KARADENIZ TECHNICAL UNIVERSITY * THE INSTITUTE OF SOCIAL SCIENCES

DEPARTMENT OF WESTERN LANGUAGES AND LITERATURE

MASTER'S PROGRAM IN APPLIED LINGUISTICS

CORPUS ANALYSIS OF P-FRAMES IN ENGINEERING EAP: THE CASE OF GEOLOGY AND MINING ENGINEERING

MASTER'S THESIS

Ayşenur HOŞOĞLU

SEPTEMBER - 2025

TRABZON

KARADENIZ TECHNICAL UNIVERSITY * THE INSTITUTE OF SOCIAL SCIENCES

DEPARTMENT OF WESTERN LANGUAGES AND LITERATURE

MASTER'S PROGRAM IN APPLIED LINGUISTICS

CORPUS ANALYSIS OF P-FRAMES IN ENGINEERING EAP: THE CASE OF GEOLOGY AND MINING ENGINEERING

MASTER'S THESIS

Ayşenur HOŞOĞLU

Supervisor: Assoc. Prof. Dr. Ali Şükrü ÖZBAY

SEPTEMBER - 2025

TRABZON

APPROVAL

Upon the submission of the dissertation, **Ayşenur HOŞOĞLU** has defended the study **"Corpus Analysis of P-Frames in Engineering EAP: The Case of Geology and Mining Engineering "** in partial fulfilment of the requirements for the degree of Master of Arts in Applied Linguistics at Karadeniz Technical University, and the study has been found fully adequate in scope and quality as a thesis by **unanimous / majority** vote on **30/10/2025.**

Committee Member		**Decision**		**Signature**
Title - Name and SURNAME	**Mission**	**Accept**	**Refuse**	
Assoc. Prof. Dr. Ali Şükrü ÖZBAY	Supervisor	☐	☐	
Prof. Dr. Ceyhun YÜKSELİR	Member	☐	☐	
Assoc. Prof. Dr. Nazan YILDIZ ÇİÇEKÇİ	Member	☐	☐	

Approval of the Graduate School of Social Sciences.

Prof. Dr. Haydar AKYAZI
Director

DECLARATION OF ORIGINALITY

I, **Ayşenur HOŞOĞLU**, hereby confirm and certify that;

- I am the sole author of this work and I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution,
- this work contains no material that has been submitted or accepted for a degree or diploma in any other university or institution,
- no data or finding in the work has been falsified or embellished,
- this is a true copy of the work approved by my advisor and thesis committee at Karadeniz Technical University, including final revisions required by them.
- I understand that my work may be electronically checked for plagiarism by the use of plagiarism detection software and stored on a third party's server for eventual future comparison,
- I take full responsibility in case of non-compliance with the above statements in respect of this work.

Signature……..
Date: 30/10/2025

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Ali Şükrü ÖZBAY, for his endless support, patience, motivation, enthusiasm, and wisdom. His guidance helps me, and his advice sheds light on my path. He has always been there to help me in any situation and encourage me to become a good researcher.

I want to thank Prof. Dr. Ceyhun YÜKSELİR for accepting to be a jury member of my thesis and his valuable comments about the design and content of my thesis. I also want to thank Assoc. Prof. Dr. Nazan YILDIZ ÇİÇEKÇİ for being one of my jury members and sharing his invaluable comments and questions about the study.

I would like to express my deepest and most sincere gratitude to my beloved family for their moral and material support throughout my academic journey. Words are not enough to express my appreciation for everything they have done for me. My heartfelt thanks go to my father, Mustafa Sabri HOŞOĞLU, my mother, Sevim HOŞOĞLU, and my brother, Ömer HOŞOĞLU, for their constant encouragement and belief in me. A very special thank you to my dear husband, Ali ÇELİK, for his academic and emotional support. His love, patience, and belief in me have been the driving force behind the completion of this thesis. He is the best person I have ever known, and I am deeply grateful for his presence.

I dedicate this thesis to my family, to whom I owe all my academic achievements and the person I have become.

September, 2025 Ayşenur HOŞOĞLU

TABLE OF CONTENTS

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

ÖZET

Akademik Amaçlı İngilizce yazımında, disipline özgü söylemi şekillendirmeye katkı sağlayan yinelenen çok sözcüklü kalıplara sıklıkla başvurulmaktadır. Bu kalıplar arasında ifade çerçevelerini akademik metinlerin deyimsel yapısında önemli bir rol oynadığı gösterilmiştir. Bu çalışma, mühendislik akademik söylemine nasıl katkı sağladıklarını incelemek amacıyla Jeoloji ve Madencilik Mühendisliği araştırma makalelerinde kullanılan ifade çerçevelerini ve bunların dağılımını araştırmayı amaçlamaktadır. Bu çalışma için her biri 133 araştırma makalesinden oluşan iki disipline özgü derlem oluşturulmuştur. İfade çerçevelerini kfNgram aracı kullanılarak çıkarılmış ve her derlemdeki en sık geçen 100 ifade çerçevesi analiz edilmiştir. Disiplinler arasında ortak ve alana özgü ifade çerçevelerini belirlemek için sıklık ve log-likelihood analizleri yapılmış; ifade çerçevelerinin sözcüksel görünümleri ve retorik işlevlerini incelemek amacıyla ek olarak bağlam ve eşdizim analizleri uygulanmıştır. Elde edilen sonuçlar, iki derlemin birkaç yüksek frekanslı ifade çerçeveleri paylaştığını ve bunun da disiplinler arası genel akademik yapılara işaret ettiğini ortaya koymuştur. Ancak bu kalıpları dolduran sözcüklerin çoğunlukla farklı olduğu ve her alanın kendine özgü kavramsal odağını ve terminolojisini yansıttığı görülmüştür. Ayrıca bazı ifade çerçevelerinin sonuç bildirme ve neden-sonuç ilişkilerini ifade etme gibi temel retorik işlevlerle yakından ilişkili olduğu bulunmuştur. Bu bulgular, Akademik Amaçlı İngilizce bağlamlarında disipline özgü deyimsel kalıplara ilişkin farkındalık geliştirmede veri güdümlü öğrenme yaklaşımlarının potansiyelini vurgulamaktadır. Sketch Engine ve kfNgram gibi derlem araçlarının, öğrencilerin daha doğru ve disipline uygun akademik yazma becerileri geliştirmelerini desteklemek için nasıl kullanılabileceğini göstermek amacıyla DDL temelli sınıf etkinliği önerileri de sunulmuştur.

Anahtar Kelimeler: Akademik Amaçlı İngilizce, Derlem Dilbilim, İfade Çerçevesi, Mühendislik, Veri Güdümlü Öğrenme

ABSTRACT

Academic writing in English for Academic Purposes (EAP) often relies on recurrent multi-word patterns, which contribute to shaping discipline-specific discourse. Among these patterns, P-frames have been shown to play a crucial role in the phraseological structure of academic texts. This study aims to investigate the use and distribution of P-frames in Geology and Mining Engineering research articles to explore how they contribute to engineering academic discourse. Two discipline-specific corpora, each consisting of 133 research articles, were compiled for this study. KfNgram was used to obtain the top 500 P-frames. Among them, the top 100 P-frames were analysed. Frequency and log-likelihood analyses were conducted to identify shared and discipline-specific P-frames, while concordance and collocation analyses were used to examine their lexical realisations and rhetorical functions. The results demonstrated that both corpora have shared high-frequency P-frames, which refers to the existence of general academic language across disciplines. However, when it comes to the types of filling these frames, they were generally different, demonstrating the unique terminology of each discipline. In addition, it has been found that some P-frames are closely related to general academic language such as reporting results and expressing cause–effect relations. Furthermore, the findings highlight the potential use of corpus-based data-driven learning (DDL) to raise students' awareness about discipline-specific phraseological patterns in EAP engineering contexts. Suggested DDL-based classroom activities are provided to demonstrate how corpus tools such as Sketch Engine and kfNgram can be used to support students in developing more accurate and discipline-appropriate academic writing skills.

Keywords: Corpus Linguistics, Data-driven Learning, EAP, Engineering, P-frames

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CL	: Corpus Linguistics
DDL	: Data–Driven Learning
EAP	: English For Academic Purposes
ELT	: English Language Teaching
Freq.	: Frequency
LL	: Log-Likelihood

CHAPTER ONE

1. OVERVIEW OF THE STUDY

1.1 Introduction

"I took the one less travelled by, and that has made all the difference."

Robert Frost

Technology has gained considerable attention over the last decade, resembling a growing avalanche that accelerates daily. These advancements have had a profound impact on various sectors, including education. In particular, language education has greatly benefited from technological developments. Advancements such as Web 2.0 tools, computers, and online/offline applications have been increasingly integrated into language classroom settings, which is transforming traditional approaches to teaching and learning. Among the innovative methods that benefit from technological progress, Corpus Linguistics (CL) has been used in language teaching. The growing interest in corpus linguistics (CL) and the increasing use of corpus tools within the field of linguistics have contributed to the rising popularity of corpus-based teaching methods. A large body of electronically stored texts directs researchers, lexicographers, terminologists, translators, and language instructors to sympathise with the use of corpus (Kunilovskaya & Koviazina, 2018). Moreover, the nature of CL can be conceptualised as an independent field of study and as a methodological tool that enhances and supports research in various disciplines (McCarthy & O'Keeffe, 2010). In other words, the flexible nature of corpus linguistics enables it to be applied across all fields.

In today's globalised academic environment, proficiency in English for Academic Purposes (Hereafter, EAP) is essential for students in different disciplines. In the academic setting of engineering disciplines that rely heavily on technical reports, research papers and academic interaction, proficiency in academic English is essential for academic success and effective professional practice in the field. However, many resources are designed with a general language perspective and often ignore the specific linguistic needs of engineering students. Therefore, a need has emerged for more focused and detailed academic language in an engineering context. In this respect, corpus-based data-driven learning plays a significant role in providing various tools and methods for learners to engage them with authentic and discipline-specific academic language. One

of the main characteristics of corpus linguistics is the ability to enable users to analyse language more profoundly. With this, engineering students have a chance to access their specific needs in their academic journey. Particularly, vocabulary knowledge, which can also be referred to as mastery of the language, is a significant part of the process. Especially, the majority of EAP learners need to produce discipline-specific written work. Therefore, it is essential for learners to be a master in technical terms in their field, as well as to raise awareness of vocabulary and frequently occurring multi-word expressions. In this context, Data-Driven Learning (DDL), a subfield of corpus-based approaches, has emerged as a potential to increase learner autonomy and raise language awareness through direct interaction with authentic language data. DDL supports learners by encouraging them to discover linguistic patterns, engage with real-life language use, and take an active role in their own learning process. As noted by Gilquin and Granger (2010), by using DDL activities, students gain three significant abilities, which are encouraged to observe corpus data, make a hypothesis and formulate rules" (p. 359). These abilities create an environment for EAP learners not only to enhance their linguistic competence but also to improve critical thinking and independent learning, which are crucial in academic contexts.

Overall, it can be suggested that the integration of corpus linguistics into English for Academic Purposes (EAP) through Data-Driven Learning (DDL) provides valuable insights for authentic language use and academic language awareness. Within the scope of EAP, especially in the case of engineering learners, they need to improve their academic and discipline-specific language. With this respect, this thesis aims to explore phrase frames (P-frames) by using corpus linguistics methods. Regarding this, corpus-based DDL applications can raise learners' language awareness and enable them to analyse discipline-specific language by themselves.

1.2 The Framework of the Study

Throughout history, language learning has been a complex and challenging process, often requiring considerable effort for learners to develop successful language skills. Over time, however, significant advancements have been made in the field of language education. Among these developments, technology and its associated tools have emerged as some of the most influential innovations, transforming both the methods and accessibility of language learning. In the field of linguistics, corpus linguistics can be among the new techniques.

Corpus methods can serve various purposes. In particular, the discovery of patterns in lexis, grammar, semantics, and pragmatics can be a leading example of areas in which corpus methods can be applied. Regarding this statement, Boulton (2010) mentioned that the potential integration of electronic corpora in language teaching and learning has gained growing interest in recent years. Furthermore, the only things that can be searched with a corpus are not limited to them. Researchers,

teachers, and students can also use the corpus to check work frequency, keywords, N-grams, P-Frames, collocations, and keywords in context (KWIC). In a nutshell, Flowerdew (2009) made clear the potential of the corpus as "the combination of frequency data regarding all the words in a corpus and the verbal environment in which these words occur." (p.328). Together with this explanation, he added that thanks to the corpus, researchers can reach out with a click to the nature of typical patterns of lexical and grammatical use.

The integration of corpus linguistics into English for Academic Purposes (EAP) has significantly enhanced our understanding of authentic academic language use and enabled more targeted and practical instruction for discipline-specific learners. In other words, corpus linguistics can be a beneficial method in English for Academic Purposes (EAP), where learners are required to master academic conventions, discipline-specific vocabulary, and formal language structures. In many respects, learners must be equipped not only with academic writing skills but also with discipline-specific vocabulary to engage with their field of study effectively. In this regard, as Gillett (2022) states, the purpose of an EAP course is to help learners acquire linguistic and cultural knowledge within an institutional and disciplinary context which is relevant to their studying or working environment, where English is the primary language.

Corpus Linguistics can be practical when integrated with Data-Driven Learning (DDL). Moreover, DDL aims to expose learners to authentic language use through corpora and encourage them to take an active role in their language learning process. One of the most distinctive features of DDL is its ability to integrate corpus data in the creation of language learning materials. By interacting directly with real language data, learners have an opportunity to notice patterns, formulate hypotheses, and develop a deeper understanding of how language functions in authentic contexts. In this regard, corpus-based DDL enables EAP learners to explore authentic academic texts that are related to their field and become familiar with real usage patterns in their academic disciplines.

1.3 Background of the Study

In recent decades, education has greatly benefited from the use of technology. With the increasing availability and affordability of technological devices, along with easy access to the internet, new opportunities have emerged for researchers and teachers to integrate technology into their work. As Flowerdew (2009) emphasised, with digital literacy, people become more open to new methods. The integration of technology, in turn, brings innovation to education and paves the way for new teaching methods. With increasing pressure on non-native writers to publish in international journals, it becomes essential to propose effective methodologies and strategies to address the linguistic challenges encountered during the publication process (Abdelmageed, 2024). This is particularly essential for students of English for Academic Purposes (EAP). Discipline-specific

standardisation becomes necessary for students in their field to express themselves effectively in the academic world. In this context, corpus linguistics emerges as a supportive tool for teaching foreign languages. Recent improvements in the availability and quality of user-friendly, freely downloadable corpus tools, supported by Data-driven learning (Hereafter, DDL), have turned out to be significant assets in promoting pedagogical approaches for EAP learners.

To draw a clear frame, the definition of the corpus can be stated as "the study of language based on large collections of 'real life' language use" (Nordquist, 2019, para.1). The main aspect to emphasise here is focusing on 'real-life use.' Corpus provides authentic language data, making it valuable not only for linguists and lexicographers but also for teachers and students. Teachers become more aware of the evident benefits of the corpus. From Boulton's (2010) perspective, thanks to the promising use of corpora in education, there is an increased interest in language teaching and learning. It offers new paths in language learning, contrary to stereotyped and conventional approaches. For example, most textbooks and dictionaries rely on fabricated and invented language examples that often encourage learners to memorise patterns, sentences, and words that are rarely used. On the other hand, the corpus provides examples from authentic resources and real-life contexts, creating an environment where learners can actively explore and engage with the language itself. This issue is significant for EAP learners, as they need strong proficiency in discipline-specific vocabulary to be successful in their academic studies. In this case, Corpus tools perform an essential role in identifying common lexical and grammatical patterns in academic texts, allowing instructors and students to focus on the language most relevant to their discipline. The process of learning and using English is shaped mainly by the learner's specific goals and communicative needs. In other words, individuals learn English not only for general proficiency but also for particular purposes, such as academic study, professional communication, or specialised fields of work. English for Academic Purposes (EAP) is one of the key areas in which students seek to develop proficiency, as it equips them with the language skills necessary for academic success. English for Academic Purposes (EAP) is a teaching and learning approach specifically designed to meet the academic needs of students in higher education. For instance, Hamp-Lyons (2011) pointed out that EAP has a "focus on teaching English specifically to facilitate learners' study or research through the medium of English" (p. 89). It places particular emphasis on academic vocabulary, critical thinking, and the conventions of academic discourse across disciplines.

Building on this perspective, recent corpus-based research has paid increasing attention to a specific type of recurrent multi-word sequence known as p-frames. In other words, even the results can vary, the majority of studies have highlighted that a substantial portion of language consists of fixed or semi-fixed multi-word combinations, which speakers and writers utilise to achieve communicative goals in diverse contexts (Römer, 2009). P-frames are structurally recurrent patterns with variable slots that serve as functional units in discourse. In the context of English for Academic Purposes (EAP), especially within engineering disciplines, p-frames are particularly valuable as they

help identify the formulaic structures commonly used to express stance, describe processes, present results, and structure arguments. Thus, examining p-frames provides insights not only into phraseological tendencies of academic texts but also into how language is shaped by disciplinary field and communicative needs.

The basic definition of DDL, as defined by Crosthwaite et al. (2021), is "the use of corpora for the purposes of language teaching and learning" (p. 1392). In other words, the DDL approach is based on the principle that "evidence is needed to answer the learner's questions and relies on the learner's intelligence to find answers" (John, 1991, p. 2). Ma et al. (2024) suggested that the main emphasis of DDL lies in language learners independently exploring corpora, rather than on the pedagogical application by teachers. Crosthwaite and Baisa (2024) emphasised that most corpus tools are commonly used for corpus-based data-driven learning, primarily designed for research rather than teaching. They added that much of the research on DDL reveals that students and teachers often fail to use DDL after the initial training due to usability issues with the tools. In this section, the general definitions and theoretical background of Data-Driven Learning are explained, with further details provided in the literature review section. As mentioned earlier, advancements in technology have opened new paths for the field of linguistics, and a range of methods for studying language have emerged. DDL is one of the innovations that contribute to language studies as a subfield of corpus linguistics. In the context of English for Academic Purposes (EAP), DDL offers significant advantages by allowing learners to access and analyse discipline-specific language used in authentic academic texts. This hands-on engagement with real academic discourse helps EAP learners develop a deeper understanding of field-specific vocabulary, collocations, and rhetorical patterns—skills that are crucial for success in academic writing and research. Therefore, the integration of corpus-based DDL into EAP instruction not only supports language acquisition but also empowers students to become more independent and critical users of academic English.

1.4 Purpose of the Study

This study aims to examine the use and distribution of P-frames in Engineering English for Academic Purposes (EAP) texts, explicitly focusing on research articles written in the fields of Geological Engineering and Mining Engineering. By comparing these two sub-disciplines, the study seeks to reveal both the shared and discipline-specific patterns of P-frames and to explore how these patterns contribute to the construction of academic discourse in engineering. Adopting a corpus-based approach, this study also aims to discuss the potential of integrating Data-Driven Learning (DDL) as a pedagogical method to help engineering students become more aware of such phraseological patterns in their academic writing.

1.5 Research Questions

1. To what extent do Geology and Mining Engineering research articles share high-frequency P-frames, and which P-frames occur significantly more often in one discipline than the other?
2. How can user-friendly corpus tools be used to raise engineering students' awareness of discipline-specific phraseological patterns in academic writing?
3. What rhetorical functions (e.g., reporting results, showing cause–effect relations) are commonly associated with the most frequent P-frames, and how can these be taught through data-driven learning (DDL) activities?

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Introduction

Academic writing in English for Academic Purposes (EAP) has increasingly drawn attention to the role of formulaic language and recurrent multiword patterns in shaping disciplinary discourse. Research based on corpus linguistics reveals that fixed and semi-fixed multi-word expressions comprise a considerable part of the English phraseology (Garner, 2016). In this regard, the study of phraseological patterns, particularly P-frames, offers valuable insights into how discipline-specific discourse is constructed. Moreover, there have been supporters of corpus linguistics as "the emergence of a new view of language" (Sinclair, 1991, p. 1). Therefore, analysing language through corpus-based approaches is considered essential for understanding how academic discourse operates.

