges based on the complexity of *you* and the contextual importance in determining the impersonal types, highlighting another use of the second person—the interpersonal—that enables the speaker to personalize and generalize the topic at the same time.

When identifying the second-person usage within COMP 101, the first step is to discuss the framework established in earlier research related to the topic. Chang and Swells (1999) rank the second person as one of the distinguishing language features of informality and motivate other researchers to rate texts and investigate the occurrence of such features. Hyland and Jiang (2017) measure a slight increase in informality, including the use of *you,* in electrical engineering and biology but a decrease in applied linguistics and sociology, thus concluding the need for further research. Another research to support the use of the second person is conducted by Leedham (2012), who finds that undergraduate writing is marked by greater informality than professional academic writing. In trying to measure the use of informal and formal features in academic writing, Liardet et al. (2019) suggest that "rather than viewing the formal/informal distinction as a language binary, it can be understood as varying degrees of more or less formal (p.148)," which seems to be an appropriate positive lens for undergraduate writing and one that this study will adopt.

Kitagawa and Lehrer (1990) move beyond the two types of using the second person, providing a precise characterization of the functions it plays in the text. The same framework is applied and extended in O'Connor's (1994) research. Kitagawa and Lehrer (1990) distinguish between three types of *you*: referential or deictic use, which refers to specific individuals identified in the context of the speech situation; vague or membership use, which applies to specific individuals who are not identifiable by the speaker; and general or impersonal use, which applies to anyone or everyone and can be replaced by the indefinite pronoun *one*.

Among these three types, the impersonal use of *you* addresses everyone or anyone, which leads to its name—generic use (Kitagawa and Lehrer 1990; Berry 2009), and it seems to be the closest one to the engagement marker of the second person that Hyland (2001a; 2005a) discusses in

relation to how writers acknowledge readers and connect with them. Discussing these involving qualities of *you*, O'Connor (1994) expands on Kitagawa and Lehrer's (1990) framework and refers to the generic use as involving, stating that it helps the writer to simultaneously generalize and personalize the topic. Such involvement allows the writer to share their viewpoint while inviting the reader to do the same: "It was like you could feel it through the skin partly but you couldn't do nothing about it" (O'Connor 1994, p.48). This use of the second person seems to be a natural way for undergraduate students, who are often new to academic writing conventions, to express their thoughts and ideas, often relying on their default practices.

Only Kitagawa and Lehrer (1990) distinguish between three different sub-functions of the generic use of the second person: structural knowledge, moral formulation, and life drama (see the examples in Table 6.1). Structural knowledge is realized using *you* and the present tense verb to describe the structure of the process or experiment. Such structural use may include the use of the second person in writing instructions or directions as part of assignments or processes the reader must follow. The moral formation, a function discussed in previous research by Laberge and Sankoff (1979), expressed through *you* and a present tense verb to communicate a generally admitted truth. Life drama is the only sub-function in the generic *you* that differs from the other two, using the present progressive rather than present tense to stress a continuous event.

This study focuses on three functions of the second-person pronoun based mainly on Kitagawa and Lehrer (1990), which include referential, membership or vague, impersonal, and involving use of the second person. Table 6.1 summarizes the characteristics of each function and lists their corresponding examples.

Table 6.1: Functions of *you*

Referential *you*	Generic *you*	Membership *you*
Based on Kitagawa and Lehrer(1990), *you* identifies specific individuals in the discoursal context. In a classroom setting, a teacher might say: *You need to submit your homework by Friday*. In this sentence, *you* specifically refers to the students in the class, making it clear that the instruction applies to them directly.	Addresses anyone and/or everyone with three sub-functions: Structural knowledge—*You react instinctively at a time like that (Kitagawa and Lehrer 1990, p.749)* NOTE: *You* used in instructions or directions. Moral formulation—*You kill to raise your kids properly (Kitagawa and Lehrer 1990, p.750)* Life drama is based on the limited scene-setting in a progressive mode—*You are in Egypt admiring the pyramids and feeling that you have really left your own world (Kitagawa and Lehrer 1990, p.751)*	Addresses specific individuals or subgroups who the speaker does not identify; also known as vague use—Kitagawa and Lehrer (1990, p.743) illustrate the function—a European addressing an American: *You're—I don't mean you personally – you're going to destroy us all in a nuclear war* In recent research (Stirling and Manderson, 2011), it is discussed as shared experiences or membership category with its authority/credibility deriving from the group.

The next step in determining the functions of *you* in this study is detailing the procedures and tools used that help uncover the patterns co-occurring with *you* in the context of COMP and the possible interpretations based on the three functions discussed in this section.

6.2 *You* in COMP101: Procedure and tools

The COMP 101 raw frequency list (see Chapter 4, Table 4.4) ranks *you* at position 19 and demonstrates its extensive use by first-year composition writers, making it an item of interest to the research. Table 6.2 excludes the punctuation marks from COMP 101 and BAWE lists and focuses only on the word frequency. Both frequency counts are normalized per 1,000,000 words. In Table 6.2, *you* is ranked at position 16. In BAWE, compared to COMP 101, the second person does not occur in the top 25 most frequent items and is in position 215 with 3297 raw frequency and 473 normalized frequency in the BAWE frequency list.

Table 6.2: *You* in the top 25 most-frequent items in COMP 101 and the BAWE (normalized per million words)

N	COMP 101	N Freq	BAWE	N Feq
1	the	56562	the	70646
2	to	32760	of	38903
3	and	29896	and	29950
4	of	23966	to	27497
5	a	21771	in	22004
6	in	17403	a	19575
7	is	17132	is	15974
8	I	13067	that	11386
9	that	12929	as	9769
10	it	11664	for	8548
11	for	9836	be	8341
12	are	9432	this	7806
13	they	7424	it	7355
14	with	7408	are	6134
15	my	7402	with	6072
16	you	7232	on	5833
17	was	7211	by	5821
18	on	6855	was	5289
19	be	6818	not	4805
20	not	6812	from	4514
21	have	6494	an	4195
22	as	6122	which	4186
23	this	6047	’s	4147
24	we	5819	have	3677
25	their	5208	can	3642

The frequency of the word *you* in the COMP 101 dataset is within the top twenty most frequently used words (7232 times per million words). This high frequency is not the case with the BAWE list. This stark contrast highlights the more conversational and informal nature of first-year composition writers, where direct address and personal engagement seem to be common. In contrast, the BAWE writers do not use the second person with the same frequency, suggesting a more objective and structured style. This study delves into the functions of the second person in COMP 101 to examine its use in the context and compare it with BAWE.

Similar to the procedure and tools in Chapter 5 that discusses the use of the first person, this section uses the collocation and concordance tools in the SketchEngine interface (Kilgarriff *et al.* 2014) to investigate the reasons for the frequency differences between COMP 101 and BAWE. The collocation tool identified verb usage with the second person and determined its functions based on Kitagawa and Lehrer (1990) by looking at the verbs in their context through the concordance tool.

The study filtered the collocate verbs based on their LogDice score, which resulted in a different number of verbs between COMP 101 and BAWE. The LogDice score ranges from 0 to 14, and a score of 7 indicates a moderate low. LogDice score is used for comparability across corpora. After identifying the verbs with a score of 7 or higher in each corpus, the study examined the individual contexts of verb-pronoun pairs. Table 6.3 shows the collocate verbs used with the second person in COMP 101 and BAWE and their LogDice scores.

Table 6.3: Collocate verbs used with the second person pronoun in COMP 101 & BAWE and their LogDice scores

	COMP 101	LogDice	BAWE	LogDice
1	are	10.6301	are	8.09226
2	re	9.66474	can	8.33454
3	be	9.68355	consider	7.85952
4	can	11.1023	do	9.49484
5	could	8.55531	feel	7.97763
6	did	7.50125	get	8.63264
7	do	10.5931	go	8.33855
8	feel	9.17935	have	8.76010
9	find	9.16322	help	7.69434
10	get	10.17530	know	9.50664
11	going	8.69439	look	8.06678
12	have	10.62440	must	7.95472
13	is	8.59145	need	8.09191
14	help	8.71645	say	7.75905

15	know	9.90243	see	8.99077
16	make	9.69622	should	8.52907
17	may	9.47018	want	8.52737
18	must	8.58762	will	7.98939
19	need	9.93345	would	8.23038
20	put	8.87072		
21	see	9.49594		
22	should	9.25853		
23	take	8.57187		
24	want	10.40568		
25	will	10.78499		
26	would	8.98377		

Table 6.3 shows that COMP 101 has seven more collocates than BAWE. In COMP 101, the frequent use of modal auxiliaries may suggest a strong focus on expressing possibility, necessity, and recommendations, while in BAWE, there are verbs such as, *consider*, *say*, and *look*, reflecting a more observational tone.

The functions of the second-person pronoun *you* in the COMP 101 corpus are based on Kitagawa and Lehrer's (1990) taxonomy, which organizes the second person into the following categories: **referential** (addressing specific individuals), **generic (structural knowledge**, **moral formulation**, and **life drama)**, and **membership** (addressing subgroups) (see Table 6.1). In the COMP 101 data, only referential and generic uses were observed, with the generic function encompassing all three sub-functions. Each observed function is described and illustrated in examples 1-4 below:

1. **Referential:** "I said, 'Hi, do you remember me?'" (direct address)

This category uses *you* to address specific individuals within the dialogic space. In this example, *you* refers to a specific individual in a dialogue. Unlike the generic function, which addresses

anyone, the referential *you* targets a defined individual in direct speech, marked by quotation marks.

2. **Generic: Structural knowledge:** "But before you begin this process, remember to make your jack-o'-lantern on a proper date." (instructional process)

This sub-function of the generic *you* uses the pronoun with present-tense verbs to describe procedural knowledge, often in instructional contexts. In this example, the writer addresses anyone who is undertaking the task described and provides directions on how to accomplish it.

3. **Generic: Moral formulation:** "The one thing that attracts many people is the thrift stores where you find clothing from different cultures." (general observation)

This sub-function uses *you* and the present tense to express generally accepted truths. In this example, *you* refers to anyone and generalizes a shared lifestyle experience.

4. **Generic: Life drama:** "The main purpose of action movies is to make you feel adrenaline while you are watching the movie."(scene-setting)

This sub-function uses *you* and the present progressive tense to create vivid and continuous scene-setting. In this example, *you* invite the reader to imagine experiencing adrenaline, enhancing the experience through the progressive tense.

The process mirrored that of Chapter 5 for first-person pronouns. The Sketch Engine concordance tool identified second-person pronouns and their collocating verbs with a LogDice score of 7 or higher (Table 6.3). Each collocate pair was contextually analyzed and classified according to Kitagawa and Lehrer's (1990) taxonomy. Accuracy and reliability were ensured through consultation with the dissertation-study supervisors, Dr. Brian Clancy and Joan O'Sullivan.

Looking at these collocations allows the study to examine the patterned usage of these grouped words while the concordance lines show their uses in the context of the corpora (Biber *et al.* 1998). The next section discusses each function in the context of COMP 101 and BAWE.

6.3 *You* in COMP 101 and BAWE

The following section focuses on the different ways writers use the second person in COMP 101 and discusses each function of *you* within the corpora contexts: *you* in reference, *you* in membership, and *you* in the following type of generic meanings: structural knowledge, moral formulation, life drama, and involving *you*. This analysis contributes to the research related to academic writing by providing a perspective of the different type of generic use that *you* has in first-year composition writing (COMP 101) and comparing the use to advanced university writing (BAWE). Table 6.4 shows the distribution of the functions of the second person in each corpus based on examining the concordance lines with their raw and normalized occurrences.

Table 6.4: Functional distribution of *you* in COMP 101 & BAWE

Functions	COMP 101	BAWE
Referential *you*	34 (raw occurrences) 180 (normalized per 1,000,000)	107 (raw occurrences) 15 (normalized per 1,000,000)
Generic *you*	511 (raw occurrences) 2707(normalized per 1,000,000)	138 (raw occurrences) 20 (normalized per 1,000,000)
Membership *you*	0	0

The results show that the second-person pronoun is used generically and referentially in COMP 101 and BAWE, but it is not used to express membership. None of the corpora reveal membership functions, which suggests that writers do not distinguish between a specific group of people and people at large when using the second person. The comparison shows that the

occurrences of the second person in BAWE are minimal compared with COMP 101, which shows that first-year composition writers rely heavily on *you* when expressing their thoughts and ideas. The differences in using the second person between the two corpora might be due to the differences in the tasks required by the students and also the differences in the competence levels between the writers. The tasks outlined in Section 3.3 specifically instruct students not to use the second-person pronouns in classification, comparison and contrast, cause and effect, and argumentative essays, yet *you* remains prevalent, which possibly indicates that some COMP 101 writers struggle to adhere to these restrictions. As a distinct feature of informality (Chang and Swales 1999; Hyland and Jiang 2017), the second person seems to characterize some of the first-year composition texts (COMP 101) as less formal and shows upper-level (BAWE) writers use it sparingly. The next section discusses each function in the context of COMP 101 and BAWE.

6.3.1 Referential use of *you*

The first function of the second person that this study discusses is the referential use, which is the first type in Kitagawa and Lehrer's (1990) categorization. This function typically occurs in conversations where the speaker refers to the addressee and constructs a social dialogue (Kitagawa and Lehrer 1990), which is an informal use. Conversations often involve the repetitive use of verbs that indicate actions (*get, do, find*), mental status (*find, know, see*), and communication (*say*) (Biber *et al.* 1999). The referential type of *you* and *your* addresses specific individuals identified in the speech situation.

Table 6.5 shows the collocate verbs that co-occur with *you* and *your* to realize the referential function in COMP 101 (*find, get, know, see*, and *want*) and BAWE (*feel, get, know*, and *see*). The

table presents the raw frequencies of the collocate verbs with the second-person pronoun. They are not normalized in the dataset because the focus is on the verb types rather than the precise comparison of these occurrences in the two corpora.

Table 6.5: Referential use of *you* in COMP 101 and BAWE

Type of Use	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Accounts for 8% of all uses in COMP 101 Accounts for 22% of all uses in BAWE	did	4	will	28
	get	4	are	18
	have	4	know	16
	see	4	help	13
	are	2	get	9
	be	2	need	8
	going	2	feel	6
	know (you and your)	2	do	3
	want	2	have	3
	will	2	're	2
	would	2	see	1
	can	1		
	do	1		
	find	1		
	take	1		

In COMP 101, all the referential uses appear in the narrative texts (NS) and demonstrate speech situations specific to each writer, as the examples in (6.1) show. The narrative essay task, which requires five paragraphs relating a meaningful event with dialogue, specific characters, and sensory details (see Section 3.3), encourages writers to use the second-person pronouns in dialogues with specific speakers, identified by name or narrative context. The first-year composition writers detail dialogues with specific interlocutors, some identified by name while others by the clues of the narrative context. None of these uses are impersonal, relating to generic use, and are not replaceable by the impersonal pronoun *one, anyone*, or *everyone*.

6.1 COMP 101

> **Narrative Subcorpus:** *Ralph, still not believing his ears at the moment,* ***"You will be so happy*** *to be up there in the mountains, Henry, I know it."*
> **Narrative Subcorpus:** *I said, "Hi,* ***do you remember*** *me?"*
> **Narrative Subcorpus:** *Meanwhile, I heard a man's voice speaking: "Hi, little girl,* ***what do you want****?"*
> **Narrative Subcorpus:** *Ann said, with a genial smile, "Sure,* ***see you in the self-study*** *at night."*
> **Narrative Subcorpus:** *I answered, "****Yes, I know you****,* ***you broke my legs*** *five years ago, and* ***you didn't apologize****."*

The uses of *you* in example (6.1) specifically identify the addressee by their name (e.g., *Henry*), adding specific details (e.g., little girl), or to the context of the story (e.g., *do you remember me* or *what do you want*). Each communicative event in the texts refers to a specific individual or situation. In narrative texts, the referential use of language is common and includes the use of activity verbs and informal grammar, such as contractions, which Biber et al. (1999) notes as features demonstrating the conversational nature of the texts. This task-driven use of *you* aligns with the narrative genre's emphasis on personal storytelling, supported by the task instructions.

In BAWE, the referential use accounts for 22 percent of the overall use of the second person. Example (6.2) lists the typical speech events writers facilitate using *you* or *your* in BAWE. The apparent similarity between COMP 101 and BAWE is the conversational use of *you* and the frequently repeating common verbs (*are, feel*, or *give*). In example (6.2), none of the addressees appear identified by name or being described by external characteristics. In contrast to COMP 101, however, the communicative events in BAWE seem to point to a specialized context—a conversation between a medical professional and a patient. The clues for the specialized context are the phrases, such as *your symptoms, you are witnessed to pucker your lips, give you some medication to help with your tremors*, or *you aren't having any further outbreaks*. In BAWE, all the referential uses are contextualized by the doctor-patient setting and demonstrate typical

conversations in the doctor's office. Thus, the informal language features, such as the second person and the contractions, demonstrate records of discipline-specific conversations that might be required as part of the writers' studies and related to task-specific instructions.

6.2 BAWE

> *There are some features of **your symptoms**, which support this diagnosis such as the fact that **you are witnessed to pucker your lips** or stare, a behavior common in some seizure types.*
> *We will **give you some medication to help with your tremors** and **you should feel** much better fairly quickly.*
> *We will continue **to see you** in the outpatient departments under the care of the dermatologists to make sure that the medication is working and that **you aren't having any further outbreaks** of blisters.*

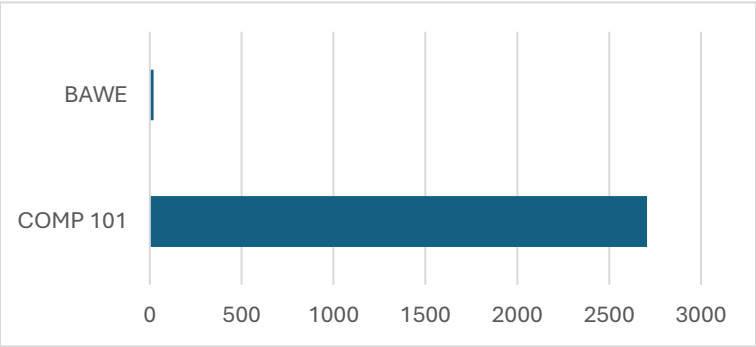
Even though the deictic use occurs in both corpora, in BAWE, it is characterized by a common purpose—writers relate conversations typical for a doctor's office—transcribing visits that reveal a professional aspect of using *you*. The referential use of *you* in COMP 101 occurs only in the narrative subcorpus and supports dialogical texts that seek to immerse the reader in the writers' experiences where the dialogical content involves specific individuals, thus aligning with the narrative genre's focus. On the other hand, the BAWE writers use the second person to convey conversations between medical professionals and patients, which reflects the discipline-specific content, such as *your symptoms*, *you are witnessed to pucker your lips*, or *give you some medication to help with your tremors*. By using the second person, the BAWE writers address particular individuals and reflect on their experiences, which most likely are part of a required practice in medical facilities.

6.3.2 Generic use of *you*

The impersonal or generic *you* is a form of address that can refer to anyone or everyone, and it informalizes students' writing by addressing the reader as *you*. Hyland (2001; 2005b) describes it as an engagement marker in academic writing, but in first-year composition writing, it seems to show a predominantly conversational style rather than engagement with the readers. In this initial stage, the second person offers three ways of expressing positions: structural knowledge, moral formulation, and life drama (Kitagawa and Lehrer 1990). Structural knowledge is conveyed using the second-person pronoun *you* and the present tense verb to describe the structure or organization of a process or experiment. This generic function uses *you* when writing instructions or directions that the reader must follow. Similarly to structural knowledge, moral formation is also facilitated with *you* and the present tense verb, but this time to communicate a generally accepted truth. The life drama function differs from the previous two sub-functions in that it uses the present progressive tense rather than the present tense with the purpose of emphasizing a continuous event (Laberge and Sankoff 1979).

Based on the concordance lines' investigation, the generic *you* is predominantly used in COMP 101 and seldom occurs in BAWE. Figure 6.1 shows a comparison between the generic use of the second person in COMP 101 and BAWE.

Figure 6.1: Generic functions of *you* in COMP 101 and BAWE based on their normalized occurrences



The writers in BAWE use the second person in a generic sense only when expressing structural knowledge (Section 6.3.2.1). In COMP 101, the writers express both structural knowledge (Section 6.3.2.1) and moral formulation (Section 6.3.2.2), and they also use the progressive form associated with the life drama (Section 6.3.2.3) in very few instances. The next three subsections will discuss these three generic functions of the second person in the context of COMP 101 and BAWE.

6.3.2.1 Structural Knowledge

Kitagawa and Lehrer (1990) explain how the generic use of *you* in language conveys structural knowledge by the speaker describing what usually occurs in certain situations. This indicates that the speaker shares these experiences with the wider community. Impersonal pronouns such as *everyone, anyone*, or *one* can often replace the use of *you.* The function typically co-occurs with present tense and modal auxiliary verbs to help identify the generality of the event. When used in this way, speakers describe what typically happens in specific circumstances based on their general experience (Stirling and Manderson 2011). Table 6.6 shows the occurrences of the structural knowledge function and co-occurring verbs in COMP 101 and BAWE.

Table 6.6: Structural knowledge use of *you* in COMP 101 and BAWE

Type of Use	COMP 101		BAWE	
Accounts for 64 % of all generic use in COMP 101 Accounts for 100 % of all generic uses in BAWE	can	57	need	35
	have	51	look	19
	will (you and your)	40	get	13
	are	34	must	13
	should	30	see	13
	need	24	can	10
	do	20	go	9
	get	17	will	8
	may	13	consider	6
	must	11	do	3

	know	9	should	2
	make (you and your)	9	want	2
	put	8	would	2
	going	5	have	1
	take	3	're	1
	be	2	say	1
	is (your)	2		
	would (your)	2		

This usage is particularly common in COMP 101, making up 64% of all instances, and is a primary generic use in BAWE. The table presents the raw occurrences of structural knowledge that is facilitated by the use of the second-person pronoun, along with the co-occurring verbs. The data is not normalized, as the emphasis is on the types of verbs utilized rather than their quantitative representation in the two corpora. Normalization adjusts the data to consider any differences in sample size or total occurrences, but in this case, the raw frequencies are presented for the purpose of examining the specific types of verbs being used. The verbs in Table 6.5 include a high frequency of modal verbs: in COMP 101, *can* is used 57 times, *should* is used 30 times, *will* 40 times, and the semi-modal *need* is used 24 times; in BAWE, *can* is used 10 times, *must* 13 times, and the semi-modal *need* 35 times. Whitley (1978) differentiates between possibility (*can*) and procedure (*make, get, have*), while Kitagawa and Lehrer (1990) treat them as a single type: structural knowledge. The modal verbs, based on Kitagawa and Lehrer (1990), characterize the function of the second person in the text and make it easily recognizable as structural knowledge. The frequent use of structural knowledge in COMP 101 suggests that students at this stage often rely on this function to demonstrate processes and address the reader directly using second-person pronouns like *you* rather than focusing the topic on objectivity. This use is understandable in the process-analysis essays where the task instructions (see Section 3.3) do not specifically prohibit the second-person pronouns, accommodating the need for the directional approach addressing the reader. However, the appearances of *you* in other essays,

such as the comparison and contrast and argumentative, suggest that writers extend the use beyond the process-analysis tasks and may have challenges in complying with formal academic conventions, indicating the need to grow in their level of writing proficiency.

The examples in (6.3) demonstrate some of the typical uses of structural knowledge in the COMP 101 texts where the writers explain different processes involving the reader using the second person pronouns and modal verbs: *you should be, you must gather*, or *a company would sell your data*. The examples in (6.3) represent procedures given by the writer to the reader when purchasing a pet, making a sandwich, explaining data selling, or using a device. The writers seem to focus more on familiar discussion topics, which usually occur in everyday conversations, rather than discipline-specific subjects.

6.3 COMP 101

> **Compare-contrast Subcorpus:** *That is why **you should be familiar** with how your pet choice **will impact you financially**, attention-drawing, and training needs.*
> **Process-analysis Subcorpus:** *First, **you must gather all your ingredients** for the sandwich.*
> **Argumentative Subcorpus:** *Finally, **a company would sell your data to another** company to make a little more capital, helping grow a business.*
> **Process-analysis Subcorpus:** *However, nine times out of ten, **you can immediately use your new device** after turning it back on.*

In the previous examples, the use of the second person serves to engage the reader and make them consider the information being presented on a personal level. This can create a sense of structural knowledge by directing the reader's attention to specific actions they should take, the potential consequences of a particular decision, or the likelihood of a particular outcome. This use of the second person can be effective in academic writing as it helps to make the information

more relevant and applicable to the reader. It can also make the writing more engaging and easier to understand by using language that is more personal and interactive.

In BAWE, writers use the second person generically primarily to facilitate structural knowledge. The examples in (6.4) show the typical contexts in BAWE with the use of structural knowledge. The function refers to procedures written by *you* and modal (*should*), semi-modal (*need*), or directional (*look*) verbs to address the readers and involve them in the text.

6.4 BAWE

> *When selecting software**, you should always look to the future** and assess how easy it is to upgrade existing software and if there are any compatibility issues.*
> *This division holds some weight **when you look closely at the changes** in artifacts such as pottery, grave goods and those connected with ritual practice.*
> *First of all, **you need to create four files** called 'names', email', mobile numbers', and 'sales records' and separate the data into each according to the name of the file.*
> *Nevertheless**, you should be aware** that the size of the market is a crucial factor influencing the decisions of publishers buying translation rights.*

The topics discussed by the BAWE writers include software, ritual practice, data classification, and market, which demonstrate more discipline-specific terms, such as *existing software, artifacts, sales records*, and *translation rights*. The overall use in BAWE does not significantly differ from the one in COMP 101. In BAWE, writers advise the readers to consider future compatibility issues when selecting software, to create certain files and separate data into them, and to be aware of the size of the market when making decisions about translation rights.

Structural knowledge is the leading generic function of *you* in COMP 101 and the main use of *you* in BAWE, making it one of the most significant functions of the second person in both corpora. In COMP 101 and BAWE, the writers use structural knowledge to convey procedures or

instructions to readers, realized mostly through modal, semi-modal, or instructional verbs (e.g., *get, put, take, look*). Even though academic writing, in its quest for objectivity, requires abstention from *you*, it seems to be a natural and fair choice in framing procedures, making the text both personal and generic. Hyland (2001a) discusses this function of the second person in professional academic writing as a means of conveying shared knowledge and engaging the reader in the cognitive or procedural perception that is recognized within the disciplinary landscape. In this way, writers construct texts that make direct and explicit calls for the reader to recognize "familiar topographic features" (Hyland 2001a, p.567) within the discipline.

6.3.2.2 Moral formulation

As a function of the second person, the moral formulation is originally discussed by Laberge and Sankoff (1979) and later included by Kitagawa et al. (1990) as one of the impersonal or generic functions of *you*. This usage is associated with an individual's reflection based on conventional wisdom and carries an evaluative connotation. When *you* is used in this way, the writer downplays the individual experience and presents it as something that anyone could say (Laberge and Sankoff 1979). Whitley (1978) also discusses this type as a viewpoint through which the speakers project their reactions, positions, and inferences to the audience. Researchers, including Whitley (1978), Laberge and Sankoff (1979), and Kitagawa and Lehrer (1990), agree on the generic nature of this usage, suggesting that it could potentially be replaced with the impersonal pronouns *one, anyone*, or *everyone*. Table 6.7 displays the occurrences of the second person in facilitating moral formulation in COMP 101 and its absence in BAWE.

Table 6.7: Moral formulation use of *you* in COMP 101 and BAWE

Type of Use	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Accounts for 34 % of all generic use in COMP 101	get	19	No occurrences in BAWE	
	can (*you* and *your*)	17		
	feel	16		
	will	15		
	have	13		
	see	12		
	're	11		
	know (*you* and *your*)	8		
	need	8		
	are	7		
	should	7		
	may	5		
	would	5		
	find	4		
	make	4		
	want	4		
	could	3		
	do	3		
	help	2		
	take	2		

As the table shows, this function accounts for 34 percent of all generic use in COMP 101, but it does not occur in BAWE. When facilitating moral formulation, the writers in COMP 101 use what Biber (1999) describes as conversational verbs—*are, do, feel, find, get, need, see, take*—to express positions, reactions, and experiences, which may refer to anyone, but instead of using the formal *one*, they gravitate towards the less formal *you*. The topics vary from life lessons, college life, and health to online business ownership—all topics that easily could be formulated by using the formal *one*. By the reoccurring use of *you*, COMP 101 writers seem to indicate that they gravitate towards the less formal expression at this stage. In COMP 101, *you,* in moral formulations, seems to fulfill the role of the engagement marker (Hyland 2017; Hyland and Jiang 2017), which involves the reader and becomes the means of expressing personal observations about life and societal issues, as the examples in (6.5) indicate. O'Connor (1994) calls this type

of *you* an opinionated *you* that positions the addressee as a learner or a novice, and in the examples below the writers pass along such life tips:

6.5 COMP 101

> **Classification Subcorpus:** ***You can let the struggles of life get to you, or you can accept*** *that there will be struggles and* ***do your best to use them to your advantage****.*
> **Compare-contrast Subcorpus:** ***Being in college, you find*** *that there are a lot of things that are just the same as high school.*
> **Argumentative Subcorpus:** ***Vaccines help you in many ways.*** *For example,* ***vaccines protect you from viruses or other illnesses.***
> **Argumentative Subcorpus:** *Not receiving a good score on the test however can limit school options and* ***what degrees you're allowed to proceed*** *to.*
> **Argumentative Subcorpus:** *Private companies are controlling the internet, and* ***you should allow it*** *with some moderation.*

The task instructions, however, for classification, comparison and contrast, and argumentative essays specifically ask students not to use *you* in these genres, which may suggest that some writers rely on informal expressions, potentially indicating developmental challenges in adopting academic conventions.

In contrast to COMP 101, the writers in BAWE do not use this function of the second person, likely due to disciplinary tasks, such as research papers, prioritizing formal third-person constructions to meet academic conventions.

6.3.2.3 Life drama

The *life drama* function of the second person, as discussed by Kitagawa and Lehrer (1990), involves the use of the present progressive tense to convey a continuous event or action in contrast to the present simple tense. This use of the second person creates a mini-tale or narrative

that involves the reader and draws them into the scene being described. The present progressive tense "constrains the scene-setting as a mini-tale whose resolution is presented in the present tense" (Kitagawa and Lehrer 1990, p.749), creating a sense of continuity and ongoing action. The term "life drama" refers to the relationship between the two tenses, with the present progressive tense showing the continued action or event and the present tense showing the resolution. This use of the second person emphasizes the continuous nature of the event and creates a dramatic sense of engaging the reader in the scene.

Table 6.8 demonstrates that the function is not common in first-year composition writing (COMP 101), accounting for less than 1 percent and completely missing from advanced writing (BAWE).

Table 6.8: Life drama use of *you* in COMP 101 and BAWE

Type of Use	COMP 101		BAWE	
	Verbs	**Occurrences**	**Verbs**	**Occurrences**
Accounts for less than 1% of all generic uses in COMP 101	are	7	No occurrences in BAWE	
	will	2		
	feel	1		

Even though the writers demonstrate a very limited usage, the study considers it important to distinguish this generic function from the others since it serves a different purpose and produces different effects on the reader by immersing the reader in the scene or mini tale. The verbs used by COMP 101 show the signature use of life drama in the use of progressive forms with the verb *to be*. This type of *you* is another example that may indicate challenges in adopting academic conventions since the tasks related to the comparison and contrast, cause and effect, and argumentative essays (see Section 3.3) specifically ask students not to use the second-person pronouns.

The examples in (6.6) show the types of scenarios writers relate with the second person in its life drama function: *committing to give shelter to a pet*, *health risks that you are taking, adding your vote*, *you are looking for a service, you watching the movie* and the resolutions: *know if the pet will be the right, inform yourself, feel nervous*, *your vote matters, Hava Java is your place or the purpose of the action movies*.

6.6 COMP 101

> **Argumentative Subcorpus:** *When you adopt a pet,* ***you are committing to give shelter*** *to a pet, and it is imperative that* ***you know*** *if the pet is the right one for you.*
> **Cause-and-effect Subcorpus:** ***Inform yourself*** *of the health risks that* ***you are taking****.*
> **Argumentative Subcorpus:** ***You are adding your vote*** *to a plethora of other voices, yes, but* ***your vote matters****.*
> **Compare-and-contrast Subcorpus:** ***If you are looking*** *for fast friendly service with a variety of coffee styles Hava Java is your place.*
> **Cause-and-effect Subcorpus:** *The main purpose of action movies is to make you feel the adrenalin* ***while you are watching the movie****.*

The progressive tenses used in the paragraphs with the second person involve the reader in various scenarios, which Kitagawa and Lehrer (1990) describe it as "creating a mini-tale" around the reader and showing the solution in the present tense. By using this technique, first-year composition students attempt to persuade on a personal level, as if they are having a conversation with the readers. This function of the second person is not present in BAWE, which supports the emphasis on objectivity that upper-level students display in their texts.

Based on the analysis of the first-year and upper-level student writers, novice writers seem to rely heavily on the second person when conveying structural knowledge, moral formulation, and sometimes life drama. Similarly, the upper-level writers also utilize the second-person pronouns when communicating procedures or structural knowledge, but their focus shifts toward

discipline-specific topics rather than the everyday conversational subjects typical of first-year students. So, while the second-person perspective often signals informal writing, its use is essential for instructional content, as it directly engages readers and proves valuable for both groups of writers.

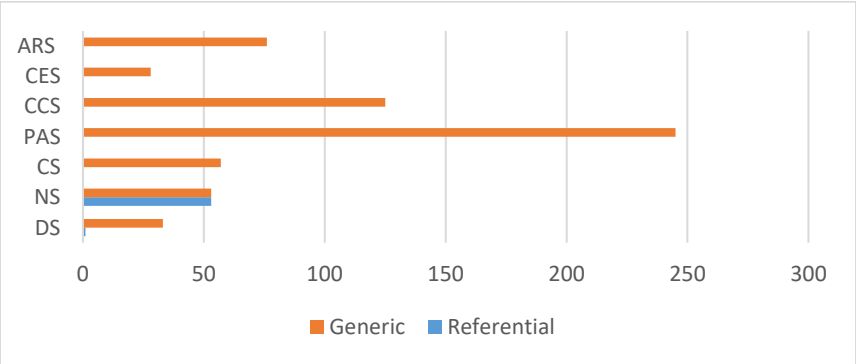
In COMP 101, students are allowed to choose their own topics for the writing tasks, and as first-year students who are not yet fully immersed in their chosen majors, they gravitate towards everyday topics. Such topics help students engage with relatable subjects and help them gain experience with the textual structures. The concerns with the second person arise when it is used to facilitate moral formulation or express opinions, as this demonstrates familiarity with the readers and a lack of objectivity. Such usage shows the importance of developing instructional strategies within the writing curriculum, specifically designed to address objectivity in academic writing and to clarify the appropriate role of the second-person perspective.

6.4 *You* in the subcorpora

To gain a better understanding of the functions performed by the second person in COMP 101, the study provides a brief discussion of these functions across the subcorpora. Figure 6.2 illustrates how their functions are distributed in the subcorpora. The vertical axis displays the subcorpora in descending order, starting with texts submitted at the end of the semester (e.g., ARS or argumentative) and progressing to texts submitted at the beginning of the semester (e.g., DS or descriptive) at the bottom. The analysis considers the influence of the essay genres or tasks—classification, comparison and contrast, cause and effect, and argumentative—required by the tasks (see Section 3.3). The task instructions explicitly require students to use the first or third person, prohibiting the second person in these genres. In the case of the process-analysis

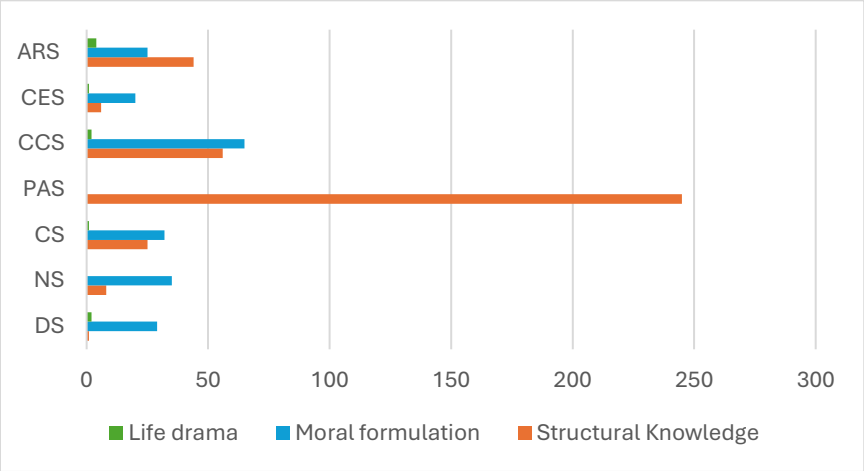
essay, the use of *you* is understood with the need for writers to use a directional approach, addressing the reader. Similarly, descriptive and narrative tasks allow second-person pronouns to convey personal experiences effectively. Based on the distribution, however, first-year students rely on *you* beyond the tasks that allow the use, which may indicate a challenge for this particular stage in their writing and a need for more focused instruction and practice.