2.2 Corpus Linguistics and Corpora

The field of Corpus Linguistics (CL) has attracted academic interest in recent decades. As technology has advanced, CL has gained importance, providing valuable insights into the use of spoken and written language in various contexts and fields. Starting in the mid-1980s, language corpora have seen great potential in computer-assisted language learning, research and teaching (Cotos, 2014). Similarly, since 1990, computers and concordance tools have become increasingly popular for language research as well as pedagogy due to their growing interest (Bao, 2021). Contrary to popular belief, corpus linguistics has existed in paper-based collections, not in electronic form, for nearly a century. Xu (2022) noted that "the utilization of corpus-based quantitative methods relying on a large body of natural texts dates back to the early 19th century" (p.12). In the past, lexicographers consistently attempted to compile written texts of language and store these texts. Since then, technological advances have considerably simplified the process of compiling and analysing extensive collections of natural language texts, which leads to potential for both research and educational purposes (Ordoñana, 2023). However, in the modern world, language samples are now often stored digitally. In other words, thanks to the widespread use of the internet, corpus linguistics is going beyond its traditional boundaries.

In the relevant literature, the corpus can be defined based on a variety of criteria. A basic definition of corpus is a large amount of text. In other words, corpus linguistics is a way of studying the language by using the vast amount of collected texts (plural 'corpora'), as Rühlemann (2019)

noted, “texts ranging from, say, 100,000 words to trillions of words.” (p.1). Therefore, there are, in fact, no maximum boundaries for the number of words in a corpus. For instance, the first electronic corpus, known as the Brown Corpus, was created in 1961 at Brown University in the United States and consisted of approximately one million words. At the time, the size of the Brown Corpus was considered remarkable as it was the first of its kind. However, the sizes of corpora, such as the British National Corpus and the Bank of English, now exceed millions of words. Similarly, Bennet (2010) pointed out that a corpus is “a large, principled collection of naturally occurring examples of language stored electronically” (p. 2). Moreover, as Leech (2007) highlighted, “continuing growth in the number and extent of text archives and other text resources” gives rise to the improvement in corpus linguistics (p.133). In this respect, it is worth noting that various types of corpora have emerged over the years.

Building a corpus involves a series of carefully planned steps to ensure the collected data is both relevant and reliable for linguistic analysis. Contrary to common belief, simply collecting large amounts of text to build a corpus is not enough. Specific steps and design criteria need to be followed. Firstly, one of the cornerstones of the corpus is that it should consist of “naturally occurring language rather than ‘edited’ language” (Rühlemann, 2019, p.1). Natural language data is the key element of a corpus. For instance, in the case of some corpora used to provide examples of how native speakers use a given language, real-life language use is necessary. Moreover, edited language is standard, as almost all dictionaries, textbooks, and educational resources contain edited language. Therefore, deciding the purpose of a corpus in the first step is significant, as a natural or authentic language corpus differs from these resources. In this sense, edited language can mislead individuals. Another essential element of building a corpus is representativeness. The corpus should be representative of that language to provide authentic data. At that point, Leech (2007) noted that researchers often prioritise representatives and claim them as a key to the corpus. Additionally, he stated that he identified representation as the holy grail, an ideal that is challenging to attain, yet he insisted that it is worth pursuing. However, many researchers overlook the importance of including representative samples during the text collection process, which can lead to biased or misleading results in the corpus analysis later.

It is essential to acknowledge that corpora have made substantial contributions to the field of language studies today. For instance, one of the most tremendous benefits of corpora is that they enable learners to conduct more systematic and evidence-based research in the field of linguistics as well as “how language is used today and how that language is used in different contexts, enabling us to teach language more effectively” (Bennett, 2010, p.7). This feature of a corpus is closely related to authenticity. A corpus provides individuals with a unique language used in a natural context without any editing process. In this way, people can observe and understand the meaning of the language and learn more about its use. Furthermore, through a corpus, individuals can gain insights

into aspects such as frequency, register and language use, as found, for example, in exercises involving adverbs of degree.

While it is essential to understand what corpus linguistics is, it is equally important to be aware of what corpus linguistics is not. Bennett (2010) drew attention to this issue by pointing out three elements of corpus linguistics that it is not. The first one, "corpus linguistics is not able to provide negative evidence", refers to the fact that the corpus cannot show what is true or what is wrong. The only information provided by a corpus is what is present in that corpus. The second one, "corpus linguistics is not able to explain why," he points out that corpus can only say "why something is the way it is, only tells us what it is. (p.3). The last one "corpus linguistics is not able to provide all possible language at one time" emphasises features of representatives of the corpus. Even if a corpus consists of many words, that does not mean that it surely represents the language itself.

2.3 Types of Corpora

As mentioned above, a corpus is an electronically stored text that helps researchers explore language and find answers to questions that might be difficult to see based on intuition. Each corpus has its speciality and is used for a specific aim. Therefore, there are various types of corpora in the field. According to Bennett (2010), a crucial consideration when selecting a corpus is asking the fundamental question: "What exactly should you look for?" (p. 12). This highlights the importance of having clear criteria and objectives during the corpus selection process. While numerous types of corpora exist, only the most well-known will be discussed in the following: generalised corpora, specialised corpora, learner corpora, pedagogic corpora, historical corpora, parallel corpora, comparable corpora, and dynamic corpora.

2.3.1 Generalised Corpora

Generalised corpora provide a general overview of the language. Primarily, it comprises a substantial number of words, including at least 10 million words. As stated, even though corpora cannot be present in every aspect of language, generalised corpora aim to provide people with as complete a picture of a language as possible (Bennett, 2010). This type of corpus consists of texts from newspapers, magazines, articles, fiction, non-fiction texts, and spoken data. In this respect, the Cambridge International Corpus, the British National Corpus (BNC), the Longman Spoken and Written Corpus, and the Corpus of Contemporary American English (COCA) can be good examples of generalised corpora.

2.3.2 Specialised Corpora

As it is clear from the name, specialised corpora encompass texts of a specific type and aim to be representative of the language. This type of corpus can be both large and small in size. The primary features of this type of corpus are to lead researchers to find answers to their specific questions. The Michigan Corpus of Academic Spoken English (MICASE) is a well-known example of a specialised corpus that contains spoken language in academic discourse.

2.3.3 Learner Corpora

Learner corpora are a type of specialised corpus that consists of written or spoken data from students. These students are mainly in the process of learning a language. This kind of corpus provides insight into learners' mutual errors or diversities. For instance, the International Corpus of Learner English (ICLE) is among the leading and popular learner corpora, which is a corpus of essay writing by upper-intermediate and advanced learners (Granger et al., 2020).

2.3.4 Pedagogic Corpora

The primary focus of the Pedagogic corpora is the language that is used in the classroom setting. This type of corpora can consist of textbooks, transcripts of classroom interactions, or any other written text or spoken data that students can be exposed to in the classroom (Bennett, 2010). Students can use a pedagogic corpus to learn specific language patterns and usage, as well as explore authentic language.

2.3.5 Dynamic Corpora

Dynamic corpora are a type of corpus that demonstrates differences in linguistic features within a language over a period. Regarding this kind of corpus, researchers can analyse how the use of language changed and what increases or decreases have been observed in word usage.

2.4 Corpus In ELT

Corpus linguistics has become an increasingly recognised approach in English Language Teaching (ELT) for providing data-driven insights into actual language use. The general framework of corpus linguistics is taken into consideration; it is noteworthy to mention that corpora can offer significant opportunities in the field of education. Advancements in technology are making it easier for everyone to access technology. Flowerdew (2009) emphasised that people become more open to new methods with digital literacy. In other words, researchers and teachers have slowly but gradually

discovered and realised the possible advantages of applying technology in the classroom. The corpus has a critical place in language teaching, providing a considerable amount of language data. Similarly, it was noted that the effects of the corpus have revolutionised both the field of linguistics and second language learning and teaching in recent decades (Römer, 2011).

As can be seen, the corpus provides numerous and diverse examples of language. To support this idea, it is noted that with the light of the corpus "the potential for finding out new facts about language is infinite" (Flowerdew, 2009, p. 329). It provides lexis, grammar, semantics, pragmatics, and textual features of the language, which are almost impossible and time-consuming to investigate one by one without a corpus. In this regard, the use of a corpus in teaching can be an effective way to enhance language learning. Specifically, the basic advantage and significance of the corpus is that it consists of 'real' language rather than ready-made, which provides an advantage in the field of language teaching. In addition to this, a corpus serves as a language that dictionaries or grammar books can hardly offer, as these sources have been deficient in fully describing the language. Therefore, the use of a corpus broadens the view of language knowledge with given examples of authentic language.

As previously mentioned, a corpus can provide a wealth of highly beneficial insights for language pedagogy. Researchers and educators have integrated corpus linguistics into various areas, including language description, curriculum design, and teaching methodology, making it a valuable resource for enhancing English language teaching practices. For instance, using corpora in the design of grammar books, dictionaries, course books, testing materials, and other educational resources is common (Bessidhoum & Ghania, 2021). To emphasise this point, it can be highlighted that the corpus provides detailed information about the nature and frequency of words, multi-word phrases, collocations, grammatical patterns, semantic and pragmatic features, and textual characteristics. Awareness of these properties and prior knowledge of them enable teachers to determine which terms to teach and when to teach them, with guidance from input on related materials.

It is worth mentioning that some key issues should be taken into consideration when applying the corpus to the classroom setting. One of the significant issues is selecting an appropriate corpus for teaching. This is particularly among the most crucial issues. To support this, Leech (2007) emphasised that "the corpora which are easiest to compile are not necessarily those which are most useful for language learning purposes." (p. 18). In other words, while creating a corpus or using a ready-made corpus can sometimes be seen as attractive, this does not mean that these corpora are suitable for use in language education. Therefore, it is essential to remember that not all corpora are appropriate for students. Another vital issue is the differences among students. Teachers should also not ignore students' language level, age, and needs when selecting a corpus. The corpus should be within the scope of their necessities and respond to their needs. For instance, as Flowerdew (2009) suggested, "a model for young learners might be child language" (p.336). Moreover, another issue is

the authenticity of corpus data and the level of the corpus. As stated in the learner needs, at some point, maybe students need more simplified language examples rather than complex ones.

Although interest in corpus linguistics has grown steadily, teachers and educators have shown a weakness in the practical application of corpora in classroom settings. While the benefits of corpus linguistics are almost undeniable, there is “still a lack of awareness of corpora and, in some cases, resistance toward corpora from students, teachers, and materials writers” (Römer, 2011, p. 206). Corpus linguistics has some limitations, such as a lack of types of equipment or a corpus that can be purchased. As Flowerdew (2009) noted, it is a bare fact that the lack of technological equipment has caused slow improvement in corpus-based teaching and learning applications in the classroom environment.

2.5 Corpora for Teaching and Learning: Data-Driven Learning (DDL)

Corpus linguistics plays a central role in the development of Data-Driven Learning (DDL), an approach that encourages learners to explore authentic language data and discover linguistic patterns independently. Rather than relying solely on rules provided by textbooks or teachers, DDL promotes learner autonomy by giving students direct access to corpora and corpus tools. In the relevant literature, the rationale of DDL is explained with a variety of definitions. Firstly, it is worth noting that the term "Data-Driven Learning" was first used by Tim Johns, also known as the father of DDL. According to him, DDL is a way of learning “whose learning needs to be driven by access to linguistic data.” (Johns, 1991, p.2). In terms of this statement, it can be stated that the ‘linguistic data’ refers to the corpus. Another well-known definition of DDL is clarified by Crosthwaite et al. (2021), who define DDL as the utilisation of corpora in the language teaching and learning process. To support this explanation, Vincent and Nesi (2018) define DDL as a method of language teaching and learning that utilises corpora in the learning process, helping students explore authentic data and apply it to their learning. Overall, the purpose of DDL is to “provide authentic usage of language through internet-based corpora” (Çelik, 2011, p.276). As is obvious, these definitions mainly emphasise the importance of using authentic and large-scale data in the language learning process. In other words, it can be called using the corpus.

To emphasise the fundamental mission of the DDL, Johns (1998) created a metaphor, “Every student is a Sherlock Holmes” (p.101). This metaphor presents the general framework of DDL because it is well known that Sherlock Holmes is a brilliant detective who always tries to solve cases and figure out who was murdered and how the crime occurred. Namely, he is always on the way to discovering and investigating. In this regard, when DDL is taken into consideration, learners also become detectives in the language learning process, analysing the language itself.

In many ways, DDL approaches differ from traditional teaching and learning methods. These points can be stated in terms of several subcategories, such as DDL, which provides potentiality by discovering the language, serving authentic language, and fostering learner autonomy. To clarify these titles, DDL provides learners with opportunities to take an active role in their learning process and discover how language works by investigating thorough corpus data. In this sense, corpus data guides learners to authentic examples rather than invented ones. In comparison, learners discover the language through online corpus tools and become autonomous learners. In this regard, it can be claimed that DDL encourages and guides learners on how to be active in their learning, and as a result, students' awareness increases. This leads to "empowering learners and making them more autonomous outside the classroom" (Bessidhoum & Ghania, 2021, p. 594). In other words, students become conscious and active inside and outside the classroom.

Besides these definitions, there are various remarks about DDL, as it "consists in using the tools and techniques of corpus linguistics for pedagogical purposes" (Gilquin & Granger, 2010, p. 359). Moreover, they emphasise two significant elements of DDL: the corpus and a tool to explore the corpus. As stated previously, the importance of which corpus is used in this process is among the key points. Also, Chambers (2010) emphasised the importance of corpus-based sources, such as dictionaries and coursebooks; however, she notes that it is worth indicating that DDL is quite effective in learning, provides an active role for students in their knowledge, and helps them familiarise themselves with the corpus.

There are several significant advantages of DDL in the teaching of language. Among them, one of the considerable benefits of DDL for students is that it provides examples of language that are hard to find in conventional learning methods, such as dictionaries, grammar books, and textbooks. Even this issue is noted by Çelik (2011) as he firstly criticises the insufficiency of traditional dictionaries. Then he adds that conventional ways "often provide vague, limited, and artificial examples for each definition, which is insufficient to fully understand an unfamiliar word." (p. 275). Thus, it is evident that DDL is different from traditional methods in terms of teaching and support. What separates DDL from other methods can be listed in two main titles (Smart, 2014). The first one is that DDL authentic language is a fundamental source of the learning process. The second one is the corpus-based DDL activities that engage students in the discovery process and enable them to become autonomous learners.

To this point, Frankenberg-Garcia (2014) has commented that, thanks to corpora and the advanced technology in the field of pedagogical lexicography, learner dictionaries have comprised authentic language examples for more than a decade. Thus, it is noteworthy to mention that exposure to the language is a significant part of the language learning process (Boulton, 2017).

Undeniably, although DDL has significant aspects in teaching language, it also has some shortcomings and limited parts. These parts are titled by Gilquin and Granger (2010) as "the logistics, the teachers' point of view, the learners' point of view, and the content of DDL." (p.366). The logistics aspects of DDL are related to the availability of equipment such as computers, internet access, purchased corpora, and corpus tools. All these have costs and paying them at some point can be challenging. The second one concerns teachers' abilities. As mentioned previously, teachers sometimes lack digital literacy and cannot use corpus tools efficiently in classroom settings. Moreover, teachers need digital literacy skills, and students should have them. Otherwise, dealing with the corpus tool would be hard for them. Additionally, some students may struggle to find the correct answers, leading to a fallacy, which can create anxiety. Therefore, some researchers suggested that DDL can be a proper tool for a specific group of students. The last one is regarding the content of the DDL. This aspect of DDL has already been mentioned, but it would not be wrong to remind ourselves of it again. The corpus should meet the requirements; in other words, it should not include irrelevant data, and it should provide answers to the questions.

Another significant point of the DDL approach is that it changes the roles of language instructors and students. This transition is part of the learning and teaching process. As students' role changes from learner to explorer, teachers' role also changes "From knowledge-transmitter to consultant, guide, coach, and/or facilitator." (Chong, 2016, sec. 9). Furthermore, Johns (1991) described language instructors' role as "who has to learn to become a director and coordinator of student-initiated research." (p.3). In other words, students become explorers or detectives who search for answers while the teacher guides them.

2.6 The Pedagogic Potential of Data-Driven Learning

In recent years, corpus linguistics has had a significant influence on English Language Teaching (ELT), particularly through the Data-Driven Learning (DDL) approach. DDL emphasises learner autonomy and discovery by encouraging students to analyse authentic language data from corpora, rather than depending solely on pre-determined rules or teacher explanations. This method enables learners to observe real-world usage patterns, including frequent vocabulary, collocations, grammatical structures, and discourse features, thereby fostering a more realistic and contextual understanding of English. Regarding the language teaching perspective, Gilquin and Granger (2010) highlighted the DDL as "using the tools and techniques of corpus linguistics for pedagogical purposes" (p.359). Similarly, Johns (1991) noted about DDL, "simply provide the evidence needed to answer the learner's questions and rely on the learner's intelligence to find answers" (p. 2). In addition, Çelik (2011) noted that using concordance programs enables users to investigate language structures in authentic contexts through a corpus. Moreover, he emphasised the importance of corpus tools and noted that students are more engaged with language, and their learning skills improve

quickly with corpus tools. Likewise, Boulton (2017) reported that, contrary to the traditional method of directly asking instructors or using prepared sources, the DDL approach enables students to access an authentic corpus in terms of their needs and to explore the answers.

One of the most evident features of Data-Driven Learning (DDL) is its potential to raise learners' language awareness. Learners are encouraged to notice patterns and make discoveries by engaging directly with authentic language data. Thus, their linguistic knowledge becomes broadened, and their understanding of how language functions in authentic contexts is improved. The DDL method is also named as an advisory service that "is likely to lead to improved language awareness and noticing" (Boulton, 2017, p. 1). Furthermore, Crosthwaite et al. (2021) expanded the definition of DDL into two versions: the first involves teacher-printed concordances prepared before the classroom by the teacher using a corpus, and the second involves the direct use of a corpus by students to gain competence and awareness in the target language.

Several benefits of this approach were stated from the pedagogical part of the DDL (Gilquin & Granger, 2010). The first benefit refers to authenticity. According to them, DDL provides numerous natural language examples, creating an excellent opportunity for learners to learn a language from real-world examples. The second one is related to the corrective features of DDL. This feature should be clarified, as students have the opportunity to compare their writing with native or non-native corpora. With this implementation, students can identify areas of deficiency, such as misuse, overuse, and underuse. It is worth noting that this second feature can be suitable for advanced learners. The last benefit of DDL is interest discovery. The discovery can make the learning process more fun and motivating for students. For that matter, it can be said that this discovery suggests that the learning and active participation of learners give rise to lasting learning. Additionally, O'Sullivan (2007) emphasised the skills acquired by students through corpus consultation. These skills are "predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying" (p.277).

It is essential to note that integrating the DDL approach requires two fundamental prerequisites: a suitable corpus and a corpus tool. The selection of an appropriate corpus is vital to this process (Gilquin & Granger, 2010). Almost all types of corpora can be suitable for DDL, including written, spoken, learner, and monolingual corpora. However, it is worth emphasising that focusing on a specific corpus may provide more benefits for particular topics. Additionally, when selecting a corpus, teachers should consider their students' needs. For instance, in the wrong corpus, students may feel like strangers due to cultural differences and struggle to adapt to it. To refer to this issue, Braun (2006) noted that if a corpus does not work for students, then "the texts themselves usually remain an 'anonymous mass' to the learners." (p.26). Furthermore, only selecting a proper corpus does not meet the requirements. The corpus should be authentic and have strict design criteria.

Another significant aspect of DDL that should be considered is manipulation. Manipulation is a part of this process. While language instructors prepare materials, they can include specific examples or choose which text to use, the amount of text, or the size of the corpus (Gabrielatos, 2005). Moreover, at some point, this manipulation is also related to concordance results. At some point, the corpus can give complex results, which can be confusing for the students. Therefore, researchers argue whether these concordance results should be edited or shared as they are. For this, Gabrielatos (2005) stated that "This manipulation should be carried out with the understanding that the adapted samples are not good guides to the frequency of a language item." (p.18). Another statement made by Boulton (2009) is that editing concordance lines causes a lack of authenticity and does not provide students with accurate language samples, and "leaving them unprepared for the realities of the authentic language we are presumably preparing them for." (p. 89). However, Nesselhauf (2004) suggested that, at some point, editing concordance results should be conducted to prevent learners from learning incorrect information. Specifically, he studied this issue in learner corpora. He noted that sometimes results should be edited because "learners might even come to the conclusion that this construction is not possible in English." (p.144).

2.7 English for Academic Purposes (EAP)

Since English has established itself as the global language of science and technology, English for Academic Purposes (EAP) has become a crucial area of language education, particularly for students in disciplines that require practical academic skills. English for Academic Purposes (EAP) is becoming an indispensable component of language education for achieving professional and academic competence in English. EAP instruction helps learners become familiar with the genre-specific features, rhetorical structures, and specialised vocabulary that characterise academic texts in their discipline.

There are numerous reasons why individuals seek to learn English, including educational, general communicative, professional, and academic purposes. In recent years, corpus tools in English for Academic Purposes (EAP) research have become increasingly widespread due to the accessibility of corpus-based tools and online platforms that offer versatile functionalities for text analysis. This development has contributed significantly to popularising data-driven approaches in language teaching and research. EAP-oriented studies have benefited from these developments, using corpora to investigate academic discourse, identify discipline-specific patterns and improve teaching practices.

The role of instructors in EAP courses is equally essential. Although EAP encourages learner autonomy and places more responsibility on learners, the guidance and support provided by instructors remain highly important. First, EAP teachers "need to be aware of different learning

preferences and approaches" (Robinson, 1991, p. 5). Additionally, Klimova (2015) stated that teachers play a vital role in analysing students' needs and creating course instruction based on students' academic and disciplinary requirements. Moreover, Robinson (1991) highlighted that, after this process, teachers may not leave students alone and need to check students' improvements and analyse the students' present performance and knowledge.

One of the main components of EAP is the understanding of specialised technical vocabulary. Students in engineering and related fields need to engage with subject-specific texts and research, so they must acquire a solid knowledge of technical terminology. In this respect, Ha and Hyland (2017) emphasised the importance of tools that can provide technical assistance to EAP instructors, particularly in identifying and teaching specialised vocabulary. Thanks to this tool, it would be possible "to select words of appropriate technicality levels from the reference discipline-specific word lists according to students' needs and abilities" (p. 36).

2.8 The Integration of Corpus into EAP Classrooms

The integration of corpus linguistics into the English for Academic Purposes (EAP) classroom has brought about a truly radical change in the way academic language is taught and learned. With the increasing popularity of corpus tools and user-friendly interfaces, both teachers and students have the opportunity to reach out directly to authentic language data. This opportunity enables students to explore real examples of academic language, identify frequently used collocations, phrases, and discipline-specific patterns. Therefore, they can gain a deeper understanding of how language functions in academic contexts. In the literature, there have been many studies about integrating corpus linguistics into language education in terms of discipline-specific students' needs.

For instance, in the study of Malakhovskaya et al. (2021), which is titled "Teaching noun-phrase composition in EAP/ESP context", it has been pointed out that there is adequate focus on noun phrase structures in EAP/ESP teaching. The students who participated in the study were taught to master the use of NNCS based on the published articles in their disciplines. The result was that this method provided students with NNP patterns to be mastered, and EAP/ESP instructors needed to create new, discipline-specific, and updated language teaching materials.

Rudneva (2020) investigated how a corpus can be presented to non-linguist students to help them understand inherent semantics, go beyond the general superficial understanding of professional texts, and develop a deeper understanding of the structures and meanings to improve students' and teachers' overall linguistic competence. In her lesson, she utilised the free EcoLexicon online corpus. The study found that using corpus-based tools shows the unreliability of automatic translation tools

and encourages novice professionals to move away from assisted translation toward independent exploration of the semantic nuances of target-specific languages.

In another related study conducted by Marinov (2017), the researcher focused on the use of corpora and data extraction, employing a data-driven learning method with twelve tourism students. The vocabulary files of students were examined and identified by referring to the problems faced, and the effectiveness attained in responding to the given exercises. The variety of approaches used to complete the assignments emerged in many areas to analyse, such as chunking, paraphrasing, coherence, paragraphing, and organising.

A workshop designed by Flowerdew (2015) focused on the discussion section of the PhD theses. The PhD students were first introduced to the discussion sections of PhD theses for analysing different rhetorical functions involved, and this was followed by the use of ConcGram for exploring the lexico-grammatical patterns employed to observe the rhetorical functions and judge their appropriateness using the corpus. The result was that they had recently noticed some familiar multi-word combination patterns. The workshop was completed with such positive results that the students easily adjusted themselves to the corpora.

Cortes (2014) designed a 14-week corpus-based academic writing course and taught the participants how to use AntConc through an experimental and control group procedure. The experimental group of students compiled discipline-specific corpora and explored various moves. The result was that the experimental group evaluated the course with more positive outcomes than the control group students.

Önder-Özdemir (2014) used a data-driven learning method to teach medical English terminology in Turkey. A needs analysis was conducted with 323 Turkish medical students to identify their needs. The study showed that 41.2% of the students agreed that learning English words with their collocations is very important, and only 0.9% considered collocations unimportant. Considering the need analysis, a specialised corpus was compiled, and each content word was identified in frequency through AntConc. Önder-Özdemir noted that using a specialised corpus is appropriate for students due to its ease of control and potential to provide relevant samples.

In their 13-week course, Lee and Swales (2006) utilised a BNC corpus and an expert corpus, employing the WordSmith tools software package. The students were also asked to compile their own corpora and conduct several independent comparative analyses. The result was that several students found using corpora, especially with discipline-specific texts, effective, giving them greater autonomy.

2.9 The Concept of Phraseology: Patterned Language

The growing confidence in corpus evidence has reshaped how linguists and educators perceive language use, especially in academic contexts. Sinclair (2004), one of the pioneers in corpus linguistics, emphasised that the ability to study language through corpus data has the potential to transform our understanding of language itself fundamentally.

> "we should strive to be open to the patterns observable in language in quantity as we now have it. The growing evidence that we have suggests that there is to be found a wealth of meaningful patterns that, with current perspectives, we are not led to expect. We must gratefully adjust to this new situation and rebuild a picture of language and meaning which is not only consistent with the evidence but also exploits it to the full. This will take some time, and the first stage should be an attempt to inspect the data with as little attention as possible to theory." (p.10).

The basic meaning of the phraseology is a branch of linguistics that covers p-frames, multi-word combinations and the patterned nature of language. In the case of English for Academic Purposes (EAP), understanding and having an idea about patterns is a significant issue. Regarding this, students may produce discipline-specific and idiomatic academic texts. By drawing on large corpora, researchers have identified common phraseological patterns across academic genres, providing new insights into how knowledge is built and integrated within specific disciplines.

Regarding the popularity of corpus linguistics, language analysis has evolved from observations at the level of single texts to a more comprehensive examination of recurring patterns in large collections of texts. With this respect, Tognini-Bonelli (2004) highlighted that "from basically observing and describing a single communicative event, a text, to accessing, with the help of a computer, a whole collection of texts, a corpus, and observing the repeated lexical and grammatical choices of a certain community of language users" (p. 11). This transformation has brought about significant consequences for the study of phraseology. These patterns are not only reflections of language but also serve as cornerstones of discipline-specific academic contexts. Thus, corpus-based approaches bring valuable insights into how phraseology contributes to English for Academic Purposes (EAP) settings.

The study of Golparvar and Barabadi (2020) aimed to identify the key P-frames structures used in the discussion of research articles. The corpus consisted of 422 articles published in 13 prestigious journals between 2013 and 2018. The size of the corpus was 503,877 words. The four- and five-word p-frames were extracted with the kfNgram software. These p-frames were compared with the British Academic Written English Corpus (BAWE) to identify those words which is specific to higher

education. A total of 58 four-word and 40 five-word specific p-frames were found in the study. It was found that most of the p-frames contained non-verb content words (nouns, adjectives, adverbs) and that these structures mostly had a referential function (structures that convey information). The study emphasised that p-frames used in discussion can contribute to students' academic writing. The results also showed the importance of preparing discipline-specific phraseology lists for EAP (English for Academic Purposes) programs. Additionally, p-frame-based materials can enhance students' language awareness and improve their writing.

The study of Casal and Kessler (2020) examined an understudied promotional genre by analysing the distribution of frequent p-frames across the rhetorical stages of 148 purpose statements from accepted Fulbright grant applications. The researchers identified 69 unique p-frames using a corpus of approximately 95,000 words comprising 148 successful application texts. They examined how these p-frames are distributed across different rhetorical phases of writing and how helpful teachers and students find these structures pedagogically. The study results showed that p-frames have strong associations with specific rhetorical functions in application essays. P-frames were found to be used with high intensity in rhetorical moves. Moreover, 10 experienced EAP teachers and 10 advanced L2 learners who participated in the study found most p-frames highly useful in teaching promotional writing.

Liu and Chen (2022) investigated the phraseological exploration of university lectures through phrase frames. While previous research has primarily focused on academic writing, this study aimed to fill an essential gap in the literature by exploring the phraseological profile of academic speech. The researchers used a 4.38-million-word corpus of 565 course transcripts from four main disciplines (Engineering, Humanities and Arts, Natural Sciences, and Social Sciences). The corpus analysis identified 117 five-word and 29 six-word p-frames. The study categorised them based on structures and functions. Finally, structural categorisation showed that most p-frames consisted of verb-based structures (62%). Functionally categorised, reference p-frames accounted for the most significant proportion of lessons (29%), indicating that the genre is focused on providing informative content. Moreover, expert teachers found 93.2% of the p-frame list to be pedagogically useful. This supports the integration of such structures as instructional materials in EAP (English for Academic Purposes) teaching.

2.10 P-Frames

In recent years, several corpus-based studies have suggested the need for lists of academic phrases for students and teachers of English for Academic Purposes (EAP), as compiling such lists and providing several pedagogically useful lists of different types of academic phrases can be beneficial. Similarly, Garner (2016) noted that corpus-based and psycholinguistic studies reveal “the

importance of acquiring a productive knowledge of English phraseology for L2 learners." (p. 32). Building on this tradition, the present study adopts Phrase frames (P-frames), which are recurrent word combinations that follow a fixed structure. The basic definition of the p-frames is a type of semi-fixed multi-word sequence in which fixed words surround an open slot (e.g. the * of the) (Stubbs 2007). Unlike fully fixed phrases, P-frames are semi-fixed and can adapt to different lexical items, which makes them particularly useful for analysing academic discourse where variation and repetition coexist.

By identifying and categorising these P-frames using corpus linguistics, the research aims to contribute to a better understanding of idiomatic behaviours in Engineering EAP. In doing so, it aligns with the pedagogical goals of helping students master repetitive academic expressions for more effective communication.

2.11 Corpus-Based EAP in Engineering Context

Corpus linguistics and the use of corpus tools played a significant role in applying many EAP-based lexicogrammatical features and identifying the genre-based lexical properties of the articles under scrutiny.

In his 2021 study in the context of civil engineering, Otto (2020) identified special-purpose words for teaching as part of data-driven learning (DDL) vocabulary activities, using a corpus approach. The word selection was made according to the level of difficulty, the order of importance, and the suitability of DDL. The results were effective in that 18-word patterns that were problematic for the students were identified. Finally, two strengths were discussed related to the functionality of the words in a civil engineering context and the identification of words that could be disregarded easily due to a lack of sound connection.

A case study conducted by Eak-in (2015) focused on corpus-based methods for learning abstract writing. He worked with 100 low-proficiency engineering students, and the scope of his research was to examine the possible effects of the corpus methodology in improving students' abstract writing skills, as well as creating awareness of the linguistic characteristics of journals and applying them to students' writing process. Eak-in (2015) collected data using a mixed-methods approach and concluded that the corpus-based method appears to have a positive effect on students' use of the DDL methodology when composing abstract sections.

Although previous studies have highlighted the significance of P-frames in identifying phraseological patterns in academic discourse, most of these studies have focused on general EAP contexts. In addition, there are very few studies that have explored the discipline-specific use of P-

frames within engineering fields. Therefore, this thesis aims to fill this gap by conducting a corpus-based comparison of high-frequency P-frames in Geology and Mining research articles. This approach will contribute to a better understanding of how P-frames operate across related yet distinct engineering disciplines, revealing both their shared and discipline-specific patterns. Moreover, the thesis also aims to explore the potential of implementing Data-Driven Learning (DDL) in terms of pedagogical materials.

CHAPTER THREE

3. METHODOLOGY

3.1 Introduction

This chapter presents the methodological framework of the thesis, which investigates the use of P-frames in Geological Engineering and Mining Engineering texts. The overall aim is to examine how P-frames function in engineering discourse and their contribution to academic language use. This chapter covers a brief introduction to the research design, detailed explanations of the data collection procedures, materials and tools, corpus completion process and extraction of P-frames.

The following parts introduce the corpus tools and the procedures used for identifying P-frames. The specific emphasis on 3-slot and 4-slot patterns. The analysis includes both qualitative and quantitative components. The study also consists of descriptive statistical data to illustrate the frequency and contexts in which certain P-frames occur across two corpora. Furthermore, this thesis suggested potential pedagogical materials from the findings, regarding the teaching of academic expressions in Engineering EAP contexts.

3.2 Research Design

The present study employs a corpus-based analysis design, combining qualitative and quantitative research approaches. The study focuses on authentic academic texts within the discipline of Engineering English for Academic Purposes (EAP) rather than examining learner output or conducting classroom-based research. Specifically, it analyses how P-frames function across a corpus which consists of research articles written in the field of Geological and Mining Engineering.

This research design enables the investigation of naturally occurring language patterns using corpus linguistics, without requiring data manipulation or involving human participants. All the analyses were conducted on a corpus compiled for this study. A quantitative analysis was conducted to examine the frequency, structural predictability, and variability of the P-frames in Geology and Mining Engineering. 3-slot and 4-slot P-frames were analysed. Regarding the analyses, the study aims to indicate phraseological aspects and discourse functions of these P-frames. In addition, a qualitative approach was adopted to interpret the potential pedagogical application of selected P-frames. In this respect, the DDL method was used. This combination of methods allowed for both statistical insight and functional features of phraseological patterns within the discipline-specific

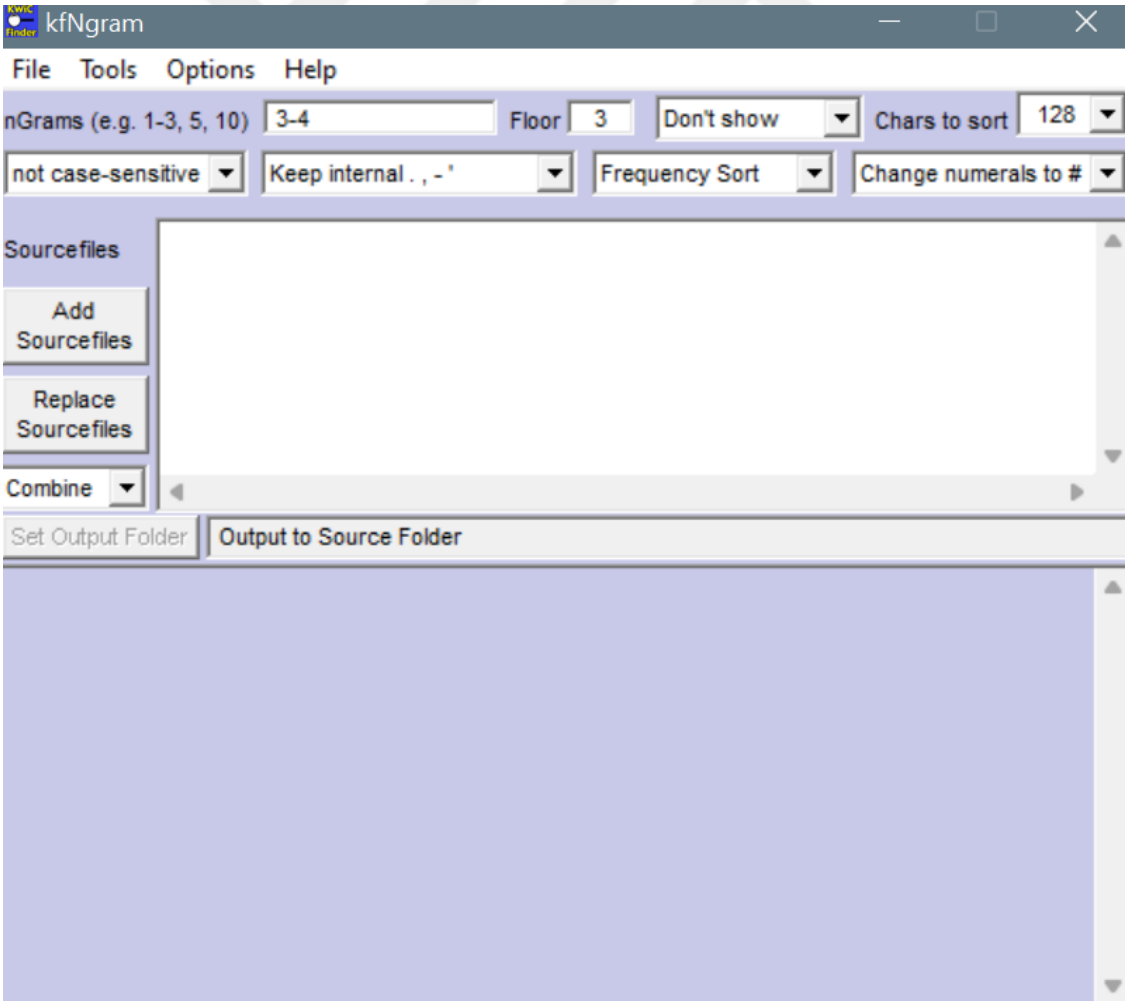
academic texts. As it is known, corpus-based studies rely on empirical data and patterns of language. Therefore, the research does not include surveys or interviews but instead focuses on language patterns in context.

3.3 Materials and Tools

3.3.1 KfNgram

In this study, the kfNgram program (Fletcher, 2012) was used as a primary tool. This program enables users to generate lists of n-grams and p-frames of varying lengths from a corpus. It is a free, downloadable, and user-friendly tool. This tool works with text files; therefore, the corpus should be converted to the txt format. It has one main screen and does not require additional page viewing or operations. By inputting the corpus files into kfNgram, frequently occurring lexical bundles with one or more variable elements (slots) were automatically retrieved. Figure 1 shows the main interface.

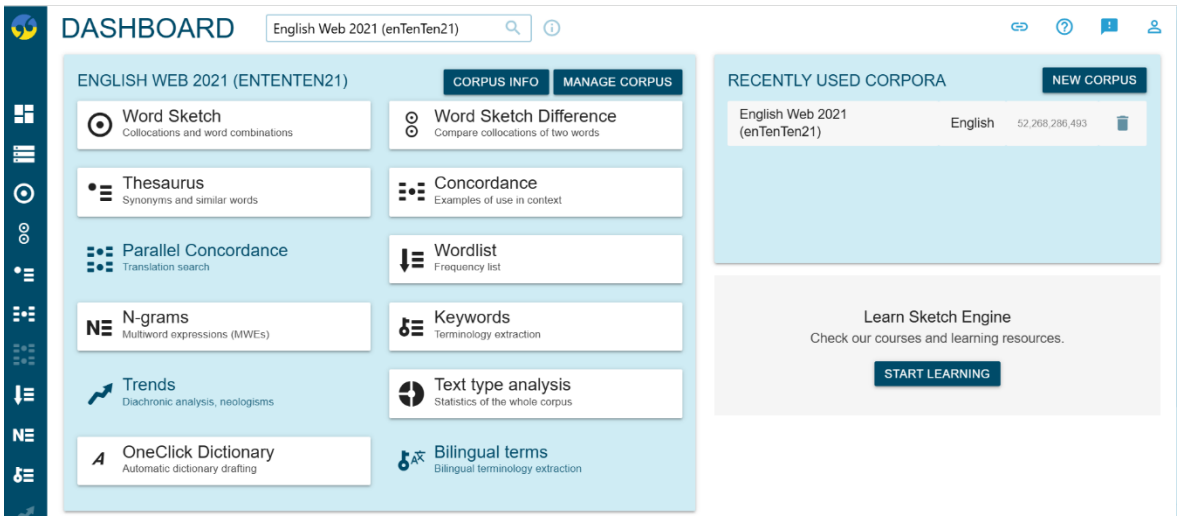
Figure 1. Main Interface of kfNgram



3.3.2 Sketch Engine

As clearly mentioned above, corpus linguistics has become an inseparable part of the field of linguistics, thanks to the opportunity for easy access to technology. Therefore, the use of corpus linguistics has been widespread among a variety of experts, such as researchers, lexicographers, terminologists, translators, teachers, and students. In this regard, it can be said that the number of people benefiting from the corpus has been increasing steadily. An area has emerged as an open space for various types of corpora, as well as corpus tools. Each corpus has its unique features and special design criteria, which enable it to address different users with varying purposes. From the perspective of the corpus tool, this situation can be more flexible. In other words, while some corpora are just suitable for certain specific groups, corpus tools can be used for many purposes. Almost all corpus tools share common characteristics such as wordlist, concordance, keyword, and n-gram options. In other words, each corpus tool also has its speciality. Moreover, it has a friendly interface with various ready-made corpora, as well as enables users to load their own corpora. Sketch Engine is a powerful and multifunctional tool that guides users to discover the language. It analyses billions of words within the authentic texts in minutes. Kilgarriff et al. (2014) divided Sketch Engine into two main parts, which are the software and the web service. The web service side of Sketch Engine provides "a large number of corpora pre-loaded and 'ready for use', and tools for creating, installing and managing your own corpora." (p.8). The main options of the Sketch Engine are Word Sketch, Word Sketch Difference, Word lists, Concordance, Thesaurus, Trends, N-grams, Keywords, and WebBootCaT. See Figure 2 for the main page of Sketch Engine.