Figure 6.2: Normalized distribution of *you* and its functions in the subcorpora



As Figure 6.2 illustrates the referential function of *you* is a feature of only the narrative texts, while the generic is used throughout the subcorpora. The generic usage peaks in the process-analysis text and decreases in the comparative (CCS), causal (CES), and argumentative texts (ARS). Throughout the subcorpora, the generic function is expressed in three different ways: structural knowledge, moral formulation, and life drama. Figure 6.3 shows the distribution of these functions across the subcorpora.

Figure 6.3: Distribution of the second person functions across the subcorpora



The dataset demonstrates that out of the three functions, the writers mostly utilize the second person to convey structural knowledge. The mode of the function is the process analysis texts, which aim to provide instructions and guidelines to the reader. The first-year composition writers use the second person to address the reader directly, as the process analysis essay permits. It is logical to assume that in instructional writing, the second person is an effective method for writers to engage with their readers about specific steps or actions within processes.

Unlike the structural knowledge that peaks in process-analysis (PAS) texts, the function expressing moral formulation does not occur in PAS texts at all but reaches its highest in compare-and-contrast (CCS) texts. This indicates a reliance on the second person in comparison analysis, as writers aim to engage their readers with their findings with the use of the second person. In cause-and-effect (CES) texts, the use of moral formulation is minimal, which might be due to the writers' focus on presenting their causal analysis objectively rather than engaging the reader on a personal level. However, in argumentative (ARS) texts, there is a slight increase in the use of this function, possibly due to the need for students to create and defend positions. It is

important to note that while the second person can create intimacy with the reader, it can also undermine the objectivity of the text and distract from the data and evidence being presented.

The "life drama" function is rarely used in subcorpora. Its highest use is in argumentative texts, which is surprising because argumentation encourages students to focus on objectivity. It seems that the "life drama" might be a more suitable choice in descriptive (DS) and narrative (NS) texts due to its ability to immerse the reader in the scenes on a personal level. However, the results of the concordance lines examination reveal a tendency for some writers to use the function in analytical writing when engaging the reader with the topic by emphasizing prolonged action and thus creating a dramatic effect. As shown in examples (6.6), the instances from the argumentative and cause-and-effect subcorpora show the prolonged actions, such as “you are adding your vote to a plethora of other voices…,” and “inform yourself of the health risks that you are taking,” emphasizing the continuous duration like voting or taking health risks. The different functions of the second person across the subcorpora may suggest some practical implications for composition instruction. First, the high use of the second person in process analysis emphasizes its effectiveness in directly addressing the reader, which can enhance engagement and clarity when providing instructions and guidelines. On the other hand, the presence of the second person in analytical writing can indicate the need for students to understand the importance of maintaining objectivity in cause-and-effect (CES) and argumentative (ARS) texts. In such texts, the use of the second person can undermine the formal tone required for presenting data and defending positions. In compare-and-contrast (CCS) texts, the second person can be used effectively to engage readers, but students should be guided on balancing this engagement with formality. It is ultimately up to the instructors to help students understand how to balance

engagement and objectivity, using specific writing techniques to enhance their overall communication skills.

6.5 Conclusion

The overall goal of this chapter was to investigate the high-frequency use of the second person in COMP 101 and to demonstrate that during the first year of composition writing, students often rely on *you* to express their viewpoints and procedures. In some of the essay genres like descriptive, narrative, and process analysis, students are permitted to use you, but in others like classification, comparison and contrast, cause and effect, and argumentative essays, instructions specifically prohibit the use. It is important to note that as students progress through the semester, that use tends to decrease, which indicates a slow shift toward objectivity. In contrast to COMP 101, writers in BAWE mainly use the second person to give instructions and directions to readers, which aligns with the structural knowledge as a function of the second person.

A close look at the distribution of the second person in the subcorpora shows that it is also important to understand the different functions the second person plays in composition texts, such as using it referentially in narrative texts or generically in multiple genres to engage the reader. Of the generic functions of the second person, structural knowledge is the most common in both COMP 101 and BAWE, indicating its effectiveness in personalizing instructions and involving readers in the context. This high use of structural knowledge suggests that writing manuals should recognize the role of the second person in creating guidelines and instructions. Also, the varied usage provides instructors with evidence that students who are inexperienced with university writing use the second person even in causal analysis and argumentative texts

that require objectivity. Instructors can help students understand the importance of a balanced use of the second person and keeping an objective voice in analytical writing.

The next chapter of the study focuses on the role of conjunctions as frequency items in COMP 101 and their role in the texts of first-year composition writers. Unlike the first and second person, which relate to engagement with readers, conjunctions pertain to sentence structure. However, alongside the pronouns discussed in Chapters 5 and 6, conjunctions contribute to the overall profile of first-year university writers, identifying the most frequently used textual features at this stage of academic writing. In this way, the frequency lists enable this corpus-based analysis and shift from engagement, expressed by first- and second-person pronouns, to an exploration of the syntax and sentence structure characteristic of first-year university students in COMP 101 and how this differs from that of upper-level students as in BAWE.

Chapter 7

The Role of Conjunctions in First-Year Academic Writing

7.0 Introduction

This chapter examines the role of conjunctions as cohesive devices in COMP 101 texts, focusing specifically on their function in linking finite and nonfinite clauses. Chapters 5 and 6 focused on the first and second-person pronouns as high-frequency features in COMP 101 and examined their role in the texts as well as how their use differs from upper-level writing, thus addressing the thesis research questions discussed in Chapter 1. This chapter continues to examine the frequency features in COMP 101 and targets the use of conjunctions (*and, that, as*), which rank among the first twenty-five most frequently used items in COMP 101 (Table 4.4 in Section 4.3.1). This high ranking indicates the significant role of conjunctions in the COMP 101 text composition. Conjunctions are critical for cohesion, as they establish logical and syntactic connections within the texts (Halliday and Hasan 1976; Carter and McCarthy 2006). While conjunctions can link various linguistic elements, such as noun phrases or adjectives, this chapter focuses on their role in clause linking due to the importance of clausal structures in shaping syntactic conmplexity and textual organization in academic writing (Biber and Gray 2010; Staples et al., 2016). Given the focus of this study on the most common linguistic features and their impact on entry-level composition writing, the role of conjunctions in clause linking becomes a subject of particular interest to this research. Additionally, the study examines and compares the COMP 101 usage of conjunctions in such clause linking within the broader context of academic writing, represented by the BAWE texts. By examining these conjunctive elements in COMP 101 and how they differ from BAWE, the chapter addresses the further need for research regarding sentence structure between novice and upper-level writers.

7.1 Analytical Framework

Conjunctions fall under cohesion, a large-scope semantic area that "refers to the relations of meaning that exist within the text, and that define it as a text" (Halliday and Hasan 1976, p.4). Cohesion is "the grammatical and lexical means by which written sentences and speakers' utterances are joined together to make texts" (Carter and McCarthy, 2006, p. 242). Carter and McCarthy (2006) distinguish cohesion from coherence, discussing the latter as a concept that transcends the lexical or grammatical properties and is realized when the textual "semantic and pragmatic meanings make sense in its real-world context to readers/listeners" (p.242). Thus, coherence refers to a text's overall logical organization and flow with its ability to make sense and be easily understood by the reader. Coherence is determined by the text's ability to present, develop, connect, and establish relationships among its various parts, guiding the reader through the discourse. While cohesion is closely connected with coherence, it stands apart by its direct association with the text's grammatical and lexical properties (Carter and McCarthy 2006). As a crucial linguistic element, cohesion facilitates text unity (Thompson 2014) and provides linguistic cues that allow the reader to make appropriate connections between the ideas in that text (Crossley *et al.* 2016, p.2). These connections can include various elements such as pronouns, articles, conjunctions, and other linking words that help to create a logical structure and flow within the text.

In its role to create unity and logical connections, cohesion is affected by genres and registers. Each genre has its own conventions, rules, and expectations for language use and organization, which can affect the cohesion of a text (Swales 1990). For example, across three grade levels of argumentative and narrative writing, Crowhurst (1987) indicates higher usage of cohesive links in narrative texts. On the other hand, register refers to a text's style or level of formality (Biber *et*

al. 1998). Based on Biber et al. (1998), different registers are used in different social contexts and for different purposes, and they can also affect the cohesion of a text. For example, a formal register might use complex sentence structures and specialized vocabulary, while a casual register might use simpler language and a more relaxed structure. Johns (1980) notes the varied use of cohesion elements in business discourse across letters, reports, and textbooks. In government documents, Trebits (2009) affirms the integral role of cohesive devices in creating organizational patterns and facilitating a meaningful discourse. Consequently, cohesive devices impact various types of texts, including first-year composition writing, which is the focus of this thesis. Texts produced by first-year composition students may use cohesive devices differently than their upper-level counterparts as they strive to adapt to formal academic writing upon entering university.

While it is important to understand the role of cohesion, this chapter does not aim to provide a comprehensive discussion of all the elements encompassing the concept. Due to their frequency in COMP101, the chapter focuses on conjunctions as one of the representative features of cohesion and explores their role in linking finite and nonfinite clauses in texts in COMP 101. The texts relate to a planned discourse within academic writing and may reveal various types of syntactic features used by novice writers. When discussing the syntactic nature of academic writing, it is important to highlight Biber and Gray's (2010) observation that it is characterized by complex syntactic structures that are compressed and synthesized with limited explicit linguistic features. Based on the compressed textual structure and lack of explicitness, academic writing challenges novice readers who cannot easily extract large amounts of information from relatively short and condensed texts. These novice readers must learn to understand the unspecified relations among the grammatical elements and digest the textual meaning. Such an

implication applies not only to novice readers but also to novice writers who struggle to use compressed style and syntactic complexity, which depends on the number, type, and depth of embedding in a text.

7.1.1 Categorization of conjunctions

One of the most comprehensive discussions of cohesion is Halliday and Hasan's work (1976), which categorizes between grammatical (reference, substitution, ellipses, and conjunction) and lexical cohesion (reiteration and collocation) and informs a wide range of research focused on cohesion and its role in promoting writing quality (Levinson 1983; McCulley 1985; Crossley 2012; Crossley et al. 2016; He 2020). Halliday and Hasan (1976) consider any semantic expression able to operate conjunctively to be classified in the category of conjunctions—compound adverbs, propositional, and linking adverbials are classified as conjunctions. On the other hand, traditionally, comprehensive English grammar textbooks (Quirk *et al.* 1985; Biber *et al.* 1999; Carter and McCarthy 2006; Carter *et al.* 2016) classify conjunctions and linking adverbials separately because of their focus on the syntactic structure of sentences, categorizing words based on their grammatical roles. Since this section seeks to analyze the role of conjunctions in COMP 101 as syntactic and grammatical elements, the study uses Carter and McCarthy's (2006) classification.

Based on Carter and McCarthy (2006), conjunctions express logical relations between phrases, clauses, and sentences, dividing them into coordinating and subordinating conjunctions. Coordinating conjunctions link identical grammatical elements: words, phrases, clauses, and sentences. Clauses attached with coordinating conjunctions carry coordinated or symmetrical relations. The main coordinating conjunctions are *and, but,* and *or* (Carter and McCarthy 2006,

p.315). Biber (1988) discusses the conjunction *and* as one of the main logical coordinators, which establishes the linking of general purpose, making it the most frequently used logical coordinator for that purpose. Halliday and Hasan (1976) classify *or* as an additive conjunction that adds an alternative opinion, interpretation, or possibility to the sentence, defining the conjunction as realizing an alternative relation. In contrast *but* has the ability to facilitate generic contrastive relations (Halliday and Hasan, 1976) and is the most frequently used contrastive conjunction.

On the other hand, subordinating conjunctions facilitate nonsymmetrical connections between clauses by making the clause being introduced subordinate to or dependent on the main clause. Some common subordinating conjunctions include *after, although, as, before, if, since, that, until, when, whereas,* and *while* (Carter et al., 2016; Beaman, 1984). In addition to the independent and dependent structures, clauses can be further classified into finite and nonfinite (Carter and McCarthy 2006).

A finite clause is a clause that contains a verb indicating a tense. This means that the verb in the clause is conjugated to show when the action or state of being takes place. Finite clauses are the most common type of clause in English and are essential for creating well-formed sentences. They can function as the subject or the predicate of a sentence and can be independent or dependent. They are typically introduced by a subject and a finite verb (e.g., *She **sings** beautifully)* and can be linked to other clauses using conjunctions (e.g., *I know **t**hat he **is** **coming***). On the other hand, a nonfinite clause is a clause that contains a lexical verb but does not indicate tense, which means that the verb in the clause is not conjugated to show when the action or state of being takes place. Nonfinite clauses are often used in academic writing and are known for their complex nature. Some typical examples of nonfinite clauses include infinitives

(e.g., *She wants **to sing***), past participles (e.g., ***Tired** from the trip, she went to be earlier*) and *ing*-participles (e.g., *Although **feeling** tired, she kept typing the text*) that are usually introduced by subordinating conjunctions (Carter and McCarthy 2006).

Nonfinite clauses are commonly used in academic writing to express complex meanings within restricted space. Research studies show that nonfinite clauses are typical for academic writing and demonstrate its complex nature. Staples et al. (2016) state that as the academic level increases, the use of finite dependent clauses decreases (2016). Biber and Gray (2010) also emphasize the role of embedded nonfinite rather than finite subordinate structures as a significant type that differentiates academic writing from other types. According to Biber et al. (1999), finite dependent clauses are more commonly found in spoken than written language, and they are particularly rare in academic writing compared to other types of writing. Hyland and Tse (2004) also investigate the role of conjunctions, specifically as a way to create logical relationships and "signal the writer's understanding of the logical relationships between ideas"(p.162). Their research examines conjunctions as part of a larger concept—metadiscourse (see Chapter 2.5.2).

It is interesting to note that Staples et al. (2016) observe that most research has focused on the expert use of academic writing rather than novice writers' language structures, which motivates the efforts of the present study to uncover the syntactic structures based on the finite and nonfinite clauses. To achieve this, the study adopts Carter and McCarthy's (2006) classification of conjunctions in clausal relationships. In addition, the research incorporates the findings of Biber and Gray (2010) and Staples et al. (2016) related to the finite and non-finite structures when analyzing these patterns in the texts produced by the entry-level composition writers (COMP 101) and advanced writers (BAWE).

7.2 Frequency list and Corpus Query Language (CQL)

The frequency list in COMP 101 ranks *and, that* and *as* in the top twenty-five positions. Table 7.1 shows the normalized frequency of the first twenty-five ranks in COMP 101 and also the comparative corpus BAWE to illustrate the difference in ranking between the two groups of writers. Both frequency lists are normalized per 1,000,000 words (N Freq).

Table 7.1: Conjunctions in the top 25 most-frequent items in COMP 101 and the BAWE (normalized per million)

N	COMP 101	N Freq	BAWE	N Freq
1	the	56392	the	70646
2	to	32662	of	38903
3	and	29806	and	29950
4	of	23894	to	27497
5	a	21706	in	22004
6	in	17351	a	19575
7	is	17081	is	15974
8	I	13028	that	11386
9	that	12890	as	9769
10	it	11629	for	8548
11	for	9807	be	8341
12	are	9404	this	7806
13	they	7401	it	7355
14	with	7385	are	6134
15	my	7380	with	6072
16	you	7211	on	5833
17	was	7189	by	5821
18	on	6834	was	5289
19	be	6797	not	4805
20	not	6792	from	4514
21	have	6474	an	4195
22	as	6103	which	4186
23	this	6029	’s	4147
24	we	5801	have	3677
25	their	5192	can	3642

The table highlights in gray ( ) the conjunctions that appear in both corpora. A quick observation shows that *and* is at position 3 in both COMP 101 and BAWE, *that* is at position 9 in COMP 101 and 8 in BAWE, and *as* at position 22 in COMP 101 and at position 9 in BAWE. The only difference between the conjunctive items in the two corpora is *which*, at position 22, appearing only in BAWE. It is important to note that *that* has high frequency because of its versatile usage and the study focuses only on its conjunctive use when it introduces subordinate clauses. The prominent ranking of the conjunctive words in both corpora motivates an inquiry into the grammatical patterns facilitated by these connectives in the first-year composition texts. Previous research (Staples *et al.* 2016) investigates the syntactic complexity across the levels of university writing and suggests the need for further investigation of the sentence structure reflected by student writers at different educational stages. The conjunctive nature of *and*, *that,* and *as* provides an opportunity to examine the sentence structures writers use during their first year of composition writing as in COMP 101, as well as those in later university stages, such as BAWE.

In order to provide an effective investigation of the patterns based on conjunction, the study examines the conjunctive words beyond the top twenty-five frequency items and considers the subsequent ones up to the cut-off raw frequency score of 100, which provides results of the most common conjunctive patterns in COMP 101 and BAWE. Once the conjunctions in COMP 101 are determined, they are matched to the corresponding ones in BAWE. Table 7.2 presents the filtered results showing the conjunctions in both corpora and their respective raw and normalized frequencies per 1,000,000 words. The study displays raw frequencies because it examines conjunctions with a minimum raw frequency score of 100.

Table 7.2: Conjunction frequencies in COMP 101 and BAWE (normalized per million)

Position	COMP 101	R Freq	N Freq	BAWE	R Freq	N Freq
1	and	5625	29806	and	208693	29950
2	that	2433	12890	that	79337	11386
3	as	1152	6103	as	68072	9769
4	or	869	4604	which	29167	4186
5	but	809	4286	or	23173	3326
6	when	691	3661	but	15259	2190
7	because	637	3375	however	12267	1760
8	so	527	2792	if	11446	1643
9	if	527	2792	when	11303	1622
10	which	429	2273	so	10362	1487
11	who	408	2162	who	7357	1056
12	after	264	1473	where	7061	1013
13	however	239	1399	because	6900	990
14	while	238	1266	after	4842	695
15	where	193	1261	although	4586	658
16	before	185	1023	since	4276	614
17	since	120	980	while	3457	496
18	until	106	636	before	3396	487
19	though	106	562	whether	3130	449

A close examination of the filtered results reveals that most of the words are common to both corpora, with the exceptions of *until* (see in grey) in COMP 101 and, *although* and *whether* (see in grey) in BAWE. The list contains coordinating conjunctions (*and, or, but)* and subordinating (*that, when, because*, or *then),* which may indicate the presence of the finite and nonfinite clausal structures between the COMP 101 texts and the ones in BAWE.

The method that allows precise extraction of grammatical patterns in large collections of texts is the corpus query language (CQL), a tool available in Sketch Engine. CQL is often used in linguistic research and text analysis to retrieve data from large text collections through syntactic algorithms similar to other programming languages. It searches using specific parameters, such as word forms, parts of speech, and grammatical structures. This study uses CQL to extract information from the COMP 101 and BAWE corpora, specifically focusing on finite clauses that

follow coordinating and subordinating conjunctions and nonfinite clauses that follow subordinate conjunctions.

To ensure replicability and clarity of results, the study provides a detailed description of each CQL algorithm utilized to search for patterns related to coordinating and subordinating finite clauses as listed below.

[lemma="and"][tag="PP.?"][tag="V.*"] and + pronoun + verb
[lemma="and"][tag="N.*"][tag="V.*"] and + noun + verb
[lemma="and"][tag="DT"][tag="N.*"][tag="V.*"] and + article/determiner + noun + verb

The same CQL searches were applied to the BAWE texts with the exception of the tag notation designating an article or determiner. The reason for that variation in the notation is that Sketch Engine uses two different tagsets: CLAWS and Tree Tagger. The BAWE corpus is tagged using the CLAWS tagset (e.g., using the tag “AT” denoting article), whereas COMP 101 is tagged based on Tree Tagger (e.g., using the tag “DT” denoting determiner). Even though Sketch Engine uses two different sets of notations for the grammatical category in the two corpora, the tagset does not change the category (Sketch Engine 2022).

CQL searches that examine the texts in the two corpora for nonfinite clauses focus on the infinitives, past participles, and *-ing* participles, some of the most typical nonfinite structures that follow subordinate conjunctions (Carter *et al.* 2016; Mala 2017). The queries used include the following configurations:

[lemma="that"][][lemma="to"][tag="V.*"] that + space + to + verb
[lemma="that"][tag="VVN"] that + past participle
[lemma="that"][tag="VVG"] that + ing-participle

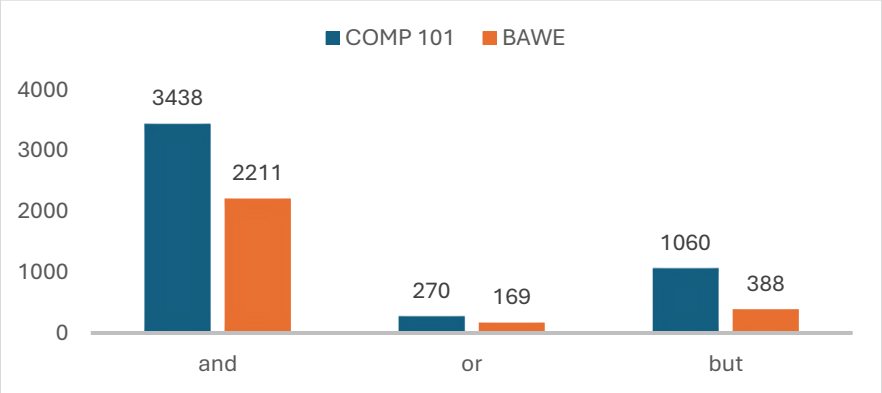
The described CQL algorithms were applied to each of the 19 conjunctions in Table 7.2. The search results were recorded in Excel, normalized per 1,000,000 words, and organized in tables to represent the use of the finite and nonfinite clauses in COMP 101 and BAWE. The same tools and processes were applied to the individual subcorpora to examine the clausal patterns across the genre texts. The next section discusses the results of the finite and nonfinite clauses in COMP 101.

7.3 Finite coordinate clauses linked by conjunctions in COMP 101

Some researchers (O'Donnell 1974; Chafe and Tannen 1987) consider clausal coordinate structures syntactically simple and more typical for spoken rather than written language. Lintunen and Makila (2014) note that coordination is regarded as the syntactic starting point for writers, which may classify it as a feature characterizing beginners' texts rather than advanced ones. Even though coordinate conjunctions appear to be more frequently used in beginners' writing (Staples *et al.* 2016), their integral and versatile role in creating meaning makes them useful in a wide range of texts. For example, Halliday (2007) credits them as potent enough to form complex ideas when used with high-content words. It is not surprising then that the central coordinators *and*, *but*, and *or* occur with high frequency in both COMP 101 and BAWE. As this chapter, considers the use of conjunctions in finite clauses, it is important to note that the task instructions did not require specific use of one clausal structure over another, but students were required to use grammatical correctness and, in some essay genres (comparison and contrast, cause and effect, and argumentative), sentence variety, beyond that, there were no specific instructions for students to show preference to particular clausal structures (see Section 3.3). As this study analyzes completed, graded assignments collected after the COMP 101 course,

differences in clausal structures, particularly the prevalence of coordination in COMP 101 texts, legitimately reflect writers' levels of proficiency rather than solely task-driven constraints. The CQL search results reveal the number of times they facilitate clausal relationships. Figure 7.1 shows the CQL results and the number of times coordinators create finite clauses.

Figure 7.1: Finite clauses with coordinating conjunctions in COMP 101 and BAWE



A quick look at Figure 7.1 shows that the finite coordinate clauses linked by the conjunctions *and, but*, and *or* are more typical for COMP 101 writers than the BAWE ones, which leads to the observation that during first-year composition writing, students' texts are characterized by a high occurrence of finite coordinate clausal structures. Coordinate clausal structures are more typical for novice writers (Beaman 1984), which correlates to these findings. Figure 7.1 indicates that the most prominent coordinator in these syntactic relationships is the additive *and,* observed in both corpora, favoring COMP 101 with 38 percent more instances than BAWE. The other two coordinators—*or* and *but*—appear less frequently in clausal relationships, which indicates that non-clausal structures support their top ranks in the two corpora word frequencies (e.g., Table 7.1).

The examples in (7.1) show the typical uses of the coordinators in the context of the COMP 101 sentences. The topics discuss subjects of general interest that are not focused on discipline-

specific content but on personal or experiential knowledge, such as the islands in a particular geographic location, shower hygiene, school procrastination, personal drama, or a certain process. In all the instances, the writers elaborate on the topics using general vocabulary rather than discipline-specific terms.

7.1 COMP 101

> **Classification Subcorpus:** *There are more islands in the East of the Caribbean, **and they consist of** St. Lucia, Barbados, Trinidad, Grenada, Tobago and others.*
> **Argumentative Subcorpus:** *The shower is a private place for all, **and it is** to be kept clean and to be smelling good.*
> **Cause-and-effect Subcorpus:** *My procrastination is usually because I get distracted, **or I do** not want to do that assignment that I need to do.*
> **Narrative Subcorpus:** *I tried to defend myself, **but tears filled** my eyes.*
> **Process-analysis Subcorpus:** *Some may think it's a simple process, **but it's** also a detailed one.*

This choice of vocabulary is closely linked to the instructional design of COMP 101, which encourages students to choose their own topics for each assignment based on different genres. Since many students are in their first year and may be undecided about their majors or are just beginning to explore their fields, they are less likely to engage with content specific to their disciplines. As a result, their writing often reflects everyday language that aligns with their current knowledge rather than using terminology specific to their fields of study.

In BAWE, on the other hand, writers use coordinating conjunctions in clausal relationships to name particular items and treat specific topics as the sentences in (7.2) show: data relationships, M.Xanthus or bacteria, and child temperament.

7.2 BAWE

Again, there is a considerable amount of scatter on the graph and it is arguably harder to detect the presence of a significant negative relationship between the two variables from the graph alone.
The way in which M. Xanthus moves may not be efficient on the agar medium, ***or it could*** *be that the bacteria were well fed and therefore had no impetus to be motile in the search for food.*
Also attachment claims to be a causal factor ***but it may*** *be the child's temperament that causes attachment and popularity therefore attachment is not the cause*

The writers in the BAWE corpus use non-complex syntactic structures, such as finite coordinate clauses, to qualify complex concepts and phenomena. Although the finite coordinate clauses are more typical of first-year composition writing, as in COMP 101, with high-context vocabulary, they can create complex ideas, as in the BAWE text. As Halliday (2007) points out using simple grammatical structures like these, along with high-level vocabulary, can effectively convey complex ideas. The writers in BAWE effectively demonstrate this by expressing challenging concepts through simple structures like coordination. The next section continues to explore the finite structures, shifting to the subordinate clauses.

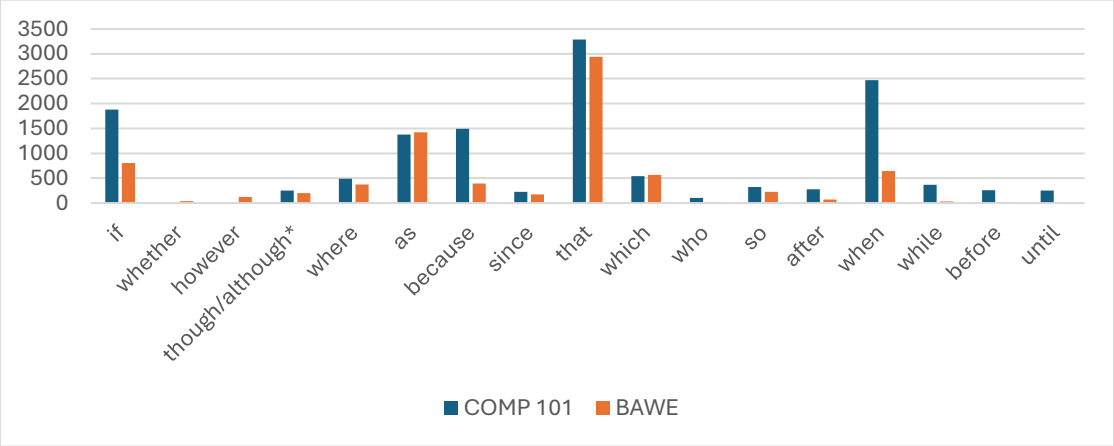
7.4 Finite subordinate clauses linked by conjunctions in COMP 101

Biber (1988) discusses subordinate structures as some of the most frequently used textual features for register comparison. In writing, subordination contributes to syntactic complexity by transforming simple or compound sentences into more complex units of grammar, which contributes to sentence variation and text dynamics (Beaman 1984; Lintunen and Makila 2014). Biber et al. (2010), in a subsequent study, observe that finite dependent clauses are more prevalent in conversation than in academic writing. Another aspect of the finite dependent clauses is highlighted by Halliday (2007) who notes that finite subordinate structures may characterize both advanced and novice writing. Advanced writing often uses high-content lexical

density with subordinate structures, while novice writing employs low-content vocabulary with these clausal structures.

This section examines the textual patterns created by the finite subordinate structures in COMP 101 and BAWE. As in the case of the task instructions related to the use of the finite coordinate conjunctions (see Section 7.3), the task instructions do not refer to specific use of some subordinate conjunctions over others but simply require students to keep to the grammatical correctness and in some essay genres (e.g., comparison and contrast, cause and effect, and argumentative) use sentence variety. As a result, the use of subordinate clauses in the COMP 101 corpus reflects individual proficiency and also preferences rather than explicit requirements. Figure 7.2 summarizes the CQL results targeting the subordinating conjunctions in their clausal roles. These conjunctions are included in Figure 7.2 as items of interest. Figure 7.2 contains the most frequently occurring conjunctions, with raw frequencies up to 100 in COMP 101, and shows their equivalents in BAWE.

Figure 7.2: Finite clauses with subordinating conjunctions in COMP 101 and BAWE (normalized per million)



As Figure 7.2 illustrates, seven of the seventeen subordinate conjunctions appear to be equally used in clausal linking in both corpora—*though/although, where, as, since, that, which,* and *so*, which may speak of their practical functions and applicability in the various levels of academic writing. The conjunctions that hardly appear in clausal relationships in the two corpora are *whether*, *however*, and *after*. Four conjunctions—*while, who, before*, and *until*—occur in clausal relationships only in COMP 101. The first two, *while* and *who*, most frequently occur in the narrative and compare-and-contrast texts. In the narrative subcorpus, w*hile* is used 516 times (normalized occurrences), and *who* is used 163 times (normalized occurrences). In the compare-and-contrast texts, while is used 520 times (normalized occurrences), and *who* is used 186 times (normalized occurrences). On the other hand, *before* (486 normalized occurrences) and *until* (648 normalized occurrences) are most frequently used in descriptive texts, indicating sequential relations.

This section will focus on the differences between *if*, *because*, and *when* as subordinate conjunctions, which occur in both corpora but demonstrate considerable differences in their use. These clausal structures are more prevalent in the first-year composition texts (COMP 101) than in upper-level writing (BAWE). Table 7.3 shows their normalized frequencies per million in each corpus to illustrate the differences in their distributions.

Table 7.3: Normalized frequencies of *because*, *if*, and *when* in COMP 101 and BAWE (normalized per million)

Conjunctions	**COMP 101**	**BAWE**
because	1494	396
if	1881	803
when	2474	645

In COMP 101, the normalized frequency of *because* is 1494, while in BAWE it is 396. Next, the normalized frequency of *if* in COMP 101 is 1881, and in BAWE is 803. Finally, *when* in COMP 101 has a normalized frequency of 2474, and in BAWE it is 645. Further examination of the subordinate clauses *if, because*, and *when* reveals distinct patterns in the two corpora. In COMP 101, the subordinate conjunctions facilitate clausal relationships between personal pronouns and verbs more frequently than in BAWE, where they are typically used with nouns. This observation aligns with the findings in Chapters 5 and 6, which examine the frequency of the first-person and second-person pronouns in developing the COMP 101 texts and also show the natural tendency of the writers to engage the readers in a conversational tone. It is important to note here that the use of the first-person is prescribed by the task instructions, while the use of the second person is prohibited in some of the essay genres (e.g., classification, comparison and contrast, cause and effect, and argumentation).

In COMP 101 *if, because*, and *when* are typically followed by personal pronouns as in (7.3) and facilitate general topics, such as watching a movie, raising grandchildren, texting while driving, popular music in France, or personal narrative. This use echoes Halliday's (2007) notion that the finite subordinate clauses coupled with low-content vocabulary demonstrate conversational use of language.

7.3 COMP 101

> **Classification Subcorpus:** *However, **if we watch** that thriller, we may want to turn on the lights until we get to our room.*
> **Classification Subcorpus:** *But they aren't aware that **when they raise** their grandchildren instead of their children, their children are repeating their story.*
> **Argumentative Subcorpus:** ***If you educate** drivers about the dangers of texting while driving, then they will act accordingly.*

Classification Subcorpus: *This music only begins to become popular in France* ***because it has*** *a bad reputation.*
Narrative Subcorpus: ***When we arrived****, there were already families in the cabin.*

Personal pronouns may reflect, in some cases, the writers' inclination towards conversational use of the clauses, even the use of the second person when task instructions specifically state that it should not be used, which further highlights everyday language use. Only 20 percent of the clausal structures in COMP101 display nouns that follow the subordinate conjunctions, as in (7.4), that identify general topics, such as school, religious content, or personal weight. The treatment of these topics does not show specialized terms or discipline-specific discussion but rather general knowledge and reflection, which is expected based on the choice that the first-year students have in selecting topics for their essays.

7.4 COMP 101

Cause-and-effect Subcorpus: ***If schools start*** *to lessen the usage of the old grading system and try to evaluate students on their other skills, it will help students not to be focused on grades all the time.*
Argumentative Subcorpus: *There's no separation of the sacred and the secular* ***because everything is*** *sacred to God.*
Cause-and-effect Subcorpus: *The opposite is* ***when someone feels negative*** *about their body weight and may compare themselves to others.*

In contrast to COMP 101, the BAWE corpus reveals a minimal usage of personal pronouns within subordinate conjunctions. The limited instances of personal pronouns in the clausal structures use the second person to communicate process directions and address the readers as demonstrated in (7.5). Every example relates to specific items, such as a particle, swallowing assessment, or request using HTTP. The terms do not seem to refer to casual topics based on general knowledge or experience that may come up in everyday conversations, which probably reflects the discipline-specific tasks that the BAWE students have to address. Despite the use of

personal pronouns, the finite subordinate clauses include high-content vocabulary to discuss discipline-specific topics.

7.5 BAWE

> ***If you measure*** *the particle before it has chosen its path then the wave properties disappear all together and no interference occurs.*
> ***Because you are*** *still having difficulty swallowing after your strokes we would like to get a swallowing assessment to make sure that you are able to get the right intake of food and drink to try and prevent this from happening again.*
> *Basically* ***when you type*** *the URL address into your web browser you sent a request using HTTP. In this case a browser is the HTTP client and a web page server is the HTTP server (see Figure 1).*

The majority of the occurrences show the subordinate conjunctions being followed by nouns that name particular concepts, items, or individuals, as in (7.6): *Parliament and water* These specific items are treated with specialized words like *hypothesis and large excess* that do not relate to everyday topics. Biber (1995) discusses the high occurrence of nouns versus pronouns as an indicator of formal register. These observations are confirmed in a later study by Hyland and Jiang (2017), noting that pronouns "deixically anchor a statement to a given context," (p.42), while nouns focus on specific content and support the formality of expression.