Figure 2: Main Interface of Sketch Engine



3.4 Corpus Compilation

The primary data source for this study is a self-compiled academic corpus of Engineering EAP texts, specifically constructed to investigate phraseological patterns, with a focus on P-frames, in expert academic writing. The texts included in the corpus were drawn from two disciplines: Geology Engineering and Mining Engineering, both of which are departments within Karadeniz Technical University (KTU). These disciplines were selected due to their shared technical language, relevance within the field of applied sciences, and availability of specialised academic publications.

To ensure the academic and linguistic quality of the data, the texts included in the corpus consist of research articles published in peer-reviewed, high-impact journals indexed in internationally recognised databases such as Scopus and Web of Science. Articles written by authors who are experts, such as scholars affiliated with engineering departments or research institutions. These articles represent a range of research topics within the respective fields and were written entirely in English for international dissemination, making them suitable examples of Engineering English for Academic Purposes (EAP). Non-research genres such as book reviews, editorials, and conference announcements were excluded to ensure genre consistency.

The corpus was compiled in accordance with fundamental corpus design principles, with particular attention to representativeness and disciplinary balance between the two selected engineering fields. Expert consultancy was used to ensure the suitability and academic relevance of the selection of articles. Each corpus consists of 133 full-text research articles, resulting in a total of 985,778 tokens in the Geology corpus and 974,261 tokens in the Mining corpus. The overview of the corpus is presented in Table 1 below. All the articles were selected from open-access sources. Then, the identified articles were downloaded from the Karadeniz Technical University's database.

All research articles were converted into plain text format, and several non-linguistic elements were systematically removed from the texts. In order to obtain more accurate data, a cleaning process was conducted. In this process, especially, author names, institutional affiliations, journal titles, tables, figures, reference lists, and appendices were excluded from the texts. Thanks to the cleaning process, the data became consistent and understandable. The corpus, consisting of the main body of the articles, provided a more uniform and reliable dataset for the extraction and examination of P-frames. Corpus sizes were calculated using Sketch Engine, which was also employed for additional corpus searches in the following analysis, such as frequency lists and concordance searches.

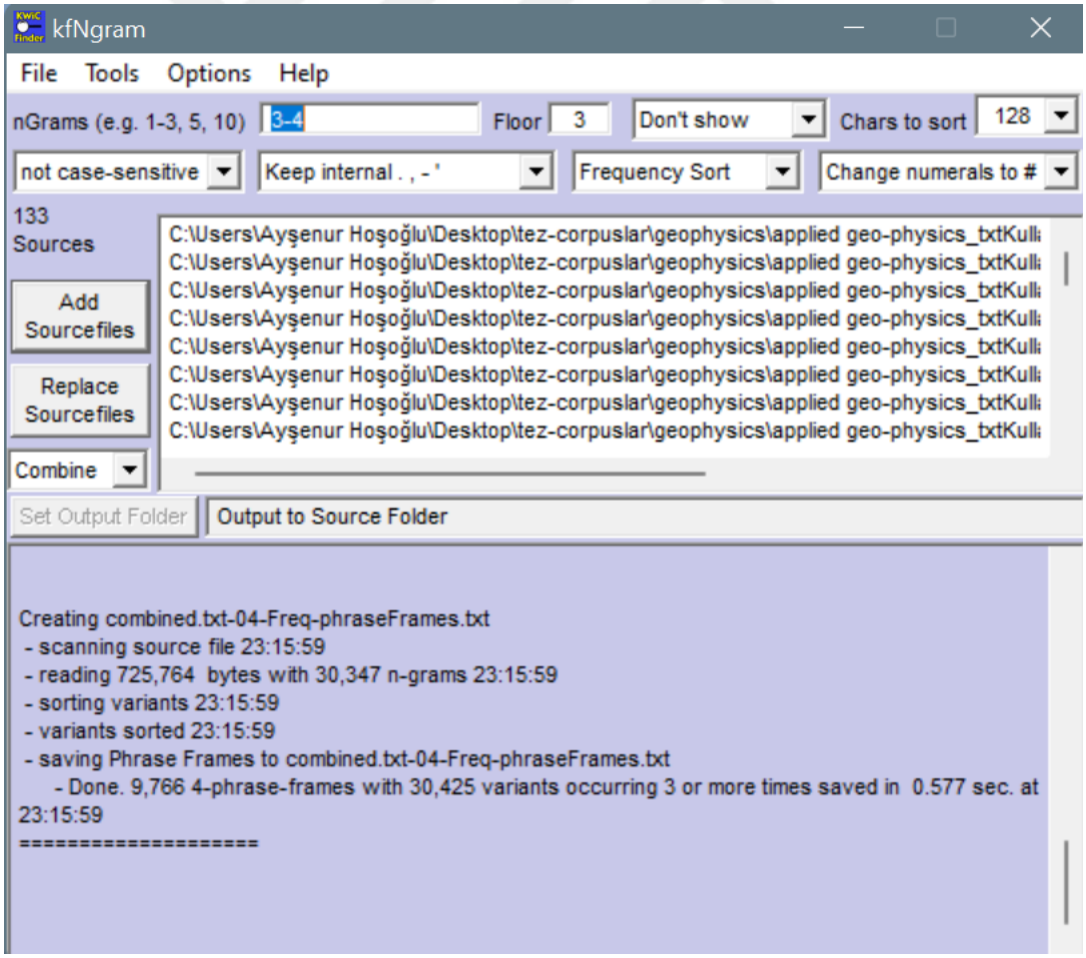
Table 1 : Corpus Info

Discipline-Specific Corpora: Geology and Mining Engineering	Number of Texts	Corpus size (Tokens)
Geology	133	985.778
Mining	133	974.261

3.5 Extraction of P-Frames Using KfNgram

P-frames were extracted using the kfNgram (Fletcher, 2012) application, which identifies n-grams containing at least one variable slot (*), (See Figure 3). KfNgram was used for the extraction of P-frames and helps to analyse n-gram sequences that include one or more variable slots in academic texts. The first 500 P-frames from each corpus were listed, and the top 100 were selected for detailed analysis to narrow the focus on high-frequency patterns.

Figure 3: The Extraction of P-frames with kfNgram



In this study, both 3-word and 4-word P-frames were extracted from both geology and mining corpora. The analysis criteria were respectively a minimum frequency threshold of 5, a maximum of one variable slot, and case-insensitive matching. This identification enables the detection of high-frequency P-frames that characterise academic language use in engineering contexts.

Furthermore, a collocation analysis was conducted with the P-frames as well. The reason was to see whether selected P-frames are combined with general academic or discipline-specific vocabulary. This analysis revealed not only the structural features of the p-frames themselves but also their functional roles in academic texts. By focusing on the collocates occurring to the left and right of the P-frames, it became possible to identify whether these patterns are seen as part of widely shared academic phraseology or whether they acquire discipline-specific meanings through their association with specialised terminology.

3.6 Ethical Considerations

All texts used in the compiling corpus were publicly available research articles. They were downloaded from the university database. Public information and other indicative information have been removed.

3.7 Validity and Reliability

In order to ensure the validity and reliability, several actions were taken. Firstly, there was an appropriate selection criterion which met the purpose of the study. These criteria consist of the type of articles, where they are published, and their authors. The selected articles were published in high-impact journals and written by experts in the relevant fields. Moreover, the P-frame extraction process was carefully completed. All extracted P-frames were manually checked to eliminate potential failures generated by the extraction tools. In addition, the analysis tools were also selected among the well-known corpus tools, which are kfNgram and Sketch Engine. Additionally, during the analysis, patterns and results were cross-checked to reduce researcher bias and verify the internal consistency of the classification process.

CHAPTER FOUR

4. RESULTS AND DISCUSSIONS

This section demonstrated the findings of the P-frame analysis conducted on the Geology and Mining corpora. The framework of the study aims to identify both shared and discipline-specific phraseological patterns, focusing on the three-word and four-word P-frames in each corpus. With this respect, frequency, type diversity, and structural similarities were examined to reveal how these frames contribute to the phraseological profile of Engineering English in different disciplines. The P-frames were extracted using the kfNgram application. As thousands of results were generated, the scope of this study was limited to the top 100 P-frames to ensure clarity of analysis and manageability. Nevertheless, the first 500 P-frames are provided in Appendix 1 and Appendix 2 for reference. Furthermore, possible pedagogical materials were also created.

4.1 Common P-Frames across Geology and Mining Corpora

In the scope of this study, the top 100 P-frames were analysed. When focusing on the shared P-frames in both the Geology and Mining corpora, significant similarities emerge. These common patterns highlight the extent to which academic writing in engineering-related disciplines relies on a shared phraseological core; in other words, they reflect general academic English discourse.

In the case of three-word P-frames, 65 out of the top 100 were shared across both corpora. These shared P-frames are listed in Table 2 below. This considerable overlap indicates that Geology and Mining academic writing largely shares phraseological similarities, reflecting the pervasive role of general academic English structures.

Table 2: Shared Three-Word P-frames

	Geology Corpus		Mining Corpus			Geology Corpus		Mining Corpus	
P-frame	**Freq**	**Type**	**Freq**	**Type**	**P-frame**	**Freq**	**Type**	**Freq**	**Type**
* and the	1429	224	1966	290	a * of	1166	103	1284	106
* as the	439	39	442	42	and * of	566	104	859	152
* at the	470	82	462	87	and * the	622	77	595	94
* based on	651	87	647	96	and the *	1403	250	1842	290
* between the	398	46	397	43	are * in	475	40	413	32
* by the	959	89	823	88	as a *	303	36	399	45
* can be	717	83	749	93	on the *	1646	208	1612	320
* for the	1697	247	186	167	of the *	9176	933	9900	908

Table 2: (Continued)

	Geology Corpus		Mining Corpus			Geology Corpus		Mining Corpus	
P-frame	**Freq**	**Type**	**Freq**	**Type**	**P-frame**	**Freq**	**Type**	**Freq**	**Type**
* from the	1029	129	873	109	of a *	438	73	402	65
* in the	5255	562	5210	613	of * and	730	112	945	141
* in this	519	76	403	58	it is *	591	57	773	75
* is the	725	79	730	92	is the *	635	111	735	111
* it is	446	61	465	54	is * to	729	62	817	74
* of a	514	90	525	93	is * in	443	46	399	40
* of the	10175	779	10719	816	in this *	883	26	756	35
* on the	1655	141	1822	147	in the *	5498	509	5269	518
* shown in	405	12	680	8	in * the	409	46	370	30
* that the	1355	96	1009	84	in * of	347	24	503	26
* the same	462	35	383	26	from the *	887	153	742	114
* this study	414	11	361	13	from * to	513	30	388	28
* to be	396	47	440	53	for the *	1652	239	1049	178
* to the	2542	178	2407	164	due to *	453	25	592	30
* used to	358	15	469	14	can be *	979	84	1000	99
* with the	1042	105	752	105	by the *	491	97	465	94
at the *	640	81	703	83	based on *	793	42	801	42
shown in *	464	9	728	19	the * and	1607	251	1506	244
that the *	762	148	527	105	the * area	531	32	512	45
the * in	610	71	633	97	the * between	348	34	356	30
the * is	475	90	518	90	the * of	8117	572	10820	714
the * the	389	76	448	71	the results *	441	23	427	27
to * the	1859	152	285	179	to be *	300	51	385	64
to the *	1911	298	1720	276	with a *	411	71	385	62
with the *	867	136	686	110					

Similarly, in the case of four-word P-frames, 46 out of the top 100 were shared across both corpora. These shared P-frames are listed in Table 3 below. As in the three-word P-frames, there is also a significant degree of overlap in the four-word P-frames. Several high-frequency structures are shared across both corpora, reflecting the reliance on standard academic phraseology.

Table 3: Shared Four-Word P-frames

	Geology Corpus		Mining Corpus			Geology Corpus		Mining Corpus	
P-frame	**Freq**	**Type**	**Freq**	**Type**	**P-frame**	**Freq**	**Type**	**Freq**	**Type**
* based on the	252	33	259	47	are shown in *	105	5	122	9
* due to the	109	20	127	24	as a * of	129	8	127	6
* one of the	90	9	116	10	as shown in *	138	7	358	15
* part of the	164	21	149	21	as well as *	133	12	120	9
* shown in fig	257	7	222	4	at the * of	119	17	172	16
* the case of	113	4	108	3	based on the *	203	37	204	40
* the number of	149	19	195	21	can be seen *	80	6	101	5
* the study area	257	9	113	7	due to the *	81	17	159	28
and * of the	158	34	132	26	in * of the	126	10	254	9
and the * of	123	27	254	51	in order to *	156	29	145	28
can be * as	142	13	112	14	in the * area	174	12	201	16
for the * of	217	37	182	32	in this study *	186	18	133	14
in the * of	700	71	838	80	is shown in *	100	5	147	10

Table 3: (Continued)

	Geology Corpus		Mining Corpus			Geology Corpus		Mining Corpus	
P-frame	**Freq**	**Type**	**Freq**	**Type**	**P-frame**	**Freq**	**Type**	**Freq**	**Type**
in this * the	179	16	126	12	it can be *	112	11	119	13
is the * of	123	23	175	25	it is * that	149	17	139	20
it is * to	140	9	275	16	of the * and	178	33	189	34
of the * area	160	10	102	10	of the * in	86	18	109	27
of the * of	281	49	262	53	on the * of	224	37	376	58
one of the *	84	7	102	9	the * between the	159	18	170	15
the * in the	184	26	151	32	the * of a	158	23	164	28
the number of *	124	21	150	27	to the * of	249	43	279	60
the results of *	129	9	133	7	used to * the	135	23	168	28
with a * of	78	14	117	16	with the * of	120	20	236	26

Overall, both three and four-word P-frames primarily consist of high-frequency prepositions and structural patterns that are widely recognised as part of general academic English. Their frequent occurrence in both text corpora indicates that, regardless of discipline focus, authors rely heavily on these traditional idiomatic units to structure arguments, describe processes, and present results. This overlap suggests that while discipline-specific variations certainly exist, a significant part of academic writing in engineering fields is built upon a shared recurrent linguistic frame. However, it should be kept in mind that differences in discipline can play a significant role as well. In some cases, even if they shared a common phrase, its contextual meaning may change depending on the discipline. Therefore, it can be emphasised that while these frameworks function as devices for constructing discourse that ensures consistency across disciplines and rhetorical clarity, the subtle lexical differences within them can still reflect the identities of different disciplines.

4.2 Comparison of the Top Ten Three-word P-frames in Geology and Mining Corpora

As stated above, this study aimed to examine the academic use of P-frames; therefore, the top ten three-word P-frames were compared across both corpora. The analysis of the top ten three-word P-frames across the Geology and Mining corpora reveals both considerable overlap and discipline-specific variations. By presenting the most frequently used P-frames side by side, it becomes apparent that these two engineering fields share a particular set of recurrent phraseological structures, mainly centred around prepositional patterns as shown in Table 4.

Table 4: Top Ten Three-Word P-frames

Geology Corpus Three-word P-frames			Mining Corpus Three-word P-frames		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
** of the*	*10175*	*779*	*the * of*	*10820*	*714*
*of the **	*9176*	*933*	** of the*	*10719*	*816*
*the * of*	*8117*	*572*	*of the **	*9900*	*908*

Table 4: (Continued)

Geology Corpus Three-word P-frames			Mining Corpus Three-word P-frames		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
*in the **	*5498*	*509*	*in the **	*5269*	*518*
** in the*	*5255*	*562*	** in the*	*5210*	*613*
** to the*	*2542*	*178*	** to the*	*2407*	*164*
to the *	1911	298	*to * the*	*2085*	*179*
*to * the*	*1859*	*152*	* and the	1966	290
* for the	1697	247	and the *	1842	290
** on the*	*1655*	*141*	** on the*	*1822*	*147*

As shown in Table 4, both the Geology and Mining corpora display a significant number of overlapping three-word P-frames, which is eight words out of ten. P-frames indicated in italics represent those shared by both corpora. The recurrence of these patterns across both disciplines demonstrates that they are not only discipline-specific but rather constitute part of the broader world of academic English. In addition, represents common phraseological structures widely used in academic writing. Moreover, the frequencies of these shared P-frames are also relatively close to each other, which emphasises their role as stable and pervasive elements of academic discourse.

Table 5: Top Ten Four-Word P-frames

Geology Corpus Four-word P-frames			Mining Corpus Four-word P-frames		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
*the * of the*	*2523*	*296*	*the * of the*	*3022*	*321*
*in the * of*	*700*	71	*in the * of*	*838*	*80*
*of the * of*	281	49	*on the * of*	*376*	*58*
* shown in fig	257	7	as shown in *	358	15
* the study area	257	9	*to the * of*	*279*	*60*
** based on the*	*252*	*33*	it is * to	275	16
*to the * of*	*249*	*43*	of the coal *	263	19
bulletin of the *	240	6	*of the * of*	262	53
*on the * of*	*224*	*37*	** based on the*	*259*	*47*
for the * of	217	37	in * of the	254	9

The findings in Table 5 reveal a significant structural similarity between Four-word P-frames in the Geology and Mining corpora. In both corpora, fundamental academic patterns such as *"the * of the, in the * of, * based on the, to the * of, on the * of"* indicate the existence of a common interdisciplinary academic language. Additionally, it may be interesting that they share the two highest frequencies in both corpora. However, while field-focused and visualisation-based structures such as *" * shown in fig, * the study area, bulletin of the * , and for the * of"* are noticeable in the Geology corpus, process descriptions and material-focused expressions such as *"as shown in *, it is * to, of the coal *, and in * of the*" are encountered in the Mining corpus. Additionally, in terms of frequency, the overall frequency of four-word P-frames is lower than that of three-word structures.

Overall, these findings highlight how P-frames can simultaneously reveal shared academic discourse while also uncovering disciplinary preferences in textual organisation and rhetorical practice.

In addition to the analyses made above, Log-likelihood tests were conducted to compare the distribution of shared three-word and four-word P-frames across the geology and mining corpora. These tests were employed to determine whether the frequency differences between the two corpora were statistically significant. The following tables present the frequencies of each P-frame in both corpora, along with their log-likelihood values and corresponding p-values.

Table 6: Log-Likelihood Test Results of Shared Three-Word P-Frames in the Geology and Mining Corpora

Shared Three-word P- frames	Geology Corpus Freq.	Mining Corpus Freq.	Log-likelihood value	p-value
* of the	10175	10719	1.79	.18
of the *	9176	9900	3.37	.06
the * of	8117	10820	39.43	< .001
in the *	5498	5269	0.41	.52
* in the	5255	5210	0.00	.95
* to the	2542	2407	0.66	.41
to * the	1859	2085	5.19	.02
* on the	1655	1822	3.61	.05

From a statistical aspect, Table 6 presents the log-likelihood test results for the most frequent three-word P-frames shared between the geology and mining corpora. The results indicate that only "the * of" (*LL = 39.43, p < .001*) and "to * the" (*LL = 5.19, p = .02*) show statistically significant differences between the two corpora. In addition, "* on the" reached a marginal level of significance (*p = .05*). No significant differences were found for the other P-frames (*p > .05*). These findings suggest that while certain three-word P-frames are more characteristic of one corpus, the majority display similar distributions across both.

Table 7: Log-Likelihood Test Results of Shared Four-Word P-Frames in the Geology and Mining Corpora

Shared Four-word P- frames	Geology Corpus Freq.	Mining Corpus Freq.	Log-likelihood value	p-value
the * of the	2523	3022	13.27	< .001
of the * of	281	262	0.37	.54
* based on the	252	259	0.15	.69
to the * of	249	279	1.62	.20
on the * of	224	376	31.22	< .001
in the * of	700	838	7.85	.005

Similarly, Table 7 presents the log-likelihood test results for the most frequent four-word P-frames shared between the geology and mining corpora. The results indicate that "the * of the" (*LL = 13.27, p < .001*), "on the * of" (*LL = 31.22, p < .001*), and "in the * of" (*LL = 7.85, p = .005*) show statistically significant differences between the two corpora. No significant differences were observed for the other P-frames (*p > .05*). These findings suggest that while certain four-word P-frames are more characteristic of one discipline, the majority exhibit similar distributions across both corpora.

4.3 Examining the Types of Shared High-Frequency P-frames

In the case of P-frames, another significant aspect to consider is their types. While frequency highlights how often particular P-frames occur, the notion of type reflects the number of different variations of a given P-frame found within the corpus. In other words, type measures the range of lexical or structural variation that a P-frame can take. This distinction is essential because a high frequency may indicate the prominence of a certain P-frame. Still, a higher number of types suggests greater variability in how that frame is employed across different contexts. By examining both frequency and type, it becomes possible to capture not only the quantitative dominance of specific patterns but also the qualitative richness of their usage within the discipline.

To provide more detailed information, the types of shared high-frequency P-frames were analysed individually in order to determine whether they show similarities or differences across the two corpora. In the following section, this analysis is presented to illustrate how shared P-frames vary not only in their frequency but also in their structural diversity.

There were eight shared three-word P-frames identified across the two corpora. To provide a clearer picture of their distribution, the top ten types of these shared P-frames are presented in the following table, arranged in order of frequency.

Table 8: Types of Shared High-frequency Three-Word P-frames

Geology Corpus		Mining Corpus	
P-frame	**Types**	**P-frame**	**Types**
* of the	*part of the* bulletin of the *one of the* *analysis of the* *results of the* most of the distribution of the effect of the *value of the* structure of the	* of the	*part of the* *one of the* characteristics of the *analysis of the* *results of the* *value of the* area of the strength of the length of the model of the

Table 8: (Continued)

Geology Corpus		Mining Corpus	
P-frame	**Types**	**P-frame**	**Types**
of the *	of the seismologic *of the study* of the model of the earth of the fault of the seismic of the earthquake of the data of the source of the tsunami	of the *	of the coal of the mining of the rock of the rib of the working of the mine of the roof of the shearer *of the study* of the main
the * of	*the number of* *the effect of* *the results of* *the presence of* *the use of* *the case of* the gulf of the distribution of the role of the value of	the * of	*the number of* *the use of* *the results of* the impact of the development of *the effect of* the stability of *the case of* *the presence of* the amount of
in the *	*in the study* in the region *in the case* in the north in the present in the south in the western in the upper *in the area* in the northern	in the *	in the mining *in the case* *in the area* in the coal *in the study* in the rock in the process in the form in the goaf in the mine
* in the	*changes in the* *used in the* *increase in the* *located in the* observed in the occurred in the variation in the data in the earthquakes in the values in the	* in the	*located in the* *used in the* *changes in the* *increase in the* included in the concentration in the shown in the change in the content in the present in the
* to the	*due to the* *according to the* *compared to the* *related to the* *close to the* applied to the similar to the attributed to the respect to the refers to the	* to the	*according to the* *due to the* *related to the* *compared to the* *close to the* perpendicular to the link to the addition to the lead to the line to the

Table 8: (Continued)

Geology Corpus		Mining Corpus	
P-frame	**Types**	**P-frame**	**Types**
to * the	*to determine the* to estimate the *to obtain the* *to calculate the* *to study the* *to evaluate the* *to improve the* to investigate the to understand the *to assess the*	to * the	*to determine the* *to improve the* *to evaluate the* *to study the* *to calculate the* *to assess the* to reduce the *to obtain the* to analyze the to predict the
* on the	*based on the* *depending on the* *study on the* *effect on the* *impact on the* *located on the* *focused on the* dependent on the data on the *information on the*	* on the	*based on the* *depending on the* *impact on the* *study on the* *effect on the* research on the mining on the *information on the* *focused on the* *located on the*

As illustrated in Table 8, certain types of shared P-frame types are found in both corpora (italicised). For instance, the P-frame *"* * *of the"* can be considered as discipline-general academic expressions, reflecting phraseological patterns widely used in academic writing. This is because the P-frame includes shared types across both corpora, indicating its broader role in academic discourse beyond disciplinary boundaries. The types *"part of the, one of the, analysis of the, results of the, and value of the"* are the same in both corpora. On the other hand, the remaining types are discipline-specific, for example, *"bulletin of the"* in the Geology corpus or *"strength of the"* in the Mining corpus, which highlight the distinctive textual practices of each field. This distinction between general and discipline-specific types provides an essential insight into how academic discourse operates at both a shared and specialised level.

In the case of *"of the *"* in Table 8, there is only one shared type, which is *"of the study"*. Beyond this overlap, however, the majority of the types were discipline-specific, reflecting the distinct thematic and terminological phrases of Geology (*e.g., of the earthquake, of the seismic*) and Mining (*e.g., of the coal, of the mine*). This indicates that although the underlying p-frame pattern is the same, the types employed tend to become discipline-specific. In other words, rather than frequency alone, it is terminological diversity that makes the disciplinary differences more visible.

As can be seen in Table 8, the P-frame *"the * of"* also shares a variety of types across both corpora. Types such as *"the number of, the effect of, the results of, the presence of, the use of, and*

the case of" can therefore be regarded as part of general academic English, reflecting phraseological patterns that are widely employed in academic writing.

As shown in Table 8, the P-frame *"in the *"* has fewer shared types across the two corpora, with only three identified in common. However, when examining the discipline-specific side, it can be observed that the Geology corpus contains field-related terms such as *region, north, south, and western*. In contrast, the Mining corpus features its own specialised terminology, including *coal, mining, rock, and goaf*.

As demonstrated in Table 8, the P-frame *"*in the"* has four shared types and *"* to the"* has five shared types across the two corpora. A noteworthy finding is that both p-frames have these types of ranks among the top in terms of frequency. It can be noted that they have a significant role in academic discourse across both disciplines.

In the case of the P-frames to "to * the" and "*on the", they account for the majority of the shared types across the two corpora. This suggests that these patterns play an essential role in academic discourse within the field of engineering.

Table 9: Types of Shared High-frequency Four-Word P-frames

Geology Corpus		**Mining Corpus**	
P-frame	**Types**	**P-frame**	**Types**
the * of the	*the results of the* the effect of the the magnitude of the the top of the the amplitude of the the accuracy of the the location of the the case of the *the size of the* the center of the	the * of the	*the results of the* the length of the the stability of the the thickness of the the area of the the height of the the end of the *the size of the* the mining of the the roof of the
in the * of	*in the case of* in the gulf of in the range of *in the vicinity of* *in the area of* *in the form of* in the frequency of *in the context of* *in the presence of* *in the middle of*	in the * of	*in the case of* *in the form of* in the process of *in the area of* *in the vicinity of* in the field of *in the context of* *in the presence of* in the order of *in the middle of*

Table 9: (Continued)

Geology Corpus		Mining Corpus	
P-frame	**Types**	**P-frame**	**Types**
of the * of	*of the effects of* of the gulf of of the order of *of the number of* of the presence of of the center of of the ratio of of the amount of of the occurrence of of the trend of	of the * of	of the impact of *of the number of* of the use of of the results of of the height of of the development of *of the effect of* of the influence of of the analysis of of the area of
* based on the	pet based on the models based on the Eurasia based on the results based on the *calculated based on the* estimated based on the *values based on the* *method based on the* *model based on the* approach based on the	* based on the	*model based on the* *calculated based on the* *method based on the* evaluated based on the *values based on the* established based on the developed based on the paper based on the determined based on the data based on the
to the * of	*to the presence of* to the north of to the study of to the south of *to the number of* to the east of *to the development of* *to the value of* to the end of *to the lack of*	to the * of	to the results of *to the presence of* to the formation of *to the lack of* to the use of to the bottom of to the influence of *to the development of* *to the number of* *to the value of*
on the * of	*on the basis of* on the simulation of *on the surface of* *on the results of* on the use of *on the analysis of* on the number of on the order of on the effect of *on the characteristics of*	on the * of	*on the basis of* *on the surface of* on the stability of *on the results of* on the type of *on the characteristics of* *on the analysis of* on the application of on the degree of on the principle of

As illustrated in Tables 9, the four-word p-frames also reveal both general and discipline-specific tendencies. Several -frames such as "*the * of the*, *in the * of*, *of the * of*, * *based on the*, *to the * of*, and *on the * of"* contain types that are shared across both corpora, for example *the results of the*, *in the case of*, *in the presence of*, *of the number of*, *based on the method*, *to the development of*, and *on the basis of*. These shared types can be regarded as part of general academic English,

reflecting common phraseological patterns frequently preferred in academic writing regardless of discipline. At the same time, many other types are field-specific. In the Geology corpus, types such as "*the magnitude of the*, *the fault of the*, or *in the range of"* are closely related to earth science terminology, whereas in the Mining corpus, types like "*the mining of the*, *the roof of the*, or *in the process of*" highlight the technical and operational aspects of mining discourse. This contrast indicates that while shared four-word p-frames demonstrate the commonality of general academic expressions, the discipline-specific types underline the distinctive lexical and conceptual needs of each field.

4.4 Collocational Tendencies of Shared High-frequency Three-word P-frames

In the previous section, it was observed that some shared p-frames also appeared as shared types across both corpora. With this respect, it can be understood that these structures play a role in general academic English beyond disciplinary boundaries. Nevertheless, frequency and type alone cannot fully capture how these p-frames function within texts. For this reason, a collocation analysis was conducted to examine whether similar patterns could also be identified in their own lexical environments. By focusing on the words that co-occur with these p-frames, this analysis provides further insight into whether they are used in connection with general vocabulary or whether they acquire more specialised meanings through associations with discipline-specific terminology. The following tables present the results of this collocational investigation. The collocation analysis of P-frames in the form of "... * ..." was not included in this study, as the collocates were examined only at the immediate left (-1) and right (+1) positions. Since four-word P-frames extend beyond this span, they were not suitable for this type of analysis and sorting. For this reason, the collocational patterns presented in the following section focus exclusively on shared high-frequency three-word p-frames.

Table 10: Discipline-specific Collocational Patterns of Shared High-frequency Three-word P-frames

Geology Corpus		**Mining Corpus**	
P-frame	**Frequent Collocates (±1 window)**	**P-frame**	**Frequent Collocates (±1 window)**
* *of the*	northern, vertical, eastern, spatial, western, seismic, different, *central*, *upper*, *lower*	* *of the*	*upper*, *lower*, maximum, mechanical, mining, stress, surface, average, water, *central*
of the *	*area*, *model*, data, region, wave, zone, field, source, function, *layer*	*of the* *	pillar, face, *area*, seam, coal, mass, *model*, rock, body, *layer*
in the *	*area*, *region*, *direction*, *zone*, domain, section, model, atmosphere, sea, *layer*	*in the* *	*area*, industry, mass, *region*, mining, *direction*, *layer*, *zone*, process, mine

Table 10: (Continued)

Geology Corpus		Mining Corpus	
P-frame	**Frequent Collocates (±1 window)**	**P-frame**	**Frequent Collocates (±1 window)**
* *in the*	seismic, earthquake, water, seasonal, data, *stress*, velocity, large, increasing, *significant*	* *in the*	mining, water, *stress*, gas, *significant*, energy, rock, environmental, metal, methane
* *to the*	very, *mainly*, probably, directly, analytic, direction, similar, mostly, lines, extended	* *to the*	appropriate, credit, directly, plane, *mainly*, carcinogenic, waste, direction, areas, calculated
* *on the*	models, *significant*, *mainly*, *calculated*, seismic, *stress*, estimated, pressure, values, highly	* *on the*	*significant*, negative, carried, mining, *calculated*, great, experimental, method, *mainly, stress*

As shown in Table 10, the collocational analysis of the shared high-frequency three-word P-frames reveals both similarities and discipline-specific tendencies. On the one hand, several collocates overlap across the two corpora, particularly those that represent general academic or descriptive functions. For example, collocates such as *"area, lower, central, significant, stress, calculated, and mainly"* appear in both the Geology and Mining corpora, suggesting that these words form part of a general academic phraseological unit commonly used in engineering-related disciplines.

On the other hand, the results also demonstrate apparent disciplinary variation. In the Geology corpus, collocates such as "*seismic, earthquake, wave, fault, zone, tsunami, and atmosphere"* reflect terminology closely tied to earth sciences and natural phenomena. Conversely, the Mining corpus is characterised by collocates such as *coal, seam, pillar, mine, methane, environmental, and waste,* which are specific to the technical and industrial context of mining. This indicates that while the underlying p-frames are shared, their immediate collocates tend to adapt to the disciplinary focus of each field.

Overall, these findings suggest that collocational analysis demonstrated an essential dimension to understanding p-frames. It highlights not only their role as general academic structures but also their capacity to integrate field-specific terminology, thereby reflecting the dual nature of academic discourse as both shared and specialised.

In order to gain a deeper understanding of how P-frames operate in authentic academic discourse, several examples from the concordance lines are provided below. While the quantitative results highlight the frequency and collocational tendencies of p-frames, concordance lines allow for a closer examination of their actual usage in context. This qualitative perspective makes it possible

to see how shared collocates may function differently across disciplines and how discipline-specific collocates contribute to shaping the distinctive discourse of each field.

Concordance Examples Extracted from Shared Collocations in Both Corpora

- "…The **upper part of the** area is a pebble block and the size of the pebbles decreases towards the sea…" *(Geology Corpus)*
- "…The most intense contamination affects the **upper part of the** old embankment to a depth of 3-4 m…" *(Mining Corpus)*

- "…which in turn provided the inter-annual variability in terms **of the geographical area** influenced by the extreme winds over…" *(Geology Corpus)*
- "…Exceeding the optimal size **of the mining area** leads to an increase in transport and ventilation costs…" *(Mining Corpus)*

- "…Therefore, daily precipitation records of 83 meteorological stations **in the Marmara region** were extracted between 00:00 and 23:59 UTC…" *(Geology Corpus)*
- "…The average thickness **in the northern region** is 13.67 m and the maximum is 35.04 m…" *(Mining Corpus)*

- "…and since 1950 onwards, a **significant decrease in the** frequency of moderate rainfall events (5–100 mm)…" *(Geology Corpus)*
- "…in the form of crystallo- hydrates contributes to a **significant increase in the** volume of the material…" *(Mining Corpus)*

- "…As the new scaling field is derived from Euler's Homogeneity equation, the scaling exponent is also revealed to be **directly related to the** structural index…" *(Geology Corpus)*
- "…The locality of a mining product is **directly related to** the extraction and logistic costs, prices, and socio-political factors…" *(Mining Corpus)*
- "…Normal faults are **mainly distributed on the** northern margin of the Jianghan Basin…" *(Geology Corpus)*
- "…Under no-tillage conditions, crop residues were **mainly distributed on the** soil surface, while under ploughing conditions…" *(Mining Corpus)*

Concordance Examples Extracted from Discipline-Specific Collocations in Both Corpora

- "…hydrostatic pressure fluctuations due to the **vertical motion of the** ocean because of the weather conditions,…" *(Geology Corpus)*
- "…Besides, the principal **stress ratio of the** principal stress field in the area within 100 m of the working front of 23130 working face…" *(Mining Corpus)*

- "…Thus, the amplitude and duration **of the seismic wave** radiated back to the half-space will be smaller…" *(Geology Corpus)*
- "…otherwise, the maximum value within the fusion result **of the target layer** is taken as the final state of the terminal…" *(Mining Corpus)*

- "…The rod was centered **in the solution domain** with its long axis parallel to the x-axis…" *(Geology Corpus)*
- "…Introducing sustainable development and reviewing environmental sustainability **in the mining industry** technological developments and safe and clean production methods…" *(Mining Corpus)*

- "…we employed a smooth interpolation scheme between the **velocity values in the** bottom layers of our profiles…" *(Geology Corpus)*
- "…The share of electricity generated by renewable **energy sources in the** total electricity generation…" *(Mining Corpus)*

- "…the second at greater depth, **probably linked to the** transition between fractured…" *(Geology Corpus)*
- "…where a large amount of mining **waste goes to the** Atrato River or its tributaries…" *(Mining Corpus)*
- "…field of **pressure variations on the** Earth surface near Moscow was constructed…" *(Geology Corpus)*
- "…Currently, the mining industry has a **negative impact on the** environment in the regions…" *(Mining Corpus)*

4.5 Pedagogical Implications of Data-Driven Learning (DDL) and Suggested Activities

The scope of this thesis may provide a basis for the application of Data-Driven Learning (DDL) in engineering education as part of its pedagogical implications. As stated above, DDL is a valuable way of integrating students into their own learning process. This process not only promotes autonomous learning but also strengthens academic writing skills by linking linguistic forms with

disciplinary contexts. Therefore, by employing user-friendly corpus tools such as Sketch Engine, students can be trained to actively explore p-frames, collocations, and concordance lines to enhance their awareness of both general academic English and discipline-specific phraseological patterns.

Regarding the integration of Data-Driven Learning (DDL) into the engineering context, several drafts of classroom activities were designed based on the findings of this study. These activities aimed to help students explore discipline-specific P-frames and understand their functional roles in academic texts. Therefore, learners are encouraged to observe recurring phraseological patterns and identify their collocational and rhetorical functions by using Sketch Engine. The following Table 11 presents three suggested DDL activities that can be integrated into language education. These activities tried to enhance students' awareness and support them in using P-frames in their academic writing.

Table 11: Data-Driven Learning Activities for P-frame Awareness

Activity	Objective	Materials	Procedure	Expected Learning Outcomes
Teacher-provided keywords	Raise awareness of frequent discipline-specific P-frames	Pre-selected list of P-frames and frequent collocates	1.Students search the given P-frames on Sketch Engine 2. Then generate concordance lines 3.And highlight collocates and discipline-related terms	Students have a chance to compare P-frame structures and their collocational behaviour in authentic texts
Cause-effect identification	Help students understand rhetorical functions (explaining results and causes)	List of P-frames used in cause/effect contexts	1.Students classify which concordance lines show cause or result 2.And rewrite some sentences using alternative P-frames	Students may develop awareness of the functional use of P-frames in academic argumentation
Guessing from mixed concordances	Enhance pattern recognition and contextual guessing skills	Mixed concordance lines with missing P-frames	1.Students predict which P-frame fits in each blank 2.Then, check with the actual concordance 3.And discuss the clues they used	Students have the ability to infer the correct P-frames from context and collocates, showing a more profound understanding of patterning

Another significant issue is how to apply P-frames in practice. In this regard, students are assigned to write a short academic paragraph that contains word combinations identified through corpus analysis. This activity allows students to understand the meaning of the discipline-specific words and integrate them into their own learning. Moreover, their awareness may be raised, and they may gain a new perspective on the use of corpus analysis.

The suggested activities begin with students' existing knowledge. In the first steps, there are no right or wrong answers. The primary goal is to enhance students' understanding of their field's terminology. Then, the activities follow a step-by-step learning process by using corpus data. In this stage, concordance, frequency, and collocation analyses, along with error identification and recognition of rhetorical functions, are the most commonly used features. There are detailed activities below.

Activity 1: Predicting Shared Collocates

This activity aims to enable students to explore how shared p-frames are used in various disciplines. Students first predict collocations based on their own background knowledge, then check these predictions by using Sketch Engine. This allows students to compare both general academic patterns and discipline-specific terminology usage.

Steps: The students are given shared three-Word P-frame structures. Then, students are asked to complete these p-frames in terms of their own fields. Students then use Sketch Engine to conduct a concordance search and check whether the selected collocations appear in the corpus. To do this search, students used a concordance search, followed by a frequency order, and then chose the location indicated by the asterisk (*). The detailed explanation of these steps is given below in Figure 3. In the final stage, students compare their predictions with the corpus results and engage in a brief discussion about cross-domain differences.