7.6 BAWE

> ***If Parliament had regarded*** *the ideas and behaviour of the 'Ranters' as a real threat to societal order, then Davis's hypothesis that the Ranters did not exist is weakened.*
> *This may be* ***because water is present*** *in such a large excess as pure water is highly concentrated that it is not having an effect on the rate.*

Overall, the finite subordinate clauses in COMP 101 are used more frequently and are often followed by personal pronouns, indicating a conversational writing style. The topics discussed in

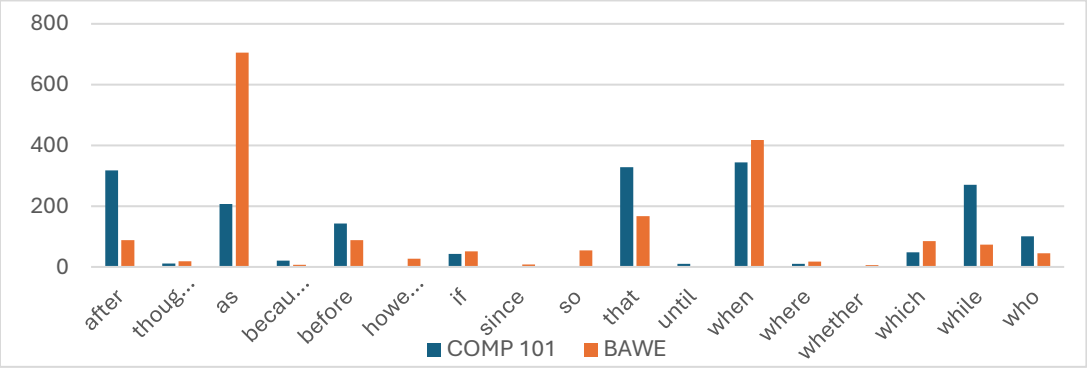
COMP 101 using these subordinate clauses tend to be general and relate to personal experiences or knowledge, which also relates to the free choice that the COMP 101 students have in selecting their own topics. In contrast, in BAWE, these subordinate clauses are used less frequently and are more often followed by nouns that refer to specific concepts, items, or individuals. Thus, the language features correspond to specialized topics and technical language, seeking to explain or discuss discipline-specific areas. The subordinate clauses in both corpora bring variety to the syntactic structure (Lintunen and Makila 2014), but the main difference between the two group of writers is the choice of words that determine specific areas of study in the BAWE corpus, whereas in COMP 101 writers show preference to general topics that relate to their current knowledge, as the tasks in COMP 101 allow them to select individual topics Perhaps, first-year writing students can gain valuable insights by studying examples that demonstrate the use of subordinate clauses to add depth and complexity to their writing, particularly in focusing on professional contexts and research studies.

7.5 Nonfinite structures linked by conjunctions in COMP 101

Biber et al. (2010) and Staples et al. (2016) note that nonfinite clausal features are more frequently used in academic writing than finite ones. In academic writing, which is typically characterized by a high level of formality and precision, writers often need to convey a large amount of information in a relatively short space. To do this, they may use nonfinite clausal features, which are verb forms that do not have a subject and do not indicate tense. Finite coordinate and dependent clauses grammatically develop from clause combining while nonfinite clauses from clause expansion (Vercellotti and Packer 2016), the latter considered a more challenging concept for novice writers (Lintunen and Makila 2014). The instructions for the COMP 101 assignment do not specifically require the use of nonfinite structures over finite ones.

Instead, they focus on grammatical correctness, sentence variety, and the use of vivid language to support a clear thesis across all seven subgenres (see Section 3.3). This absence of explicit guidance on clausal structures suggests that students' use of nonfinite clauses reflects their individual proficiency and preference choices rather than task-driven requirements. Thus, the observed patterns in the use of nonfinite clauses may reflect developmental stages. This is particularly relevant for COMP 101 writers, who are first-year students and often experience greater challenges in navigating the complexities of clause expansion due to limited experience with academic conventions compared to BAWE writers. Since the CQL algorithms are based on subordinators, it is reasonable to demonstrate their distribution across COMP 101 and BAWE. Figure 7.3 shows the normalized occurrences of the subordinators in their role to link nonfinite clauses.

Figure 7.3: Nonfinite clauses in COMP 101 and BAWE (normalized per million items)

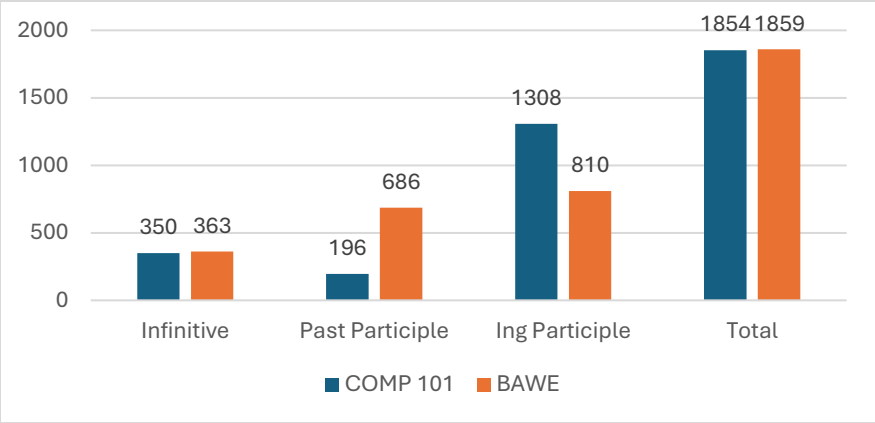


Overall, the results indicate that nonfinite clauses linked by subordinating conjunctions are utilized in both COMP 101 and BAWE. In fact, certain subordinators appear to be more frequently used with nonfinite clauses in COMP 101 than in BAWE, and vice versa. For instance, the subordinator *after* is three times more common in COMP 101 compared to BAWE,

while *as* is three times more prevalent in BAWE than in COMP 101. Four subordinators—*however*, *since*, *so*, and *whether*— create nonfinite clauses only in BAWE.

The subordinators are further grouped into the most common nonfinite structures, such as infinitives, past participles, and *-ing* participles (Carter *et al.* 2016; Mala 2017) corresponding to COMP 101 and BAWE. Figure 7.4 summarizes these CQL search results on infinitive, present, and past participle when they are linked by subordinating conjunctions. The figure shows the normalized occurrences per million words and provides a standardized comparison between the two corpora.

Figure 7.4: Types of nonfinite clauses in COMP 101 and BAWE (the occurrences are normalized per million items)



Based on Figure 7.4 totals, it would be misleading to assume that the BAWE writers have only a slight advantage in the nonfinite clausal use because out of the three types, the past participle peaks in BAWE, while it is barely represented in COMP 101. On the other hand, the COMP 101 writers show a significant preference for using the *ing*-participle rather than the infinitive and the past participle. The past participle is the least represented category in COMP 101, which suggests that this nonfinite structure seems to be the most challenging type for first-year writers. The next

three sections discuss the -ing participles first as the prevalent nonfinite structure in COMP 101, then the infinitives as the second most popular choice in COMP 101, and finally, the past participles as the least used.

7.5.1- ing participles

The most frequently used type of non-finite clauses linked by subordinating conjunctions in COMP 101 and in BAWE are the *-ing* participles (see Figure 7.4). This section focuses on the subordinate clauses used with *-ing* participles to create nonfinite clauses. The distribution of the subordinates with *-ing* participles is summarized in Figure 7.5 and includes COMP 101 and BAWE corpora.

Figure 7.5: Ing-participles with subordinates in COMP 101 and BAWE (the occurrences are normalized per million items)

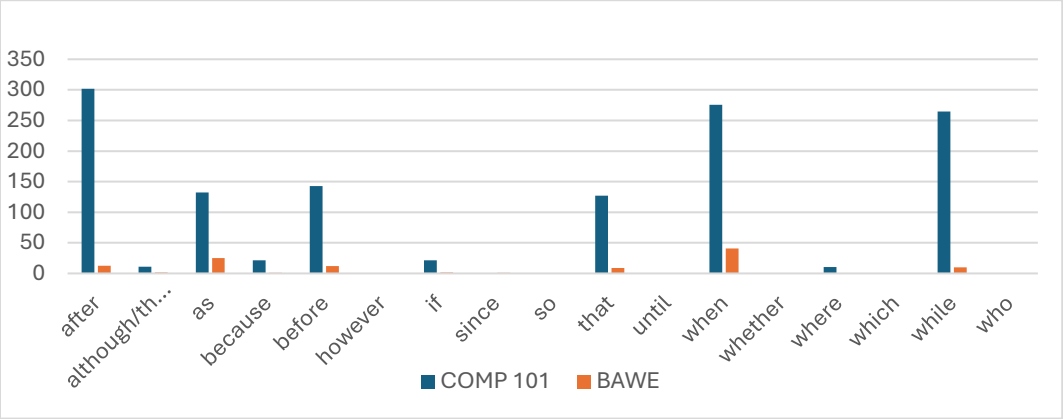


Figure 7.5 shows that –*ing* participles are used with a larger selection of subordinators in COMP 101 than in BAWE. In COMP 101, six subordinators co-occur with *-ing* participles over 100 times per million words: *after, as, before, that, when*, and *while*. In contrast, in BAWE, it is

difficult to find equivalent patterns with such high co-occurrence frequencies. *As* and *when* are the only two subordinators with more than ten co-occurrences with *-ing* participles in BAWE. In regard to the use of *-ing* participles, Mala (2017) indicates that this type of nonfinite clause is the most frequent in general corpora, such as Brown and Frown, which represents general language use and the high frequency of *-ing* participles, suggesting that this grammatical structure is common in everyday communication. First-year writers, who are still developing their writing skills, are likely to use these common patterns in everyday language, as the previous section on finite clauses demonstrated. In this case, *-ing* participles might be seen as a reflection of this familiarity with common language structures.

In COMP 101, the writers use the *ing*-participles to refer to particular situations or show the duration of an activity in progress as in (7.7).

7.7 COMP 101

> **Classification Subcorpus:** ***When administering*** *medication intravenously, it is important to note if the medication is to be given all at once in a push or if the medication is to be given over a period of time through a pump.*
> **Compare-and-contrast Subcorpus:** *Most times,* ***when living*** *on campus, in cases where the college does not allow a commute option, it is an option to walk right to classes, which saves time and gas money.*
> **Cause-and-effect Subcorpus**: *Doing these things repeatedly* ***while running*** *can greatly increase leg-eye coordination.*

In BAWE, the writers show a similar use of the *ing*-participles, which clarifies meaning through examples and contextualization. In (7.8), the samples exhibit similar patterns, but this time, they are focused on specialized political or philosophical theories utilizing high-content vocabulary. This high-level vocabulary appears to be the distinctive characteristic of the BAWE writers.

7.8 BAWE

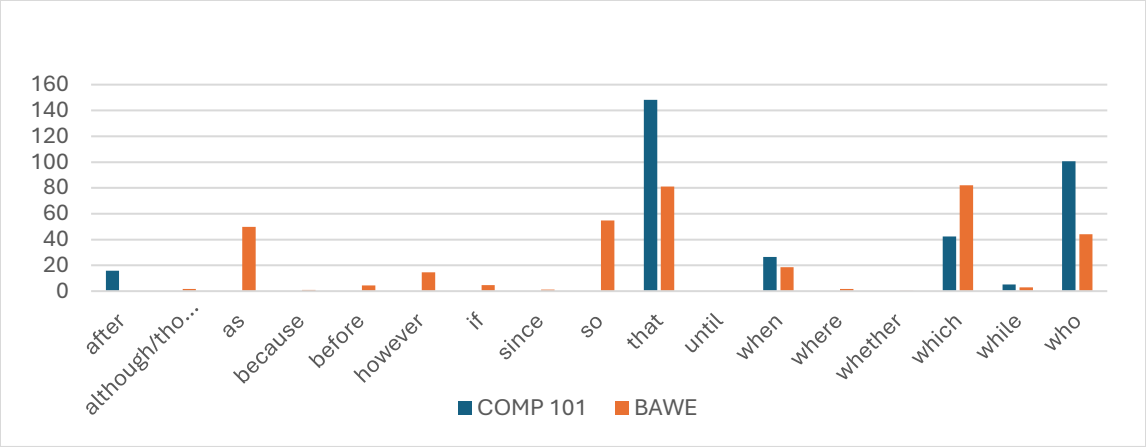
> *Eastern block socialism, which has been suggested **as imposing** some form of a state-directed will upon its citizens, has been severely criticized.*
> *Having a theoretical starting point **when conducting** research is vital because researchers who perform data collection and then generate their theories (through inductive reasoning or the grounded theory approach) inevitably reproduce the stereotypes and assumptions of everyday life (May, 2001; 31).*

The -*ing* participles are typical for both COMP 101 and BAWE writers and are useful ways for both groups to refer to events, situations, or processes. In COMP 101, writers use -*ing* participles to talk about everyday topics, indicating that these nonfinite structures are not limited to higher-level writing. -*Ing* participles are also commonly used in general corpora, as discussed by Mala (2017), which may suggest that these frequently used structures in everyday language are also reflected in first-year composition writers. The vocabulary used in COMP 101 relates to general, everyday topics that reflect the individual interests of the students, as they are allowed to choose their own subjects. When BAWE writers use -*ing* participles , the difference lies in the discipline-specific topics and technical terms that reflect the various majors of upper-level students.

7.5.2 Infinitives

The infinitive structures based on the CQL algorithms are summarized in Figure 7.6 and show the normalized occurrences in COMP 101 and BAWE. Unlike other sections of this chapter, where clauses are directly linked to subordinators, the infinitive clauses discussed here are dependent on finite verbs, which are, in turn, dependent on subordinators.

Figure 7.6: Infinitives with subordinators in COMP 101 and BAWE (the occurrences are normalized per million items)



A close look at the figure shows that they occur with only five subordinators in COMP 101 (*after, that, when, which*, and *who*) and nine subordinators in BAWE (*as, before, however, if, so, that, when, which*, and *who*. The normalized frequencies for the infinitives show a similar rate of occurrence between COMP 101 and BAWE. In COMP 101: the infinitives are used 339 times and in BAWE 363 times. The main difference is that the writers in COMP 101 use a narrower selection of subordinators, which may suggest a lower level of competence in this pattern.

Examining the differences in the use of the *to*-infinitives, the study focuses on the bars that show apparent differences: *that, which*, and *who*. The examples in (7.9) show some of the typical use of the *to*-infinitive clauses functioning mainly as post modification clauses. In the cases *that* followed by *to*-infinitives, the context varies from personal to general as in the first two examples—the first focused on personal emotions, and the second describing cultural celebrations and events.

7.9 COMP 101

> **Cause-and-effect Subcorpus:** *I treated my emotions like sin, but my emotions were a gift* ***that needed to be*** *restored and transformed into a godly perspective.*

Process-analysis Subcorpus: *Culture is the food **that people choose to eat** and how people celebrate holidays.*

The other common subordinators in COMP 101 are *which* and *who* used in contexts describing people or objects, as illustrated in (7.10).

7.10 COMP 101

Classification Subcorpus: *The Enneagram describes nine different personality types and maps each of these types on a nine-pointed diagram **which helps to illustrate** how the types relate to one another.*
Cause-and-effect Subcorpus: *Usually happens to people **who tend to be** more emotionally attached to people, memories, experience, and even things.*

In the occurrences of *which* and *who*, the COMP 101 writers distance themselves and focus on outside items, such as the different personality types and circumstances that happen to particular people. Trying to describe items outside the personal scope demonstrates a sense or attempt towards objectivity, a purpose in academic writing. The infinitive clauses are not directly linked to the subordinators but are embedded within structures where finite verbs facilitate the connection, like the infinitive clause *to be restored and transformed* depends on the finite verb *needed*, which is dependent on the subordinator *that* or the infinitive clause *to illustrate* depends on the finite verb *helps* linked to the subordinator *which*.

In BAWE, the *to*-infinitive clauses, similar to COMP 101, function as postmodifiers but are surrounded by technical vocabulary and seek to create specialized meaning such as specific methodology, spacers, or a group of individuals. This specialized vocabulary is expected, as BAWE students have had more exposure to research and their chosen fields of study. The examples in (7.11) demonstrate the typical occurrences in discipline-specific areas, such as *TMGA*, *circular cross-section*, and *new markets*.

7.11 BAWE

> *In TMGA Archer develops a methodology **that seeks to examine** the interplay between structure and agency over time in order to account for 'why things are so and not otherwise' in society.*
> *They were held apart by a set of seven spacers, **which were to be** solid bars of circular cross-section.*
> *Sony did not convince those **who wanted to share** their loud music but did create new markets among youngsters, female and adults population.*

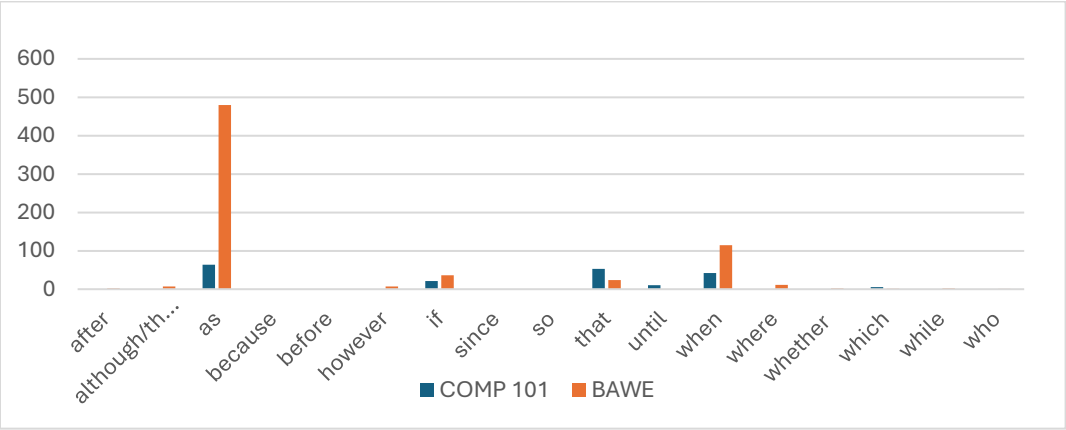
The analysis indicates that both groups commonly use nonfinite structures involving infinitives. The differences in usage are primarily related to the complexity of vocabulary used by the two groups. This supports Staples et al.'s (Staples *et al.* 2016) findings that there are no significant differences between entry-level and upper-level writers in the number of nonfinite structures with infinitives, but the key distinction lies in the use of discipline-specific vocabulary. Simply including infinitives in nonfinite structures does not necessarily indicate complexity in academic writing; the context in which these structures are used should be considered to assess the writer's engagement with the topics and possibly their chosen fields of study.

7.5.3 Past participles

The past participle structures are the most infrequent nonfinite clauses (see Figure 7.4) only for first-year composition writers. They account for roughly 11 percent of the overall nonfinite structures in COMP 101. In her corpus study, Mala (2017) notes that past participles are the least commonly used type of nonfinite clause in general corpora like Brown and Frown. This suggests that past participles are not as prevalent in everyday language and first-year composition writers, like the COMP 101 writers, may not be familiar with these structures yet. In academic writing, Biber and Gray (2010) observe that past participles in nonfinite clauses are more typical for academic writing than conversations. The study examines COMP 101 and BAWE for the

occurrence of the past participles in nonfinite structures linked by subordinating conjunctions, using the CQL searches, and summarizes their normalized occurrences in Figure 7.7.

Figure 7.7: Past participles with subordinates in COMP 101 and BAWE (the occurrences are normalized per million items)



The texts in COMP 101 have a low frequency of past participle occurrences linked by subordinating conjunctions, with fewer than 100 instances per million words that co-occur with three subordinators: *as, that*, and *when*. In contrast, the texts in BAWE show a higher frequency of past participle usage, with a notable peak in usage with the subordinator *as* and approximately 100 occurrences per million words with the subordinator *when*. The frequency patterns created by *as* and *when* demonstrate the differences between the two groups of writers and provide points of comparison regarding the contextual use of these two patterns.

Both *as* and *when* used with past participles, provide a reference for a particular context or situation to illustrate that given information is based on previously established factors, which in academic writing marks an effective strategy since it seeks to compress information. The examples in (7.12) show the use of *as* and *when* being used with past participles in COMP 101 to qualify or contextualize information and focus on the topic rather than the speaker.

7.12 COMP 101

> **Compare-and-contrast Subcorpus:** *Additionally, **as mentioned** before, Sleeping Beauty shows Maleficent as a wicked person and the clear villain in the story.*
> **Classification Subcorpus:** ***When paired** with fantasy it becomes a truly mesmerizing with the different colors and the vibrance of the new world.*
> **Cause-and-effect Subcorpus:** *They help to calm the owner **when needed**.*

The examples in (7.13) show very similar use of the past participles co-occurring with the subordinators *as* and *when* in BAWE, but this time, the context is characterized by high-context vocabulary and specialized topics, such as *production profiteering*, particular *data*, and *carbon dioxide*.

7.13 BAWE

> *More sinisterly, the data could suggest production profiteering, **as captured** in the illustration below (albeit in 1935).*
> *The data supports this, **as exemplified** by recent figures from Cambridge University showing that around half of its students are still from private fee paying schools.*
> ***When compared** to carbon dioxide, this amount is minimal, but nitrous oxide is thought to be responsible for at least 6% of the greenhouse effect so far2, with levels of this gas rising about 0.25% per year1.*

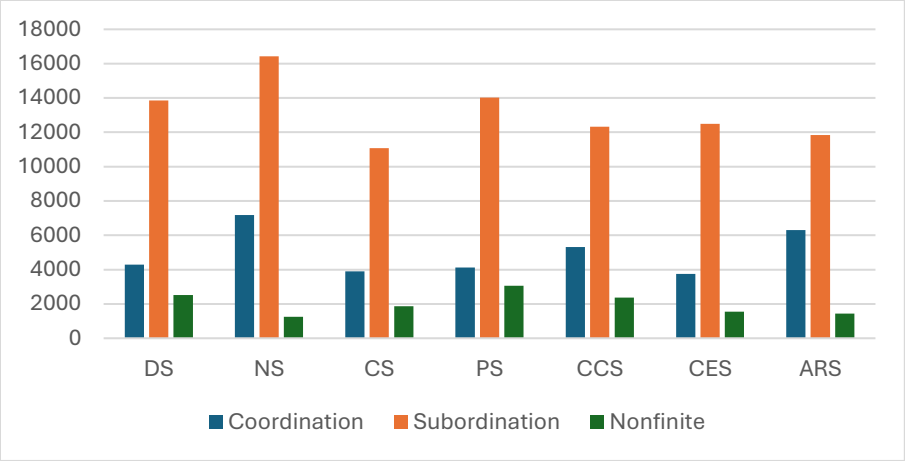
The infrequent use of the past participle in COMP 101 suggests that while the first-year writers are beginning to incorporate them similarly to the BAWE writers, they still lack confidence, especially in their use of discipline-specific content. This low usage is understandable since many first-year students are new to their chosen fields of study or still looking for a particular field. The most frequent patterns in this type of nonfinite clauses are *as* and *when* with past participles, which provide writers with effective ways to show context or a reference for a statement and compress information. The low occurrence of past participles in nonfinite structures in COMP 101 and its moderately high use in BAWE confirm Biber and Gray's (2010) and Staples et al. (2016) research that these structures characterize academic writing and are

nonfinite structures are features of higher levels of academic writing at university. The increased use of past participles and discipline-specific vocabulary reflects the choices made by upper-level writers. In contrast, the COMP 101 writers use fewer past participles along with simpler vocabulary, which may indicate less familiarity with these grammatical structures and a lack of necessity to engage in discipline-specific content, as the COMP 101 tasks allow freedom of choice for topic selection. This analysis of the nonfinite structures contributes to the wider research of academic writing by demonstrating how subordinators are used with the past participles in COMP 101 and BAWE, showing that first-year composition writers display usage of such structures but are not as familiar with them as with *-ing* participles and infinitives.

7.6 Finite and nonfinite clauses linked by conjunctions in the COMP 101 subcorpora

Staples et al. (2016) investigate the clausal structures and their distribution in university writing, as well as the patterns of development mediated by genre and discipline. In similar ways, this study seeks to examine the finite and nonfinite structures used in COMP 101 and note their movements across the subcorpora represented by different genres: descriptive subcorpora (DS), narrative subcorpora (NS), classification subcorpora (CS), process-analysis subcorpora (PAS), compare-and-contrast subcorpora (CCS), cause-and-effect subcorpora (CES), and argumentative subcorpora (ARS). This section provides a brief overview of the distribution and patterns that the finite and nonfinite clauses create in the subcorpora. Figure 7.8 shows a summary of the finite coordinate and subordinate clauses as well as the nonfinite clauses when they are linked by conjunctions and their distribution based on the normalized occurrences across the subcorpora.

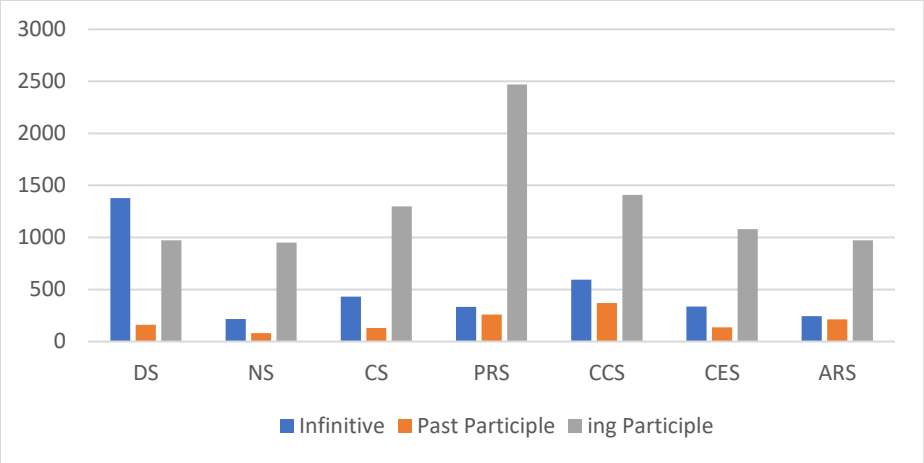
Figure 7.8: Summary: finite and nonfinite clauses in the subcorpora (the occurrences are normalized per million items)



The bars, representing each subcorpus, are ordered in chronological order of text completion in the semester, starting with the DS and ending with the ARS. The dataset reveals a clear dominance of subordinated syntactic structures across all genres, indicating a prevalent use of complex sentences, using dependent clauses. Coordination structures follow consistently as the second most frequent grammatical unit and suggest a balanced use of compound sentences. In contrast, the nonfinite clauses are the least frequent, illustrating a lesser reliance on such structures. The distribution of clauses does not show an increase in nonfinite structures towards the end of the semester, but rather a combination of all three types of clausal structures, with a preference for subordinators.

Figure 7.9 illustrates the distribution of nonfinite clauses linked by subordinating conjunctions across the subcorpora, beginning with the descriptive texts (DS) at the start of the semester and ending with the argumentative texts submitted at the end of the semester.

Figure 7.9: Summary: distribution of nonfinite clauses in the subcorpora (the occurrences are normalized per million items)



The data shows that *-ing* participles are distributed normally across the subcorpora, with the highest usage found in process-analysis texts (PRS). Infinitives, on the other hand, do not follow the same pattern and are more prevalent in descriptive texts (DS) with relatively equ

al usage across the other subcorpora. Past participles are most common in process-analysis (PRS), compare-and-contrast (CCS), and argumentative (ARS) texts, indicating that as students progress through the semester, the complexity of nonfinite structures increases. This increase in complexity is not as evident in Figure 7.8, as nonfinite clauses are grouped together. However, the specific distribution in Figure 7.9 provides more data to support the increase in complexity as Staples et al. (2016) observe and confirm Biber and Gray's (2010) findings that past participles in nonfinite clauses characterize advanced academic writing and are challenging for entry-level writers. These findings highlight the distinctive clausal features of the first-year university students at the top genre level in COMP 101 and demonstrate the differences setting them apart from the upper level writers, as well as the distribution of these features across the genre subcorpora.

7.7 Conclusion

This chapter has explored the role of conjunctions as cohesive devices in linking finite and nonfinite cluses in COMP 101, addressing the primary research question, aiming to identify the most common features in first-year composition texts, as seen in COMP 101, and to examine how these features differ from those in upper-level writing, such as BAWE. Additionally, the chapter addresses the secondary question, which investigates how these features are distributed across different genres within COMP 101 and provides further insight into the writing patterns of first-year composition students. The focus of this chapter is the high frequencies of conjunctions, specifically their role in linking clausal structures, making them an item of interest in the study. The conjunctions are used in both finite and nonfinite clauses, providing syntactic information about the writers' ability to create text. Clausal structures have been examined by other researchers (Beaman 1984; Biber and Gray 2010; Lintunen and Makila 2014; Staples *et al.* 2016) as a means to investigate the syntax complexity in various levels of academic writing. This research contributes to this research and confirms the findings (Biber and Gray 2010; Staples *et al.* 2016) that as writers progress through university, they show increased use of past participles in facilitating nonfinite clauses.

The chapter also highlights that while first-year composition writers rely more on subordinators and coordinators in in clausal structures when constructing texts (Beaman 1984; Lintunen and Makila 2014), these structures should not be considered as the only measure of writing complexity and must be examined in light of the context. The use of concordance lines enables the observation of vocabulary differences in the context. Writers in COMP 101 typically concentrate on everyday topics and utilize straightforward language, which also reflects the tasks' instructions that allow students to choose their own topics, whereas BAWE writers employ

discipline-specific terminology that might be less accessible to a broad audience and cater to particular readers. BAWE writers incorporate complex vocabulary alongside coordinated clausal structures, exhibiting sophisticated meanings. This aligns with Halliday's (2007) observation that simple syntactic structures are able to demonstrate complexity when used with high-content vocabulary.

The findings regarding nonfinite clauses linked by conjunctions indicate that first-year writers and upper-level writers use nonfinite clauses at similar rates. However, when the nonfinite structures linked by conjunctions are examined separately, such as -ing particples, infinitives, and past participles, the differences in their use become more apparent. Out of the three types of nonfinite clauses, past participles present the greatest challenge for first-year composition writers, and this supports previous research (Biber and Gray 2010; Staples *et al.* 2016) that past participles are characteristic of academic writing. As with the finite clauses, the chapter highlights that nonfinite structures alone do not measure syntactic complexity, but their context should also be considered.

To continue the exploration of the most frequent patterns in COMP 101, the next chapter focuses on analyzing the most frequent punctuation marks as part of the top 25 frequency items in COMP 101. Along with personal pronouns and conjunctions in both finite and non-finite structures, the usage of punctuation provides additional insights into the writing choices made by first-year university students in COMP 101 and across the genre subcorpora.

Chapter 8

Punctuation

8.0 Introduction

This chapter continues the primary focus of this study, which is identifying the most common characteristics of first-year university writing in COMP 101 based on the corpus data. Building on previous chapters, Chapter 8 analyzes the most frequently occurring features in COMP 101 texts and their emerging patterns in first-year composition writing. This chapter studies the patterns that relate to the use of punctuation marks in COMP 101. It also compares the frequency of these marks in COMP 101 to those in upper-level texts, such as in BAWE. In the COMP 101 frequency list, three punctuation marks—full stops, commas, and quotation marks—rank among the top twenty-five, providing reference points for the corpus investigation to better address the main research question. This section targets the use of these marks and examines how they are used across different genres within the COMP 101 subcorpora, thus providing insights into the genre's influence on punctuation usage and addressing the secondary question of how the most frequent features are represented across different genres. Section 8.3 discusses each punctuation mark at the top level across COMP 101 and compares it to BAWE, as well as its usage across the subcorpora.

8.1 Previous Literature

The primary purpose of punctuation marks has been to make extracting information from written texts easier for the readers. Their usage evolved in stages based on the changing patterns of literacy motivated by the readers' demands (Parkes 1993). Bayraktar et al. note that "true understanding of written language will be impossible if punctuation marks are not taken into account." (1998, p.1). The historical development of punctuation marks reflects their essential role in conveying and understanding written information for readers. Discussing these

developmental stages, Parkes (1993), notes the importance of classical Latin writing, which focuses mainly on public speaking and uses punctuation marks to identify the length of pauses in the text. In the 8th century, the Irish scribes adapted ancient punctuation marks to the needs of copying texts, placing more emphasis on the "visual impact of punctuation and layout" (Parkes 1993, p.25). Mulvey (2015) notes that the Irish monks found it helpful to separate the individual words with a point, a practice that also proved useful for other scribes. However, not all scribes followed consistent practices in using punctuation marks, which resulted in variations between the copies of the same text. Standardization in using punctuation marks occurred only when printing emerged, and writers gradually became more particular about their sentence structure (Parkes 1993, pp.54-55). This concise historical account underscores the pragmatic role of punctuation marks, serving as tools to facilitate various functions and satisfy the demands of various reader groups. Also, punctuation marks evolved or were adapted as text users felt the necessity to preserve the meaning and integrity of the text.

Regarding this historical trajectory, Schou (2007) traces two main types of punctuation mark usage: one focused on rhetorical functions and the other on syntax and grammar. Both functions strike important purposes in punctuation usage. Rhetorical functions express the emphasis that readers need to note regarding a particular text segment (Sun and Wang 2019) or create illusions and "streams of consciousness" (Parkes 1993, p.87) that are embedded inside the text. Even though the rhetorical functions of punctuation are related to language use and important in communicating authors' attitudes about texts, recent research has focused on punctuation as "grammatical marking" and "visual devices" that facilitate syntactic relationships (Schou 2007, p.195). Through analyzing these syntactic relationships, Schou (2007) underscores the grammatical nature of punctuation and its related practices, expressed in style guides and

grammar textbooks (Strunk and White 1999; Carter and McCarthy 2006; Fowler and Aaron 2016).

One of the comprehensive discussions on punctuation is presented by Nunberg et al. (2002), who consider the topic as a linguistic subsystem that is both separate and related to grammar. According to Nunberg et al. (2002), punctuation marks are classified as segmental units that identify the "grammatical structure and/or meaning of stretches of written text" occupying a "position in the linear sequence of written symbols" (p.1724). In addition to the punctuation marks, these segmental units also include the space separating the individual items. On the other hand, non-segmental indicators include italicized text, capitalization, boldface, or small capitals. The punctuation marks are in a linear sequence because they occur within a sentence but outside the individual words. The two punctuation marks that are classified as word-internal include the apostrophe and the hyphen. In his survey of English punctuation, Mulvey (2015) takes a more practical approach and discusses the full stop, the semicolon, the colon, the comma, the slash, the hyphen, the parenthesis, the exclamation, the apostrophe, the quotation mark, and the question marks as the twelve fundamentals of English punctuation. These are the main punctuation marks listed by Carter and McCarthy (2006) and are defined as indicators of the boundaries and relations between sentence segments. In COMP 101, the full stop, comma, and quotation marks are rated within the top 25 frequencies, which makes them an item of interest to this study and raises the question of how first-year university students use them in their texts. In this way, the study examines the patterns in the use of punctuation marks based on prescriptive practices and, at the same time, provides a descriptive discussion of those practices.

Say and Akman (1996) note the need for descriptive treatment of punctuation marks in written text that investigates their functions and realizations in the text. Their research surveys the role

and use of punctuation marks in computational linguistics. For example, in some early natural language processing (NLP) projects, punctuation marks are tagged to mark a prosodic transcription or are treated as lexical categories and integrated into grammar as a component of machine translation systems. Such examples underline the important role punctuation marks play in the creation of text and "information packaging" (Say and Akman 1996, p.464). Another research in computational linguistics investigated the most frequent uses of commas to classify syntactically annotated sentences in a corpus containing material only from the *Wall Street Journal* (Bayraktar *et al.* 1998). Based on the Bayraktar et al. (1998) classification, commas appear most frequently in appositives, followed by elements in a series and sentence-initial elements, and sentence-final elements and quotations at the end. Bayraktar et al. (1998) indicate the need for further investigation of marks in other types of texts, such as fiction or scholarly writing, where punctuation might show variety.

Sun and Wong (2019) provide such an analysis of punctuation across various texts. The study uses frequency lists to investigate the synchronic and diachronic use of punctuation patterns in COCA (Corpus of Contemporary American English), COHA (Corpus of Historical American English), BNC (British National Corpus), GloWbe (Global Web-Based English). According to the study's findings (Sun and Wang 2019), the most used punctuation marks include the full stop, comma, colon, semicolon, hyphen, question mark, exclamation mark, parenthesis, and apostrophe. Based on the study, academic writing exhibits the lowest occurrence of periods, question marks, and exclamation marks while showing the highest occurrence of parentheses and semicolons. This illustrates that sentences in the academic domain are longer than in other types of writing and vary between 19 and 25 words on average (Li *et al.* 2023). Compared to full stops, question and exclamation marks are rarely observed in academic English. Another common mark

in academic writing includes the parentheses, which occur with citations and oftentimes with quotations. Both quotations and parentheses speak of the intertextuality or the multiple shared voices and borrowed authority in the academic community (Lombardi 2021). Occasionally, parentheses are used to indicate explanation, but that usage is rare. Sun and Wang's (2019) large-scale analysis of punctuation frequencies claims that they are not typical outside of academic English.

Unlike academic writing, Sun and Wang's (2019) research shows that fiction has the highest frequency of periods and exclamation marks. This increased usage of periods indicates a tendency for novel writers to use brief sentences. Exclamation marks are popular in fiction since they are widely used to communicate emotions or establish the atmosphere. The lowest frequencies in fiction are related to using hyphens, while newspaper texts show the highest frequency. Another high-frequency item typical in newspapers is the apostrophe. Both hyphenations and apostrophes facilitate concise language, a feature promoted in good journalism that stresses the importance of brevity and clarity of information while effectively conveying the story. This tendency of concise language might be the reason for the extensive use of abbreviations, so apostrophes and hyphens are used between words to facilitate the shortcuts in communicating the information.