Figure 4: Detailed Explanation of Sketch Engine Searching Steps

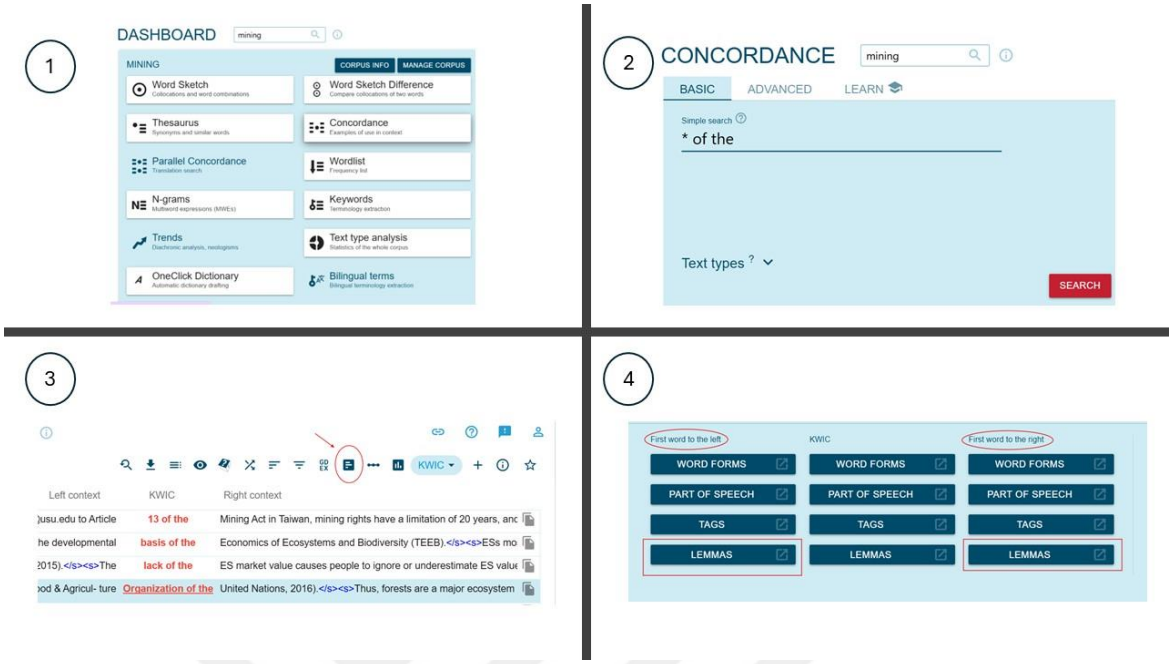


Table 12: Activity 1

* of the	• …… • …….. • ……..
of the *	• …… • …… • ……
in the *	• …… • …….. • ……..
* in the	• …… • ……... • ……

Table 12: (Continued)

* to the	• …… • …….. • ……..
*to * the*	• …… • …….. • ……..
* on the	• …… • …….. • ……..

Activity 2: Discovery of P-frames

The purpose of this activity is to familiarise students with the P-frame templates used in academic writing for explanation/justification, and to observe how these templates are used in Geology and Mining engineering disciplines. The teacher gives the first 50 four-word P-frame lists to the students. Students are then asked to find patterns in this list that serve an explanatory/justification function. After that, students are required to examine the usage of the p-frames that they have identified in the list. In this section, they should use the Sketch Engine application and select the concordance option.

Table 13: Activity 2.1

Four-word P-frames in Geology		
P-frame	**Explanation/justification**	**Not sure**
the * of the		
in the * of		
of the * of		
* shown in fig		
* the study area		
* based on the		
to the * of		
bulletin of the *		
on the * of		
for the * of		

Table 13: (Continued)

based on the *		
in this study *		
the * in the		
in this * the		
of the * and		
To be continued…		

Table 14: Activity 2.2

Four-word P-frames in Mining		
P-frame	**Explanation/justification**	**Not sure**
the * of the		
in the * of		
on the * of		
as shown in *		
* science and technology		
science and technology *		
of mining * and		
of mining science *		
* mining science and		
to the * of		
To be continued…		

Overall, through such activities, Sketch Engine, with an easy-to-use interface, can be effectively integrated into a DDL approach, enabling engineering students to discover and practice both general and discipline-specific language patterns. As a result, students have the opportunity to develop greater linguistic awareness, which especially contributes to improved academic writing competence within their respective fields.

4.6 Discussion

This thesis aims to examine the use and distribution of P-frames in Engineering English for Academic Purposes (EAP) texts, explicitly focusing on research articles written in the fields of Geological Engineering and Mining Engineering. The study tries to explore how P-frames contribute to the construction of academic discourse in the engineering field. Adopting a corpus-based approach, this study also aims to discuss the potential of integrating Data-Driven Learning (DDL) as a pedagogical method to help engineering students become more aware of such phraseological

patterns in their academic writing. This section discusses the research questions presented in the introduction chapter of this thesis. The research questions which pave the way for this thesis are as follows:

1. To what extent do Geology and Mining Engineering research articles share high-frequency P-frames, and which P-frames occur significantly more often in one discipline than the other?

2. How can user-friendly corpus tools be used to raise engineering students' awareness of discipline-specific phraseological patterns in academic writing?

3. What rhetorical functions (e.g., reporting results, showing cause–effect relations) are commonly associated with the most frequent P-frames, and how can these be taught through data-driven learning (DDL) activities?

4.6.1 Discussion of the First Research Question

The analysis of the two discipline-specific corpora revealed that both Geology and Mining Engineering articles share several high-frequency three-word and four-word P-frames. This suggests that specific phraseological frames are widely used as part of general academic discourse in engineering-related fields. This overlap can be referred to as recurrent frames, which constitute part of a general academic register within engineering, providing authors with ready-made structural patterns for academic writing.

At the same time, the log-likelihood results showed that some of these shared P-frames were significantly more frequent in one corpus than the other, indicating a degree of disciplinary preference. This finding supports previous research arguing that while academic writing heavily relies on recurrent phraseological frames, the lexical realisations within these frames often become discipline-specific (Garner, 2016). Therefore, these results highlight that shared P-frames serve as a common structural base. At the same time, their discipline-specific lexical fillings reflect the distinctive conceptual focus and discourse practices of each field.

4.6.2 Discussion of the Second Research Question

The findings of this study also provide implications for how user-friendly corpus tools, such as Sketch Engine and kfNgram can be integrated into academic writing instruction to raise students' awareness of discipline-specific phraseological patterns. Although this study did not directly assess learners' performance, the corpus-based approach demonstrates that such tools can act as effective mediators between disciplinary discourse and student awareness of phraseological patterns.

Moreover, the corpus-based DDL activity design suggests that these tools can support students in discovering recurrent P-frames and their collocational behaviour in authentic disciplinary texts. Through the analysis of frequency lists, collocational networks, and concordance lines, learners can independently explore how academic meaning is constructed in their discipline, which is a key principle of Data-Driven Learning (DDL) (Johns, 1991). This can be related to previous studies that emphasise the value of data-driven learning in promoting learner autonomy and inductive learning (Boulton & Cobb, 2017). Similarly, unlike traditional top-down teaching methods, where instructors present ready-made examples, DDL encourages a discovery role for students. By observing authentic language data, students construct their own generalisations about how academic texts work. This process aligns with the principles of learner autonomy and noticing, both of which are central to contemporary EAP pedagogy (Flowerdew, 2015).

As a result, corpus-based DDL can promote more discipline-sensitive writing awareness, fostering both transferable academic skills and specialised discourse competence. Students can observe how discipline-specific terminology is embedded within general academic frameworks by exploring concordance lines and frequency lists. Such hands-on exploration enables students to move beyond direct memorisation and gain a deeper understanding of the language used in their field.

4.6.3 Discussion of the Third Research Question

The analysis of concordance lines and collocational behaviour revealed that several of the most frequent P-frames are systematically associated with specific rhetorical functions, particularly in relation to reporting results, describing procedures, and expressing cause–effect relationships. These findings confirm that P-frames are not only structural elements but also functional in constructing disciplinary argumentation. To support students in recognising and applying these rhetorical patterns, corpus-based data-driven learning (DDL) tasks can be helpful.

From a pedagogical view, these insights demonstrate how corpus-based Data-Driven Learning (DDL) activities can bridge the gap between linguistic form and rhetorical function. By exploring authentic concordance lines, students can observe how particular P-frames perform communicative roles within disciplinary contexts. In this case, students can explore authentic concordance lines, classify the rhetorical roles of P-frames, and apply them to their own writing. Such activities can bridge the gap between linguistic form and rhetorical function, as well as enabling students to use P-frames more deliberately in disciplinary contexts. Therefore, it can be noted that P-frames can be viewed not only as frequent lexical templates but also as cognitive and rhetorical scaffolds that support disciplinary knowledge construction.

CONCLUSIONS AND RECOMMENDATIONS

This study investigated the use and distribution of P-frames in Geology and Mining Engineering research articles. Regarding these disciplines, the study explores the P-frames and their potential contribution to academic writing in engineering fields. The results revealed that both disciplines share several high-frequency words in both three-word and four-word P-frames, which appear to function as general academic structures. Moreover, log-likelihood analyses demonstrated that some of these shared P-frames occurred significantly more often in one corpus than the other. This refers to the discipline-specific preferences. Additionally, the concordance and collocation analyses showed that the lexical items filling types of these P-frames usually reflected field-specific terminology. In other words, when types were analysed one by one, it can be seen. In the case of pedagogical application, several activities have been created. These activities may be seen as a starting point. The activities can be developed.

Pedagogical Implications

The findings of this study highlight the potential of integrating data-driven learning (DDL) into English for Academic Purposes (EAP) engineering contexts. By using user-friendly corpus tools such as Sketch Engine, students have a chance to explore authentic examples of discipline-specific P-frames and understand how they are employed in real academic texts. This approach can help learners to gain an awareness of phraseological patterns. Moreover, thanks to activities, students can use them more accurately and purposefully in their own writing. The corpus-based DDL activities presented in this study may serve as a starting point for integrating corpus-based methods into engineering education. Further developments should be done in order to vary.

Limitations of the Thesis

This thesis has several limitations. Firstly, the analysis was conducted on only two corpora, which may limit the generalisability of the findings to other academic disciplines. Therefore, additional corpora from different fields could have provided a broader picture of the use of p-frames in academic texts. Secondly, the thesis specifically focused on P-frame analysis. While this offers valuable insights into phraseological patterns, other corpus-based approaches were not taken into consideration. Therefore, expanding the scope of study to include broader corpus-based analysis might have brought a more comprehensive understanding of disciplinary variation in academic English.

Further Suggestions

This thesis focused on P-frames in two disciplines, which are Geology and Mining Engineering. As stated in the findings, the analysis was limited to three-word and four-word P-frames, and their types and collocational patterns. Moreover, the study tried to provide pedagogical DDL materials. Further research may expand this study by adding different engineering fields or other academic disciplines to compare similarities and differences in the use of p-frames. Moreover, the scope of analysis can be extended beyond p-frames to provide a more comprehensive analysis of discipline-specific features in academic texts.

Another worth noting suggestion, which should be taken into consideration, is the pedagogical application of corpus-based findings. Tools such as Sketch Engine can be used in English for Academic Purposes (EAP) classrooms to raise students' awareness of both general academic expressions and discipline-specific language patterns. By analysing concordance lines and collocational patterns derived from authentic texts, learners are encouraged to notice how specific phraseological units function in academic discourse and how their meanings shift depending on context. In this way, corpus-based data-driven activities can not only enhance students' understanding of academic English but also contribute to the development of teaching resources and materials for specific disciplines.

REFERENCES

Abdelmageed, Mohamed A. (2024), "Data-Driven Learning: A Practical, Theory-Oriented Approach to Academic Writing", **SSRN**, 1-32. http://dx.doi.org/10.2139/ssrn.4996157

Bao, Kai (2021), "Should We Use It in Our Classrooms: An Analysis of Data-Driven Learning Research", **English Linguistics Research,** 10(3), 66-76. DOI:10.5430/elr.v10n3p66

Bennett, Gena R. (2010), "An Introduction to Corpus Linguistics", In **Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers,** (1-22), University of Michigan Press ELT.

Bessidhoum, Mahfoudh and Ghania, Ouahmiche (2021), "Corpus-Based Instruction in The EFL Classroom: The Case of Data-Driven Learning: 'Spreading the Word", **Djoussour El-maarefa**, *7*(3), 594-613.

Boulton, Alex (2009), "Data-Driven Learning: Reasonable Fears and Rational Reassurance", **Indian Journal of Applied Linguistics**, 35 (1), 81-106.

_____ (2010), "Data-Driven Learning: Taking the Computer Out of The Equation", **Language Learning**, 60(3), 534-572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

_____ (2017), "Data-Driven Learning and Language Pedagogy", Steven Thorne, Stephen May (Eds.), In **Language, Education and Technology: Encyclopedia of Language and Education**, (1-12), Springer.

Boulton, Alex and Cobb, Tom (2017), "Corpus Use in Language Learning: A Meta-Analysis: Meta-Analysis of Corpus Use in Language Learning" **Language Learning,** 67(2)**.**

Braun, Sabine (2006), "ELISA - A Pedagogically Enriched Corpus for Language Learning Purposes" Sabine Braun, Kurt Kohn, Joybrato Mukherjee, Thomas Kohnen (Eds.), In **Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods**, (25-47), Peter Lang.

Casal, J. Elliott and Kessler, Matt (2020), "Form and Rhetorical Function of Phrase-Frames in Promotional Writing: A Corpus- And Genre-Based Analysis", **System**, 95, 1-13. https://doi.org/10.1016/j.system.2020.102370

Chambers, Angela (2010), "What is data-driven learning?", Anne O'Keeffe, Michael McCarthy (Eds.), In **the Routledge Handbook of Corpus Linguistics,** 1st Edition (345–358), Routledge, London.

Chong, Chia Suan (2016), "Ten Innovations That Have Changed English Language Teaching", **British Council**, https://www.britishcouncil.org/voices-magazine/ten-innovations-have-changed-english-language-teaching

Cotos, Elena (2014), "Enhancing Writing Pedagogy with Learner Corpus Data", **ReCALL**, 26(2), 202–224. DOI:10.1017/S0958344014000019

Cortes, Viviana (2014), "Genre Analysis in the Academic Writing Class: With or Without Corpora?", **Quaderns de Filologia-Estudis Lingüístics**, 16, 65-80.

Crosthwaite, Peter, et. al., (2021), "Exploring Language Teachers' Lesson Planning for Corpus-Based Language Teaching: A Focus on Developing TPACK For Corpora And DDL", **Computer Assisted Language Learning**, 1-29. https://doi.org/10.1080/09588221.2021.1995001

Crosthwaite, Peter and Baisa, Vít (2024), "A User-Friendly Corpus Tool for Disciplinary Data-Driven Learning: Introducing Corpusmate", **International Journal of Corpus Linguistics**, 29(4), 595 – 610. https://doi.org/10.1075/ijcl.23056.cro

Çelik, Serkan (2011), "Developing Collocational Competence Through Web Based Concordance Activities", **Novitas-ROYAL (Research on Youth and Language),** 5(2), 273-286.

Eak-in, Suparada (2015), "Effects of A Corpus-Based Instructional Method on Students' Learning of Abstract Writing: A Case Study of an EAP Course for Engineering Students", **Journal of English Studies**, 10, 1-41.

Flowerdew, Lynne (2015), "Using Corpus-Based Research and Online Academic Corpora To Inform Writing of The Discussion Section of A Thesis", **Journal of English for Academic Purposes**, 20, 58 – 68. https://doi.org/10.1016/j.jeap.2015.06.001

Flowerdew, John (2009), "Corpora in Language Teaching", Michael H. Long, Catherine J. Doughty (Eds.), In **the Handbook of Language Teaching**, (327 – 350), Wiley-Blackwell.

Fletcher, William H. (2012). **KfNgram** (Version 1.3.1), [Computer Software], Annapolis, MD: USNA. http://kwicfinder.com/kfNgram/kfNgramHelp.html

Frankenberg-Garcia, Ana (2014), "The Use of Corpus Examples for Language Comprehension and Production", **ReCALL,** 1-19. DOI: 10.1017/S0958344014000093

Gabrielatos, Costas (2005), "Corpora and Language Teaching: Just A Fling or Wedding Bells?", **Teaching English As a Second or Foreign Language**, 8(4).

Garner, James (2016), "A Phrase-Frame Approach to Investigating Phraseology in Learner Writing Across Proficiency Levels", **International Journal of Learner Corpus Research**, 2(1), 31–67. https://doi.org/10.1075/ijlcr.2.1.02gar

Gillett, Andy J. (2022), "What is EAP?", **UEfAP,** https://www.uefap.org/eap-in-general/what-is-eap

Gilquin, Gaëtanelle and Granger, Sylviane (2010), "How Can Data-Driven Learning Be Used in Language Teaching?", Anne O'Keeffe, Michael McCarthy (Eds.), In **The Routledge Handbook of Corpus Linguistics**, 1st Edition (359-370), Routledge, London.

Granger, Sylviane et. al., (2020), **The International Corpus of Learner English**, Version 3, Presses universitaires de Louvain, Louvainla-Neuve.

Golparvar, Seyyed Ehsan and Barabadi, Elyas (2020), "Key Phrase Frames in The Discussion Section of Research Articles of Higher Education", **Lingua**, 236, 1-15. https://doi.org/10.1016/j.lingua.2020.102804

Ha, Althea Ying Ho and Hyland, Ken (2017), "What Is Technicality? A Technicality Analysis Model for EAP Vocabulary", **Journal of English for Academic Purposes**, 28, 35-49. https://doi.org/10.1016/j.jeap.2017.06.003

Hamp-Lyons, Liz (2011), "English for Academic Purposes", Eli Hinkel (Ed.), In **The Handbook of Research on Second Language Learning and Teaching**, Vol. II (89-105), Routledge.

Johns, Tim (1991), "Should You Be Persuaded: Two Examples of Data-Driven Learning", **English Language Research Journal**, 4, 1–16.

_____ (1998), "Contexts: The Background, Development and Trialling of A Concordance-Based CALL Program", Anne Wichmann, Steven Fligelstone (Eds.), In **Teaching and Language Corpora**, (100-115), Routledge, London.

Kilgarriff, Adam et. al., (2014), "The Sketch Engine: Ten Years On" **Lexicography**, 1, 7-36.

Klimova, Blanka Frydrychova (2015), "Designing an EAP Course", **Procedia - Social and Behavioral Sciences**, 191, 634-638. https://doi.org/10.1016/j.sbspro.2015.04.329

Kunilovskaya, Maria and Koviazina, Marina (2018), "Sketch Engine: A Toolbox for Linguistic Discovery", **De Gruyter Open,** 503 – 507. https://doi.org/10.2478/jazcas-2018-0006

Lee, David and Swales, John (2006), "A Corpus-Based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-Compiled Corpora", **English for Specific Purposes**, 25(1), 56-75. https://doi.org/10.1016/j.esp.2005.02.010

Leech, Geoffrey (2007), "New Resources, Or Just Better Ones? The Holy Grail of Representativeness", Marianne Hundt, Nadja Nesselhauf, Carolin Biewer (Eds.), In **Corpus Linguistics and the Web**, (133-150), Rodopi, Amsterdam.

Liu, Chen-Yu and Chen, Howard (2022), "A Phraseological Exploration of University Lectures Through Phrase Frames", **Journal of English for Academic Purposes,** 58, 1-13. https://doi.org/10.1016/j.jeap.2022.101135

Ma, Qing et. al., (2024), "Teacher Paths for Developing Corpus-Based Language Pedagogy: A Case Study", **Computer Assisted Language Learning**, 37(3), 461-492. https://doi-org.ezproxy.eduhk.hk/10.1080/09588221.2022.2040537

Malakhovskaya, Maria et. al., (2021), "Teaching Noun-Phrase Composition In EAP/ESP Context: A Corpus-Assisted Approach to Overcome a Didactic Gap" **The Journal of Teaching English**

for Specific and Academic Purposes, 9(2), 257−266. https://doi.org/10.22190/JTESAP2102257M

Marinov, Sanja (2017), "Using Data-Driven Learning with Students of Tourism: Content- And Vocabulary-Oriented Exercises", **In II. International Conference from Theory to Practice in Language for Specific Purposes**, 153- 168.

McCarthy, Michael and O'Keeffe, Anne (2010), "What Are Corpora and How Have They Evolved?", Anne O'Keeffe, Michael McCarthy (Eds.), In **the Routledge Handbook of Corpus Linguistics**, 1st edition, (3-13), Routledge, London.