Finally, in terms of spoken English, Sun and Wang (2019) identify question marks and colons as the most frequently utilized punctuation marks. People commonly employ questions or exclamation marks in social media and texting, indicating an informal style suitable for practical conversations. Colons serve multiple purposes: announcement, explanation, apposition, parallelism, and more. Consequently, colons are extensively employed to convey complex situations when transcribing spoken language into written form. This is why the frequency of

colons in transcribed spoken English exceeds that in other forms of communication. Newspaper texts prioritize concise language to convey immediate meaning, and hyphenation facilitates these qualities by integrating several words into one unit. Another punctuation mark typical for newspaper texts is the apostrophe due to its ability to reduce abbreviations. In conversational English, the question marks reach their peak usage and characterize the dialogical interaction between users.

As the frequency results in the survey of Sun and Wang (2019) show, punctuation marks have a variety of usage across registers and deserve the attention of linguists in tracing the distinction between the functions of marks and practices. Discussing the present and future of punctuation marks, Mulvey (2015) remarks that “serious modern punctuators do not invent punctuation marks, they apply them with rigor and style” (p.47) and also establish the guidelines that others must follow. Academic English is one of the areas where both scholars and students strive or attempt to use punctuation marks with rigor and style based on the grammatical guidelines. First-year composition students are an entry-level group in academic English. They need to familiarize themselves with the conventions and specific content related to their chosen major. As shown in Chapters 6 and 7, they use conversational English across different genres of writing, relying heavily on first and second-person pronouns. This chapter raises the question of how punctuation defines writing at this stage and what the major patterns of punctuation usage are revealed in the text. To examine these patterns, the study uses frequency lists and concordance lines, which are discussed in the following section.

8.2 Corpus tools to investigate the punctuation patterns

Punctuation mark usage varies depending on the individual text producers. Sun and Wang (2019) emphasize the importance of frequency in punctuation and the quantitative role of corpus linguistics in identifying patterns based on punctuation frequency. Likewise, this study uses the frequency list to understand the role punctuation marks play in first-year composition writing as represented in COMP 101 and how that role compares to more experienced writing as in the BAWE texts. Apart from the frequency list, the study also utilizes randomized concordance lines to examine the patterns created by punctuation marks in the texts. This section discusses the use of the frequency list and the concordance lines as the two tools utilized to examine the patterns.

8.2.1 Frequency list

Three prominent punctuation marks—comma, period, and quotation marks—are among the top twenty-five most frequently used items in COMP 101. Table 8.1 shows the normalized frequency of the first twenty-five ranks in COMP 101 and BAWE per 1,000,000.

Table 8.1: Punctuation marks in the top twenty-five frequencies in COMP 101 (raw frequencies)

N	COMP 101	R Freq	N Freq	BAWE	R Freq	N Freq
1	the	10644	56562	the	492270	70646
2	,	9998	53129	,	391643	56205
3	.	9947	52858	.	313580	45002
4	to	6165	32760	of	271079	38903
5	and	5626	29896	and	208693	29950
6	of	4510	23966	to	191604	27497
7	a	4097	21771	in	153326	22004
8	in	3275	17403	a	136398	19575
9	is	3224	17132	is	111307	15974
10	I	2459	13067	)	91843	13181
11	that	2433	12929	(	90538	12993

12	it	2195	11664	that	79337	11386
13	for	1851	9836	‘	72584	10417
14	are	1775	9432	as	68072	9769
15	they	1397	7424	for	59564	8548
16	with	1394	7408	be	58120	8341
17	my	1393	7402	this	54393	7806
18	“	1367	7264	it	51248	7355
19	you	1361	7232	“	47283	6786
20	was	1357	7211	:	47060	6754
21	on	1290	6855	are	42739	6134
22	be	1283	6818	with	42310	6072
23	not	1282	6812	on	40642	5833
24	have	1222	6494	by	40564	5821
25	as	1152	6122	was	36855	5289

Looking at the punctuation marks only, the comma is at rank 2 in COMP 101, followed by the full stop or period at rank 3, and the quotation marks at rank 18. Similarly, in BAWE, punctuation marks have high rankings. The comma and the full stop share the same rankings as in COMP 101. The quotation marks rank 19, one below the COMP 101 ranking. The differences are in the high frequency of the parenthesis at ranks 9 and 10, single quotation mark at 13, and colon at 20 in BAWE, which do not occur in the top COMP 101 rankings.

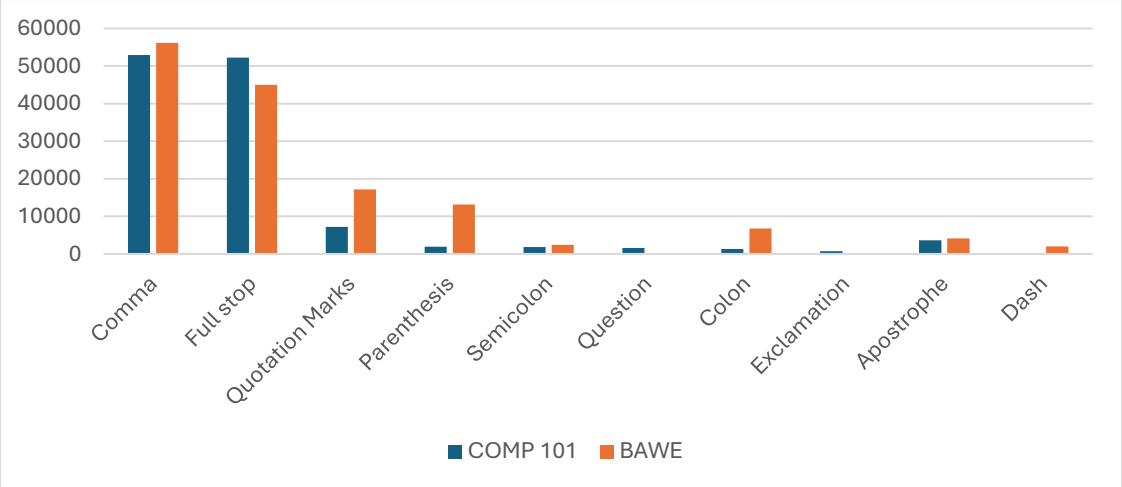
To thoroughly investigate the most used punctuation marks, this study looks into punctuation marks beyond the top twenty-five frequently used items. The current research considers the subsequent lines up to the cut-off raw frequency score of 100 in COMP 101 and compares them with the corresponding items in BAWE. The filtered results in Table 8.2 show the punctuation marks found in both corpora and their respective raw and normalized frequencies per 1,000,000 words.

Table 8.2: Punctuation frequencies in COMP 101 (normalized per 1,000,000 words)

N	COMP 101	R Freq	N Freq	BAWE	R Freq	N Freq
1	, (comma)	9995	52953	, (comma)	391643	56205
2	. (period)	9867	52275	. (period)	313580	45002
3	“ (double quotation marks)	1364	7226	() (parenthesis)	91843	13181
4	’ (apostrophe)	689	3650	‘(single quotation mark)	72584	10417
5	() (parenthesis)	364	1928	“ (double quotation marks)	47283	6786
6	; (semicolon)	354	1875	: (colon)	47060	6754
7	? (question mark)	296	1568	’ (apostrophe)	28898	4147
8	: (colon)	244	1293	; (semicolon)	16685	2394
9	! (exclamation mark)	130	689	- (dash)	13701	1966

To avoid confusion in distinguishing the marks, Table 8.2 lists each mark with its name, followed by the raw and normalized frequency. For a quick summary of these results, Figure 8.1 presents a visualization of the rankings based on their normalized frequencies in COMP 101 and BAWE. The quotation marks for BAWE include both singles and doubles.

Figure 8.1: Most commonly used punctuation marks in COMP 101 & BAWE



The two lists of COMP 101 and BAWE broadly align with Sun and Wang’s (2019) findings regarding the most commonly used punctuation marks, namely the full stop, comma, colon,

semicolon, question mark, exclamation mark, parenthesis, and apostrophe. It is interesting to observe the presence of exclamation and question marks in COMP 101. These two punctuation marks are not typical in academic writing (Sun and Wang 2019). Both occur at high frequency only in COMP 101 but not BAWE. This chapter examines the usage of all the punctuation marks displayed in COMP 101 and their most frequent patterns demonstrated by novice writers. As discussed in previous chapters, it is crucial to emphasize the significance of task instructions regarding punctuation in COMP 101. These instructions require the correct use of punctuation but do not dictate specific punctuation patterns. This flexibility enables students to make individual choices based on their proficiency and preferences. In some cases, the choices made by students may reveal areas that require improvement or reflect the influence of the genre.

8.2.2 Concordance lines in COMP 101 and BAWE

The punctuation marks were classified based on their syntactic and grammatical function, following Carter and McCarthy (2006). The COMP 101 corpus emphasizes nine frequently used punctuation marks: **comma**, **period**, **quotation marks**, **apostrophe**, **parenthesis**, **semicolon**, **question mark**, **colon**, and **exclamation mark** (see Table 8.2). Each category below illustrates the functions of the punctuation marks with examples from COMP 101 1-9:

1. **Comma:** The comma plays versatile functions in COMP 101 (Section 8.3.1, Figure 8.3) that include marking clause boundaries, separating list items, introductory elements, and references. Each function is illustrated with examples for clarity:

   Coordination: “ Her grandmother was lightly snoring just down the hall, but she was disabled with fear.”

Subordination: “One requires physical contact between the GM plant and another plant, however no such limit exists for the other methods.”

Elements in series: “Usually, it happens to people who tend to be more emotionally attached to people, memories, experience, and even things.”

Introductory elements: “Without saying it verbally, my upbringing taught me to ‘fake it until I made it.’”

References: “Lifewire, 5 Dec. 2019, www.lifewire.com/what-do-http-and-https-stand-for-3482375. Horowitz, Michael.”

2. **Full stop:** The full stop’s primary function is to mark sentence endings, which is also the main function in COMP 101 (Section 8.3.2, Figure 8.6). It is important to note that full stops also appear in references. Each of these functions is illustrated with examples from the COMP 101 dataset:

   Sentence ending: Moving on to my roommate’s side of the dorm room, which is much cuter than mine, you see her beautiful collage made up of old magazines and polaroid pictures she made that covers her entire wall.

   References: Hensley, Scott. “Poll: Americans Show Support for Compensation of Organ Donors.” NPR, 16 May 2012, www.npr.org/sections/health-shots/2012/05/16/152498553/poll-americans-show-support-for-compensation-of-organ-donors.

3. **Quotation marks:** The quotation marks can enclose direct quotes from speech or writing, indicate titles and headings, or highlight a word or phrase used ironically, often called scare quotes. In COMP 101, they are used in scare quotes, titles, foreign words, quotes, and direct speech (Section 8.3.3; Figure 8.8). The examples below illustrate each of these functions:

   Quotes: Additionally, “this cycle will only continue to get worse: If you don’t sleep enough at night, your body boosts its levels of stress hormones.”

   Direct speech: “This is like a geologist’s playground,” croaked Anthony.

   Foreign words: “In France, people celebrate the day of the first bottle of wine of the year, and the name of the event is ‘Beaujolais’”

Scare quotes: “When an officer ‘mistakenly’ shoots a black man, his first response is that the black man was reaching for a gun, because a black man with a gun fits stereotypes.”

Title: “An example of this genre is the movie ‘Miracles From Heaven’.”

4. **Apostrophe:** This mark is used mainly to indicate shortening or contractions in words and also possessive forms, which are also the two functions displayed in the COMP 101 dataset (Section 8.3.4, Figure 8.11). The examples below illustrate each of these functions:

   Contraction: “Opening up themselves to new people, it’s not easy, but it doesn’t block the ability to be nice towards others.”

   Possession: “Galaxy's display shares 3040 x 1440 and the Iphone-11’s resolution is 2436 x 1125, but just because there’s a dissimilarity in them, they both give the same type of vibrant colors creating a realistic look on the screens.”

5. **Parenthesis:** Parentheses are used to indicate in-text citations, provide additional information, and clarify abbreviations for lengthy titles or technical terms. In COMP 101 (Section 8.3.5, Figure 8.14), they are used in references, additional information, and abbreviations as illustrated by these examples:

   Additional information: “Some favorites were Santa Claus, Snow Buddies, Halloween Town, the Night Before Christmas (played during both Halloween and Christmas, thankfully).”

   Abbreviation: “These are some of the philosophies of those who are in favor of the practice of genetically modified organisms (GMO).”

   References: “Genetically modifying them for commercial use began in the 1990's (Green America, 2013).”

6. **Semicolon:** The semicolon separates items in complex lists or closely related independent clauses, especially when the second clause explains the first. In COMP 101 (Section 8.3.3, Figure 8.17), the mark is used between clauses as the example below illustrates:

Between clauses: “The latter description would perfectly apply to the Brazilian hot chocolate; it is rich and velvety.”

7. **Question mark:** Indicates interrogative sentences or signals declarative sentences that need to be read as questions. In COMP 101 (Section 8.3.7, Figure 8.19), it is used in titles, direct speech, and as an engagement marker. The following examples illustrate these uses:

   Title: “Are Vaccinations Beneficial Or Harmful?”

   Engagement markers: “When a doctor or nurse is told they have to kill a patient, how does that make them feel?”

   Direct speech: “Hey, Ann! Can I study with you tonight?”

8. **Colon:** The primary function of a colon is to introduce lists, subtitles, subdivisions in texts, references, and quoted speech. In some cases, colons can amplify a phrase or explain an idea. In COMP 101 (Section 8.3.8, Figure 8.22), it is used to introduce direct speech, explanation, divisions, lists, and is also used in references and titles as illustrated below:

   Direct speech: “Meanwhile, I heard a man’s voice speaking: "Hi, little girl, what do you want?”

   Explanation: “Last but not least, the Uyghur music: the traditional music of Uighur inherits the artistic traditions of Guoz music…”

   Divisions: “ Voyage to Freedom Part 1: The Need for Freedom”

   Lists: “Mongolian folk songs are divided into two types: ‘Urdu’ songs (long-toned songs) and ‘Wu Hue’ songs (short-key songs).”

   References: “HYPERLINK: http://www.bushcenter.org/publications/resources-reports/reports/immigration.html”

   Titles: “Cause and Effect Essay: Dress Code Effects”

9. **Exclamation mark:** This mark emphasizes exclamatory sentences, conveying strong emotion or surprise. In COMP 101 corpus (Section 8.3.9, Figure 8.25), it is used in direct speech and exclamatory sentences.

   Direct speech: “These customers can be heard saying, ‘I know right, that plate is heavy!’ Haha!”

   Exclamatory sentences: “Another miracle beverage I drink most mornings is coffee!”

Unlike the pronoun-verb pairs used in Chapters 5 and 6 and the CQL algorithms used in Chapter 7, the process in this chapter relied on concordance lines to examine the main patterns of the punctuation marks. After the high frequencies of the punctuation marks were established, as the previous section demonstrated, the study used the concordance lines in Sketch Engine. To observe the punctuation marks in the context, the study gradually increased the concordance lines with each punctuation mark up to 50, and added an additional set of 50 lines that were examined. This two-phase sampling approach helped confirm the consistency of the occurring functional categories across the marks. This sampling strategy was supported by the Central Limit Theorem (CLT), which states that with a sufficiently large sample size, typically around 30, the distribution of sample means approximates a normal distribution. This provided a statistical basis for the reliability of the observed patterns. In addition to that, accuracy and reliability were ensured through consultation with the dissertation supervisors, Dr. Brian Clancy and Dr. Joan O’Sullivan. Figure 8.2 shows ten randomized concordance lines that focus on using the comma in COMP 101 in SketchEngine.

Figure 8.2: Use of the comma in COMP 101

us updates in relationships or institutes of education	,	and media sharing.</s><s>These photos and videos are general
ntally and physically for adulthood.</s><s>However	,	by abusing the trust they have over their children by going to gre
Institute, a US- based reproductive health non-profit	,	the abortion rate is 37 per 1000 people in countries where aborti
/s><s>If the writer can put her/himself in their shoes	,	the writer will be able to connect with the audience more.</s><s>
><s>Religion and Politics At Oral Roberts University	,	religion is one of the main focuses.</s><s>This gets one thinking
ve.</s><s>A Dream City Once you visit Barranquilla	,	you will never want to leave.</s><s>From the moment you step
experiencing restaurant ambiances, various cuisines	,	and late-night snacks.</s><s>As a result of my love for stimulati
er grandmother was lightly snoring just down the hall	,	but she was disabled with fear.</s><s>Alisa loved visiting her gra
eople have gotten the virus and beaten it.</s><s>So	,	why are people being forced to make these huge accommodatio
e down to the point where I felt like I failed everyone	,	even my mom who put so much time into the army to get me to h

Some of the observations that can be made about these randomized lines are the patterns of the comma use in the following scenarios: lists, introductory elements, appositives, subordination, and coordination. The study extracts 100 randomized concordance lines of the comma and the other punctuation marks, transferring the results to Excel, labeling each instance, and summarizing the common patterns of each mark.

Since this study aims to discuss the most frequent textual features in entry-level composition writing, which fall in the broad context of academic writing, it is important to focus on the grammatical use of the punctuation marks as described in grammar and style textbooks. Thus, the purpose of the next section is to examine the most common patterns in punctuation use in COMP 101 and BAWE based on conventionally considered practices.

8.3 Punctuation marks in COMP 101

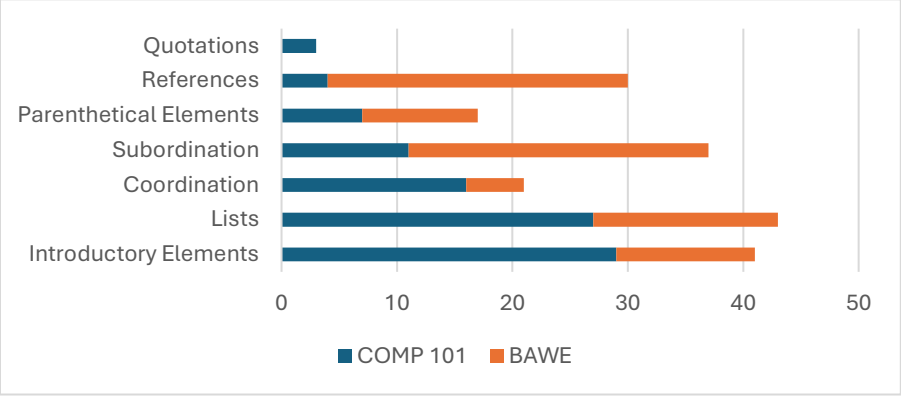
This section explores each punctuation mark identified in Table 8.2 and provides a detailed discussion based on their regular usage as described in English grammar and demonstrated use in the 100-randomized concordance lines.

8.3.1 Comma

Based on the frequency list in Figure 8.1, the comma is the most frequent punctuation mark in COMP101. According to Bayraktar et al. (1998), its high frequency is understandable based on its versatile role both in segmenting sentence elements and references. In sentences, commas mark the clause boundaries, such as main clauses, main and subordinate clauses, and relative clauses. Also, commas separate items in a list sequence, indicate a comment with adjuncts, mark tags, interjections, indirect speech, and address. In academic writing, commas are used to separate certain elements in references and in-text citations (Carter and McCarthy 2006).

This section discusses the most common patterns of comma usage in COMP 101 based on the 100-randomized concordance lines and compares its use to an equivalent randomized sample from BAWE. The study independently analyzed 100 randomized concordance lines in each corpus to identify the most common patterns. Based on the concordance examination, Figure 8.3 shows the most common patterns in COMP 101 and BAWE.

Figure 8.3: Summary COMP 101 & BAWE comma usage



Both groups use commas to signal references, parenthetical elements, subordination, coordination, lists, and introductory elements. Based on the examination of the concordance

lines, quotations introduced by a comma surfaced only in COMP 101, but with a very limited representation--only four instances. The fact that the randomized lines did not reveal examples of quotations in BAWE does not suggest that the corpus lacks such cases, but only that they might not be as frequent as the other comma patterns and may require further study. In COMP 101, three patterns—coordination, lists, and introductory elements—occurred more frequently than in BAWE. On the other hand, two patterns—references and subordination—occurred more frequently in BAWE.

It is probably not surprising that the commas signaling the coordinating relationships are more frequent than those indicating subordinate ones since previous literature (Beaman 1984; Biber and Gray 2010; Lintunen and Makila 2014) observes that authors using syntactically complex styles use longer sentences with more subordinate clauses, thus indicating higher complexity than entry-level writers. Similar to this finding is the high frequently used coordinating conjunctions in COMP 101 discussed in Chapter 7. Examples of the commas signaling coordinating and subordinating relationships in COMP 101 are included in (8.1).

8.1 COMP 101 (Coordination & Subordination)

> **Narrative Subcorpus:** *Her grandmother was lightly snoring just down the hall**, but she was disabled with fear.*** **(Coordination)**
> **Compare-and-contrast Subcorpus:** *Even though there are some similarities in the forms of communication**, there are also some obvious differences.*** **(Subordination)**

In COMP 101, the writers use coordination and subordination to talk about personal topics, opinions, or reflections. In BAWE, in examples (8.2), the writers use coordination in specialized contexts to relate information, discoveries, or discipline-specific content. It has been observed in the previous chapters that the COMP 101 students tend to use everyday topics because the curriculum allows them to choose their own topics of interest, which often results in general

subject areas. In contrast, the more specialized vocabulary used by BAWE writers likely reflects the discipline-specific tasks, which is understandable since the texts come from various fields, reflecting the specialized content.

8.2 BAWE (Coordination & Subordination)

> *Ewes are dragged**, so the ewes are clean before shearing, and so the possibility of fly strike is minimized*.** **(Coordination)**
> *One requires physical contact between the GM plant and another plant**, however no such limit exists for the other methods*.** **(Subordination)**

The comma usage related to coordination and subordination in COMP 101 mirrors the findings on coordinating and subordinating conjunctions discussed in Chapter 7. On average, first-year university writers use coordination almost twice as frequently as upper-level writers. The upper-level writers also show the use of coordination, but in all the examined instances, the topics relate to discipline-specific areas, reflecting the field of the chosen studies. The same tendency occurs when upper-level writers use subordinating clauses – they continue to demonstrate their professional interest. Thus, the findings in using commas align with the results and discussion in Chapter 7, adding to the insight related to first-year university writers.

The elements in series or lists in COMP 101 make one of the largest categories of 27 examples out of the 100 randomized lines, contrasting with the 16 cases of series in BAWE. The examples in (8.3) show the use of commas to separate items in a list in COMP 101.

8.3 COMP 101 (Elements in series)

> **Cause-and-effect Subcorpus:** *Usually, it happens to people who tend to be more emotionally attached to **people, memories, experience, and even things.***

The entry-level writers demonstrate familiarity with this use of the comma and do not show significant differences with the upper-level writers, as the example (8.4) shows in regard to BAWE. One difference between the use of the comma in lists between COMP 101 and BAWE is underlined by the differences between American and British English related to commas in lists, specifically the comma in front of *and*, also known as the Oxford or serial comma. In American English, the comma is required, while in British English, the comma is not considered necessary (Carter and McCarthy 2006).

8.4 BAWE (Elements in series)

> *According to (Nigel, Stuart, Robert, 2004), the company founder needs to consider the following aspects when setting up mission:* ***values, environment, customers, profitable and public image.***

The next comma pattern that indicates more occurrences in COMP 101 than BAWE is in identifying introductory elements. In both corpora, the introductory elements vary from a single word to a phrase. The example in (8.5) shows some of the uses in COMP 101.

8.5 COMP 101 (Introductory elements)

> **Cause-and-effect Subcorpus:** ***Without saying it verbally,*** *my upbringing taught me to "fake it until I made it."*
> **Argumentative Subcorpus: So,** why are people being forced to make these huge accommodations that are hurting them worse than the virus ever could.
> **Compare-and-contrast Subcorpus:** ***Therefore,*** *one can't deny or really hide the truth from his or her parents.*
> **Argumentative Subcorpus:** ***For example****, in 2016, there were more than 300 million bottles of champagne sold out globally (Insee).*

The examples show a variety in the use of introductory elements, which is a very positive feature for first-year composition writers as they elaborate on their thoughts and provide clues for the reader whether these clues relate to personal attitudes towards the topic, showing an example, or

serving as a discourse marker (e.g., *So, why are people…*). These introductory elements contribute to the text coherence and help the readers find the connections between the textual elements.

The writers in BAWE also demonstrate use of introductory phrases, as illustrated by the examples in (8.6), but the occurrences based on the 100 randomized concordance lines are three times less than COMP 101 (see Figure 8.3).

8.6 BAWE (Introductory elements)

> ***Breaking formally with Spain in 1821,*** *postcolonial Peru would witness significant changes in the state's approach to its indigenous majority and the "Indian problem".*
> ***However,*** *staff cuts are also an HRM issue because this is the department concerned with the welfare and management of employees.*
> ***Similarly,*** *a lot of the prospective benefits of labor mobility have been curbed due to the restrictions imposed by most of member states (with the exception of the UK, Sweden and Denmark).*

The main difference between the use of introductory phrases between the two groups is the level of complexity the writers bring to the topic, but the punctuation used on both sides shows a good understanding of these phrases.

The last category that shows underlying differences in the use of commas between the two corpora is in the use of references. In COMP 101, this use is underrepresented, with only four occurrences out of the 100 randomized lines. Example (8.7) shows one of the references in COMP 101.

8.7 COMP 101 (References)

> **Classification subcorpus:** ***Lifewire,*** *5 Dec. 2019****, www.lifewire.com/what-do-http-and-https-stand-for-3482375****. Horowitz, Michael.*

The use at the entry-level does not show a particular style but the random use of commas to separate the different elements in the bibliographic reference. References are one of the signature characteristics of academic writing, where writers borrow the authority of experts to establish causal relationships or build arguments (Fowler and Aaron 2016). The small number of references in COMP 101 corresponds to the tasks' expectation in COMP 101. While outside research is encouraged, it is not specifically required. In the argumentative essay, students are expected to provide evidence from outside sources as the task instructions require them to follow the MLA style when using citations (see Section 3.3). Still, there are some references that show the beginning awareness of writers about crediting their sources. The examples in (8.8) show the references used in BAWE.

8.8 BAWE (References)

> ***Crawley, M., J. Brown, M, S. Heard, M, S. and Edwards, G, R. (1999)*** *Invasion resistance in experimental grassland communities: species richness or species identity?*
> ***Wallace, Gavin, Stevenson, Randall. ed.,*** *The Scottish Novel Since the Seventies.*
> ***Jaffery, F.N., Chawla, G., Kakkar, P.et al. (1989)*** *Toxicology Data Handbook, Vol. III. pp. 262267.*

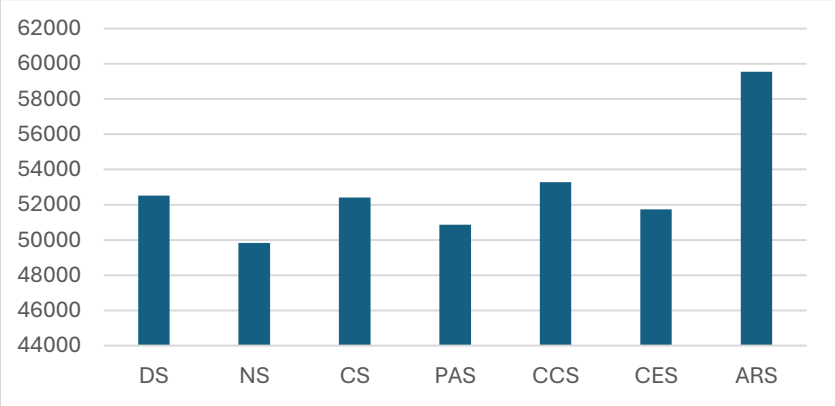
In BAWE, based on the concordance lines examination, writers appear to show consistency in the comma use in references separating the author's last name from the first initials and the other authors. However, these writers may indicate inconsistencies in documenting resources due to being at different university stages and still unfamiliar with some of the specific academic styles. Although these writers are upper-level students, they may not always demonstrate flawless documentation of sources. The overall implication, based on the use of commas, is that upper-level writers use commas in references five times more often than first-year writers.

The main differences in comma usage between COMP 101 and BAWE involve coordination, subordination, elements in series, introductory elements, and references. Coordination, lists, and introductory elements were more prevalent in COMP 101, while references and subordination appeared more frequently in BAWE. The next subsection shows the use of the commas in COMP 101 subcorpora.

8.3.1.1 Comma in COMP 101 subcorpora

The distribution of commas across the subcorpora is visualized in Figure 8.4. As the bar chart shows, the usage peaks in argumentative texts and remains relatively consistent across the other subcorpora.

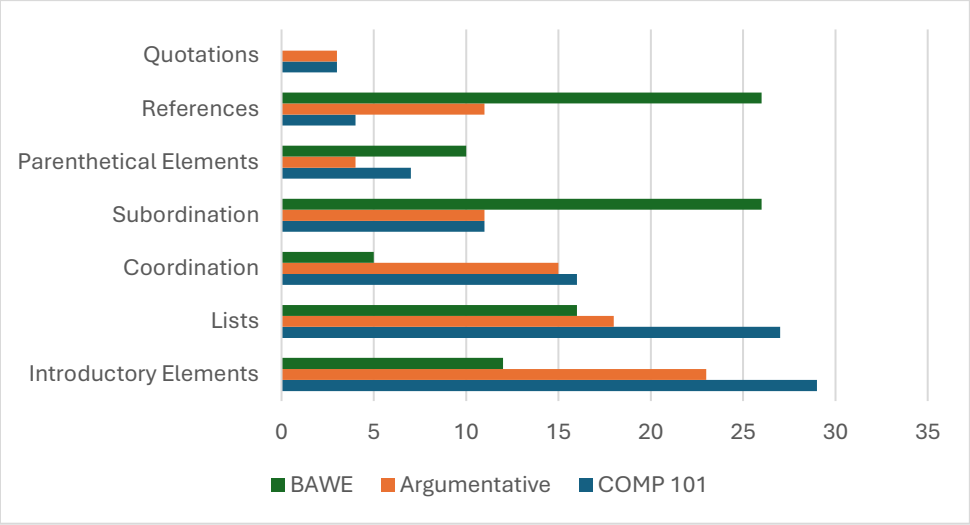
Figure 8.4: Distribution of the comma in the COMP 101 subcorpora



The bar chart shows the subcorpora in order of the text collection during the semester, starting with the descriptive texts (DS) and ending with the argumentative texts (ARS). In the argumentative texts, the writers show the highest usage of commas. This increased use of commas can be attributed to their integration of references. For example, the normalized comparison between ARS, COMP 101, and BAWE in Figure 8.5 shows the distribution of

commas across the main categories, such as quotations, references, parenthetical elements, subordination, coordination, lists, and introductory elements.

Figure 8.5: Comparison of comma usage in COMP 101, BAWE and ARS



The distribution of commas across ARS, in comparison with COMP 101 and BAWE, shows the argumentative texts to be closer to BAWE in reference usage and lists. Compared to COMP 101, the texts show similarities in the use of subordination, coordination, introductory elements, and quotations. The comparison suggests that first-year writers continue to follow the COMP 101 structure, which emphasizes coordination and subordination, but they are also beginning to pay more attention to external resources by incorporating research similar to that found in BAWE. The use of references in argumentative texts highlights how task requirements and essay genres influence the role of commas in citing sources. The incorporation of references in the ARS is intentional and aligns with the task instructions that guide students to cite sources correctly (see Section 3.3). While upper-level writers in BAWE show consistent progress in utilizing literature, first-year writers tend to demonstrate their use of research towards the end of the semester in their argumentative texts.

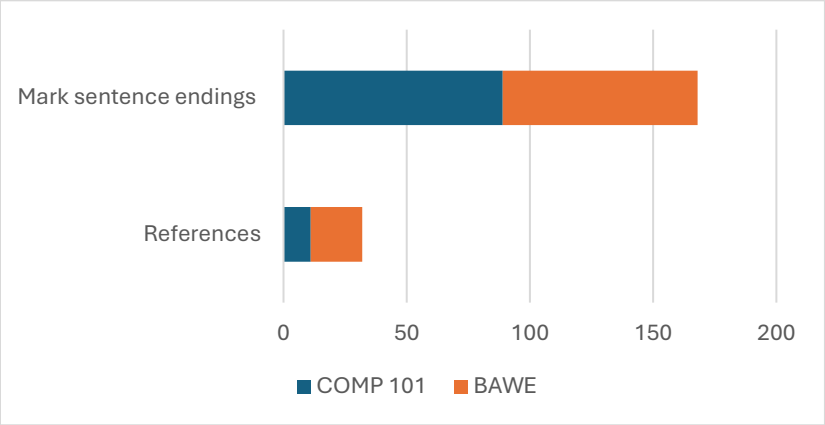
8.3.2 Full stop

This is one of the most straightforward punctuation marks in the text since it marks the end of the sentence. In American English, it is called a period, while in British English, it is known as a full stop. Typically, a sentence is expected to include one or more clauses, but sometimes full stops may close a sentence containing only one word for a dramatic effect in advertisements or dialogues. Another practical use of full stops includes marking individual letters in abbreviations and the three consecutive periods in ellipsis to indicate omitted text in quotations or sentences (Carter and McCarthy 2006).

In the frequency lists of COMPS 101 and BAWE, full stops take the second rank of the most common punctuation (see Table 8.2). The normalized frequency places COMP 101 with a slight increase in the normalized frequency of 52,275 versus 45,002 in BAWE. SketchEngine calculates the average sentence length by dividing the number of words in a corpus by the number of full stops. The analysis counts all the full stops, both for sentence-ending punctuation and those in references or web addresses (e.g., examples 8.9 and 8.10). Calculating the mean sentence length shows 18.6 words per sentence for COMP 101 and 22 words per sentence for BAWE. This average for BAWE aligns with the findings of Li et al. (2023), which reported the average sentence length in journal articles, indicating that the BAWE writers meet the average standard.

Upon examination of two sets of 50 randomized concordance lines in SketchEngine for each corpus, the study observed two main uses of the full stops: marking the end of the sentence and signaling separate elements in references. Figure 8.6 shows the summary of the findings from the concordance lines examination.

Figure 8.6: Summary of the full stop use in COMP 101 and BAWE



The summary shows that the primary use of the full stops, as expected, is to mark the end of sentences. The second type of usage pertains to references, which is a distinctive characteristic of academic writing, as it draws evidence for support from research. In COMP 101, the full stop in references is three times less frequent than in BAWE. This observation aligns with the fact that the COMP 101 tasks do not prioritize research at this stage. However, emerging research indicates a positive trend, showing that students engage with outside sources to meet the expectations for the tasks that do include research, such as the argumentative essay (see Section 3.3).

A review of the references in the examined concordance lines in COMP 101 and BAWE shows two different types of sources between COMP 101 and BAWE. In COMP 101, in example (8.9), the writers gravitate towards sources available on the Internet that allow students to simply search by using a keyword related to their topic of interest.

8.9 COMP 101

> **Compare-and-contrast Subcorpus:** *Understanding the American Education System. www.studyusa.com/en/a/58/understanding-the-american-educn-system*.

Cause-and-effect Subcorpus: *Sources Jaret, Peter. "Eating Disorders and Depression." WebMD, 30 July 2010, www.webmd.com/mental-health/eating-disorders/features/eating-disorders#1.*
Argumentative Subcorpus: *Hensley, Scott. "Poll: Americans Show Support for Compensation of Organ Donors." NPR, 16 May 2012, www.npr.org/sections/health-shots/2012/05/16/152498553/poll-americans-show-support-for-compensation-of-organ-donors.*

In BAWE, in example (8.10), the writers focus their research on journals or specific books, demonstrating a very deliberate approach toward evidence in discipline-specific literature.

8.10 BAWE

Arango, Sebastian and Nadiri Ishaq. 1981. "Demand for money in open economies." Journal of Monetary Economics, 7, pp. 69-83.
Cunningham, I., Hyman, J. and Baldry, C. (1996): Empowerment: the power to do what? Industrial Relations Journal, 27:2, pp143-154.
Secondary SourcesBreebaart, A. B. (1971), 'Plutarch and the Political Development of Pericles', Mnemosyne, Ser. 4, 24, 260-272.s

Examining the use of full stops in references may also reveal that students use different ways in entering the bibliographical information. For example, in (8.10), the three references show three different methods to cite the year of publication. In the first example, the year is listed without regular parentheses and is separated from the title by a complete stop. However, in the other two references, the year is separated by a colon or a comma instead of a full stop. These differences in the reference documentation suggest the use of different styles, which indicates that further investigation may show the most common patterns displayed by students in references.

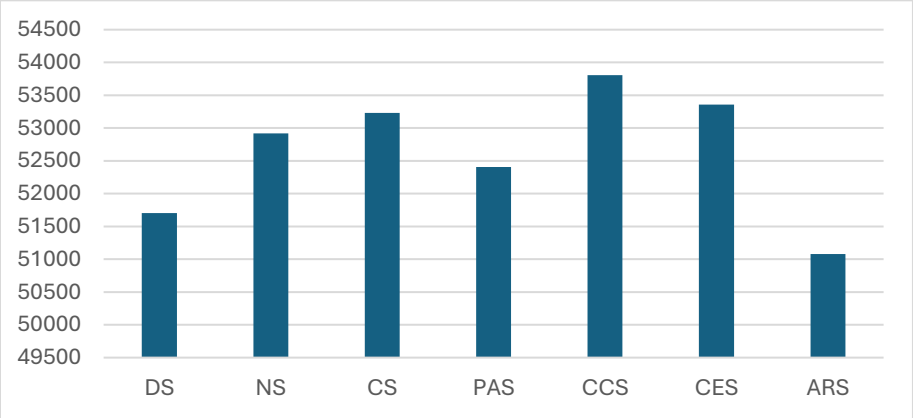
Apart from being used in references, the main usage of full stops is marking the end of sentences, as this section briefly discussed. To gain a deeper understanding of the role that full stops play in determining sentence length, this research conducted a study using a sample of 100 randomized lines from COMP 101. Some of these lines included references. The lines were

downloaded as a separate file and uploaded to SketchEngine in a corpus format. A second file was created that excluded the references, which was also uploaded to SketchEngine in corpus format. This approach allowed for easy calculation of word count and full stops. The file that included references contained 17,714 words and 977 full stops, while the file without references had only 7,553 words and 398 full stops. Based on this data, the average sentence length in the file with references was approximately 18 words per sentence, while the average sentence length in the file without references was about 19 words. Although this approximation does not yield significant results, it may provide useful insights for future research, which could be conducted on a larger scale for greater accuracy.

8.3.2.1 Full stop in the subcorpora

The chart in Figure 8.7 shows the frequency distribution of full stops across subcorpora, based on the normalized figures from Table 8.3. The mark appears frequently in almost all subcorpora but is less common in argumentative texts. This finding may suggest progress among first-year composition students, as they are not required to keep a specific sentence length, but appear to produce longer sentences in the argumentative genre. This shift could reflect both a progression in students' writing skills towards the end of the semester and the demands of the argumentative writing, which may encourage more complex and longer sentence structures for persuasive effect.

Figure 8.7: Distribution of the full stop in the COMP 101 subcorpora



Having fewer full stops in ARS suggests that argumentative texts should have longer sentences than the other subcorpora. When the average sentence length across the subcorpora is calculated, the argumentative texts show an average sentence length of 19.6, slightly higher than the average COMP 101 sentence length (18.6) and closer to the BAWE (22). Considering the timeline and knowing that the texts are submitted towards the end of the semester in the argumentative subcorpus, it might be considered that students begin to use longer sentences, which is more representative of academic writing (Sun and Wang 2019).

It is worth noting that descriptive texts, like argumentative ones, tend to use fewer full stops. However, unlike argumentative texts, these are usually submitted early in the semester. By using the previously discussed formula, it is estimated that the average sentence length in descriptive texts is 19.3, which is very similar to that of argumentative texts. The question is why students tend to write longer sentences at the beginning stage of their writing.

It is possible that students tend to write longer sentences at the beginning of the semester when they are working on descriptive writing because the texts are about describing the characteristics of places, people, or objects, which allows the use of figurative language and a less rigid

academic style. Descriptive texts also tend to be longer due to their focus on narrative style and fluid explanatory details, as shown in the examples in (8.11).

8.11 Descriptive subcorpus

> **Descriptive language:** *You look up around and see the sky is a reddish rose color with several moons covered in craters from the local asteroid belt.*
> **Complex ideas:** *Soon after the tiger feels the last throbbing heartbeat of its prey, he carries it in a prideful manner to share the evening meal with his cubs.*
> **Specific details:** *Moving on to my roommate's side of the dorm room, which is much cuter than mine, you see her beautiful collage made up of old magazines and polaroid pictures she made that covers her entire wall.*
> **Lists:** *I have had many injuries including, dislocating my foot and double high ankle sprain, all at the same time, four broken fingers (both thumbs, and both pinky's), torn meniscus, torn labrum, five concussions, and lastly, broken my back three times.*

Although the sentences in this genre may be longer, they still provide students with opportunities to express their thoughts and observations about a subject without the need for research or gathering sources. The language used in this type of writing relates to general experiences and is not specialized in fields such as engineering, biochemistry, or psychology. The descriptive writing style gives students opportunities to write about life in general, which suggests that this genre may be easier for students to handle than analytical academic writing.

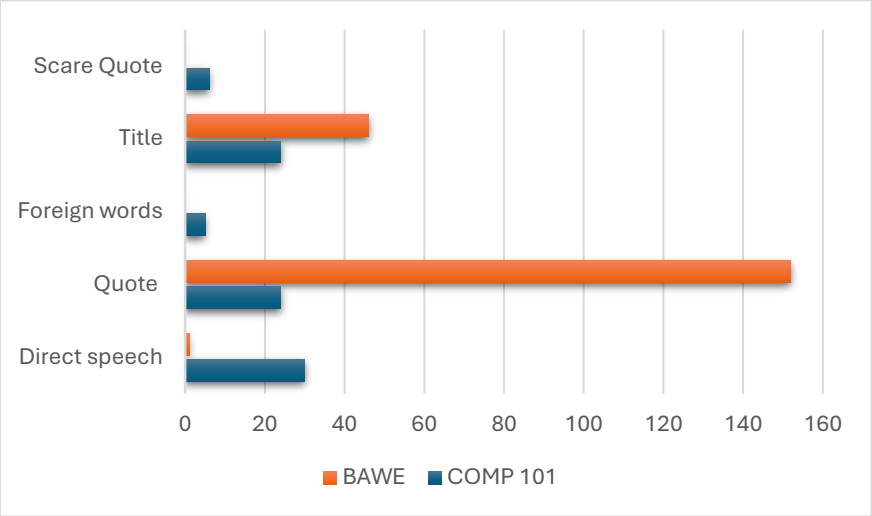
Considering the examples in the descriptive texts and reviewing Figure 8.7, it becomes understandable that the distribution of full stops among the subcorpora follows a normal distribution. The lower use of full stops in earlier texts in the semester might be based on the personal aspect of writing, allowing more freedom to students, whereas later on, students are required to write in genres such as classification, process, or cause-and-effect, which demand analytical skills. In the middle of the semester, the high frequency of the full stops may speak to shorter sentences, but towards the end of the semester, writers demonstrate a more complex writing style with longer sentence structures.

8.3.3 Quotation marks

Quotation marks, also known as quote marks, are used to enclose direct quotations from speech or writing. They also indicate headings and titles or draw attention to a word or phrase used with irony to show a different or opposite meaning, commonly referred to as scare quotes (Fowler and Aaron 2016). They can be either single or double, depending on the reference style. This variation in the notation of quotation marks is perhaps the reason Lester (2018) does not distinguish between double and single quotation marks when discussing their usage. SketchEngine accommodates this difference by listing the single quotations as (‘) and double quotations as (“).

In Table 8.2, the double quotation marks are ranked at position 3 in COMP 101 and at position 5 in BAWE. In BAWE, the frequency list shows double quotation marks at rank 5, but it also includes single quotation marks at rank 4. When their raw frequencies are combined, their total reaches 119,867, which enables them to rank at position 3, currently occupied by parentheses (e.g., 91,843). When examining the concordance lines, the study does not treat single quotation marks in a separate category but looks at both as quotes. Figure 8.8 summarizes the quotation marks usage in COMP 101 and BAWE based on the 100 randomized concordance lines in each corpus. The main patterns include quotes and titles in both corpora and direct speech, foreign words, and scare quotes in COMP 101 only.

Figure 8.8: Summary of the quotation marks usage in COMP 101 and BAWE



The first observation based on the summary is that both groups of writers utilize quotations to some degree and demonstrate intertextuality, a distinct characteristic of academic writing (Lombardi 2021). Based on the lower usage of direct quotations in COMP 101, it might be noted that intertextuality is emerging, which is also understandable for entry-level writers since most tasks do not require a research component. The exception is the argumentative essay, where the use of outside sources is expected to provide support; however, even in this case, the primary goal is not research itself, but the development of a coherent and persuasive argument. . Compared to COMP 101, BAWE writers use quotations 5 times more than entry-level writers, showing that writers focus more on outside sources at later stages, which should be expected as students progress through their majors. However, the need for evidence and research is seen even at this entry-level of university writing, which shows that students begin to recognize the role of the larger academic or professional community in constructing their texts. Academic writing is inherently connected to the community of scholars and develops from their work. It is fascinating to see how first-year composition writers take their first steps in exploring this area. The examples in (8.12) show some of the typical ways the students integrate outside sources in

their work. At this stage, it is common to see short introductions for the quotes included in the text, and these introductions may consist of one word or phrase.

8.12 COMP 101 (Quotes)

> **Argumentative Subcorpus:** *Additionally, "this cycle will only continue to get worse: If you don't sleep enough at night, your body boosts its levels of stress hormones"*
> **Process-analysis Subcorpus:** *Moreover, "knowing that someone else expects (the person trying to change) to be better is a powerful motivator."*
> **Compare-and-contrast Subcorpus:** *Carbonation is "when highly pressurized carbon dioxide is dissolved into a solution."*

By including outside sources, the writers acknowledge the importance of evidence, examples, and the opinions of others that they use to strengthen their position. It is important to note that while process-analysis and comparison-and-contrast tasks do not require the use of outside sources, students are encouraged to use sources and often recognize their value in strengthening their writing. In the first-year composition texts, the steps to integrate outside sources do not show the same confidence as in the BAWE texts. For instance, in example (8.13), the introductions are longer, with more details and a positive attitude toward the sources. However, it is encouraging to see the growing role of research in COMP 101 texts.

In BAWE, the examples in (8.13) show the typical ways the writers integrate research into their texts and the higher level of integration that does not only provide outside information but also shows the writers' attitude towards the source.

8.13 BAWE (Quotes)

> *Also, of Virgil's Fame and Homer's Discord: "These figures in painting would be clear enough, but I fear they might become ridiculous."*
> *This, however, was only included as obiter, with Justice Collins acknowledging that "whether in due course a court would find that there was any breach of the law in that regard is another matter"*

*The ultimate result of these circumstances was that they **"**served to disillusion workers with the possibilities for peaceful change under the tsarist system and drove them toward more drastic and radical solutions to labor problems.**"***

The quotations' frequencies, along with the concordance lines, not only demonstrate the increasing role of research in university writing at the entry level but also reveal two different patterns of integrating external sources. In COMP 101, writers use brief introductions for their quotes to underscore the importance of research in their writing. However, they are not as familiar with literature as the upper-level writers in BAWE. The quotes highlight an area for improvement for first-year composition writers, indicating the need to learn more about individual sources and engage with them. Even though research may not be the primary focus of the instructional strategy in COMP 101, instruction could begin to emphasize individual sources, such as short journal articles that students may explore throughout the semester and learn how to engage with their content.

The next interesting difference between COMP 101 and BAWE is the use of direct speech in COMP 101 and its limited presence in BAWE. In COMP 101, direct speech relates to the narrative genre includes dialogues as part of the storyline and shows task-based writing. The examples in (8.14) show some of the typical uses of direct speech as dialogues in narratives.

8.14 COMP 101 (Direct speech)

Narrative Subcorpus: "*This is like a geologist's playground,"* *croaked Anthony.*
Narrative Subcorpus: *I said, "Hi, do you remember me?"*

The dialogues show a variety in the quotation marks usage and students' ability to incorporate dialogues that comply with the grammatical regulations of indicating direct speech appropriately

in texts. On the other hand, the examined concordance lines in BAWE showed only one occurrence of direct speech by the upper-level writers. The example in (8.15) shows the direct speech in BAWE. The example shows a correctly punctuated line expressed by someone in a narrative.

8.15 BAWE (Direct speech)

"I really don't have the time, detective," she replied resolutely.

This single instance of direct speech in BAWE is only present in the 100 randomized concordance lines in BAWE examined by the study. It does not necessarily indicate that such uses are extremely rare, but it does show that the examined lines come, for the most part, from texts that are analytical in nature.

Another striking difference between the use of the quotations between COMP 101 and BAWE is their use to indicate foreign words in COMP 101. This use is only 5 percent of the overall use of quotation marks, but it shows to some degree the international background of the writers, which is 20 percent of all the participants (see Chapter 2 for details on the demographics). Example (8.16) shows some of the foreign words used in COMP 101 to reflect the cultural background of the writers.

8.16 COMP 101 (Foreign words):

Argumentative Subcorpus: *In France, people celebrate the day of the first bottle of wine of the year, and the name of the event is "Beaujolais".*
Classification Subcorpus: *According to the characteristics of its music, Mongolian folk songs are divided into two types: "Urdu" songs (long-toned songs) and "Wu Huer" songs (short-key songs).*

Writers set apart foreign words from the rest of the text using quotation marks. These foreign words are related to various topics connected with the cultural backgrounds of the writers, indicating the multicultural backgrounds of the text producers. Spotting the foreign words through the quotation marks as part of the analysis of the frequency items in COMP 101 provides more evidence of the effectiveness of this approach in providing information about the writers.

One final difference in the use of quotes between COMP 101 and BAWE is the use of scare quotes, which are similar to the foreign words category and are around 5 percent of the overall quotation marks' usage in COMP 101. According to Predelli (2003), this type of quotation mark use is not well accepted in academic writing based on their unsuitability to the context. Unsuitability is viewed as the non-standard usage of words within a formal register or implying meaning instead of explicitly explaining the concept. In COMP 101, scare quotes show the personal style of some writers by using irony to underscore a deeper meaning. In example (8.17), the writer places *mistakenly* in quotes to suggest a regular practice some seek to excuse by using stereotypes.

8.17 COMP 101 (Scare quotes)

> **Argumentative Subcorpus:** *When an officer **"**mistakenly**"** shoots a black man, his first response is that the black man was reaching for a gun, because a black man with a gun fit stereotypes.*

In the words of Predelli (2003), such use adds metalinguistic content that "contributes to the presentation of certain information pertaining to the expression they (the quotation marks) flag" (2003, p.4). Even though scare quotes are not accepted practice in academic writing, this study reveals some of the personal styles of expression used by first-year composition writers, which is

an interesting characteristic to uncover with the help of frequency analysis. By identifying their presence, instructors may show the reason that academic writing avoids them and help students find other ways to add to the metalinguistic content or perhaps use the scare quotes with a measure of caution in strategic places for emphasis only.

Overall, the analysis of the quotation marks shows some of the most revealing patterns between the two groups of writers, which provides distinguishable features for first-year composition writers and also suggests a wealth of information for future research. The features characterizing first-year composition writers include emerging intertextuality, effective use of dialogues, foreign words, and scare quotes.

8.3.3.1 Quotation marks in the subcorpora

In the subcorpora, the use of quotation marks varies. Figure 8.9 summarizes the usage of quotation marks across different genres, based on the examination of 100 random lines from each subcorpus. The results indicate that the argumentative subcorpus has the highest frequency of quotation marks, followed by the classification and narrative subcorpora. In the argumentative and classification texts, quotation marks are used to denote citations, while in the narratives, they indicate direct speech. The use of quotation marks in the argumentative genre aligns with the task expectations that require some engagement with sources, but the other essay genres only encourage research, which demonstrates that students are beginning to recognize the value of outside sources as support for their ideas, even when not instructed to do so.

Figure 8.9: Distribution of the quotation marks in the COMP 101 subcorpora

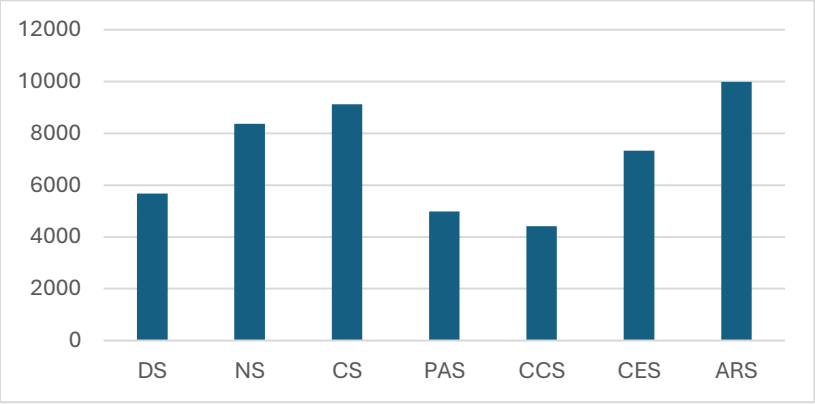
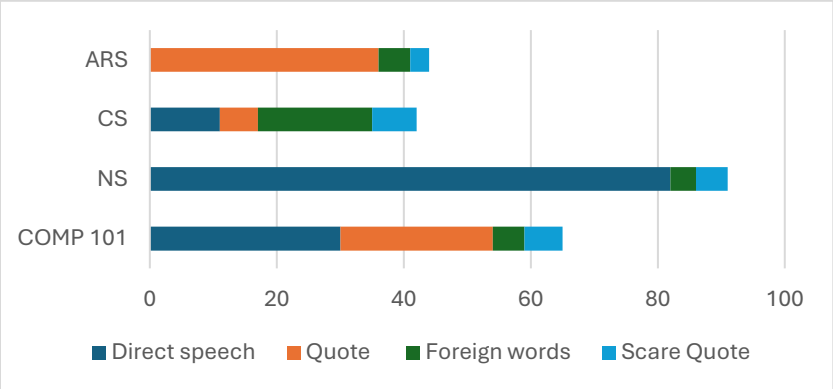


Figure 8.10 provides additional information about the ways the quotation marks are used in these three subcopora, showing a comparison of them and COMP 101.

Figure 8.10: Functions of the quotation marks in ARS, CS, NS, and COMP 101



Based on the comparison in Figure 8.10, the use of quotation marks in direct speech is primarily found in narrative texts. This is understandable since personal stories often incorporate dialogues to present characters or advance plots. On the other hand, argumentative texts, which are usually submitted towards the end of the semester, utilize quotes the most. Entry-level writers tend to use outside sources to support and enhance their positions in argumentative writing. Classification

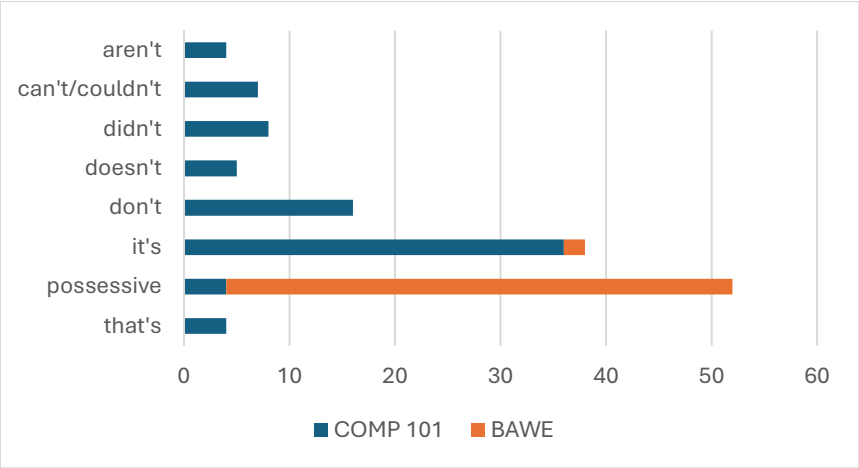
texts, on the other hand, contain the most titles related to sources and foreign words or terms distinguishing different categories within the genre topics. Lastly, scare quotes are distributed almost evenly across the three subcorpora and signify writers' tendency to use informal features. Based on this distribution of marks, the function of quotation marks appears to follow a trajectory across the subcorpora that, in some cases, may reflect a growth process, while in others it aligns closely with the demands of the essay genre. For example, early in the semester, students are assigned a narrative essay based on personal experience, and the use of quotation marks in this context aligns with task expectations, marking direct speech. Similarly, the presence of quotation marks in argumentative essays corresponds with the expectation to incorporate outside sources. However, evidence of growth emerges in students' use of quotations for support even in genres where research is not explicitly required, indicating a developing understanding of how to strengthen their arguments through intertextual references. .

8.3.4 The apostrophe

The apostrophe has two primary uses: showing possession using *-s* and indicating that something has been shortened. When apostrophes facilitate contractions, they indicate missing letters. The apostrophes are often used to show the contracted punctuation of words in verbs, such as *be*, *have*, or *will*, as in *I'm* (*I am*), *she's* (*she has*), or *we'll* (*we will*) (McCarthy 2017). According to Lester (2018), one of the biggest problems with contractions is not their misspelling but inappropriate usage in formal contexts. Typically, the apostrophe shows the shortened words in fiction, where writers communicate sounds and spoken language, but it is not considered acceptable in nonfiction writing, especially when this writing is academic (Lester 2018; Dixon 2022).

The frequency in COMP 101 ranks the apostrophe at position 4 in the punctuation list and position 7 in BAWE (See Table 8.2). The apostrophe in COMP 101 is used more than 95 percent for shortening the forms of the verb *be, do,* or *could* (e.g., *it's, that's*, *don't, didn't,* or *couldn't*) and less than 5 percent in possessive forms, while in BAWE, almost all instances reveal possessive uses. Figure 8.11 categorizes the main uses of the apostrophe observed in the examined 100 concordance lines in each corpus.

Figure 8.11: Summary of the apostrophe usage in COMP 101 and BAWE



In COMP 101, writers use apostrophes predominantly to indicate contractions, which is more typical for fiction writing (Lester 2018) than academic writing. The task instructions do not specifically address the use of contractions but emphasize the overall correct punctuation (see Section 3.3). This suggests that contractions are not explicitly featured in the curriculum, which may indicate an area in need of more targeted instruction and practice. The most common shortcuts are related to the verb *be* when used with the third person singular *it*, the auxiliary *do*, and the modal *can/couldn't*. The sentences in (8.17) show examples of the contracted and possessive uses of the apostrophe in COMP 101.

8.17 COMP 101

> **Cause-and-effect Subcorpus:** *Opening up themselves to new people,* ***it's*** *not easy, but it* ***doesn't*** *block the ability to nice towards others*. **(Contraction)**
>
> **Compare-and-contrast Subcorpus:** *Galaxy's display shares 3040 x 1440 and* ***the Iphone-11's resolution*** *is 2436 x 1125, but just because there's a dissimilarity in them, they both give the same type of vibrant colors creating a realistic look on the screens.* **(Possession)**

The contractions indicate that first-year composition writing gravitates towards informality, contrary to academic writing requirements. At this stage, writers do not demonstrate significant compliance with the conventional requirements in spelling out the complete forms but rather utilize their contracted equivalents. The contracted expressions also show that novice writers feel more comfortable with the spoken language when expressing their thoughts and reflecting on life and experiences. On the other hand, the small percentage of possessive uses shows that entry-level writers are familiar with them and use them effectively, as the example in (8.17) demonstrates.

In BAWE, on the other hand, the possessive use of the apostrophe is the predominant one, which leaves the contractions at a minimum and indicates more mature academic writing. The examples in (8.18) illustrate some of the apostrophes used in the corpus.

8.18 BAWE

> *Referring to* ***Eddington's trip*** *to view a solar eclipse and the curvature of space.* **(Possession)**
>
> *There is no mention of religion, blackness and specifically no reference to* ***Toussaint's*** *military success*. **(Possession)**
>
> *Once someone is identified as different,* ***it's hard*** *for them to be accepted* (Royal psychology association, 2005). (**Contraction)**

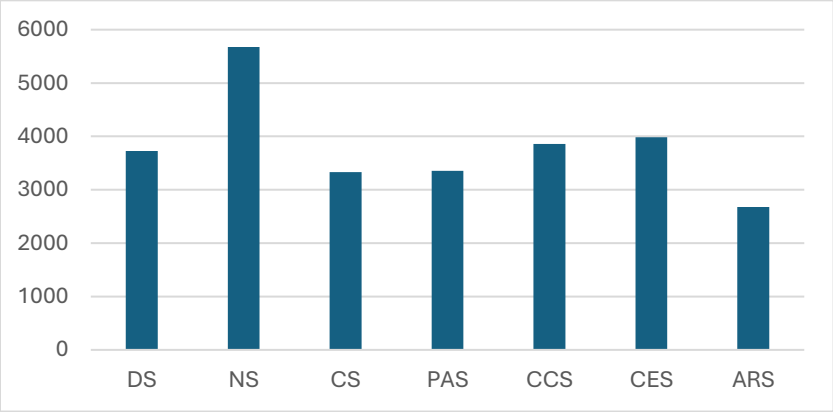
At this stage of writing, writers focus on complying with academic conventions by using possessive forms that demonstrate a legitimate use of the apostrophe. Although a few instances of contracted forms were observed in the concordance-lines examination, writers generally avoid using shortened forms, as the last example in (8.18) shows.

Analyzing the frequency of apostrophe usage in both groups of writers provides valuable insights into assessing writing proficiency and skills at different stages, specifically in COMP 101 as novice writers. The findings suggest that entry-level writers are not as familiar with and used to the regulated usage of the apostrophe and might need emphasized instructions in class to understand the distinction between the possession and contraction with this punctuation mark and demonstrate better compliance with the academic conventions. Unfortunately, there is not sufficient research on the apostrophe in the texts written by first-year composition students apart from this study to confirm these findings, and this might be a promising area for future research.

8.3.4.1 Apostrophe in the subcorpora

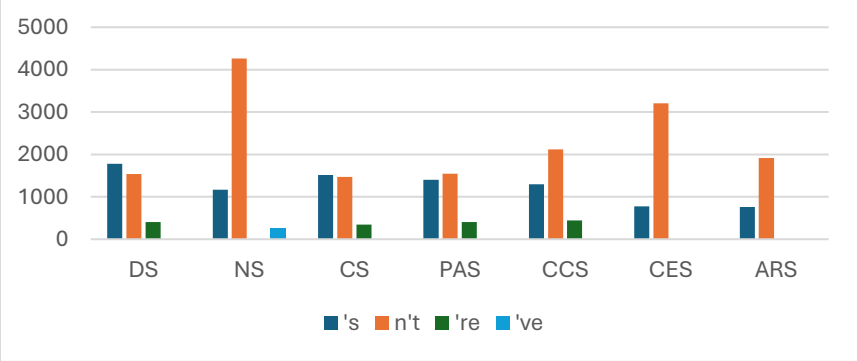
This section summarizes the usage of the apostrophe in different subcorpora, as demonstrated in Figure 8.12. The subcorpora are listed in chronological order of the text collection, starting with the descriptive texts (DS) and ending with the argumentative (ARS). The narrative texts have the highest apostrophe usage, while the argumentative subcorpus has the lowest.

Figure 8.12: Distribution of the apostrophe across the subcorpora



The distribution shows the overall use of the apostrophe but not the different categories it is observed in COMP 101. Figure 8.13 provides a normalized comparison between the types of apostrophes use across the subcorpora. The bar graphs show that the mark is most commonly used for contractions in narrative texts, particularly for the contractions *–'s, n't,* and *'ve*. This increased use of contractions might be influenced by the nature of the narrative genre, which draws on personal experience. As students describe their own stories, they tend to adopt a more natural and conversational tone, which can result in a higher frequency of contractions.The other subcorpus that shows a high frequency of the apostrophe is the cause-and-effect (CES), where the contractions are mostly negative forms, and the verb *is*.

Figure 8.13: Apostrophe in contractions across the subcorpora



The high frequency of the apostrophe in contractions suggests that entry-level writers keep their tendency towards informal writing throughout the semester, even after being consistently reminded that academic writing typically avoids contractions. By identifying the main contraction patterns, the study provides the main patterns of apostrophe use by first-year composition students that may help instructors focus on helping students understand the appropriate usage of apostrophes in academic writing. Such focus can identify specific areas in formal and informal writing features.

8.3.5 Parenthesis

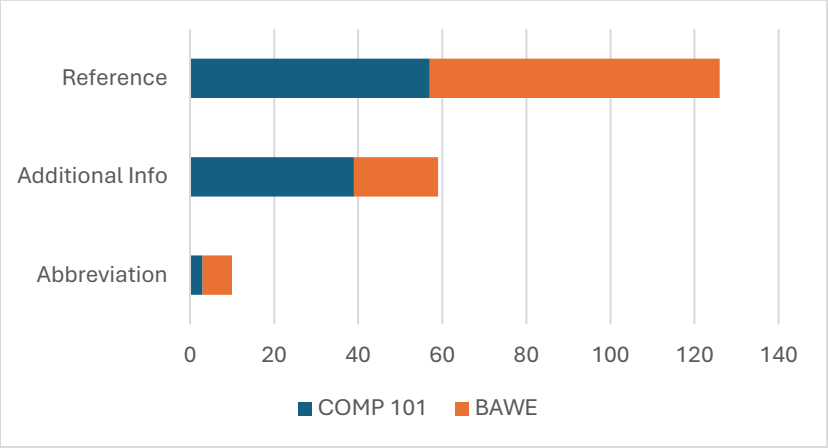
Sun and Wang's (2019) study shows that parenthesis is mainly used in academic writing. This usage is understandable since one of the main functions of parenthesis is to indicate in-text citations, a defining feature of academic texts that all university students need to learn during their first year of university writing. Another function of parenthesis is to include additional information in the text, which is not required, but writers consider it interesting for the overall textual idea. Often, writers use parenthesis to provide clarification of abbreviations related to long titles and technical names to make texts more readable (Lester 2018).

COMP 101 uses parenthesis as a high-frequency punctuation mark. In SketchEngine, each parenthesis is counted separately. Table 8.1 shows two parentheses ranked at 10 and 11 in BAWE, with raw frequencies of 91,843 (opening) and 90,538 (closing), respectively. Based on SketchEngine, the difference in the frequency count is attributable to cases where the parenthesis is used to mark items in a list—such as (iv) or 2)—rather than serving purely as parenthetical punctuation. This study uses the raw frequency of 91,843 as the enlistment frequency for normalizing this punctuation mark in BAWE. In COMP 101, the opening parenthesis mark

appears 364 times, while the closing parenthesis mark appears 361 times. In the filtered punctuation results of the raw frequencies in Table 8.2, parenthesis rank at position 5 in COMP 101 and at position 3 in BAWE.

Based on the examination of the 100 randomized concordance lines in COMP 101 and BAWE, the three primary uses of parenthesis in the texts are to indicate references and additional information and provide clarification of abbreviations related to long titles. Figure 8.14 summarizes the three main functions as they pertain to each corpus.

Figure 8.14: Summary of the parenthesis' usage in COMP & BAWE



Both groups of writers use parentheses in references at roughly the same rate, as shown in Figure 8.14. This finding reflects the proportional distribution of the parentheses used in references based on the 100 randomized lines from each corpus, rather than the overall frequency across the two entire corpora. However, they differ in using parentheses to provide additional information and abbreviations—two differences discussed in this section.

The proportional use of parenthesis for additional information in the sampled lines is almost twice in COMP 101 than in BAWE, suggesting that first-year composition writers tend to insert

more information within the parenthesis rather than integrating or omitting the information from the sentence. This difference may be influenced by audience and task considerations. or example, COMP 101 writers typically create texts for instructors or peers who may not be experts in the subject matter. This audience may motivate first-year writers to use parentheses to clarify concepts or provide additional context for general readers. In contrast, BAWE writers usually write for subject matter experts in their respective disciplines. They often integrate information directly into their sentences or may omit it completely to align with the expectations of academic tasks. The examples in (8.19) show the typical patterns related to inserting additional information in sentences by COMP 101 writers.

8.19 COMP 101 (Additional information)

> **Compare-and-contrast Subcorpus:** *Some favorites were Santa Claus, Snow Buddies, Halloween Town, the Night Before Christmas (played during both Halloween and Christmas, thankfully.)*
>
> **Process-analysis Subcorpus:** *To extract honey you will need a few tools: a handheld smoker, jacket, veil, gloves, a hive tool (small crowbar like device) and a horsehair bee brush.*
>
> **Classification Subcorpus:** *According to the characteristics of its music, Mongolian folk songs are divided into two types: "Urdu" songs (long-toned songs) and "Wu Huer" songs (short-key songs).*

In the first sentence in (8.19), the writer inserted information in the parenthesis that could have been incorporated in the sentence and provided context to the titles of the favorite movies. Also, the writer could have expressed the attitude of gratitude, expressed in the word "thankfully" for being able to watch the movies during the listed events and further personalize the experience. The second two examples illustrate the usage of additional information enclosed by the parenthesis that clarifies the terminology used by the writers. Since the terms in quotation marks

are foreign words the information included by the parenthesis is essential for the reader to understand the meaning of the terms and could have been included in the sentence. The tendency to include additional or explanatory details in parentheses may be based on COMP 101 writers considering the audience's need for more context, unlike BAWE writers, who may omit details because they address expert readers.

The writers in the BAWE use parentheses to enclose additional information in their writing, but they do so only one-third of the time compared to the writers in COMP 101. This might indicate that BAWE writers are using footnotes instead. This study, however, focuses only on using the most frequently used punctuation marks and does not analyze the functions of footnotes, leaving room for future research in this area. Examples in (8.20) show how BAWE writers use parentheses to include additional information in a sentence.

8.20 BAWE (Additional information)

> *Therefore, highly dense material (such as dense lamella containing many crystals) is not required, saving the body on resources.*
>
> *Furthermore, the exercises (mainly matching of forms, recombination and 'fill in the gaps') are designed to present and practice single linguistic rules, reflecting the traditional approach of synthetic syllabi.*
>
> *The car was 4m (13.32 ft) long, 1.6m (5.08 ft) wide and 1.5m (4.92 ft) high.*

In this example, the additional information in the first sentence could have been included in the sentence, but instead, the writer chose to include the information in parentheses. The second sentence shows additional information clarifying the preceding word "exercises," most probably with the intent of being concise. The last sentence shows a standard use of parentheses to add extra information about an item, in this case, the Imperial type of measurement to show the equivalent to the metric system. Both groups are familiar with the function of parenthesis to

indicate additional information, and one area that both groups may need to improve is recognizing when the additional information might be included in the sentence and not left as an inserted element. In the BAWE corpus, the writers use parentheses sparingly and selectively since their audience is expected to have specialized knowledge in the discussed topic, reducing the need for explanatory information.

Another main difference in the use of parenthesis between COMP 101 and BAWE is in the use of abbreviations in the text. In the 100 lines examined in each corpus, entry-level writers used abbreviations within parentheses only 3 percent of the time, while upper-level writers used them 7 percent of the time. In COMP 101, writers use abbreviations in two different ways, as illustrated by example (8.21). They spell out the complete term and follow it by its abbreviation or use the abbreviation first and then provide the complete term. The second example does not comply with the standard practice that requires the complete terms or forms and then including their abbreviations, which helps the reader understand the meaning of the shortened form (Lester, 2018). Although the COMP 101 task instructions and curriculum do not emphasize the use of parentheses beyond their standard role in in-text citations and bibliographies, student writing demonstrates an emerging awareness of their broader function—particularly for adding additional information. While parentheses are not always used correctly in this context, their presence suggests that first-year writers are beginning to understand their purpose, marking a positive step in their writing development.

8.21 COMP 101 (Abbreviation)

> **Argumentative Subcorpus:** *These are some of the philosophies of those who are in favor of the practice of* ***genetically modified organisms (GMO).***
>
> **Cause-and-effect Subcorpus:** *With* ***ESA's (Emotional Support Animals)*** *they know what's going on.*

Although the abbreviation occurrences are limited in COMP 101, the examples above show that students will benefit of focused instruction on the standard use of abbreviations. On the other hand, the writers in the BAWE corpus exhibit consistent usage of abbreviations, as shown in the example (8.22). They introduce the term first and then enclose its abbreviation in parentheses.

8.22 BAWE (Abbreviation)

> *The Commission on Sustainable Development (CSD) was also established in Rio and its purpose*
>
> *The sequences of the 70kDa heat shock protein 5 (HSP70).*
>
> *Aminopeptidase N (CD13) Is a Molecular Target of the Cholesterol*

The examples in BAWE show the standard use of the complete terms followed by their abbreviations and demonstrate consistency in this type of usage.