Nesselhauf, Nadja (2004), "Learner Corpora and Their Potential for Language Teaching", John Sinclair (Ed.), In **How to Use Corpora in Language Teaching,** (125-152), John Benjamins Publishing Company.

Nordquist, Richard (2025), "Definition and Examples of Corpus Linguistics", **ThoughtCo**, https://www.thoughtco.com/what-is-corpus-linguistics-1689936

O'Sullivan, Íde (2007), "Enhancing A Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy", **ReCALL**, 19(3), 269-286. DOI: 10.1017/S095834400700033X

Ordoñana, Carlos (2023), **Exploring Alternate Vehicles for DDL İntegration İnto Teaching Practices: A Case for Pedagogic Natural Language Processing Resources**, Unpublished PhD Thesis, University of Murcia.

Otto, Philippa (2021), "Choosing Specialized Vocabulary to Teach with Data-Driven Learning: An Example From Civil Engineering", **English for Specific Purposes**, 61, 32-46. https://doi.org/10.1016/j.esp.2020.08.003

Önder Özdemir, Neslihan (2014), "Using Corpus Data to Teach Collocations in Medical English", **Journal of Second Language Teaching & Research**, 3(1), 37 – 52.

Robinson, Pauline C. (1991), **ESP Today: A Practitioner's Guide**, Prentice Hall, London.

Römer, Ute (2009), “The Inseparability of Lexis and Grammar: Corpus Linguistic Perspectives”, **Annual Review of Applied Linguistics**,7, 140–162.

_____ (2011), “Corpus Research Applications in Second Language Teaching”, **Annual Review of Applied Linguistics**, 31, 205–225. DOI: 10.1017/S0267190511000055

Rudneva, Maria (2020), “Fostering Learning with The Ecolexicon Corpus in The ESP Classroom”, Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, Sylvie Thouësny (Eds), In **CALL For Widening Participation: Short Papers from EUROCALL 2020,** (304-308), Research-publishing.net.

Rühlemann, Christoph (2019), **Corpus Linguistics for Pragmatics: A Guide for Research**, Routledge.

Stubbs, Michael (2007), “An Example of Frequent English Phraseology: Distributions, Structures, and Functions”, Roberta Facchinetti (Ed.), In **Corpus Linguistics 25 Years On**, (89-105), Rodopi, Amsterdam.

Sinclair, John (1991), **Corpus, Concordance, Collocation**, Oxford University Press, Oxford.

_____ (2004), **Trust The Text: Language, Corpus and Discourse**, Routledge, London.

Smart, Jonathan (2014), “The Role of Guided Induction in Paper-Based Data-Driven Learning”, **ReCALL**, 26(02), 184-201.

Tognini-Bonelli, Elena (2004), “Working with Corpora: Issues and Insights”, Caroline Coffin, Ann Hewings, Kieran O'Halloran (Eds.), In **Applying English Grammar**, (11-24), Routledge, London.

Vincent, Benet and Nesi, Hilary (2018), “The BAWE Quicklinks Project: A New DDL Resource for University Students”, **Lidil,** 58, 1-18.

Xu, Jiajin (2022), “A Historical Overview of Using Corpora in English Language Teaching”, Reka R. Jablonkai, Eniko Csomay (Eds.), In **The Routledge Handbook of Corpora and English Language Teaching and Learning**, (11-25), Routledge.

APPENDICES

Appendix 1: List of Geology Corpus Top 500 P-Frames

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* of the	10175	779	the * of the	2523	296
of the *	9176	933	in the * of	700	71
the * of	8117	572	of the * of	281	49
in the *	5498	509	* shown in fig	257	7
* in the	5255	562	* the study area	257	9
* to the	2542	178	* based on the	252	33
to the *	1911	298	to the * of	249	43
to * the	1859	152	bulletin of the *	240	6
* for the	1697	247	on the * of	224	37
* on the	1655	141	for the * of	217	37
for the *	1652	239	based on the *	203	37
the * and	1607	251	in this study *	186	18
on the *	1464	208	the * in the	184	26
* and the	1429	224	in this * the	179	16
and the *	1403	250	of the * and	178	33
* that the	1355	96	in the * area	174	12
a * of	1166	103	is * to the	165	18
* with the	1042	105	the study area *	164	20
* from the	1029	129	* part of the	164	21
can be *	979	84	of the * area	160	10
* by the	959	89	the * between the	159	18
from the *	887	153	and * of the	158	34
in this *	883	26	the * of a	158	23
with the *	867	136	in order to *	156	29
based on *	793	42	in the * and	155	28
that the *	762	148	the * part of	151	18
of * and	730	112	it is * that	149	17
is * to	729	62	* the number of	149	19
* is the	725	79	in the study *	146	5
* can be	717	83	in this * we	146	8
* in fig	658	36	* in this study	145	21
* based on	651	87	can be * as	142	13
at the *	640	81	it is * to	140	9
is the *	635	111	as shown in *	138	7
and * the	622	77	a * of the	136	19
the * in	610	71	used to * the	135	23
it is *	591	57	the * intensity bin	135	5
and * of	566	104	of the study *	134	4
the * area	531	32	as well as *	133	12
* in this	519	76	the results of *	129	9
* of a	514	90	the * and the	129	30
from * to	513	30	as a * of	129	8
the study *	502	19	in * of the	126	10
in * study	492	5	the effect of *	124	11
by the *	491	97	the number of *	124	21
* applied geophysics	479	6	and the * of	123	27

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
are * in	475	40	is the * of	123	23
the * is	475	90	with the * of	120	20
* the study	470	17	at the * of	119	17
* at the	470	82	is * in the	116	16
shown in *	464	9	in the * region	115	18
* the same	462	35	* the case of	113	4
due to *	453	25	it can be *	112	11
* it is	446	61	with regard to *	112	2
we * the	444	57	can be * to	112	7
is * in	443	46	* due to the	109	20
the results *	441	23	on the other *	107	2
* as the	439	39	on the * hand	107	2
of a *	438	73	are shown in *	105	5
over the *	414	43	are * in table	102	6
* this study	414	11	* the effect of	100	17
with a *	411	71	is shown in *	100	5
in * the	409	46	part of the *	100	18
* shown in	405	12	is * by the	98	19
* geophysical research	404	18	* results of the	95	5
* between the	398	46	* summer monsoon rainfall	94	6
* to be	396	47	in the * part	92	15
the * the	389	76	as * as the	92	3
of * research	369	6	where * is the	91	20
data * the	361	26	* the presence of	91	11
* used to	358	15	* one of the	90	9
is * by	356	44	the results * the	88	4
the * data	356	51	* the results of	88	17
* analysis of	350	32	the gulf of *	87	5
of geophysical *	349	5	be * to the	86	7
during the *	349	48	of the * in	86	18
the * between	348	34	indian summer monsoon *	85	5
in * of	347	24	* that of the	85	7
the data *	340	35	one of the *	84	7
and applied *	340	7	the * distribution of	83	5
* with a	331	61	order to * the	82	16
are * to	330	40	due to the *	81	17
number of *	328	49	* the gulf of	81	5
* study area	327	6	can be * that	81	6
* due to	323	51	can be seen *	80	6
this study *	322	33	* gulf of aden	79	4
the * region	317	41	shows a * correlation	78	2
that * the	313	35	with a * of	78	14
society of *	312	8	results * that the	77	8
used to *	308	47	the * of this	77	11

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* the data	308	29	* changes in the	77	10
* the model	307	20	* bay of bengal	77	5
is a *	303	65	can be * by	76	14
as a *	303	36	used in this *	75	5
the model *	300	41	we * that the	75	10
to be *	300	51	can be used *	74	4
the same *	298	42	be * that the	74	6
this * is	296	32	is * on the	73	5
* using the	295	46	* increase in the	73	3
* distribution of	295	19	* spatial distribution of	73	4
analysis of *	293	30	in the present *	72	4
between * and	291	27	* a depth of	72	3
be * to	291	21	are * in the	72	15
in a *	290	58	* of the study	72	12
* over the	289	54	the spatial * of	72	4
part of *	289	8	summer monsoon rainfall *	71	16
* the fault	288	19	an increase in *	69	6
of * in	284	43	the presence of *	69	9
as * as	284	11	* the use of	69	12
for * the	284	34	* this study the	68	2
* the results	281	35	* related to the	68	8
in * and	277	45	is * of the	68	3
which is *	273	35	* distribution of the	68	6
the fault *	271	28	* effect of the	68	3
* changes in	270	23	a case study *	68	5
* in a	269	50	in * to the	68	6
according to *	266	13	* can be seen	67	2
changes in *	265	26	of * in the	67	14
* shows the	264	18	can be * in	67	5
the number *	263	2	* an increase in	67	7
model * the	263	17	in * study the	66	2
it * be	260	9	a depth of *	65	6
* part of	258	32	is * that the	65	10
this * the	253	24	* the present study	65	4
as * in	250	16	* correlation with the	65	5
as the *	250	54	* intensity bin shows	64	5
the seismic *	248	30	* analysis of the	64	9
* results of	247	11	* consistent with the	64	7
* the seismic	246	17	* be used to	63	5
of * data	245	20	of the * model	63	13
results * the	245	11	the * number of	63	8
* effect of	243	5	* than that of	62	7
* to a	242	29	intensity bin shows *	62	3
there is *	240	7	* of the earth	62	4

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
of this *	240	24	the western * of	62	6
the time *	238	25	* bin shows a	62	3
with * to	237	7	* freezing and thawing	61	6
the * society	234	7	* there is a	61	8
between the *	233	38	freezing and thawing *	61	4
the * method	232	40	* most of the	61	9
as well *	230	3	by the * of	61	10
has been *	230	39	are presented in *	60	6
distribution of *	229	25	intensity bin * a	60	2
* during the	229	47	* the west coast	60	4
* the two	229	17	than that * the	60	3
of * study	226	3	that the * of	59	15
using the *	226	46	in the * study	59	3
shows a *	224	21	in the * direction	59	10
compared to *	223	10	* located in the	59	7
by * the	222	32	is located * the	59	6
to * a	221	31	* shows that the	59	10
depth of *	220	13	is based on *	59	3
* depth of	218	12	used in * study	58	2
values * the	218	13	bin shows a *	58	5
* as a	217	40	of this * is	58	6
the * value	217	27	* amplitude of the	58	4
the * part	217	22	with respect to *	58	3
the * for	216	34	* standard deviation of	58	5
* the other	214	12	than * of the	57	3
up to *	214	13	than that of *	57	6
* the source	213	19	over the * coast	57	2
was * to	212	23	* magnitude of the	57	3
of * the	211	35	the effects of *	57	8
* shows a	211	18	the magnitude of *	57	2
* the north	211	11	the study * the	57	6
obtained * the	211	11	correlation with the *	56	8
along the *	210	37	of the earth *	56	3
results of *	209	16	with * to the	56	3
* the earthquake	207	15	the results * that	56	7
used * the	207	8	an * in the	55	2
* research letters	206	7	of the fault *	55	8
study area *	206	27	may be * to	55	5
used in *	203	5	* the same as	55	7
* increase in	203	10	in the frequency *	55	4
* the upper	199	14	in terms of *	55	7
* time series	198	17	in the region *	54	11
an * of	197	30	in the * zone	54	7
time series *	197	19	the case of *	54	4
and * in	197	40	in the * the	54	12
* the region	197	7	* it can be	54	8

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
the upper *	196	16	* obtained from the	54	11
the source *	195	28	in this case *	53	7
* value of	192	18	* observed in the	53	7
is * the	191	25	the * proposed by	53	4
is * as	191	25	of the model *	52	10
* than the	189	19	* value of the	52	7
study * the	188	16	in the north *	52	8
* terms of	188	2	at the * station	51	3
be * as	188	16	as a result *	51	2
the * time	188	25	* depth of the	51	6
the earthquake *	187	31	of * of the	51	9
the * to	187	33	* used in the	51	8
* the time	187	19	* the bay of	51	5
spatial * of	185	14	in the upper *	51	7
* used in	184	20	* the effects of	50	9
* the indian	184	13	the west coast *	50	6
the indian *	184	11	is * than that	50	9
located * the	183	10	of the seismic *	50	10
parameters * the	183	11	the distance * the	50	4
the * zone	182	16	the use of *	49	8
* in table	182	8	in * with the	49	5
the north *	181	18	* time of the	49	6
* has been	181	25	with * increase in	49	2
the inversion *	179	16	* geophysical research letters	49	11
* intensity bin	179	7	* this study we	48	2
* the most	178	17	are * to the	48	11
the * surface	178	15	bay of bengal *	48	8
* such as	178	31	* top of the	48	2
* the observed	178	22	in the western *	48	7
* of this	177	24	* in order to	48	8
* the south	177	11	shows a positive *	48	2
the * intensity	176	10	for * of the	48	6
have been *	176	29	in the south *	48	7
and its *	175	29	the other hand *	48	3
observed * the	174	12	* in the study	47	12
* that of	173	10	can be expressed *	47	2
is * that	173	20	* the fault zone	47	3
effect of *	173	16	* found that the	47	7
* seismic data	172	21	the location of *	47	3
the other *	172	12	* of the most	46	2
we * that	172	17	in this * is	46	7
for a *	171	31	a function of *	46	5
* the maximum	171	20	physics of the *	46	2
for each *	171	33	* a positive correlation	46	2
the most *	169	25	the amplitude * the	46	2

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* case of	168	4	earth and * science	46	2
this * we	168	11	* the ground surface	46	6
* the following	167	16	* accuracy of the	46	2
* effects of	167	7	* structure of the	46	9
is used *	167	6	at * depth of	46	2
* by a	166	29	the top * the	46	2
* carried out	166	13	pet based on *	45	2
* study the	166	7	the extreme intensity *	45	2
analysis * the	166	9	* show that the	45	5
* the surface	165	13	a large * of	45	3
the presence *	165	2	of the earthquake *	45	11
increase in *	164	18	the same as *	45	6
one * the	164	4	increase in the *	45	8
the * are	163	24	the northern * of	45	6
is * for	163	24	from the * of	45	9
the vertical *	163	23	that there is *	44	3
* which is	163	32	a * number of	44	4
the maximum *	162	24	* parameters of the	44	7
to a *	162	33	* of the earthquake	44	6
there * a	162	4	the present study *	44	7
one of *	162	3	before and after *	44	4
changes * the	160	8	at the same *	44	2
of * of	160	27	of the * the	44	11
distribution * the	160	8	* pet based on	44	3
time * the	159	13	can be * from	44	6
* the vertical	159	18	* be noted that	44	2
the correlation *	158	13	* the magnitude of	44	10
the * distribution	158	20	is * with the	44	3
* the number	158	20	* attributed to the	44	3
* comparison of	158	10	* applied to the	44	8
* the west	157	12	* the analysis of	43	6
value * the	157	8	* caused by the	43	10
* associated with	156	22	* found to be	43	4
in * with	156	15	the other * the	43	2
order to *	156	29	the difference * the	43	3
we * a	155	24	very high * extreme	43	2
* use of	155	4	* parts of the	43	9
* data and	155	18	of the indian *	43	6
the case *	155	6	and * in the	43	11
the following *	154	26	of the * is	43	10
values of *	154	17	* a negative correlation	43	2
* values of	153	19	changes in the *	43	11
is * than	152	11	* the upper atmosphere	42	5
the region *	152	22	inversion of * data	42	3

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
of seismic *	152	23	* idriss and boulanger	42	6
less than *	152	12	* the indian ocean	42	5
the * results	151	18	most of the *	42	11
associated with *	151	10	for the period *	42	4
structure * the	151	11	* of interest the	42	2
* study of	151	9	the changes in *	42	5
* have been	150	24	* be attributed to	42	2
area * the	150	15	the location * the	41	2
value of *	150	13	* compared to the	41	4
be * that	150	9	* in this paper	41	7
of earthquake *	150	21	* shows a positive	41	6
intensity bin *	149	13	* location of the	41	2
the area *	149	16	in the * section	41	4
* pure and	148	15	in the area *	41	3
there are *	148	22	* variation in the	41	3
* the area	147	9	compared to the *	41	10
* used for	146	13	* large number of	41	2
be * by	146	22	* the area of	40	4
* the inversion	145	10	to the * and	40	8
* related to	144	14	the performance of *	40	3
the * wave	144	16	* correlation coefi cient	40	6
gulf of *	144	6	is * to be	40	6
in addition *	144	8	* a result of	40	2
the period *	143	10	* the top of	40	4
* the tsunami	143	15	* to determine the	40	5
* the case	142	7	of the * zone	40	4
* the average	142	22	in this work *	40	4
* inversion of	142	16	* the changes in	40	4
was * by	142	24	used in the *	40	8
* gulf of	142	5	the ratio * the	39	2
characteristics of *	141	12	* the range of	39	2
carried out *	141	16	it * possible to	39	2
the average *	141	21	in the * domain	39	5
the frequency *	141	10	over the indian *	39	4
the south *	141	13	* the location of	39	8
* the western	140	10	* result of the	39	3
* the main	140	20	of the data *	39	7
this paper *	140	16	* noted that the	39	5
* there is	139	12	* comparison of the	39	4
the present *	138	6	* to note that	39	2
the observed *	138	17	* earth and planetary	39	3
presented in *	138	8	the * value of	39	7
effect * the	137	5	can be obtained *	39	5
* applied to	137	12	* size of the	39	3
* be used	137	9	is located in *	39	2

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
are shown *	136	6	the data * the	39	5
data and *	136	17	* close to the	38	6
related to *	136	4	very high intensity *	38	4
* and its	135	29	* spatial and temporal	38	4
the two *	135	28	* the development of	38	8
is shown *	134	7	* of earth sciences	38	4
was used *	134	7	positive correlation * the	38	2
well as *	134	12	* the distribution of	38	8
* bulletin of	134	26	* the study region	38	4
* the northern	134	12	a negative correlation *	38	8
* the present	133	6	the fault zone *	38	9
may be *	133	20	in the east *	38	5
the west *	133	8	as * in the	38	9
be * in	133	16	it * be noted	38	2
the western *	133	17	* in the northeastern	38	8
most of *	132	4	* of the fault	38	11
is not *	132	26	the area of *	38	5
of * seismic	132	7	* evolution of the	38	5
than that *	132	7	in the range *	38	2
* the earth	131	6	* compared with the	38	8
the * values	131	24	was * in the	38	9
results * that	130	10	seismic * in the	37	6
be used *	130	4	a positive correlation *	37	3
an * in	130	6	spatial distribution of *	37	5
seismic data *	130	21	carried out * the	37	7
that of *	130	10	was carried out *	37	8
* the total	130	14	* line represents the	37	4
the tsunami *	130	19	shows the * of	37	8
the initial *	129	17	* function of the	37	5
the spatial *	129	12	* values of the	37	3
* the second	128	10	the values of *	37	3
* range of	128	9	it is possible *	37	2
the * function	127	14	* the time series	37	6
of * earth	127	3	the * area is	37	3
* the ground	127	13	the * area the	37	3
the surface *	126	16	* be due to	37	3
* seismic hazard	126	10	* each of the	37	3
region * the	126	16	it was * that	36	5
correlation * the	126	6	* would like to	36	4
* monsoon rainfall	126	4	in the northern *	36	4
the * study	126	6	of the * data	36	7
* characteristics of	125	13	* associated with the	36	5
magnitude of *	125	8	* occurred in the	36	5

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
of * model	125	6	is a * of	36	7
* the east	125	8	* very high intensity	36	2
to determine *	125	4	in * case the	36	3
seismic hazard *	125	11	is * to note	35	2
* solid earth	125	3	of this study *	35	6
to obtain *	124	7	according to the *	35	10
* series of	124	4	the depth of *	35	3
* the period	124	6	the distribution of *	35	3
the * coast	124	8	the relationship between *	35	2
the * was	124	27	the low intensity *	35	2
the * parameters	123	21	in the middle *	35	3
* ground motion	123	13	is * as the	35	6
* the frequency	123	12	is * for the	35	8
are * by	123	22	the relationship * the	35	2
the analysis *	122	11	a * study of	35	3
the amplitude *	122	8	the calculated pet *	35	2
* presented in	122	8	the frequency of *	35	7
the effects *	122	2	magnitude of the *	35	4
of * earthquake	122	7	* be related to	34	5
* compared to	122	15	the main * of	34	10
caused by *	121	4	* shows a negative	34	4
comparison * the	121	4	* results and discussion	34	2
* for a	121	29	* as well as	34	8
the proposed *	121	12	a total of *	34	3
the * that	121	17	is observed * the	34	3
* the eastern	121	10	of the * wave	34	4
is * on	120	7	with an * of	34	5
* role of	120	3	oil and gas *	34	6
were * to	120	18	of the * region	34	5
for * of	120	19	* listed in table	34	3
close to *	120	7	* end of the	34	5
* the correlation	120	15	the frequency * of	34	5
* the initial	119	15	results are * in	34	4
solid earth *	119	3	are * to be	34	6
applied * the	119	6	is * than the	34	5
* of earthquake	119	18	the value of *	34	3
a * correlation	119	6	there is a *	33	6
an increase *	119	2	of the * function	33	4
* caused by	118	22	seasonal * in the	33	2
* magnitude of	118	7	are * from the	33	8
* determine the	118	4	* the critical angle	33	4
* the central	118	10	a * correlation with	33	3
depth * the	117	8	the * structure of	33	7
and a *	117	25	* to estimate the	33	7