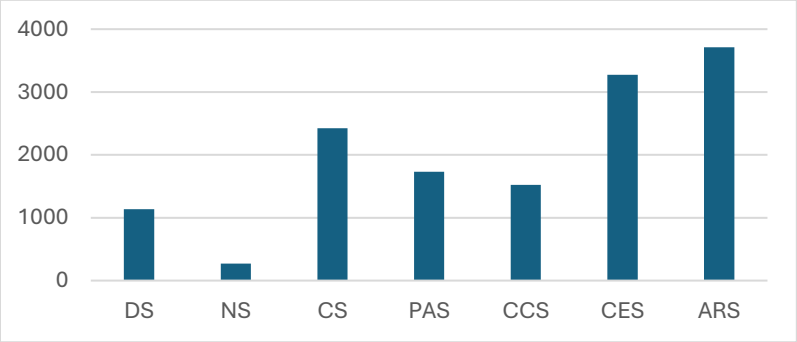
Based on the patterns demonstrated by the two groups in the use of parentheses, it can be observed that both novice and advanced writers tend to use them to present additional information that could have been integrated into the sentence. Instead, they choose to enclose it in parentheses. This may suggest that both groups could benefit from learning how to evaluate information and decide what to leave out, integrate, or include as additional information.

8.3.5.1 Parenthesis in the subcorpora

The distribution of parentheses across the subcorpora based on the normalized frequency is displayed in Figure 8.15. The bar graph shows a right-skewed distribution, corresponding with

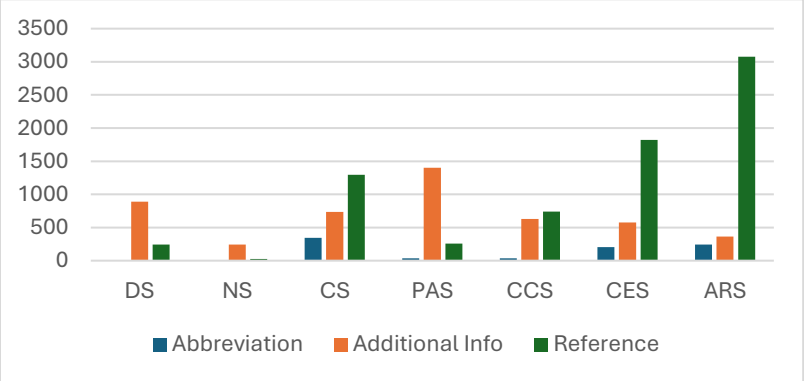
analytical writing and the latter part of the semester. In narrative texts, the usage of parentheses is minimal, while in cause-and-effect and argumentative writing, it reaches its peak.

Figure 8.15: Distribution of the parenthesis across the subcorpora



The study also presents a summary of the three main functions related to the use of parenthesis in the texts across the subcorpora in Figure 8.16. The marks facilitate additional information at the beginning of the semester and are later used predominantly in references.

Figure 8.16: Functions of the parenthesis across subcorpora



Two main trends are observed in Figure 8.16: the use of parentheses for enclosing additional information and for identifying research. The purpose of using parentheses in descriptive and narrative text is to provide supplementary information. The challenge lies in making the right decision about when to enclose the additional details within the parentheses without overusing

them. It is interesting to observe that the additional information peaks in the process analysis writing (PAS) as students describe processes or provide instructions. In a way, the parenthesis seems to become a feature of process writing in COMP 101.

The second significant function of parentheses is their use in references, which is clearly indicated by the positive trend of references in classification (CS), process analysis (PAS), cause-and-effect (CES), and argumentative texts (ARS). Similar to quotation marks, which signal the use of outside sources and show increased frequency in the ARS subcorpus, parentheses are also commonly used in ARS references, aligning with task expectations. However, their appearance in other genres—such as classification, comparison-and-contrast, and cause-and-effect essays—where references are not required, reflects a positive trend in students' developing awareness of academic conventions. This trend suggests a growing understanding among students of the importance of outside sources in academic writing for discussing various topics and developing positions. Based on the distribution of parentheses across the subcorpora, it is evident that towards the end of the semester, when students engage in writing argumentative texts, the use of references increases significantly. In this regard, COMP 101 writers begin to emulate the characteristics of upper-level writers in BAWE, who incorporate a substantial number of references not only based on their use of parentheses, but also their use of commas, full stops, and quotation marks.

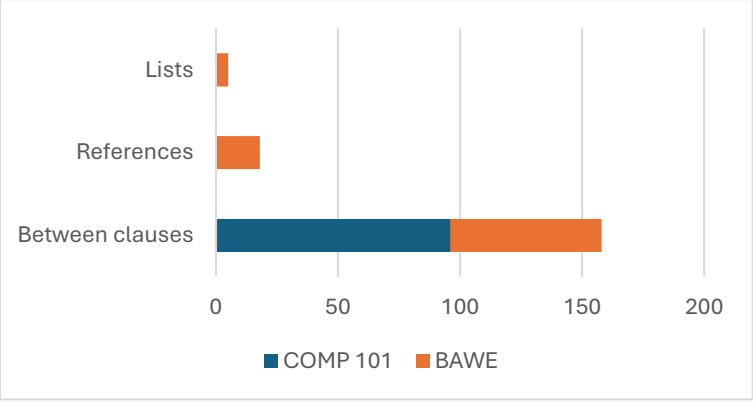
8.3.6 Semicolon

Carter and McCarthy (2006) discuss the semicolon as the punctuation mark used to separate items in a list sequence or in place of full stops to set off main clauses. The semicolon seems a convenient replacement for a full stop in cases where the two clauses are closely related, and the second clause explains the first clause. Lester's (2018) recommendation for correctly using

semicolons between clauses is that writers should use them only if they can be replaced by a full stop to create two separate sentences and caution writers to use them sparingly. The principle behind this economic use is the same principle of the close connection between clauses that Carter and McCarthy (2006) discuss. In COMP 101, the use of the semicolon is not explicitly addressed in the task instructions, but it falls under broader expectations for correct sentence structure and grammar. This absence of explicit direct instruction makes the semicolon occurrences interesting to observe and discuss.

In the filtered results featuring only the punctuation marks in Table 8.2, the semicolon is ranked at position 6 in COMP 101 (raw frequency 354 and normalized frequency 1875) and position 8 in BAWE (raw frequency 16685 and normalized frequency 2394). Figure 8.17 summarizes the observed occurrences of semicolon usage in COMP 101 and BAWE. Again, the occurrences are based on the 100 randomized concordance lines examination in the two corpora and show the main use patterns between the two writers' groups.

Figure 8.17: Summary of the semicolon usage in COMP & BAWE



It is evident from the figure that both groups use the semicolon between clauses, which is one of its primary functions (Carter and McCarthy2006; Lester 2018). Comparing the semicolon usage

between clauses shows that COMP 101 writers use the mark approximately 20 percent more than the BAWE ones. The significant differences are in the semicolon usage in lists and references that characterize only the BAWE writers, a trend observed with other punctuation marks (Sections 8.3.1, 8.3.2, 8.3.3, and 8.3.5). Another interesting observation about the semicolon usage in both groups relates to misconception or incorrect use. It is important to note that this study does not specifically examine errors in punctuation marks. However, the study found patterns of incorrect usage in both groups of writers in the concordance lines. These patterns were related to a confusion between the semicolon with the colon and are discussed later in this section.

Using semicolons between the clauses in (8.23) demonstrates correct usage in connecting two related clauses. In both examples, the independent clauses are closely related, and the second clause explains the preceding one, which is the recommended use by style guides (Carter and McCarthy 2006; Lester 2018).

8.23 COMP 101 (Between clauses)

> **Process-analysis Subcorpus:** *The latter description would perfectly apply to the Brazilian hot* ***chocolate; it*** *is rich and velvety.*
> **Classification Subcorpus:** *This year is* ***critical; every*** *vote is needed.*

In other instances, COMP 101 writers do not demonstrate correct usage of the semicolon, as in example (8.24), where the writer moves from one step of the process (e.g. “you have to take some dough…”) to another (e.g. “the balls can be the size …”) using a semicolon between two clauses that are not closely related but show two different sides of the process.

8.24 COMP 101 (Incorrect usage between clauses)

> **Process-analysis Subcorpus:** *You have to take some dough and make a* ***ball; the*** *balls can be the size of the palm, the ball is opened and filled with cheese and beans, or with cheese and pork rinds*.

Such cases show overreliance on semicolons, which is an incorrect use of the punctuation mark, and suggests the need for classroom instruction focused on appropriate semicolon usage.

The BAWE corpus also demonstrates a considerable use of semicolons between clauses and instances where a full stop should replace the semicolon. For example, in (8.25), the semicolon should be replaced by a full stop since the second clause adds to the previous concept with a different detail.

8.25 BAWE (Incorrect usage between clauses)

> *Koleman Strumpf shows that piracy has "no statistically significant effect"* ***on CD sales; Aram Sinnreich*** *even claims that downloads boost sales by acting as complementary goods to CDs (indeed, while downloading led 65% of file-sharers in Strumpf's study to not purchase an album, 80% bought an album after online sampling).*

Even though the two clauses in (8.25) are related, they address two aspects of the topic—"CDs"—and should be expressed in two separate sentences. Such instances in the upper level of writing suggest that this group of writers may also benefit from targeted instruction on the appropriate usage of the semicolon.

Another pattern of semicolon usage is seen in lists and references, which is only observed in the BAWE corpus. This pattern shows an advanced referencing skill in the use of the semicolon in references and lists in the example (8.26).

8.26 BAWE (References & Lists)

> *(Raby 2005:155/6**;** Cohen P 1972:91**;** Hodkinson 2004:137)* **(References)**
>
> *M.D.A. Freeman, 'Lloyd's Introduction to Jurisprudence', 2001, 7th ed., p.1465**;** 'Theories of Adjudication', R. Dworki" </s><s> Dworkin, 'Law's Empire' (1986)* **(References)**
>
> *gl.glColor3d (0.3, 0.7, 0.5) hexagon**;** hexagon gl.glColor3d (0.7, 0.9, 0.4) hexagon**;** gl.glColor3d (0.1, 0.3, 0.4) hexagon**;** gl.glColor3d (0.9, 0.4, 0.1) hexagon**;** hexagon Figure 29 - Hexagons with color* **(Lists)**

Both the references and the list show complex uses of semicolons. This is evident in the extended referential structures, which demonstrate a variety of evidence competency, and the compression of technical abbreviations in lists separated by semicolons.

The last observed pattern of semicolon use suggests that writers may confuse semicolons with colons, indicating a possible area of future study regarding writers' use of the semicolon. In example (8.27), an entry-level writer uses a semicolon instead of a colon to introduce an explanation of the quoted numbers related to the educational system.

8.27 COMP 101 (Semicolon in place of a colon)

> **Compare-and-contrast Subcorpus:** *The education system in Kenya uses the 3-8-4-4 system**;** three years in pre-school, eight years in primary school, four years in high school, and four years in college*.

This confusion is also evident in the BAWE texts, as shown in example (8.28). In both cases, a colon is needed to introduce the listed items instead of a semicolon.

8.28 BAWE (Semicolon in place of a colon)

> *The recent position statements include the following issues**;** avian influenza, environmentally sustainable transport and forestry, access and countryside recreation, renewable energy, and climate change.*

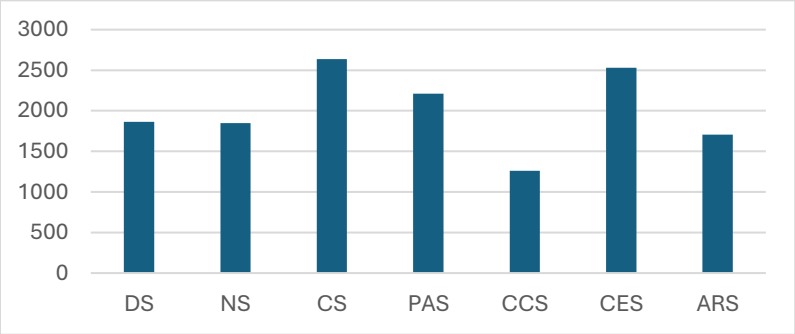
The death of the badger should provide some sort of poetic resolution; the poem's climax and subsequent end.

Inconsistent use of semicolons may indicate the need for further research to determine whether writers are also confused about colon usage and comma placement.

8.3.6.1 Semicolon in the subcorpora

The distribution of this mark across the subcorpora is illustrated in Figure 8.18. It resembles a normal distribution and is similar to the distribution of full stops across the subcorpora, listed in Figure 8.7 in Section 8.3.2. One of the reasons for the high frequency of the semicolon across the subcorpora is that students often use the semicolon in place of a full stop, a tendency frowned upon by grammarians (Lester 2018).

Figure 8. 18: Distribution of the semicolons across the subcorpora



The study found no additional uses for semicolons beyond their use between clauses in the subcorpora. Example (8.29) shows two instances of the semicolon between clauses in the narrative (NS) and compare-and-contrast (CCS) subcorpora.

8.29 Semicolon in the subcorpora

> **Narrative Subcorpus:** *Billy did not always work for the CIA**;** before he started working at Techno, he used to go to Harvard this is where he studied how to fix computers, but he got framed by his best friend for something he didn't do and was expelled*.
> **Compare-and-contrast subcorpora:** *The level of play between the NBA and WNBA is drastically different**;** the stats in the NBA are at least double the stats in the WNBA and the talent is only increasing in the NBA.*

In both of the above examples, the semicolon should be replaced by a full stop. This use resembles the previously listed examples of incorrect semicolon usage (e.g., 8.24 and 8.25) and shows students' inclination to use the semicolon in place of the full stop. In the argumentative texts in example (8.30), writers show the appropriate use of the mark.

8.30 Argumentative Subcorpus

> *Another pro is they can be produced faster**; thus,** quickly getting food to those who need it.*
> *For some, they may use both streaming devices**; however,** only willing to pay for one while borrowing a password for the other*.

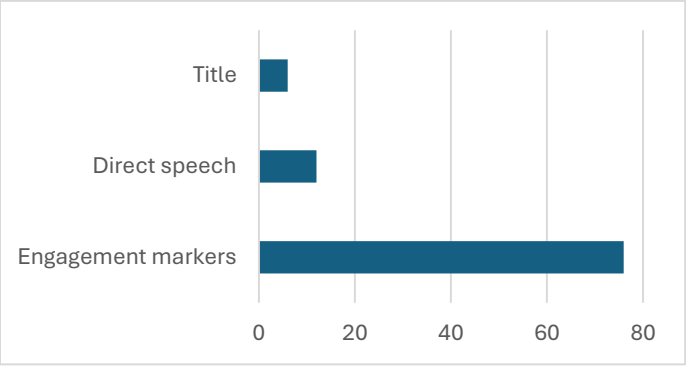
The semicolon is used in (8.29) to link ideas and create coherence by using a subordinate conjunction after the semicolon. While this is a rare occurrence, it demonstrates one of the appropriate uses of the mark. It is difficult to prescribe semicolon usage to a specific essay genre, as it is utilized across all types of texts. However, the presence of incorrect usage highlights the need for further guidance and instruction. This incorrect usage is observed not only among first-year students but also in upper-level student writing, as seen in the BAWE corpus. These instances indicate that students at both levels would benefit from targeted instruction on understanding when to use semicolons between closely related clauses and when it would be preferable to use full stops or to rewrite the sentence. Similar to other punctuation marks, semicolons should be used appropriately and sparingly to effectively connect sentences.

8.3.7 Question mark

The question mark is used for interrogative sentences and can also be placed in declarative sentences to indicate that they should be read as questions (Carter and McCarthy 2006). This punctuation mark is most common in social media texts that demonstrate the interaction between people and casual conversations but are rare in serious topics such as those in academic writing (Sun and Wang 2019). Questions are also used as engagement markers in academic blogs, seeking reader's responses and participation in the topic (Mur-Dueñas 2021). The task instructions in COMP 101 do not specifically address the use of question marks but only require correct punctuation, which leaves room for students to interpret the use of question marks based on their overall understanding and preference.

In COMP 101, this mark ranks 7th with a raw frequency count of 296 and a normalized count of 1568 (see Table 8.2), but it does not appear on BAWE's most frequently used punctuation marks. Figure 8.19 summarizes the use of question marks in COMP 101. They are primarily used as indicators of engagement and then in direct speech, as well as in titles.

Figure 8.19: Summary of the question marks' usage in COMP



Question marks are predominantly used to engage the reader in the text and suggest that in entry-level writing, students use questions to show engagement and conversation with the reader.

According to Sun and Wang (2019), questions often appear in social media, and it is reasonable to assume that first-year students participate to some degree in social media, which may impact their expressions in academic writing. They may not realize that using questions in academic writing is not suitable. For example, (8.31) shows some common questions in COMP 101 that are meant to prompt the reader to think of possible answers.

8.31 COMP 101

> **Narrative Subcorpus:** ***Will I*** *forever be known as the girl who killed her cat?*
> **Process-analysis Subcorpus:** ***What is the most magical*** *place on earth? </s> <s>* ***Is it your*** *backyard?*
> **Compare-and-contrast Subcorpus:** *Now,* ***why would going to*** *sleep after a movie matter?*

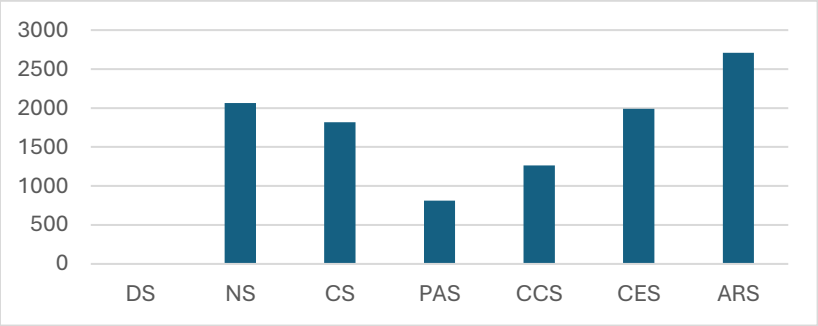
This use of the question mark is similar to how it is used in a direct conversation with someone. For instance, in example (8.31), the word "now" is utilized as a discourse marker to help transition to the next point, like in speech, and encourage the reader to reflect on the subject by posing a question. In academic writing, prompting the reader to think is usually demonstrated through the exposition of concepts, data, and carefully constructed arguments.

Discussing engagement in academic writing, Hyland (2001a) highlights the effectiveness of using question marks to engage the reader in academic writing, particularly in journal articles. However, first-year students are still responsible for understanding the proper role of questions as engagement markers in academic writing. It is crucial for students to use question marks to capture the reader's interest, rather than as a substitute for providing evidence and detailed explanations.

8.3.7.1 Question marks in the subcorpora

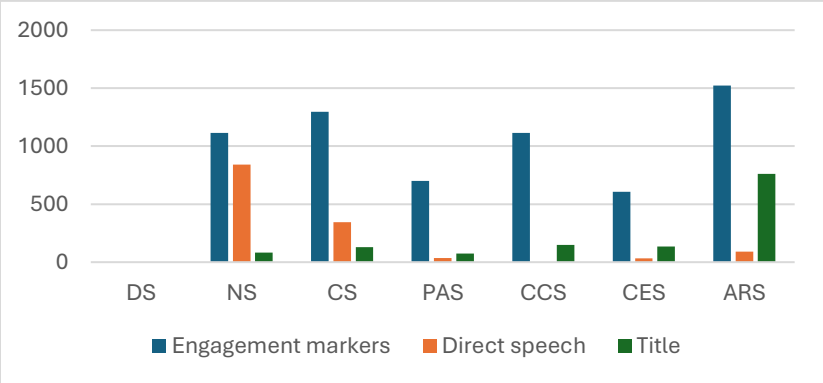
Figure 8.20 compares the normalized frequency of the question mark usage across the subcorpora.

Figure 8.20: Distribution of the question marks across the subcorpora



Based on Figure 8.20, every subcorpus in COMP 101 except the descriptive uses the question mark in some sort of capacity. The study examines the concordance lines in each of these sub-corpora and categorizes them into three categories: engagement markers, direct speech, and title. These functions include engagement markers, direct speech and titles. Figure 8.21 provides a summary of the subcorpora functions.

Figure 8.21: Functions of the question mark across subcorpora



It is not surprising to find the use of question marks in direct speech within narrative texts, as they often depict conversations between characters. Another frequent use of question marks is in titles, which are mostly present in argumentative texts. For example, in (8.32), the titles are meant to catch the reader's attention, focus on the topic, and invite the reader to the debate.

8.32 COMP 101

> **Argumentative Subcorpus:** *Should Euthanasia be Legal?*
> **Argumentative Subcorpus:** *Are Vaccinations Beneficial or Harmful?*
> **Argumentative Subcorpus:** *Should college athletes be paid for being on sports teams?*

Using questions in titles is not an unusual practice in academic writing since journal articles also use question marks in their titles (Zijlmans *et al.* 2015; Murphy *et al.* 2019), which implies that titles posing questions are also suitable for first-year composition students. The surprising use of the question mark is in its frequent role as an engagement marker, observed across different types of texts except for descriptive texts (DS). The high normalized frequencies are found in classification (CS), compare-and-contrast (CCS), and argumentative (ARS) texts. This indicates that students use questions throughout the semester as they work on analytical and argumentative writing, not just narrative (NS). Questions can potentially engage the reader, as Hyland (2001) notes in his analysis of journal articles. However, at this stage, students most likely do not intentionally recognize these features of the question mark and use it in its conversational capacity.

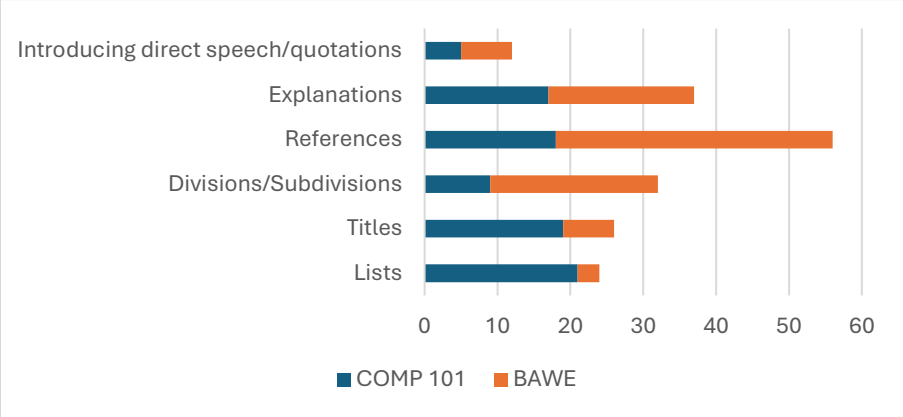
8.3.8 The colon

The primary function of the colon is to introduce lists, sub-titles, sub-divisions in texts, references, and quoted speech (Carter and McCarthy 2006; McCarthy 2017). In some cases, colons might amplify a phrase or explain an idea. This latter use is not as common and might be

challenging for some writers (Lester 2018). In COMP 101 task instructions, the colon, like other punctuation marks, is not addressed individually but falls under the broader category of correct punctuation.

The colon is at position 8 in COMP 101 with 244 raw frequency and 1293 normalized frequency. In BAWE, it is at position 6 with 47060 raw frequency and 6754 normalized frequency (see Table 8.2). Figure 8.22 summarizes the colon usage in the two corpora based on the randomized concordance lines.

Figure 8.22: Summary of the colon usage in COMP & BAWE



Both corpora show multiple uses and cover all described functions, such as lists, subdivisions in topics and titles, references, quoted speech, amplifying, and explaining (Carter and Macarthy 2006; Lester 2018).

The two groups of writers differ in their usage of the colon mainly in four categories: references, subdivisions, titles, and lists. The writers of COMP 101 tend to use the colon more in constructing lists and titles, whereas the writers of BAWE favor the mark in references and subdivisions. Example (8.33) shows the typical use of colons in lists and titles by the COMP 101writers.

8.33 COMP 101

> **Process-analysis Subcorpus:** *To get to the root of the problem, it is crucial to understand the stages of a bad habit: cue, routine, and reward*. **(Lists)**
> **Cause-and-effect Subcorpus:** *Divorce: The Positive and Negative Effects on Children* **(Titles)**

At this stage, first-year composition students feel comfortable with lists. The lists help the writers to focus on particular details within a given topic and itemize the essential parts. Also, the use of the colon in titles shows a good grasp of using the mark to highlight certain aspects and focus their texts.

In BAWE, the colon is predominantly used in references and identifying subdivisions. The use of references indicates student engagement with research, a pattern also observed with the comma, full stop, quotation marks, parenthesis, and semicolon (Sections 8.3.1, 8.3.2, 8.3.3, 8.3.5, 8.3.6). Example (8.34) shows some of the typical uses of the colon to reference literature.

8.34 BAWE (References)

> *(Gellerman 1965: 168)*
> *'Canada's Money Supply'/Publications and Research. Available from: [Accessed 15th Feb 2007] Bank of Canada. (2006).*
> *Vol.8, No.2: pp.175-196*

Another common use of the colon in BAWE is marking divisions or subdivisions in table or figure entries, explaining the data, and separating the item's name with its description, as example (8.35) illustrates.

8.35 BAWE (Divisions)

> *Figure 2.3: 16-QAM Argand diagram64-QAM (6 bits per symbol)*
> *Table 3. Stature: The stature of individuals depends strongly on the environmental conditions, geographical location, socio-economic status and genetic features*.

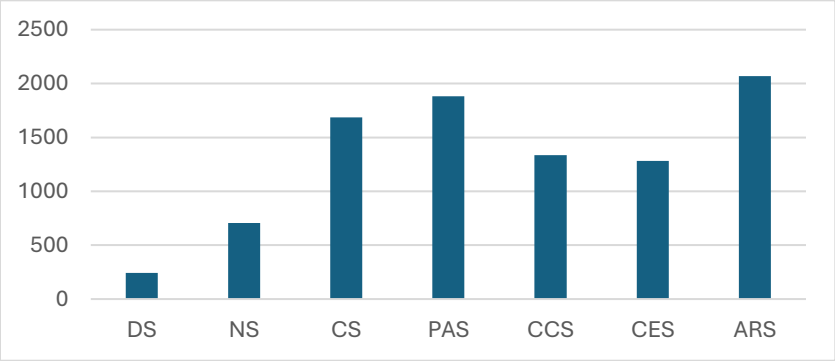
Such tables and figures commonly appear in research and academic writing to display data, which suggests that the use of colons in BAWE represents a higher level of skill in integrating data as illustration in the text. This higher-level usage is not intended as a critique of the absence of tables and figures in COMP 101, but rather as an observation specific to the BAWE writers. The COMP 101 tasks do not require the inclusion of tables or figures, which aligns with the findings that show no visual elements.While this skill is typically beyond the scope of first-year writing, it can be advantageous for students to investigate the role of the colon in such contexts as they delve deeper into their chosen fields of study.

In general, the use of colons by these two groups of writers highlights a clear difference in writing expectations between those observed in BAWE, where data integration and research are more evident in the text and their role in itemizing text elements and highlighting given aspects in titles in entry-level writing in COMP 101.

8.3.8.1 Colon in the subcorpora

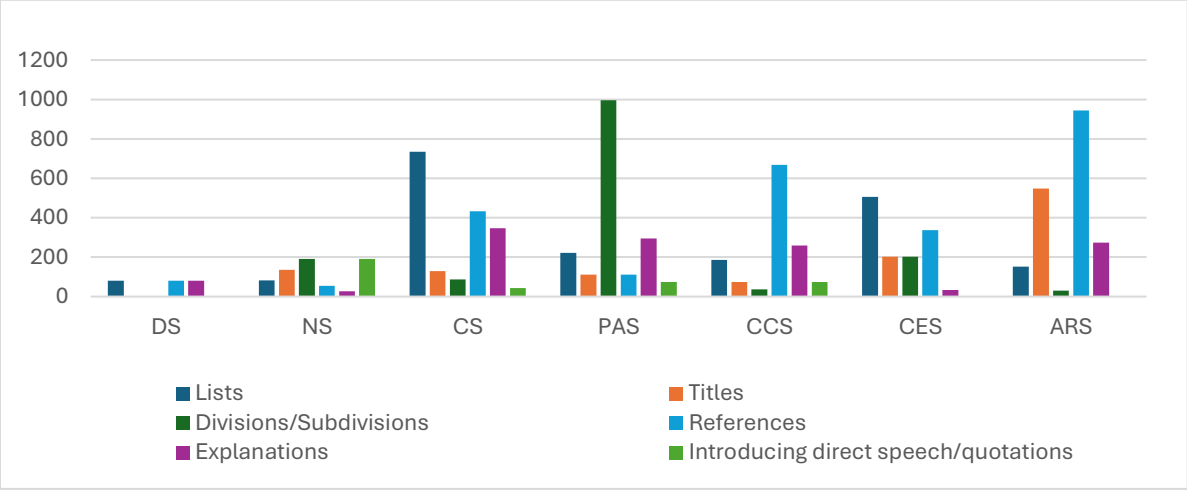
Figure 8.23 displays the distribution based on the normalized frequencies. Similar to the distribution of parenthesis, the colon has a somewhat skewed distribution to the right, having most of its occurrences in the second part of the semester, more precisely in the analytical texts, such as classification, process analysis, and argumentative.

Figure 8.23: Distribution of the colons across the subcorpora



The colon is used less frequently in descriptive and narrative texts. One possible reason is that it may not be introduced until later in the semester. Additionally, as discussed in Section 8.3.6, the colon is sometimes confused with the semicolon, resulting in the overuse of the semicolon and affecting the frequency of the colon. The colon, like the comma, serves various functions throughout the subcorpora. Figure 8.24 provides a summary of the functions found in each subcorpus.

Figure 8.24: Functions of the colon across subcorpora



Based on the bar chart, the classification texts (CS) have the highest number of lists related to the use of colons. This might indicate some of the genre features in the classification texts, where identifying items in series is part of the categorization scheme. On the other hand, the process-analysis texts (PAS) exhibit the highest frequency of functions related to division or identifying steps, a distinctive feature of the process-analysis genre. The argumentative texts (ARS) use colons mostly in references, which shows the positive tendency of bringing-in research in constructing debates and positions, which also reflects the positive role of the task expectations and students diligence in meeting them. A similar trend with the first-year composition writers is observed in the use of commas (Section 8.3.1.1) and parentheses (as discussed in Section 8.3.5.1).

In summary, the examination of the colon usage across various types of writing suggests that it might be beneficial to introduce this topic earlier in the semester and help students understand the difference between the use of the colon and the semicolon. Another helpful insight of this examination is its ability to highlight the characteristics of text genres. For example, classification texts tend to have a high frequency of lists, while process texts often include step-by-step instructions. In argumentative texts, there may be more references to outside sources. Knowing the most common functions of the colon in different genres can help instructors target these uses in such texts and explore appropriate applications.

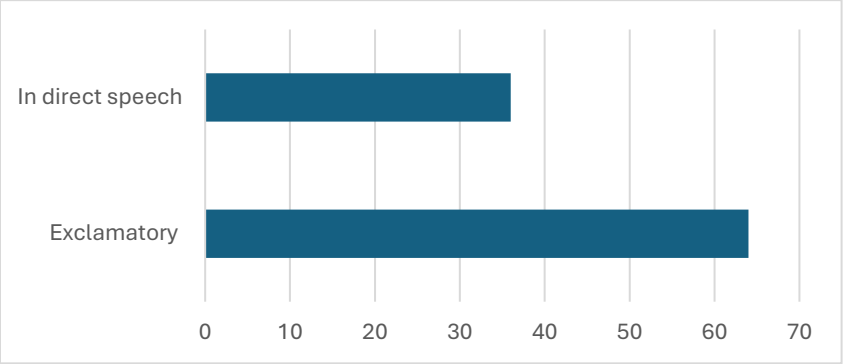
8.3.9 Exclamation mark

The exclamation mark is used in exclamatory sentences to demonstrate emphasis or communicate strong emotion or surprise (Norhaida and Tan 2018). According to McCarthy (2017), there is a trend of using exclamation marks excessively just to impress others, which is

not recommended. Fowler and Aaron (2016) also discourage the overuse of this punctuation mark in academic writing for the reason that it can dilute the meaning and make writing appear overly emotional, likening it to “crying wolf”. On the other hand, exclamation marks are frequently used in fiction and social media in their capacity to convey emotions and surprise, but not common in academic writing (Sun and Wang 2019). The exclamation mark, like other punctuation marks, falls under the broader expectation of correct usage as outlined in the task instructions in COMP 101.

The exclamation mark is ranked at position 8 in COMP 101 with a raw frequency count of 130 and a normalized frequency 689. It is the last mark in COMP 101 that meets the criterion of a cut-off score of 100 or higher, as shown in Table 8.2. The exclamation mark only appears frequently in entry-level students' writing and not in upper-level student writing, such as in BAWE. Figure 8.25 summarizes the two main uses of the exclamation mark observed in the examined 100 randomized lines in COMP 101. The two main uses of the exclamation mark are in direct speech as part of narrative accounts and in exclamatory sentences that aim to appeal to readers’ emotions and grab their attention.

Figure 8.25: Summary of the exclamation marks’ usage in COMP



Out of the 100 lines, 36 percent of the exclamation marks are used in direct speech, and the remaining 64 percent of the marks are used in exclamatory sentences. Example (8.36) illustrates how the writers convey intense emotions through their storylines in specific moments.

8.36 COMP 101

> **Narrative Subcorpus:** *Ralph getting somewhat agitated at his friend for never going anywhere "Come on man it has been way too long**!** You need to get out of your house for once**!**"*
> **Classification Subcorpus:** *These customers can be heard saying, "I know right, that plate is heavy! Haha**!**"*
> **Narrative Subcorpus:** I yelled **"NAKIMA! NAKIMA!"**

The second use of the exclamation mark is even emphasized through the short expression of amusement, "Haha!" to intensify the writer's feelings about the particular type of customers being described. Also, the use of all capital letters in "NAKIMA!" can relate to the writer's shocking expression in the given moment. All the examples in (8.34) show the exclamation mark used in direct speech to intensify the moment and the specific feeling the writer is experiencing. Even though the exclamation mark is not typical in academic writing, its use within the context of relating direct speech is understandable aligns with narrative essay expectations that include dialogue as a potential story feature.

Example (8.37) shows the second type of usage. This type shows writers' attitudes that could be expressed in elaborate declarative sentences, but instead, writers use the exclamation mark to emphasize points, shocking experiences, or daily routines.

8.37 COMP 101

> **Narrative Subcorpus:** *Most days I woke up early and went to bed late. A very important thing to know about me is I am not a morning person**!***

Narrative Subcorpus: *I open the machine and with a loud meeeoooowww my cat falls out! My cat was in the dryer**!** I repeat my cat was in the dryer**!***
Process-analysis Subcorpus: *After adding the leave-in conditioner, the sections should be completely combed out to make the second part a little more manageable. This process usually takes 1-2 hours. After you've successfully washed your hair, now it's time to moisturize**!***

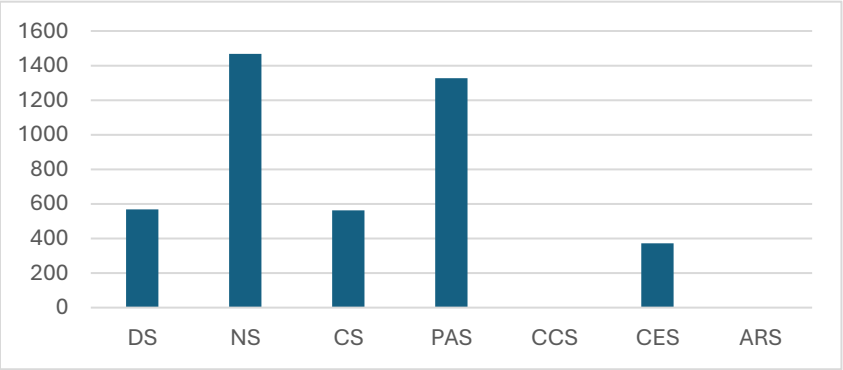
The examples in (8.37) demonstrate a conversational use of the mark, which is more typical in social-media discourse, as observed by Sun and Wang (2019). The exclamation marks are used two times more in similar exclamatory sentences, which are outside of direct speech and not suitable in the texts. Such use seems repetitive and overly appealing to the reader. In the words of Fowler and Aron (2016), it produces a style of "crying wolf" occurrences.

One possible interpretation of this excessive use of exclamation marks in writing by first-year composition students can be attributed to the increasing acceptance of informal language in social media and public spaces. At the same time, this trend highlights the importance of educators addressing emotional expressions in academic writing. Educators can guide students to understand that while exclamation marks convey emotions, they can be substituted with detailed descriptions of experiences. Encouraging students to avoid exclamation marks and explain their emotions can expand their vocabulary and writing skills, enabling them to transition from novice, more-social-media-oriented writers to articulate and proficient academic writers.

8.3.9.1 Exclamation mark in the subcorpora

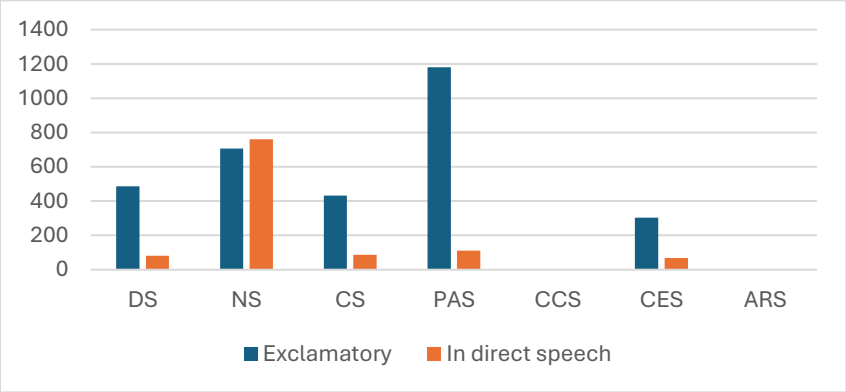
Figure 8.26 shows the distribution of exclamation marks across the subcorpora. It is only used in five subcorpora: descriptive, narrative, classification, process, and cause-and-effect, with no occurrences in the argumentative and compare-and-contrast subcorpora.

Figure 8.26: Distribution of the exclamation marks across the subcorpora



In the narrative texts, the mark is used mainly in direct speech, but in the other subcorpora, such as DS, CS, PAS, and CES, it is used in its exclamatory capacity to relate the writer's feelings about the given topic, which is a suitable use given the narrative essay genre. Figure 8.27 shows the two main ways the exclamation marks are distributed across the subcorpora: exclamatory sentences and direct speech.

Figure 8.27: Distribution of the exclamation marks across the subcorpora



The mode of the exclamation mark as an exclamatory indicator is in the process-analysis texts. The writers show their strong attitudes related to various processes, such as cooking, packing for trips, or decorating a Christmas tree, as illustrated in example (8.38).

(8.38) Process-analysis Subcorpus

> *Refrigerate any remaining cookie dough you have not baked. Enjoy**!***
> *It would not be till the next day that we unpack the car and unpack our stuff to be used for another time. Disneyland is magical**!***
> *In the beginning, when your first picked up this paper, you had no clue how to build and decorate a Christmas tree but not that you're finished reading. You now know how to build one**!***

One potential reason behind the extensive use of exclamation marks by novice writers is that they may frequently engage with how-to procedures, often found in social media. For entry-level writers, such processes may be easier to communicate through more informal features, such as exclamation marks, rather than fully expressing the ideas in words. In contrast to the process-analysis texts, the absence of exclamation marks in argumentative texts indicates that writers distance themselves from their use and focus more on the articulation of the ideas.

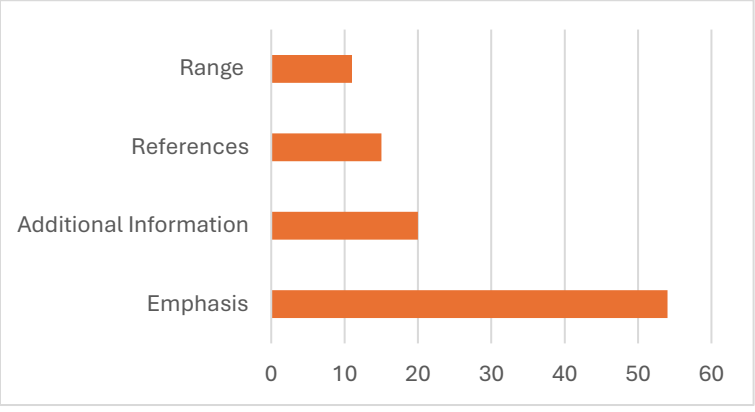
The absence of exclamation marks in argumentative and compare-and-contrast subcorpora likely reflects the genres' emphasis on objectivity and logical reasoning over emotional expression. In argumentative texts, novice writers may avoid exclamation marks because they are focused on building and defending their argument rather than displaying strong emotions by using the exclamation mark. Similarly, compare-and-contrast texts require analysis of similarities and differences, prompting writers to articulate ideas through their words rather than expressive punctuation. This pattern, observed in the COMP 101 data, suggests that novice writers adapt to the objectivity required by these genres, possibly due to an awareness of formal conventions or challenges in integrating emotional emphasis within these specific analytical frameworks.

8.3.10 Dash

Dashes can be used in three ways to emphasize or clarify words: providing additional information, indicating a range between numbers, and expressing emphasis. They are typically placed at the end of the sentence when used for emphasis. Conversely, when used for explanations, they surround the additional information on both sides (Lester 2018). Regarding their frequency status, they do not appear as one of the most common punctuation marks in Sun and Wang's (2019) study. Contrary to the dash, the hyphen is listed as one of the most frequently used punctuation marks by Sun and Wang (2019). In this study, however, only the dash makes it to the filtered list of the most common punctuation marks, but only in one of the corpora—BAWE, which seems to suggest that novice writers are largely unfamiliar with the dash uses.

After examining 100 randomly selected concordance lines in BAWE and considering the standard use of the dash, the study summarizes the main patterns in Figure 8.28. The most common usage of the dash is for emphasis, followed by providing additional information, occasional appearance in references, and identifying a range.

Figure 8.28: Summary of the dash usage in BAWE



Example (8.39) illustrates the four different patterns observed in the texts.

(8.39) BAWE

> *The symptoms of depression have been identified in the history so far, stated that he likes to splash out his money on his friends* **-** *this could indicate a degree of disinhibited behavior indicative of hypomania***. (Emphasis)**
>
> *The Jew is foreign and consequently - particularly in the 1800's* **-** *would have been regarded as a strange and distant figure in the minds of English audiences*. (**Additional information)**
>
> *Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991; 325:293* **-** *302*. (**Range)**
>
> *University of* **-** *MA in International Relations Submission date: 03.11.2003 Word length: 2491* (**References)**

As described by Lester (2018), when writers facilitate emphasis in texts, they use the dash towards the end of the sentences, as observed in the BAWE texts. The additional information usage of the dash is similar to parenthesis and accounts for 20 percent of the 100 randomized occurrences. This is a similar percentage to using parenthesis to facilitate additional information, discussed in Section 8.3.5. With the dash, however, the writer indicates the additional details with two dashes instead of enclosing them with parentheses. The dash in the range of numbers is understandable since referring to pages in academic writing is a common practice, which is demonstrated by the BAWE writers.

In summary, it is important to observe that the dash does not appear in the study of Sun and Wang (2019) discussing the most frequently used punctuation marks across registers and genres. However, the BAWE corpus, featuring more advanced academic writing, ranks dash at position 9, indicating its frequent use in academic writing. This finding suggests that teaching the use of the dash in entry-level classrooms may be beneficial.

8.4 Conclusion

This chapter analyzes the frequency and patterns of punctuation mark usage in COMP 101, comparing them with those observed in upper-level writing from the BAWE corpus. In addition to the primary analysis, the chapter also examines the distribution of punctuation marks within the COMP 101 subcorpora, highlighting patterns that offer meaningful points for comparison and potential classroom instruction. The findings reveal that COMP 101 writers use shorter sentences, fewer references, and a more conversational tone, as indicated by the high frequency of questions and exclamation marks. The fewer references align with the task instructions, which do not require extensive research. Only the argumentative essay calls for the use of outside sources. However, the inclusion of quotation marks and parentheses in references in other essay genres demonstrates students' understanding of the importance of integrating external sources into their writing. On the other hand, BAWE writers demonstrate longer sentences and advanced reference skills, exemplified by their semicolon use. However, both groups need further instruction in using semicolons, which are often incorrectly used and sometimes confused with colons.

One of the main challenges in discussing these findings is the limited research on punctuation in undergraduate writing available for comparison. Using punctuation marks can reveal incorrect usage, genre features, and stylistic tendencies, thus offering valuable insights into the writers' skills and instructional needs. This chapter underscores the gaps in current research and the need for further exploration. For example, future studies could examine the role of full stops in estimating average sentence length in entry-level writing, such as COMP 101, and compare this with BAWE data while excluding references. Such an analysis would provide more accurate

insights into sentence structure and writing development. Additionally, the role of quotation marks in incorporating outside sources is significant within academic writing. Future research could analyze the patterns of quotation use among first-year students and track their evolution throughout their academic progression.

Chapter 9

Conclusion

9.0 Introduction

This chapter revisits the research questions introduced in Chapter 1 and addresses them by summarizing the findings from the corpus analysis of COMP 101 and the theoretical framework. The findings reveal that first-year composition writers frequently use: (1) first-person singular and plural pronouns, as highlighted in Chapter 5; (2) second-person pronouns, as outlined in Chapter 6; (3) coordinating and subordinate conjunctions, as discussed in Chapter 7; and (4) the most commonly used punctuation marks, as explored in Chapter 8. Additionally, the chapter highlights the potential implications for classroom instruction, reviews the study's limitations, and suggests future research directions.

9.1 Research questions and summary of results

The two research questions introduced in Chapter 1 and addressed by the study are classified as one main question and one sub-question:

(1) What are the key linguistic features that characterize first-year composition writing, and how do they differ from upper-level or later stages of writing at university?

The sub-question that aids the investigation of the language features in the main research question is:

(2) How are these key linguistic features distributed in the descriptive, narrative, classification, process analysis, compare-and-contrast cause-and-effect, and argumentative texts?

In order to address the research questions, this study utilizes the concept of genre as its theoretical framework, which is discussed in Chapter 3. The primary reason for choosing genre as the central component of the theoretical framework is its ability to provide insight into the

lexico-grammatical features employed by writers and to compare these features with those of other groups of writers, which relates to the study's research questions. To conduct the research, the study uses a corpus, COMP 101, which consists of essays written by first-year university students to offer insights into the introductory level of writing at university (refer to Section 4.1 for specific details about COMP 101). The upper level or subsequent stages of writing at university are represented by BAWE, a larger corpus than COMP 101, containing a wide variety of academic writing. BAWE allows for a comparison with the features seen in first-year composition writing, intending to interpret the data (Section 4.3.1 discusses the comparative design and its significance). Using corpus linguistics as the methodology, the study identified first-person pronouns, second-person pronouns, conjunctions, and specific punctuation marks as high-frequency elements, constructing the textual features that characterize first-year composition writing. Table 9.1 summarizes these features, highlighting their performance in COMP 101 and how they differ from those used by upper-level students in BAWE.

Table 9.1 Summary of language features in COMP 101 and BAWE

Language Features	COMP 101	BAWE
First-person pronouns (Chapter 5)	Students primarily use the representative role and slowly move towards the opinion-holder towards the end of the semester.	Students use the representative the least and guide the most with a balanced use of the architect and the opinion-holder.
Second-person pronouns (Chapter 6)	Students rely strongly on the second person to relate structural knowledge and moral formulation; the use of the second person decreases towards the end of the semester.	Students use the second person generically to relate structural knowledge (instructions).

Conjunctions (Chapter 7)	Students rely on coordinate and subordinate clauses. The non-finite clauses are mainly constructed with infinitives and gerunds. Past participles are used on some occasions but not very frequently.	Students use fewer coordinate clauses. The nonfinite clauses are mostly infinitives and gerunds, but the past participles are also frequently used.
Punctuation (Chapter 8)	Students write shorter sentences: 18.6 words per sentence; use, on average, 20 percent references based on the overall comma, full stop, parenthesis, and colon use in references; and show conversational style based on the use of questions and exclamation marks.	Students write longer sentences: 22 words per sentence; use, on average, 40 percent of references based on the overall comma, full stop, parenthesis, and colon use in references; and show no conversational style based on the absence of high frequency of questions and exclamation marks.

Table 9.1 shows that first-person pronouns are used to facilitate authorial roles in COMP 101. These roles progress gradually from the representative to the higher authorial power throughout the semester, which also correlates with the degrees of the authorial presence in Tang and John's (1999) framework. Also, as students move through the semester, they begin integrating a wider selection of roles in their writing. This role adoption aligns with the essay genres that at the beginning of the semester focus on personal experiences, like descriptive and narrative texts, and later on causal and argumentative texts. On the other hand, in BAWE, students demonstrate a balanced use of the authorial roles and gravitate towards the role of the guide, which most clearly reflects the writer's intention to direct the reader through the text. Understandably, even at the upper level, opinion-holder roles are not frequent since it is more typical for advanced writers (Hyland 2001a; Hyland 2002a), such as researchers and experts, to express opinions and positions. This comparison suggests that first-year composition writing is an appropriate stage for students to be introduced to the various roles of authorial power. Understanding these roles

can help students engage their readers and recognize the various forms of authorial expression beyond the mere representative.

With regard to second-person pronouns, the analysis reveals that first-year students in this study rely heavily on them, especially in facilitating generic ideas to address structural knowledge, moral formulation, or viewpoints (Kitagawa and Lehrer 1990). This generic function of *you* is described as one of the markers of informality in academic writing (Hyland and Jiang 2017; Liardet *et al.* 2019). The use of the second-person pronouns in structural knowledge is aligned with the process-analysis essay directional approach, accommodating the use of *you* to address the reader, but in other tasks, such as the comparison and contrast, cause and effect, or argumentative, it is not permitted. Still, first-year students gravitate towards the use of phrasing their thoughts. However, as students approach the end of the semester, the use of the second person begins to decrease, which shows a positive trend in beginning to understand the importance of focusing on the topic rather than the reader. In contrast, in BAWE, the students use the second person to relate their conversations between medical professionals and patients and facilitate instructions to the reader, but not as a means to frame topics related to viewpoints. The findings indicate that first-year students may benefit from focused instruction on suitable uses of the second person, such as directions and instructions, but focus on text objectivity when expressing viewpoints and understanding the difference.

The next important frequency feature revealed by the analysis is the role of conjunctions. The first-year students use coordinate conjunctions 38 percent more than BAWE and use 50 percent more subordinators such as *if*, *because*, and *when* than upper-level students. However, the findings also indicate that syntax is not the only measure of complexity. For instance, in COMP 101, students use simple syntax structures with low-context vocabulary, while in BAWE, these

simple structures are used with high-context vocabulary to express complex ideas, showing that vocabulary also plays an important role in demonstrating complexity (Halliday 2007). The COMP 101 low-context vocabulary reflects the nature of the tasks, which allow students to choose their own topics. As a result, students select subjects grounded in general knowledge, rather than discipline-specific content. It is understandable that students gravitate towards general subjects since they are very new to their chosen fields of study. Both first-year students and upper-level students use nonfinite structures such as infinitives and gerunds. The key difference between the two groups is the use of past participles in nonfinite structures. Compared to upper-level students, COMP 101 students use only one-third of the structures. This infrequent usage of past participles in nonfinite structures in first-year composition writing supports the findings of Biber and Gray (2010) as well as Staples et al. (2016), indicating that participles are characteristic of higher levels of academic writing in university.

The last feature based on the high-frequency items in COMP 101 is punctuation. This is one of the areas where the study could not locate sufficient research showing frequency analysis of punctuation marks in university writing for comparison purposes. The research indicates that first-year COMP 101 students use punctuation differently compared to upper-level BAWE students. COMP 101 writers tend to use shorter sentences, contractions, questions, and exclamation marks that strongly suggest a tendency toward simple structures and less formal writing (Liardet *et al.* 2019). In COMP 101, the limited use of references or outside sources indicated by certain punctuation marks (e.g., quotation marks and parentheses) relates to tasks that do not prioritize research at this stage. On the other hand, BAWE writers demonstrate longer sentence structure, more commas, full stops, and parentheses in references, and no indication of high use of question and exclamation marks, which speaks of a tendency towards complex

structures, more integration of references, and formal writing. Another finding regarding punctuation is that both groups show incorrect use of semicolons, suggesting a need for specific instruction on semicolons in the classroom to encourage best practices.

To address the sub-question, the study analyzed the role of the high-frequency features in the subcorpora. Table 9.2 displays the findings related to each of the features regarding the subcorpora.

Table 9.2: Summary of language features across the subcorpora

Language Features	DS	NS	CS	PS	CCS	CES	ARS
First-person pronouns (Chapter 5)	Texts use primarily the representative role.	Texts use mainly the representative (75%), but the architect begins to emerge as well as the opinion holder (13%).	Roles increase and include the guide (5%), architect (10%) and opinion holder (20%) besides mainly the representative that decreases .	Roles become more balanced: guide and the opinion holder increase (41%).	Roles continue to stay balanced; the opinion holder decreases (20%).	All roles stay present; the opinion holder increases (25%).	All stay present; the opinion holder peaks (34%).
Second-person pronouns (Chapter 6)	*You* is used mainly in its generic sense (11 %).	*You* is used mainly in its referential sense.	*You* is used mainly in a generic sense (10%): mostly in moral formulation.	*You* peaks in its generic sense (38%) structural knowledge (directions)	*You* is used mainly in a generic sense (20%): structural knowledge and moral formulation.	*You* drops in its generic sense (4%): mainly in moral formulation.	*You* stays low with a slight increase (10%).
Conjunctions (Chapter 7)	Sentence structure is mainly supported by finite clauses. Participles are at their lowest, 6 %.	The sentence structure is still supported mainly by finite clauses. Participles are still at 6%.	The sentence structure is still supported mainly by finite clauses. Participles are still at 7%.	The sentence structure continues to be supported by mainly finite clauses. Participles grow to 8%.	The sentence structure continues to be supported by mainly finite clauses. Participles grow to 16%.	The sentence structure continues to be supported by mainly finite clauses. Participles drop to 9%.	The sentence structure continues to be supported by mainly finite clauses. Participles peak to 15%.
Punctuation (Chapter 8)	Second to the argumentative texts in the least number of periods. Texts tend to use longer	The highest number of apostrophes and exclamation marks.	The highest number of periods and semicolons. Based on the high number of periods,	Second to the narrative texts to have the highest number of exclamation marks.	Contain no exclamation marks and have the least number of semicolons.	Second, to the argumentative texts in the number of parentheses.	Contain the highest number of parenthesis and the fewest periods.

	sentences, like argumentative texts.		presumably have the shortest sentences.				Contain no exclamation marks.

Some important trends based on the summary in Table 9.2 can be categorized using the functions of personal pronouns, sentence structure, and punctuation. Regarding the use of the first person, there is a gradual decrease from the role of the representative, which is at full 100 percent in descriptive writing to a partial replacement by other roles, such as the guide, architect, and opinion holder. These roles, as noted in Chapter 5, also align with the task instructions. For example, the descriptive and narrative essays call for personal experiences, corresponding closely to the role of the representative. However, as students begin to engage with more analytical writing, such as classification, comparison and contrast, and others, they start to adopt roles related to the textual navigation. The opinion holder, which expresses the highest degree of authorial power, gradually increases throughout the semester, suggesting a growth in confidence when stating viewpoints. Another trend is the gradual decrease of the second person, which reaches its lowest usage in cause-and-effect texts, indicating a more formal writing style by the end of the semester. Interestingly, in argumentative texts, the use of the second person increases slightly, engaging the reader in the discussion. Only the descriptive, narrative, and process analysis tasks permit the use of the second-person, but even beyond these genres, COMP 101 writers continue to use it when framing their thoughts and engaging the reader.

Examining the trends in sentence structure, specifically through the use of conjunctions, two main features emerge: the predominant use of finite clauses and the gradual increase of the participles. The dominant characteristic across the subcorpora is the use of finite structures that remain the main mode of syntax. Since the tasks do not prescribe one clausal structure over

another, this pattern likely reflects students' natural syntactic preferences—favoring finite structures. The feature that tracks steady growth is the use of participles within nonfinite structures, where infinitives and gerunds are predominantly utilized. The trajectory of the use of participles starts at 6 percent and steadily increases to 15 percent by the end of the semester. Finally, the punctuation marks that track growth in academic writing include question and exclamation marks, which are prevalent at the beginning of the semester but gradually disappear by the end. The full stop is another punctuation mark that characterizes the shortest sentences in classification texts and the longest sentences in argumentative texts.

These findings are based on and supported by the genre-based theoretical framework, which offers lens for examining the overall features of the essay genre in COMP 101 and analyzing the distribution of these features across the subgenres within the subcorpora. A summary of these findings suggests that the writing style of the average first-year student at the beginning of the semester, when students begin with descriptive and narrative writing, is primarily informal and conversational, which is also supported by the essay genres at this time. The second-person is an indicator of the conversational style in essays that require the first or third person, such as classification, comparison and contrast, cause and effect, and argumentative essays. This tendency is also demonstrated in the use of questions and exclamation marks. The first-person roles seem to align with the task expectations, and students use the role of the representative in the more personal aspect of writing, like the descriptive and narrative, but move to a wider range in the more analytical essays, such as the comparison and contrast, cause and effect, and argumentative. The sentence structure is mostly supported by the finite clauses, but students use nonfinite constructions as well, like the infinitives and present participles. The past participles are the least common and may require targeted instruction and more practice. The punctuation

examination highlights the overly used questions and exclamation marks outside of narrative contexts that include dialogues. This suggests that students may not fully grasp the academic writing conventions related to these punctuation marks. On a positive note, there is an encouraging trend in the use of punctuation marks like quotation marks and parentheses, particularly in references, even in tasks that do not require outside sources. This demonstrates that first-year students are beginning to recognize the value of external support..

9.2 Teaching implications of the findings

This study was motivated by my professional desire to continue my education, gain a specific understanding of my students, and become a better instructor. I believe that this research contributes to the broader field of university writing (Staples *et al.* 2016; Monsen and Rørvik 2017; MacIntyre 2019; Nesi 2021) and specifically to first-year composition writing (Aull 2015; Eckstein and Ferris 2019; Lee *et al.* 2019). This work has informed my understanding and practice with entry-level students and motivated me to reflect on the teaching implications of the findings. The following paragraphs address the possible teaching implications of the research findings, suggesting teaching strategies. These strategies are not meant to be comprehensive but represent the initial step in exploring my findings and probably mark the follow-up stage of my studies beyond the dissertation work.

The study indicates that first-year students typically use the first person in the representative role, which strongly suggests that at this stage, students need to explore other roles, such as the guide, architect, and opinion holder. This means understanding the concepts behind each role and selecting the appropriate words to engage the readers. For example, students need to understand

the wide range of verbs used with the guide to navigate the reader through the text, such as *allow, consider, provide, remind,* or *lead,* as described in Chapter 5. Based on the findings of this study, the range of verbs at this entry-level of academic writing, as represented by COMP 101, includes limited choices, such as *be, can, have, see,* and *will*. Students can improve their writing skills by intentionally selecting their authorial stance based on the context and using a wider range of vocabulary. This can be achieved by observing texts written by advanced academic writers and guiding students to learn from such choices.

Understanding the use of the authorial roles in the COMP 101 subcorpora provides insights into their genre usage. For example, students primarily use the representative role in descriptive and narrative writing. These genres enable writers to narrate their own experiences by adopting the role of a representative. However, in analytical essays such as classification, cause-and-effect, or argumentative, the writer's voice may navigate the reader through the textual parts and express positions. Students need to differentiate between the genres and know how to approach them through the various authorial roles. Currently, the course syllabus and content do not focus on authorial roles but on the essay parts, such as the introduction, body, and conclusion within the different genres. While these textual components are essential as building blocks at this initial stage of academic writing, students also need to recognize the importance of developing their voice using the first-person pronouns.

In contrast to writing manual recommendations, the use of the second person in COMP 101 highlights another important aspect of classroom instruction. One of the main reasons for excluding the second person from academic writing is the conversational nature of *you*, which is more suited to engaging the reader in a relaxed talk rather than a serious academic discussion. This study indicates that first-year students facilitate structural knowledge and moral formulation

through the generic *you*. In cases when writers formulate instructions or structural knowledge, then the second person is an effective choice to address the reader (Hyland 2001a), and such uses of *you* account for all the generic uses in BAWE.

The subcorpora findings also support the role of the second person in expressing structural knowledge. For instance, the second person is most prominent in the process-analysis genre, which focuses on instructions or structural knowledge. The concern with the second person in COMP 101 occurs when students utilize it to convey moral concepts in genres, such as compare-and-contrast, cause-and-effect, and argumentation. In this regard, instructors can help students understand that even though the second person is effective in framing instructions and engaging the reader in process-analysis writing, it should be avoided in other genres where writers should focus on the topic and evidence. Also, students should recognize that while the second person may be beneficial for drafting and organizing their thoughts, it is best to avoid it in final drafts.

The use of conjunctions is another area of classroom instruction that can help students become more intentional about their writing. Chapter 7 shows that COMP 101 texts predominantly demonstrate low-content vocabulary with finite structures. The low-content vocabulary is based on students' choices of topics, which typically leads to writing on general knowledge subjects they can easily engage with and discuss. In this regard, Halliday (2007) points out that combining high-content lexical density with finite subordinate structures demonstrates advanced writing. This suggests that students may benefit from discipline-specific concepts, expand their word choices, and move toward professional topics, reflecting their majors. Such instruction can help students understand the significance of using technical vocabulary with finite coordinates and subordinate clauses. It suggests the importance of entry-level students understanding this concept and being able to enhance their writing. Another critical aspect of the results in Chapter

7 indicates that students can benefit from understanding non-finite structures and effectively balancing the usage of infinitives, gerunds, and participles. The findings of this study show that entry-level writers gravitate towards gerunds and demonstrate a limited use of past participles. Dedicating classroom time to teaching past participle structures can help students improve their writing by emphasizing results over actions.

The findings of the subcorpora confirm those of the general COMP 101, showing that students tend to use finite subordinate structures with low-context vocabulary, which aligns with the task instructions, but may also suggest that students can also benefit of tasks that incorporate discipline-specific areas and engage in research. Novice writers can enhance their grammatical structures by incorporating sentence variety and creating concise texts. For example, structures like *as mentioned*, *when paired*, or *when needed* are more concise than *those I mentioned earlier*, *when it is paired*, or *when it is necessary*. The findings show that at this stage, students use past participles in nonfinite structures, mostly in compare-and-contrast (16%), argumentative (15%), and cause-and-effect writing (9%) (see Table 9.2 for details). Focusing on these structures can help students become intentional not only in their use but also in recognizing the variations in building nonfinite structures and being able to identify them in texts at large.

Chapter 8 highlights several aspects of the punctuation marks that are relevant for classroom instruction in entry-level writing classes. One of these key aspects demonstrated by commas, full stops, parenthesis, and quotation marks is the small number of references used by COMP 101 writers compared to BAWE. At entry-level university writing, students are not required to conduct intense research, but at this point, students should begin to incorporate outside sources and reveal discipline-specific topics. Once again, the task instructions might include targeted practice and assignments with external sources to help first-year students in developing essential

skills to meet these higher expectations. Another important area revealed by the punctuation marks is the incorrect usage of contractions, semicolons, and colons. The frequent occurrence of contractions in COMP 101 indicates a need for classroom instruction on the proper usage. Also, both COMP 101 and BAWE writers tend to overuse semicolons and often mistake them for colons, and underline the importance for students to learn to differentiate them and apply them correctly. Finally, in COMP 101, students frequently use question marks and exclamation marks to convey emotions and engagement with their topics. Classroom instruction should focus on conventional writing practices and help students learn to express their thoughts and feelings through words rather than relying on punctuation marks.

9.3 Review of the limitations of the study

This study has three main limitations: (1) the challenge of controlling the amount of data going to the study corpus, (2) the relatively small size of the individual subcorpora, . This section discusses these limitations.

One of the main challenges in the study was maintaining a consistent data flow in corpus design based on the participants' choices. For instance, the texts in COMP 101 were gathered over three consecutive semesters: Fall 2019, Spring 2020, and Fall 2020. In Fall 2019, out of the 50 students enrolled in the module Composition I, only 27 agreed to participate in the research. The participation outcomes were similar in other semesters, with an average of 50 percent of the students participating. While having participants decide on their involvement in the study gives everyone an equal opportunity to participate in the research, it does not guarantee a steady flow of data in the corpus design.

The size of the COMP 101 corpus and its subcorpora, such as descriptive, narrative, classification, or cause-and-effect, depends on the amount of data collected. The COMP 101 corpus contains 188,828 words and is designed to provide researchers with specific insights into the writing of entry-level students. However, each of the subcorpora consists of a smaller number of words. For instance, the classification subcorpus has 23,126 words, and the cause-and-effect subcorpus has 29,630 words. It is important to note that some students fail to submit assignments, which negatively impacts data collection, even though submitting assignments is part of the module requirements. On the positive side, the relatively small size of the individual subcorpora allows easier data management and processing, especially with the corpus-driven nature of this study. This consideration relates back to the issue of corpus size (see Section 4.1.3).

The next limitation of this study is its focus on empirical language analysis, which looks at the data objectively, not the participants' experiences. This quantitative approach has limitations when it comes to understanding the individual experiences of the learners and their language development stages before their admission to undergraduate courses. According to Jones and Waller (2015), corpus data cannot provide information about how language users process word choices that shape language patterns, which results from corpus investigation. Since the COMP 101 texts are produced by entry-level writers new to the academic setting, their motivation behind the word choices demonstrated in the empirical analysis could have provided further insight into their backgrounds and writing competencies.

9.4 Directions for Future Research

Chapters 5 and 6 of the study provide in-depth information and analysis about using first-person and second-person pronouns in writing. However, they do not explain why writers choose to use these pronouns. For instance, in COMP 101, over 60 percent of the first-person pronouns are used to facilitate the role of the representative, while less than 10 percent are used to play the role of the guide. Similarly, when it comes to the second-person pronoun, writers often use it to frame concepts of moral formulation, which do not require directional writing but objective accounts. In COMP 101, moral formulation accounts for 34 percent of all generic use (Chapter 6.3 discusses this use in detail), and it is frequently used even in later parts of the semester in genres such as compare-and-contrast and argumentative writing. By combining student interviews with corpus analysis, it may be possible to gain valuable insights about writers, frequent language patterns, and tips for classroom instruction.

Research on the development of clausal structures in composition writing, discussed in Chapter 7, can be aided by longitudinal studies beyond entry-level writing, including specific information on subordinating and coordinating clauses, gerunds, infinitives, and past participles. Such research may require an individual researcher teaching various levels of composition writing and collecting data about the language users through several years of university experience or soliciting composition instructors to provide data from their students as part of a study corpus.

The analysis of the frequency patterns related to the punctuation marks in COMP 101 reveals the need for a large scope of research in using full stops, commas, apostrophes, semicolons, colons, questions, and exclamation marks (discussed in Chapter 8). For example, full stops provide essential information about sentence length. Still, references must be excluded from the main text to focus the calculations on the sentence length without counting in full stops in references.

Next, commas in COMP 101 are primarily used in introductory elements and lists, while BAWE writers use them mostly in references and subordination. It is interesting to research the reasons that motivate entry-level writers' choices in relation to the frequent use of lists and introductory elements.

Semicolons and colons require further research, especially in the inconsistent use of semicolons and the confusion between colons, commas, or periods. Not only do COMP 101 writers indicate a misunderstanding in using semicolons, but BAWE writers also need help with it (see Chapter 8.3.6 for details). Another area that requires further research related to punctuation is entry-level writers' frequent use of questions and exclamation marks. This indicates similarities with conversational and social-media writing, which raises questions about why writers gravitate towards these atypical marks for academic writing. This suggests a need for future research to understand their choices.

The insights obtained from the study of entry-level writers can be used as a valuable database for comparative analysis in future research. One area of interest for comparison could be the impact of artificial intelligence (AI) tools on first-year university writing and beyond. AI tools have gained immense popularity since November 2022 (Kumar *et al.* 2023), enabling users to produce language content effortlessly and quickly. However, while these tools are designed to mimic human-like language and can create content in various registers and genres, they need the nuances and experiences of human writers. Therefore, focused research is required to understand how the integration of AI input and output affects authorial choices, the direct inclusion of the reader, grammatical complexity, and variations in punctuation usage. Such research can help educators and linguists to better understand the impact of AI tools on language variation and developmental stages in university writing.

The primary purpose of this study was to investigate the frequency features of first-year composition writing through a corpus-driven approach and within a genre-based theoretical framework. Despite its limitations, this research contributes significantly to the ongoing discussion on the academic writing of first-year university students. It sheds light on the textual characteristics displayed by students at this stage within a typical university classroom setting, where both native and non-native English speakers converge to acquire fundamental writing skills, laying the groundwork for a successful academic journey.

References

ACT College & Career Readiness Standards - CCRS-English Standards.pdf (2020), available: https://www.act.org/content/dam/act/unsecured/documents/CCRS-EnglishStandards.pdf [accessed

Aull, L. (2015) *First-Year University Writing*, Palgrave Macmillan.

Aull, L. (2019) 'Linguistic markers of stance and genre in upper-level student writing', *Written Communication*, 36(2), 267-295, available: http://dx.doi.org/https://doi.org/10.1177/0741088314527055.

Aull, L.L. (2020) *How Students Write*, New York: The Modern Language Association of America.

Aull, L.L., Bandarage, D. and Miller, M.R. (2017) 'Generality in student and expert epistemic stance: A corpus analysis of first-year, upper-level, and published academic writing', *Journal of English for Academic Purposes*, 26, 29-41, available: http://dx.doi.org/10.1016/j.jeap.2017.01.005.

Aull, L.L. and Lancaster, Z. (2014) 'Linguistic markers of stance in early and advanced academic writing', *Written Communication*, 31(2), 151-183.

Baker, P. (2008) *Using Corpora in Discourse Analysis*, London: Continuum.

Barrass, R. (2002) *Scinetists must write: A guide to better writing for scientists, engineers and students*, Routledge.

Bawarshi, A. and Reiff, M.J. (2010) *Genre: An Introduction to History, Theory, Research, and Pedagogy*, West Lafayette, Indiana: Parlor Press.

Bayraktar, M., Say, B. and Akman, V. (1998) 'An analysis of English punctuation: the special case of comma', *International Journal of Corpus Linguistics*, 3, available: http://dx.doi.org/10.1075/ijcl.3.1.03bay.

Beaman, K. (1984) 'Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse' in Tannen, D., ed., *Coherence in Spoken and Written Discourse*, New Jersey: Ablex Publshing Corporation, 45-80.

Bennett, K. (2009) 'English academic style manuals: A survey', *Journal of English for Academic Purposes*, 8, 43-54.

Berry, R. (2009) 'You could say that: the generic second-person pronoun in modern English', *English Today 99*, 25(3), available: [accessed June 16, 2021].

Bhatia, V.K. (1993) *Analysing Genre: Language Use in Professional Settings*, London: Longman.

Bhatia, V.K. (2004) *Worlds of Written Discourse*, Bloomsbury, available: https://read.amazon.com/.

Bhatia, V.K. (2012) 'Critical reflections on genre analysis', *Iberica*, 24, 17-28.

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge, Great Britain: Cambridge Universiy Press.

Biber, D. (1993) 'Representativeness in corpus design', *Literary and Linguistic Computing*, 8(4), 243-257.

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*, Cambridge, UK: Cambridge University Press.

Biber, D. (2006) *University Language*, Amsterdam, Philadelphia: John Benjamins Publishing Company.

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

Biber, D., Conrad, D. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press.

Biber, D. and Conrad, S. (2009) *Register, Genre, and Style*, Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2004) 'If you look at . . .: Lexical Bundles inUniversity Teaching and Textbooks', *Applied Linguistics*, 35(3), 371-405.

Biber, D. and Gray, B. (2010) 'Challenging stereotypes about academic writing: Complexity, elaboration, explicitness', *Journal of English for Academic Purposes*, 9(1), 2-20, available: http://dx.doi.org/10.1016/j.jeap.2010.01.001.

Biber, D. and Gray, B. (2015) 'Phraseology' in Biber, D. and Reppen, R., eds., *The Cambridge Handbook of English Corpus Linguistics*, Cambridge University Press, 125-145.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *The Longman Grammar of Spoken and Written English*, Harlow, England: Pearson Education Limited.

Biber, D. and Reppen, R. (2015) 'Introduction' in Biber, D. and Reppen, R., eds., *The Cambridge Handbook of English Corpus Linguistics* Cambridge University Press, 1-8.

Bolinger, D. (1979) 'To catch a metaphor: you as norm', *American Speech*, 54(3), 194-209.

Bonelli-Tognini, E. (2001) *Corpus Linguistics at Work*, John Benjamins, available: https://ebookcentral.proquest.com/lib/mic-ebooks/reader.action?docID=680385.

Braine, G. (2001) 'When professors don't cooperate: a critical perspective on EAP research', *English for specific purposes (New York, N.Y.)*, 20(3), 293-303, available: http://dx.doi.org/10.1016/S0889-4906(00)00011-9.

Brezina, V. (2018) *Statistics in corpus linguistics: a practical guide*, Cambridge: Cambridge University Press.

Cabral-Cardoso, C. (2021) 'The Englishisation of higher education, between naturalisation and resistance', *Journal of applied research in higher education*, 13(4), 1227-1246, available: http://dx.doi.org/10.1108/JARHE-05-2020-0116.

Cambridge English Corpus available for academic use | Cambridge Language Sciences (2019), available: https://www.languagesciences.cam.ac.uk/news/cambridge-english-corpus-available-academic-use [accessed

Canagarajah, S. (2024) 'Decolonizing Academic Writing Pedagogies for Multilingual Students', *TESOL Quarterly*, 58(1), 280-306, available: http://dx.doi.org/10.1002/tesq.3231.

Carrió-Pastor, M.L. (2013) 'A contrastive study of the variation of sentence connectors in academic English', *Journal of English for Academic Purposes*, 12(3), 192-202, available: http://dx.doi.org/10.1016/j.jeap.2013.04.002.

Carter-Thomas, S. and Rowley-Jolivet, E. (2008) 'If-conditionals in medical discourse: From theory to disciplinary practice', *Journal of English for Academic Purposes*, 7(3), 191-205, available: http://dx.doi.org/10.1016/j.jeap.2008.03.004.

Carter, R. and McCarthy, M. (1995) 'Grammar and the Spoken Language', *Applied Linguistics*, 16(2), 141-158.

Carter, R. and McCarthy, M. (2006) *Cambridge Grammar of English*, Cambridge: University Press.

Carter, R., McCarthy, M., Mark, G. and O'Keeffe, A. (2016) *English Grammar Today* Cambridge, United Kingdom: Cambridge University Press.

The CEFR Levels (2020), available: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions [accessed 2022/09/17/].

Chafe, W. and Tannen, D. (1987) 'The Relation between Written and Spoken Language', *Annual Review of Anthropology*, 16, 383-407.

Chang, Y. and Swales, J. (1999) 'Informal elements in English academic writing: Threats or opportunities for advanced non-native speakers?' in Candlin, C., Hyland, K. and Candlin, C., eds., *Writing: Texts, Processes and Practices*, London: Longman.

Charles, M. (2007) 'Argument or evidence? Disciplinary variation in the use of the Noun that pattern in stance construction', *English for Specific Purposes*, 26(2), 203-218, available: http://dx.doi.org/10.1016/j.esp.2006.08.004.

Chen, Y.-H. and Baker, P. (2010) 'Lexical Bundles in L1 and L2 Academic Writing', *Language Learning and Technology*, 14(2), 30-49.

Cheung, Y.L. and Lau, L. (2020) 'Authorial voice in academic writing: A comparative study of journal articles in English Literature and Computer Science', *Iberica*, 39, 215-242.

Collins, L.C. (2019) *Corpus Linguistics for Online Communication: A Guide for Research*, Taylor & Francis Group, available: https://ebookcentral.proquest.com/lib/mic-ebooks/reader.action?docID=5732096.

Connors, R.J. (1989) 'Rhetorical history as a component of composition studies', *Rhetoric review*, 7(2), 230-240, available: http://dx.doi.org/10.1080/07350198909388858.

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

Crismore, A. (1989) *Talking with Readers: Metadiscourse as Rhetorical Art*, New York: Peter Lang.

Crismore, A., Markkanen, R. and Steffensen, M. (1993) 'Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students', *Written Communication*, 101(1), 39-71.

Crossley, S.A. (2020) 'Linguistic features in writing quality and development: An overview', *Journal of Writing Research*, 11(3), 415-443.

Crossley, S.A., Kyle, K. and McNamara, D.S. (2016) 'The development and use of cohesive devices in L2 writing and their relations to judgement of essay quality', *Journal of Second Language Writing*, 32, 1-16, available.

Crowhurst, M. (1987) 'Cohesion in argument and narration at three grade levels', *Research in the Teaching of English*, 21(2), 185-201.

Crystal, D. (1997) *English as a global language*, Cambridge: Cambridge University Press.

Culpeper, J. (2009) 'Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's Romeo and Juliet', *International Journal of Corpus Linguistics*, 14(1), 29-59.

Culpeper, J. and Demmen, J. (2015) 'Keywords' in Biber, D. and Reppen, R., eds., *The Cambridge Handbook of English Corpus Linguistics*, Cambridge University Press, 90-105, available: https://www-cambridge-org.libraryproxy.mic.ul.ie/core/search.

Dixon, T. (2022) 'Proscribed informality features in published research: A corpus analysis', *English for Specific Purposes*, 65, 63-78.

Dou, A.Q., Chan, S.H. and Win, M.T. (2023) 'Changing visions in ESP development and teaching: Past, present, and future vistas', *Frontiers in psychology*, 14, 1140659-1140659, available: http://dx.doi.org/10.3389/fpsyg.2023.1140659.

Durrant, P. (2017) 'Lexical Bundles and Disciplinary Variationin University Students' Writing: Mappingthe Territories', *Applied Linguistics*, 38(2), 165-193.

Durrant, P., Moxley, J. and McCallum, L. (2019) 'Vocabulary sophistication in First-Year Composition assignments', *International Journal of Corpus Linguistics*, 24(1), 33-66, available: http://dx.doi.org/10.1075/ijcl.17052.dur.

Eckstein, G. and Ferris, D. (2019) 'Comparing L1 and L2 texts and writers in first-year composition', *TESOL Quarterly*, 52(1), 137-162, available: http://dx.doi.org/10.1002/tesq.376.

Eggins, S. and Martin, J.R. (1997) 'Genres and registers of discourse' in van Dijk, T., ed., *Discourse as structure and process*, London: Sage, 230-56.

Fairclough, N. (2001) *Language and power*, Essex: Harlow: Longman.

Fang, Z., Cao, P. and Murray, N. (2020) 'Language and meaning making: Register choices in seventh-and-ninth-grade students' factual writing', *Linguistics and Education*, 56, 1-12, available: http://dx.doi.org/10.1016/j.linged.2020.100798.

Farahani, M.V. and Sabetifard, M. (2017) 'Metadiscourse Features in English News Writing among English Native and Iranian Writers: A Comparative Corpus-based Inquiy', *Theory and Practice in Language Studies*, 7, 1249-1260.

Ferguson, C. (1994) 'Dialect, register and genre: Working assumptions about conventionalization' in Finegan, E. and Biber, D., eds., *Sociolinguistic perspectives on register*, New York: Oxford University Press.

Firth, A. (1996) 'The discursive accomplishment of normality : On 'lingua franca' English and conversation analysis: Conversation analysis of foreign language data', *Journal of Pragmatics*, 26(2), 237-259.

Flowerdew, J. (2004) 'The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings' in Upton, T. and Connor, U., eds., *Discourse in the Professions: Perspectives from Corpus Linguistics*, John Benjamins, 11-33, available: https://www.google.com/books/edition/Discourse_in_the_Professions/WJ8vrhXfPDEC?hl=en.

Flowerdew, J. (2011) 'Action, content and identity in applied genre analysis for ESP', *Language Teaching*, 48(1), 516-528.

Flowerdew, J. (2016) 'English for Specific Academic Purposes (ESAP) Writing: Making the case', *Writing & Pedagogy*, 8(1), 5-32, available: http://dx.doi.org/10.1558/wap.v8i1.30051.

Flowerdew, J. (2019) 'Power in English for Academic Purposes' in Hyland, K. and Wong, L. L. C., eds., *Specialised English : New Directions in ESP and EAP Research and Practice*, Oxford, UNITED KINGDOM: Taylor & Francis Group, 50-62.

Foster, J. (2005) *Effective Writing Skills for Public Relations*, 3rd ed., London: Kogan Page.

Fowler, H.R. and Aaron, J.E. (2016) *The Littel Brown Handbook*, United States: Pearson.

Friginal, E. and Hardy, J. (2014) *Corpus-Based Sociolinguistics: A Guide for Students*, Taylor & Francis Group, available: https://ebookcentral.proquest.com/lib/mic-ebooks/reader.action?docID=1582627.

Gablasova, D., Brezina, V. and McEnery, T. (2017) 'Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information: Exploring Learner Language Through Corpus', *Language Learning*, 67(S1), 130-154, available: http://dx.doi.org/10.1111/lang.12226.

Gablasova, D., Brezina, V. and T, M. (2017) 'Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence', *Language Learning*, 67, 155-179, available: https://onlinelibrary-wiley-com.libraryproxy.mic.ul.ie/doi/pdfdirect/10.1111/lang.12225.

Gere, A.R., Aull, L., Escudero, M.D.P., Lancaster, Z. and Vader Lei, E. (2013) 'Local Assessment: Using Genre Analysis to Validate Directed Self-Placement', *College Composition and Communication*, 64(4), 605-633.

Gil, N.N. and Caro, E.M. (2019) 'Lexical bundles in learner and expert academic writing', *Bellaterra Journal of Teaching and Learning Language and Literature*, 12(1), 65-90.

Goulart, L., Biber, D. and Reppen, R. (2022) 'In this essay, I will …: Examining variation of communicative purpose in student written genres', *Journal of English for Academic Purposes*, 59, 101159, available: http://dx.doi.org/10.1016/j.jeap.2022.101159.

Halliday, M.A.K. (2007) 'Differences between spoken and written language: some implications for literacty teaching (1979)' in Webster, J., ed., Bloomsbury Academic, 63-80.

Halliday, M.A.K. and Hasan, R. (1976) *Cohesion in English*, New York: Routledge.

Halliday, M.A.K. and Hasan, R. (1985) *Language, Context, and Text: Aspects of language in a social-semiotic perspective*, Oxford, UK: Oxford University Press.

Halliday, M.A.K. and Matthiessen, C. (2014) *An Introduction to Functional Grammar*, fourth ed., Routledge, available: https://www.google.com/books/edition/An_Introduction_to_Functional_Grammar/JM3KAgAAQBAJ?hl=en&gbpv=1.

Harris, Z. (1959) 'The transformation model of language structure', *Anthropological Linguistics*, 1(1), 27-29.

Harwood, N. (2005) "Nowhere has anyone attempted ... In this article I aim to do just that': A corpus-based study of self-promotional I and we in academic writing across four disciplines', *Journal of Pragmatics*, 37(8), 1207-1231.

Hempel, S. and Degand, L. (2008) 'Sequencers in different text genres: Academic writing, journalese and fiction', *Journal of Pragmatics*, 40(4), 676-693, available: http://dx.doi.org/10.1016/j.pragma.2007.02.001.

Heylighen, F. and Dewaele, J.M. (1999) *Formality of language: Definition, measurement and behavioral determinants*, Center "Leo Apostel", Free University of Brussels.

Hinkel, E. (2003) 'Simplicity without elegance: Featrues of sentences in L1 and L2 academic texts', *TESOL Quarterly*, 37, 275-301.

Ho, V. and Li, C. (2018) 'The use of metadiscourse and persuasion: An analysis of first year university students' timed argumentative essays', *Journal of English for Academic Purposes*, 33, 53-68.

Hyland, K. (1998) 'Persuasion and context: The pragmatics of academic metadiscourse', *Journal of Pragmatics*, 30(4), 437-455.

Hyland, K. (1999) 'Academic attribution: citation and the construction of disciplinary knowledge', *Applied Linguistics*, 20(3), 341-367, available: http://dx.doi.org/10.1093/applin/20.3.341.

Hyland, K. (1999) 'Academic attribution: Citation and the construction of disciplinary knowledge', *Applied Linguistics*, 20(3).

Hyland, K. (2001a) 'Bringing in the Reader', *Written Communication*, 18(4), 549-574.

Hyland, K. (2001b) 'Humble servants of the discipline? Self-mention in research articles', *English for Specific Purposes*, 20, 207-226.

Hyland, K. (2002a) 'Authority and invisibility:authorial identity in academic writing', *Journal of Pragmatics*, 34, 1091-1112.

Hyland, K. (2002b) 'Options of identity in academic writing', *ELT Journal*, 56(4), 351-358.

Hyland, K. (2003) 'Genre-based pedagogies: A social response to process', *Journal of Second Language Writing*, 12(1), 17-29, available: http://dx.doi.org/10.1016/S1060-3743(02)00124-8.

Hyland, K. (2004) *Disciplinary Discourses*, University of Michigan.

Hyland, K. (2005a) *Metadiscourse*, London: Continuum.

Hyland, K. (2005b) 'Stance and Engagement: a Model of Interaction in Academic Discourse', *Discourse Studies*, 7(2), 173-192, available: https://journals-sagepub-com.libraryproxy.mic.ul.ie/doi/pdf/10.1177/1461445605050365.

Hyland, K. (2006) *English for academic purposes: an advanced resource book*, First ed., Boca Raton, FL: Routledge, an imprint of Taylor and Francis.

Hyland, K. (2008) 'The British Academic Written English (BAWE) corpus', *Journal of English or Academic Purposes*, 7(4), 294.

Hyland, K. (2009) *Academic discourse: English in a global context*, London: Continuum.

Hyland, K. (2010) 'Metadiscourse: Mapping Interactions in Academic Writing', *Nordic Journal of English Studies*, 9(2), 125-143.

Hyland, K. (2015a) 'Corpora and written academic English' in Biber, D. and Reppen, R., eds., *The Cambridge Handbook of English Corpus Linguistics*, Cambridge, United Kingdom: Cambridge University Press, 292-307.

Hyland, K. (2015b) 'Genre, discipline and identity', *Journal of English for Academic Purposes*, 19, 32-43.

Hyland, K. (2017) 'Metadiscourse: What is it and where is it going?', *Journal of Pragmatics*, 113, 16-29.

Hyland, K. and Jiang, F. (2017) 'Is academic writing becoming more informal?', *English for Specific Purposes*, 45, 40-51, available: http://dx.doi.org/10.1016/j.esp.2016.09.001.

Hyland, K. and Jiang, F. (2018) 'Lexical bunldes', *International Journal of Corpus Linguistics*, 23, 383-407, available: http://dx.doi.org/https://doi.org/10.1075/ijcl.17080.hyl.

Hyland, K. and Tse, P. (2004) 'Metadiscourse in academic writing: a reappraisal', *Applied Linguistics*, 25(2), 156-177.

Ivanic, R. (1998) *Writing and Identity: The discoursal construction of identity in academic writing*, Amsterdam: John Benjamins Publishing Company.

Ivanic, R. and Camps, D. (2001) 'I am how I sound: Voice as self-representation in L2 writing', *Journal of Second Language Writing*, 1-2, 3-33, available: http://dx.doi.org/https://doi.org/10.1016/S1060-3743(01)00034-0.

Jeffrey, R. (2016) *About writing : a guide*, Place of publication not identified: Open Oregon Educational Resources.

Johns, A. (1980) 'Cohesion in written business discourse: Some contrasts', *The ESP Journal*, 1(1), 35-43.

Jones, C. and Waller, D. (2015) *Corpus Linguistics for Grammar*, Abingdon, Oxon: Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014) 'The Sketch Engine: ten years on', *Lexicography*, 1(1), 7-36, available: http://dx.doi.org/doi:10.1007/s40607-014-0009-9.

Kim, J.-E. and Nam, H. (2019) 'How do textual features of L2 argumentative essays differ across proficiency levels? A multidimensional cross-sectional study', *Reading and Writing*, 32(9), 2251-2279, available: http://dx.doi.org/10.1007/s11145-019-09947-6.

Kirp, D. (2019) *The College Dropout Scandal*, Oxford University Press.

Kitagawa, C. and Lehrer, A. (1990) 'Impersonal uses of personal pronouns', *Journal of Pragmatics*, 14, 739-759.

Koester, A. (2010) 'Building Small Specialized Corpora' in O'Keeffe, A. and McCarthy, M., eds., *The Routledge Handbook of Corpus Linguistics* Routledge, 66-79.

Kostrova, O. and Kulinich, M. (2015) 'Text Genre 'Academic Writing': Intercultural View', *Procedia, social and behavioral sciences*, 206, 85-89, available: http://dx.doi.org/10.1016/j.sbspro.2015.10.032.

Kress, G. (2012) 'Genre as social process' in Cope, B. and Kalantzis, M., eds., *The powers of literacy: A genre approach to teaching writing*, Routledge 22-37.

Laberge, S. and Sankoff, G. (1979) *Discourse and syntax*, New York: Academic Press, available.

Lancaster, Z. (2014) 'Exploring Valued Patterns of Stance in Upper-Level Student Writing in the Disciplines', *Written Communication*, 31(1), 27-57, available: http://dx.doi.org/10.1177/0741088313515170.

Lea, M.R. and Street, B.V. (1998) 'Student writing in higher education: An academic literacies approach', *Studies in higher education (Dorchester-on-Thames)*, 23(2), 157-172, available: http://dx.doi.org/10.1080/03075079812331380364.

Lee, D. (2001) 'Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle', *Language Learning & Technology*, 5, 37-72.

Lee, J., Bychkovska, T. and Maxwell, J. (2019) 'Breaking the rules? A corpus-based comparison of informalfeatures in L1 and L2 undergraduate student writing', *ScienceDirect*, 80, 143-153.

Lee, J. and Deakin, L. (2016) 'Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays', *ScienceDirect*, 33, 23-34.

Leech, G. (1991) 'The State of the Art in Corpus Linguistics' in Aijmer, K. and Altenberg, B., eds., *English Corpus Linguistics*, London: Routledge, 8-29.

Leech, G. (1998) 'Preface' in Granger, S. and Leech, G., eds., *Learner English on Computer*, Abbingdon, UK: Routledge, XIV-XX.

Leech, G. (2011) 'Frequency corpora and language learning' in Meunier, F., De Cock, S. and Gilquin, G., eds., *A Taste for Corpora. In honour of Sylviane Granger.*, Amsterdam: John Benjamins, 7-31.

Leedham, M. (2012) *Chindese students' writing in English: Implications from a corpus-driven study*, Oxford: Routledge.

Lester, M. (2018) *McGraw-Hill Education Handbook of English Grammar and Usage*, New York, UNITED STATES: McGraw-Hill Education.

Li, Y., Gao, Y. and Lu, X. (2023) 'Effects of Word Limit on Sentence Length and Clause Length in Academic Journal Article Abstracts: A Synergetic Linguistic Perspective', *Journal of quantitative linguistics*, 1-21, available: http://dx.doi.org/10.1080/09296174.2023.2263249.

Liardet, C.L., Black, S. and Bardetta, V.S. (2019) 'Defining formality: Adapting to the abstract demands of academic discourse', *Journal of English fro Academic Purposes*, 38, 146-158.

Lintunen, P. and Makila, M. (2014) 'Measuring Syntactic Complexity in Spoken and Written Learner Language: Comparing the Incomparable?', *Research in Language*, 12(4), 377-399.

Lombardi, A. (2021) 'More is more: Explicit intertextuality in university writing placement exam essays', *Journal of English for Academic Purposes*, 50, 100955, available: http://dx.doi.org/10.1016/j.jeap.2020.100955.

MacIntyre, R. (2019) 'The Use of Personal Pronouns in the Writing of Argumentative Essays by EFL Writers', *RELC journal*, 50(1), 6-19, available: http://dx.doi.org/10.1177/0033688217730139.

Macmillan, K. and Weyes, J. (2007) *How to write essays and assignments*, Pearson Education.

Mala, M. (2017) 'A corpus-based diachronic study of a change in the use of non-finite clauses in written English', *Prague Journal of English Studies*, 6(1), 151-166.

Martin, J.R. (1985) 'Process and text: two aspects of human semiosis' in Benson, J. D. and Greaves, W. S., eds., *Systemic Perspectives on Discourse*, Norwood, NJ: Ablex, 248-274.

Mauranen, A. (2019) 'Academically speaking: English as the lingua franca' in Hyland, K. and Wong, L. L. C., eds., *Specialised English: new directions in ESP and EAP research and practice*, Abingdon, Oxon; New York, NY;: Routledge, 9-21.

Mauranen, A., Hynninen, N. and Ranta, E. (2010) 'English as an academic lingua franca: The ELFA project', *English for Specific Purposes*, 29(3), 183-190, available: http://dx.doi.org/10.1016/j.esp.2009.10.001.

McCarthy, M. (2017) *English grammar: your questions answered*, Cambridge: Prolinguam.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics Method, Theory and Practice*, Cambridge, available: https://ebookcentral.proquest.com/lib/mic-ebooks/reader.action?docID=807166#.

McIntyre, D. and Walker, B. (2019) *Corpus stylistics: theory and practice*, Edinburgh: Edinburgh University Press.

Miller, C.R. (1984) 'Genre as a Social Action', *Quartely Journal of Speech*, 70(2), 151-167.

Monsen, M. and Rørvik, S. (2017) 'Pronoun Use in Novice L1 and L2 Academic Writing', *Academic Language in a Nordic Setting - Linguistic and Educational Perspectives*, 9(3), 93-109, available: http://dx.doi.org/https://journals.uio.no/osla/article/view/5849.

Mulvey, C. (2015) 'The English project's history of English punctuation', *English Today*, 32(3), 45-51, available: http://dx.doi.org/10.1017/S0266078416000110.

Mur-Dueñas, P. (2021) 'Engagement markers in research project websites: Promoting interactivity and dialogicity', *Poznan Studies in Contemporary Linguistics*, 57(4), 655-676, available: http://dx.doi.org/10.1515/psicl-2021-0023.

Murphy, M. and Dyrenfurth, M. (2006) 'Understanding The European Bologna Process', 11.1364.1-11.1364.9.

Murphy, S.M., Vidal, M.C., Hallagan, C.J., Broder, E.D., Barnes, E.E., Horna Lowell, E.S. and Wilson, J.D. (2019) 'Does this title bug (Hemiptera) you? How to write a title that increases your citations', *Ecological entomology*, 44(5), 593-600, available: http://dx.doi.org/10.1111/een.12740.

Narvaez-Berthelemot, N. and Russell, J.M. (2001) 'World distribution of social science journals : A view from the periphery', *Scientometrics*, 51(Conference Proceedings), 223-239, available: http://dx.doi.org/10.1023/A:1010581131779.

Nelson, M. (2010) 'Building a Written Corpus' in O'Keeffe, A. and McCarthy, M., eds., *The Routledge Handbook of Corpus Linguistics* Routledge, 53-65.

Nesi, H. (2011) 'BAWE: An Introduction to a new resource' in Frankenberg-Garcia, A., Fowerdew, L. and Aston, G., eds., *New Trends in Corpora and Language Learning*, London: Continuum, 213-228.

Nesi, H. (2021) 'Sources for courses: Metadiscourse and the role of citation in student writing', *Lingua*, 253, 103040, available: http://dx.doi.org/10.1016/j.lingua.2021.103040.

Nesi, H. and Gardner, S. (2018) 'The BAWE corpus and genre families classification of assessed student writing', *Assessing Writing*, 38, 51-55, available: http://dx.doi.org/10.1016/j.asw.2018.06.005.

Nikolaev, S., Sukhomlinova, M. and Nikolaeva, S. (2021) 'Systemic genre-and-style relations in the english-language student academic essay: paradigmatics, sintagmatics, epidigmatics', *E3S Web of Conferences*, 273, 12165, available: http://dx.doi.org/10.1051/e3sconf/202127312165.

Norhaida, A. and Tan, L. (2018) *The Nuts and Bolts of English Grammar*, SG, SINGAPORE: Marshall Cavendish International (Asia) Private Limited.

Northedge, A. (2005) *The good study guide*, Open University Press.

Nunan, D. (2008) 'Exploring genre and register in contemporary English', *English Today*, 24, 56-61, available: http://dx.doi.org/10.1017/S0266078408000217.

Nunberg, G., Briscoe, T. and Huddleston, R. (2002) 'Punctuation' in Huddleston, R. and Pullum, G., eds., *The Cambridge Grammar of the English Language*, Cambridge: Cambridge UP, 1723-1764.

O'Connor, P.E. (1994) "You could feel it through the skin': Agency and positioning in prisoners' stabbing stories', *Text*, 1, 45-75, available: http://dx.doi.org/https://doi.org/10.1515/text.1.1994.14.1.45.

O'Donnell, R. (1974) 'Syntactic Differences between Speech and Writing', *American Speech*, 49, 102-110.

O'Keeffe, A.M.M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, 1 edition ed., Cambridge ; New York: Cambridge University Press.

O'Keeffe, A.M.M.a.C.R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, 1 edition ed., Cambridge ; New York: Cambridge University Press.

Ortega, L. (2015) 'Syntactic complexity in L2 writing: Progress and expansion', *Journal of Second Language Writing*, 29, 82-94.

Paltridge, B. (1995) 'Working with genre: A pragmatic perspective', *Journal of Pragmatics*, 24, 393-406.

Paltridge, B. (1996) 'Genre, text type, and the language classroom', *ELT Journal*, 50(3), 237-243.

Paltridge, B. (2004) 'Academic writing', *Language Teaching*, 37, 87-105, available: http://dx.doi.org/10.1017/S0261444804002216.

Parkes, M.B. (1993) *Pause and Effect: An Introduction to the History of Punctuation in the West*, available: https://read.amazon.com/?asin=B01MXUWK05&ref_=kwl_kr_iv_rec_1.

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.

Perez del Aguila, R. (2014) 'Exploring the best ways to support first year university students' academic writing skills', *RIDU*, 8(1), 112-124.

Pinker, S. (2010) *The language instinct*, Harper Collins e-books, available.

Predelli, S. (2003) 'Scare Quotes and Their Relation to Other Semantic Issues', *Linguistics and philosophy*, 26(1), 1-28, available: http://dx.doi.org/10.1023/A:1022278209949.

Profile, E. (2011) *The English Profile booklet*, available: https://www.englishprofile.org/images/pdf/theenglishprofilebooklet.pdf [accessed June 20].

Quirk, R., Greenbaum, S., Leech, G. and Svartik, J. (1985) *A Comprehensive Grammar of the English Language*, New York: Longman.

Ramoroka, B.T. (2017) 'The use of interactional metadiscourse features to present a textual voice: A case study of undergraduate writing in two departments at the University of Botswana', *Journal of the Reading Association of South Africa*, 5, 1-11, available: http://dx.doi.org/https://rw.org.za/index.php/rw/article/view/128.

Rausova, V. (2018) 'Pragmatic functions of I in academic discourse: linguistic research articles', *Linguistica Pragensia*, 28(2), 168-183.

Rayson, P. (2015) 'Computational tools and methods for corpus compilation and analysis' in Biber, D. and Reppen, R., eds., *The Cambridge Handbook of English Corpus Linguistics* Cambridge University Press, 32-50.

Rippen, R. (2010) 'Building a Corpus' in O'Keeffe, A. and McCarthy, M., eds., *The Routledge Handbook of Corpus Linguistics* Routledge, 31-37.

Rose, H., Curle, S., Aizawa, I. and Thompson, G. (2020) 'What drives success in English medium taught courses? The interplay between language proficiency, academic skills, and motivation', *Studies in higher education (Dorchester-on-Thames)*, 45(11), 2149-2161, available: http://dx.doi.org/10.1080/03075079.2019.1590690.

Rychlý, P. (2008) 'A lexicographer-friendly association score', *RASLAN*, available: https://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf.

Rys, J.V.M., Verne; VanderMey, Randall; Sebranek, Patrick (2019) *The College Writer: A Guide to Thinking, Writing, and Researching*, 6th ed., Cengage.

Say, B. and Akman, V. (1996) 'Current Approaches to Punctuation in Computational Linguistics', *Computers and the humanities*, 30(6), 457-469, available: http://dx.doi.org/10.1007/BF00057941.

Schou, K. (2007) 'The syntactic status of English punctuation', *English Studies*, 88(2), 195-216, available: http://dx.doi.org/10.1080/00138380601042790.

Scott, M. and Tribble, C. (2006) *Textual Patterns: Key Words and Corpus Analysis in Language Education*, Amsterdam: John Benjamins.

Seidlhofer, B. (2005) 'English as a lingua franca', *ELT Journal*, 59(4), 339-341, available: http://dx.doi.org/10.1093/elt/cci064.

Sinclair, J. (2003) *Reading Concordances*, Pearson Education: Pearson Education.

Sinclair, J. (2005) 'Corpus and Text: Basic Principles' in Wynne, M., ed., *Developing Linguistic Corpora: a Guide to Good Practice* Oxford:Oxbow Books, 1-16.

Sinclair, J. and Carter, R. (2004) *Trust the Text: Language, Corpus and Discourse*, Taylor & Francis Group, available: https://ebookcentral.proquest.com/lib/mic-ebooks/detail.action?docID=200411.

Sketch Engine (2020) *Sketch Engine | language corpus management and query system*, available: https://www.sketchengine.eu/ [accessed

Sketch Engine (2022) 'Sketch Engine Support'.

Staples, S., Egbert, J., Biber, D. and and Gray, B. (2016) 'Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre', *Written Communication*, 33(2), 149-183, available: http://dx.doi.org/10.1177/0741088316631527.

Staples, S. and Reppen, R. (2016) 'Understanding first-year L2 writing: A lexico-grammaticalanalysis across L1s, genres, and language ratings', *Journal of Second Language Writing*, 32, 17-35.

Steen, G. (1999) 'Genres of discourse and the definition of literature', *Discourse Processes*, 28(2), 109-120, available: http://dx.doi.org/10.1080/01638539909545075.

Stirling, L. and Manderson, L. (2011) 'About you: Empathy, objectivity and authority', *Journal of Pragmatics*, 43, 1581-1602, available.

Strikwerda, C. (2019) 'Faculty members are the key to solving the retention challenge', *Inside Higher Ed*.

Strunk, W. and White, E.B. (1999) *The elements of style*, 4th ed., London, Boston: Allyn & Bacon [Pearson].

Stubbs, M. (1996) *Text and corpus analysis: Computer assisted studies of language and culture*, Oxford, UK: Clarendon.

Sun, K. and Wang, R. (2019) 'Frequency distributions of punctuation marks in English', *English Today*, 35(4), 23-35, available: http://dx.doi.org/10.1017/S0266078418000512.

Swales, J. (1990) *Genre Analysis: English in academic and research settings*, Cambridge, UK: Cambridge University Press.

Swales, J. (1997) 'English as Tyrannosaurus rex', *World Englishes*, 16, 373-383, available: http://dx.doi.org/10.1111/1467-971X.00071.

Swales, J. (2019) 'The futures of EAP genre studies: A personal viewpoin', *Journal of English for Academic Purposes*, 38, 75-82.

Syrewicz, C.C. (2022) 'How do expert (creative) writers write? A literature review and a call for research', *New writing (Clevedon, England)*, 19(2), 196-224, available: http://dx.doi.org/10.1080/14790726.2021.2005631.

Tang, R. and John, S. (1999) 'The 'I' in identity: Exploring writer identity in student academic writing through the first person pronoun', *English for Specific Purposes*, 18, S24-S39, available: http://dx.doi.org/https://doi.org/10.1016/S0889-4906(99)00009-5.

Taylor, H. and Goodall, J. (2019) 'A preliminary investigation into the rhetorical function of 'I' in different genres of successful business student academic writing', *Journal of English for Academic Purposes*, 38, 135-145.

Testa, J. (2009) 'The Thomson Reuters Journal Selection Process', *Transnational corporations review*, 1(4), 59-66, available: http://dx.doi.org/10.1080/19186444.2009.11658213.

Thompson, G. (2001) 'Interaction in academic writing: Learning to argue with the reader', *Applied Linguistics*, 22(1), 58-78.

Thompson, G. (2014) *Introducing Functional Grammar*, available.

The TOEFL Family of Assessments (2020), available: https://www.ets.org/toefl [accessed

Trebits, A. (2009) 'Conjunctive cohesion in English language EU documents -- A corpus-based analysis and its implications', *English for Specific Purposes*, 28, 199-210.

Tribble, C. (2009) 'Writing academic English—a survey review of current published resources', *ELT Journal*, 63(4), 400-417, available: http://dx.doi.org/10.1093/elt/ccp073.

Tribble, C. (2015) 'Writing Academic English Further along the Road. What Is Happening Now in EAP Writing Instruction?', *ELT Journal*, 69(4), 442-462, available: http://dx.doi.org/10.1093/elt/ccv044.

Vande Kopple, W.J. (1985) 'Some explanatory discourse on metadiscourse', *College Composition and Communication*, 36, 82-93.

Vassileva, I. (1998) 'Who am I/who are we in academic writing? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian', *International Journal of Applied Linguistics*, 8(2), 163-190.

Vaughan, E. and Clancy, B. (2013) 'Small corpora and pragmatics', *Yearbook of Corpus Linguistics and Pragmatics*, 1(1), 53-73.

Vercellotti, M.L. and Packer, J. (2016) 'Shifting structural complexity: The production of clause types in speeches given by English for academic purposes students', *Journal of English for Academic Purposes*, 22.

Viana, V. and O'Boyle, A. (2022) *Corpus linguistics for English for academic purposes*, London: Routledge.

Wales, K. (1996) *Personal Pronouns in Present-day English*, Cambridge: University Press, available.

Wang, S.-p.T.W.-T. (2021) 'To We or Not to We: Corpus-Based Research on First-Person Pronoun Use in Abstracts and Conclusions', *SAGE Open*, 11(2), 1-18, available: http://dx.doi.org/10.1177/21582440211008893.

Wardle, E. (2009) '"Mutt Genres" and the Goal of FYC: Can We Help Students Write the Genres of the University? - ProQuest', *College Composition and Communication*, 60(4), 765-789.

Whicker, J.H. (2022) '"Types of Writing," Levels of Generality, and "What Transfers?": Upper-Level Students and the Transfer of First-Year Writing Knowledge', *Across the disciplines*, 18(3-4), 284-304, available: http://dx.doi.org/10.37514/ATD-J.2022.18.3-4.05.

Whitley, M.S. (1978) 'Person and Number in the Use of We, You, and They', *American Speech*, 53(1), available: https://www.jstor.org/stable/455337.

Whong, M. and Godfrey, J. (2020) *What is good academic writing?: insights into discipline-specific student writing*, 1st ed. ed., London [England]: Bloomsbury Academic.

Wood, D. (2015) *Cover Image Fundamentals of formulaic language: an introduction*, London: Bloomsbury Academic.

Xiao, R. (2015) 'Collocation' in Biber, D. and Reppen, R., eds., *English Corpus Linguistics*, Cambridge University Press, 106-124, available: https://www-cambridge-org.libraryproxy.mic.ul.ie/core/services/aop-cambridge-core/content/view/F3D394FD5B2D3952110F4AD8B8A52A9B/9781139764377c6_p106-124_CBO.pdf/collocation.pdf.

Yang, B., Tang, H. and Mou, L. (2021) 'Research on Higher English Internationalization Education Model and Evaluation Index System Based on Multi-Source Information Fusion', *Computational intelligence and neuroscience*, 2021, 1599007-8, available: http://dx.doi.org/10.1155/2021/1599007.

Zacharof, M.-P. and Charalambidou, A. (2018) 'An exploration of the sub-register of chemical engineering research papers published in English', *Publications*, 6(3), 1-19, available: http://dx.doi.org/10.3390/publications6030030.

Zijlmans, M., Otte, W.M., van't Klooster, M.A., van Diessen, E., Leijten, F.S.S. and Sander, J.W. (2015) 'Do clinicians use more question marks?', *JRSM open*, 6(5), 2054270415579027-2054270415579027, available: http://dx.doi.org/10.1177/2054270415579027.

Appendices

Appendix A: Supplementary Table

Table A.1: Comparative Word Frequencies in COMP 101, BAWE, and CAE

COMP 101				BAWE			CAE		
Number	**Item**	**Freq**	**N Freq**	**Item**	**R Freq**	**N Freq**	**Item**	**R Freq**	**N Freq**
1	the	10644	56562	the	492270	70646	the	225266	71205
2	,	9998	53129	,	391643	56205	,	162363	51321
3	.	9947	52858	.	313580	45002	of	126475	39978
4	to	6165	32760	of	271079	38903	.	125696	39731
5	and	5626	29896	and	208693	29950	to	89413	28263
6	of	4510	23966	to	191604	27497	and	88667	28027
7	a	4097	21771	in	153326	22004	in	77119	24377
8	in	3275	17403	a	136398	19575	a	64811	20486
9	is	3224	17132	is	111307	15974	is	47496	15013
10	I	2459	13067	)	91843	13181	that	42355	13388
11	that	2433	12929	(	90538	12993	)	38161	12062
12	it	2195	11664	that	79337	11386	(	38114	12047
13	for	1851	9836	'	72584	10417	'	33983	10742
14	are	1775	9432	as	68072	9769	as	32987	10427
15	they	1397	7424	for	59564	8548	for	27691	8753
16	with	1394	7408	be	58120	8341	this	27212	8601
17	my	1393	7402	this	54393	7806	be	24863	7859
18	"	1367	7264	it	51248	7355	"	24839	7851
19	you	1361	7232	"	47283	6786	with	22419	7086
20	was	1357	7211	:	47060	6754	it	22416	7085
21	on	1290	6855	are	42739	6134	are	19806	6260
22	be	1283	6818	with	42310	6072	by	19546	6178
23	not	1282	6812	on	40642	5833	on	17528	5540
24	have	1222	6494	by	40564	5821	was	17119	5411
25	as	1152	6122	was	36855	5289	not	16754	5296