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* along the	117	31	the time * of	33	4
* the previous	117	8	the vertical * of	33	5
institute of *	117	18	* calculated pet based	33	2
than * of	117	4	was * by the	33	4
* than that	116	9	from the * model	33	6
* which the	116	7	* of this study	33	5
* is not	116	18	* the role of	33	5
the main *	116	26	* in terms of	33	9
* a new	116	14	* supported by the	33	3
* obtained from	115	15	* study area the	33	2
used for *	115	17	it * observed that	33	2
* represents the	115	19	that of the *	33	8
* estimate the	115	4	* into account the	33	4
* maps and	115	3	in the central *	33	4
* the top	115	11	is less than *	33	7
of * city	115	6	during the * of	33	5
maps and *	115	3	the time series *	33	4
the * source	115	16	on the * surface	32	5
the * layer	114	19	of the total *	32	7
earthquakes * the	114	9	to * depth of	32	2
* apparent resistivity	114	10	the study * is	32	2
is * of	114	8	* study of the	32	4
the top *	114	13	in addition to *	32	2
* function of	114	9	the analysis of *	32	3
of earth *	114	8	* high intensity bin	32	2
* found that	114	9	the * of an	32	5
* data set	113	15	ground motion prediction *	32	2
* the original	113	11	spatial and temporal *	32	6
inversion of *	113	13	* change in the	32	4
as * the	113	9	to the * in	32	7
associated * the	113	2	are * with the	31	5
regard to *	113	2	in addition * the	31	2
the magnitude *	113	7	at the top *	31	2
consistent with *	113	6	* below ground level	31	2
which * the	112	17	* trend in the	31	4
this is *	112	14	* generated by the	31	6
in * area	112	4	the role of *	31	6
zone * the	112	12	* gravity and magnetic	31	5
effects of *	112	16	* earthquakes in the	31	6
shows that *	111	3	* results show that	31	3
data set *	111	18	as compared * the	31	2
* evolution of	111	11	northern * of the	31	4
increase * the	111	3	* time series of	31	5
* the proposed	111	12	* this study is	31	3
* area of	111	9	of the royal *	31	4

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
obtained from *	110	6	* study area is	31	2
* climate change	110	13	of the upper *	31	4
of * fault	110	4	to * that the	31	6
the northern *	110	15	* number of events	31	5
the east *	110	10	* effects of the	31	4
* observed in	110	11	* this article is	31	3
the depth *	109	12	* difference between the	31	2
the focal *	109	10	* used in this	30	5
the * with	109	22	* and west up	30	3
* time of	109	13	be * by the	30	6
* less than	109	13	is * from the	30	6
* change in	109	10	* the simulation of	30	3
* the lower	109	11	high and extreme *	30	3
* the city	109	9	* the increase in	30	5
which * be	108	5	* the frequency of	30	2
does not *	108	21	the range of *	30	6
the total *	108	18	over the * region	30	4
ground motion *	108	15	from the * to	30	6
the * period	108	19	the * for the	30	8
* the local	108	16	the average * of	30	6
the gravity *	107	13	also * that the	30	7
found * the	107	7	the dry intensity *	30	2
the * hand	107	2	western * of the	30	4
* the current	107	11	which can be *	30	7
* other hand	107	2	the * to the	30	6
for this *	107	17	* the end of	29	4
* the effect	107	18	be * in the	29	4
line * the	107	10	as compared to *	29	2
with an *	107	12	in the west *	29	4
the second *	107	17	in * agreement with	29	2
more than *	107	10	the * from the	29	5
* the three	107	6	the same * the	29	3
note that *	107	10	the * line represents	29	4
* as follows	106	12	* east of taiwan	29	5
* indian summer	106	8	the * component of	29	5
magnitude * the	106	6	structure of the *	29	8
a * study	106	6	* is used to	29	6
located in *	106	3	time series * the	29	5
* set of	106	9	the inversion of *	29	4
the mean *	105	16	as a * the	29	3
* consistent with	105	9	in the same *	29	6
* negative correlation	105	6	the total * of	29	3
similar to *	105	6	* note that the	29	3
the ground *	105	7	the * of earthquakes	29	4
* obtained by	104	11	the ratio of *	28	3

Appendix 1: (Continued)

Geology Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* one of	104	10	* used for the	28	4
the thermal *	104	10	of different rainfall *	28	2
* of earthquakes	104	17	the end of *	28	2
very high *	104	8	spatial * of the	28	2
well * the	104	3	that * of the	28	5
are used *	103	7	western gulf of *	28	3
* research solid	103	2	* according to the	28	4
such as *	103	9	the time of *	28	2
* all the	103	11	* oil and gas	28	5

Appendix 2: List of Mining Corpus Top 500 P-Frames

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
the * of	10820	714	the * of the	3022	321
* of the	10719	816	in the * of	838	80
of the *	9900	908	on the * of	376	58
in the *	5269	518	as shown in *	358	15
* in the	5210	613	* science and technology	357	4
* to the	2407	164	science and technology *	346	15
to * the	2085	179	of mining * and	319	7
* and the	1966	290	of mining science *	295	3
and the *	1842	290	* mining science and	290	2
* on the	1822	147	to the * of	279	60
to the *	1720	276	it is * to	275	16
on the *	1612	230	of the coal *	263	19
the * and	1506	244	of the * of	262	53
a * of	1284	106	* based on the	259	47
* for the	1086	167	in * of the	254	9
of mining *	1062	57	and the * of	254	51
for the *	1049	178	shown in fig *	246	10
* that the	1009	84	with the * of	236	26
can be *	1000	99	* shown in fig	222	4
* of mining	952	64	* the working face	211	12
of * and	945	141	based on the *	204	40
the mining *	908	67	in the * area	201	16
* from the	873	109	in the mining *	196	11
* the mining	868	36	* the number of	195	21
and * of	859	152	of the * and	189	34
* by the	823	88	of the * pillar	182	4
is * to	817	74	for the * of	182	32
based on *	801	42	* the coal seam	176	11
it is *	773	75	is the * of	175	25
in this *	756	35	of * coal seam	174	10
* with the	752	105	at the * of	172	16
* can be	749	93	of the mining *	171	22
from the *	742	114	the * between the	170	15
is the *	735	111	* proceedings of the	169	2
* is the	730	92	in proceedings of *	169	2
shown in *	728	19	used to * the	168	28
* coal seam	714	39	the * of a	164	28
at the *	703	83	of the * face	162	6
the coal *	691	33	* the rock mass	159	5
with the *	686	110	due to the *	159	28
* shown in	680	8	of the rock *	152	13
of coal *	670	44	the * in the	151	32
coal seam *	648	54	the number of *	150	27
* based on	647	96	the working face *	149	14

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* the coal	644	29	* part of the	149	21
the * in	633	97	the coal seam *	148	22
of * coal	607	38	is shown in *	147	10
mining * and	600	38	in order to *	145	28
and * the	595	94	* the rib pillar	145	3
due to *	592	30	heavy metal loid *	144	2
according to *	551	19	according to the *	140	33
of * mining	531	25	shown in figure *	140	2
of * in	529	87	it is * that	139	20
that the *	527	105	of * in the	137	28
* of a	525	93	of the rib *	136	3
* of coal	525	84	the results of *	133	7
* number of	519	22	in this study *	133	14
* coal mining	519	37	and * of the	132	26
the * is	518	90	* shown in figure	130	4
in * mining	518	29	* the mining industry	130	6
the * area	512	45	* j rock mech	129	2
in * of	503	26	* the use of	128	16
* used to	469	14	of * working face	128	3
as * in	469	17	as a * of	127	6
by the *	465	94	* due to the	127	24
* it is	465	54	* according to the	127	21
* at the	462	87	in this * the	126	12
* working face	453	18	as * in fig	125	2
coal mining *	449	44	the rib pillar *	125	14
the * the	448	71	* of the coal	125	26
* as the	442	42	of the * seam	124	5
science and *	441	6	are shown in *	122	9
* to be	440	53	of the * coal	121	13
* and technology	430	4	as well as *	120	9
the results *	427	27	as a result *	119	4
* science and	425	10	it can be *	119	13
are * in	413	32	with a * of	117	16
in * to	410	10	* one of the	116	10
in fig *	410	11	of the working *	114	2
* in this	403	58	at a * of	114	8
of a *	402	65	the * of mining	114	18
and technology *	400	18	of * rock mass	114	3
is * in	399	40	* as shown in	114	20
as a *	399	45	china university * mining	113	2
* between the	397	43	* the study area	113	7
of * rock	396	16	can be * as	112	14
from * to	388	28	china university of *	111	3
the * face	388	11	* of the rock	111	16
to be *	385	64	china coal soc *	111	2

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
science * technology	385	2	of the * is	110	20
with a *	385	62	of the * in	109	27
* rock mass	384	17	* the case of	108	3
* the same	383	26	in the case *	108	2
as shown *	379	2	at the same *	106	3
* the rock	378	20	in fig * the	105	7
in * the	370	30	of heavy metal *	104	4
the rock *	368	27	of the * area	102	10
mining * the	366	34	one of the *	102	9
* this study	361	13	* of the mining	102	21
the * between	356	30	can be seen *	101	5
working face *	355	28	university of * and	101	3
* the mine	347	19	university of mining *	101	2
the * coal	347	34	* in this study	100	19
used to *	346	53	the * and the	100	23
was * to	343	38	from the * of	100	15
in * study	337	7	in the * and	99	21
in a *	336	57	the * of this	98	16
* the main	334	30	of the * mass	97	2
* due to	334	59	on the * hand	97	2
as * as	332	9	the * value of	96	15
of * science	330	5	the * part of	96	13
to * a	327	32	* results of the	95	4
* is a	326	30	of mining and *	95	7
* in fig	326	13	the stability of *	94	6
number of *	322	55	it * possible to	93	2
that * the	321	33	rock mech min *	93	4
* such as	319	52	a * of the	93	11
the number *	315	3	the use of *	93	15
mining science *	312	3	* a result of	93	3
* mining and	311	23	* of the rib	92	11
the * to	303	56	in the * industry	92	4
the working *	302	13	* bairin left banner	92	5
* mining science	299	2	as * as the	91	3
* mining area	295	17	min sci technol *	91	3
is * by	295	36	the rock mass *	89	17
mining and *	294	28	as * result of	89	2
* use of	292	10	of * and technology	89	2
rock mass *	289	40	can be * that	89	9
this study *	286	32	is * by the	88	14
of * the	283	48	* the results of	88	12
be * to	281	28	* the mining area	87	3
the * pillar	280	4	the total * of	87	10
* with a	280	63	is * to the	87	15
it * be	278	10	of the * the	87	18
the * mining	278	42	can be * to	86	5

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
in order *	276	3	* the mining-geology-petroleum engineering	86	2
the * seam	274	11	the results * the	86	2
in * and	274	42	mining-geology-petroleum engineering bulletin *	86	2
the study *	274	17	the * of coal	85	16
* the working	273	15	it is possible *	85	2
and * in	271	51	as * in figure	84	2
the mine *	270	39	* can be seen	84	2
* coal mine	269	35	engineering * and the	84	2
between the *	268	50	can be used *	84	4
* analysis of	268	23	* value of the	84	9
* as a	267	49	* analysis of the	84	9
the main *	267	46	* of mining and	84	5
* the most	266	24	under the * of	83	11
the most *	266	30	the * coal seam	83	8
as the *	265	53	a case study *	83	6
the same *	264	31	in * mining area	82	4
* the surface	263	14	the impact of *	82	8
* according to	262	43	are * in the	81	7
mining * in	261	30	part of the *	81	17
the * model	260	29	* strength of the	80	7
under the *	258	24	the mining industry *	79	13
* results of	256	14	in the coal *	79	10
* the roof	256	18	this is an *	79	2
as well *	255	2	the distance * the	78	4
* the study	253	16	creative commons licence *	78	2
* in a	251	36	that the * of	78	18
part of *	251	6	can be * by	78	9
which is *	251	35	* creative commons licence	78	3
value of *	251	24	is * of the	77	2
* value of	250	20	* the development of	77	9
in figure *	250	5	is * in fig	76	2
of this *	248	24	than * of the	75	4
an * of	248	24	the * working face	75	6
* characteristics of	248	29	a total of *	75	8
the * stress	248	24	* increase in the	75	5
of rock *	248	28	* the drawing body	74	4
part * the	247	6	* the stability of	74	14
* development of	247	15	bairin left banner *	74	8
characteristics of *	247	19	a large * of	74	3
is a *	241	47	* the impact of	74	13
* carried out	240	12	* area of the	74	8

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
there is *	239	13	* into account the	73	4
mining area *	238	28	* included in the	73	3
* heavy metal	238	21	* located in the	73	7
the * for	238	35	be * that the	73	8
is * the	236	33	metal loid * in	72	2
area * the	236	19	* than that of	72	7
by * the	235	29	* of heavy metal	72	8
heavy metal *	234	13	of the roof *	72	13
coal mine *	233	36	* length of the	72	4
the roof *	233	25	* shown in table	72	3
when the *	232	43	* the hydraulic support	71	5
this * the	231	24	* the area of	71	5
such as *	229	28	carried out * the	71	8
the surface *	227	27	in this paper *	71	10
during the *	227	35	results * that the	70	6
the upper *	226	21	* thickness of the	69	5
this * is	224	26	of the * rock	69	12
in * coal	223	15	* height of the	69	5
carried out *	223	17	the main * of	69	13
are * to	223	35	than that * the	69	4
coal * in	222	19	* be used to	68	5
analysis of *	221	22	during the * of	68	10
in uence *	220	11	the development of *	68	10
in table *	219	6	* of the total	68	5
one * the	219	7	of the * body	67	5
* the results	219	21	the results * that	67	6
* mining industry	218	9	* this is an	67	4
of * mine	217	9	are * in table	67	5
* part of	216	27	in the area *	66	6
* the upper	214	17	in * to the	66	5
* the total	213	19	* the presence of	65	9
the total *	213	28	the mining * the	65	4
the maximum *	212	36	* a depth of	65	4
there are *	211	25	* top coal caving	64	5
the * surface	210	23	with the increase *	64	2
results * the	208	9	the length of *	64	2
results of *	208	10	the area of *	64	6
the * mass	207	3	in * rock mass	64	2
coal * the	207	19	is * in the	63	8
the * value	206	26	the relationship between *	63	4
coal * and	205	20	in the process *	63	2
* the number	205	22	the drawing body *	63	5
this is *	204	14	* the process of	63	2
* impact of	203	8	* of the working	62	16
the * was	202	40	in the rock *	62	3

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* in figure	202	10	order to * the	62	15
* amount of	199	10	strength of the *	62	7
* assessment of	199	17	* the mining of	62	7
has been *	198	37	of the shearer *	61	7
were * to	197	20	* stability of the	61	2
* area of	197	14	included in the *	61	4
* hydraulic support	195	9	* related to the	61	8
in addition *	195	12	* characteristics of the	61	7
was used *	194	7	* in the mining	61	11
* the other	194	10	environ monit assess *	61	2
* effect of	193	11	can be * in	60	6
* of this	192	30	a * number of	60	4
in * case	192	5	the case of *	60	6
the * industry	191	5	with an average *	60	7
* case study	189	10	is located * the	60	6
* the following	188	22	where * is the	60	12
study * the	188	17	was carried out *	60	9
were * in	187	34	the upper * of	59	8
is shown *	186	5	coal * in the	59	10
which * the	186	29	the effect of *	59	9
the lower *	186	15	* mining of the	59	5
used in *	186	11	the area * the	58	4
of heavy *	185	3	* coal seam no	58	3
development of *	185	22	* the range of	58	4
area of *	185	20	of * hydraulic support	58	3
* the area	184	9	the study area *	58	10
* study area	183	10	of the mine *	58	11
* the average	183	20	of coal seam *	57	5
is an *	182	9	* of the shearer	57	9
land use *	181	18	in * with the	57	6
the area *	181	15	of mining technology *	57	3
rock mech *	181	4	* fully mechanized mining	57	5
mining of *	180	16	of science and *	57	3
* the lower	180	16	creative commons * and	57	2
in * area	179	6	the * number of	56	8
* under the	179	21	* that of the	56	4
used * the	178	6	at * depth of	56	2
* than the	178	13	the height of *	56	4
the rib *	178	3	* relationship between the	56	2
distance * the	178	9	* it can be	56	8
* distribution of	177	19	* the increase of	56	2
in mining *	177	19	* this study the	55	3

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* using the	177	27	* the end of	55	3
the average *	177	26	during the mining *	55	3
* land use	177	16	the * coal sample	55	4
for * the	176	23	the mining of *	55	4
* be used	176	7	an increase in *	55	3
* the two	175	10	of coal mining *	55	8
located * the	175	10	is * that the	55	11
* in table	174	9	the mining area *	54	8
is not *	172	19	were * in the	54	13
related to *	172	13	* coal seam is	54	7
* shows the	172	8	than that of *	54	4
* in uence	171	15	a depth of *	54	9
* the maximum	171	22	the creative commons *	54	3
be * as	171	20	* the creative commons	54	3
model * the	171	13	* the effect of	54	12
have been *	170	28	top coal caving *	53	4
* mining of	170	11	of the study *	53	4
analysis * the	170	6	* an increase in	53	7
rib pillar *	170	18	on the * surface	53	7
* this paper	169	6	* in this paper	53	7
* used in	169	16	on the surface *	53	5
assessment of *	169	20	an * in the	53	3
caused by *	168	14	a result of *	53	4
in coal *	168	10	the same * the	52	6
* is an	168	10	in the * part	52	9
the impact *	168	2	height of the *	52	9
use of *	168	24	* the return corner	52	3
of * study	168	4	* of science and	52	2
is * than	165	9	was * in the	51	8
the case *	165	5	the mechanical * of	51	6
the development *	164	4	* risk assessment of	51	6
* terms of	164	2	this article is *	51	4
between * and	164	23	coal seam * the	51	8
the * process	164	29	the * strength of	51	5
* coal seams	162	24	* of coal seam	51	11
* metal loid	162	4	* end of the	51	2
the * method	162	27	* model of the	51	10
height of *	162	13	* size of the	50	3
is * as	161	22	it was * that	50	4
the * body	161	9	fully mechanized mining *	50	8
the * rock	160	18	the increase * the	50	2
* thickness of	160	8	* this article is	50	3
* coal mines	160	18	* width of the	50	2

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* to a	160	24	characteristics of the *	50	11
be * in	159	21	* there is a	49	8
was * by	159	26	the other hand *	49	7
* mining in	159	23	can be obtained *	49	4
* increase in	158	7	is * on the	49	4
it can *	158	2	* result of the	49	2
* related to	158	18	in terms of *	49	5
the * time	158	12	by the * of	49	12
* there is	158	17	* side of the	49	8
case study *	157	12	the value of *	49	5
the data *	157	25	of the * international	48	6
the water *	157	27	of the * model	48	7
* range of	157	10	of * of the	48	11
of * of	156	29	* oil and gas	48	5
be used *	156	5	* found that the	48	6
* and mining	156	20	the roof * the	48	3
the * part	156	16	* of this study	48	7
* of rock	155	27	in this * are	48	6
and mining *	155	22	the degree of *	48	10
model of *	155	16	be * to the	48	5
the ore *	155	17	the * to the	48	8
* the model	154	11	with regard to *	47	3
* the case	154	7	* working face is	47	4
* the water	154	19	the increase of *	47	5
less than *	154	14	is * than the	47	5
metal loid *	154	2	of the * river	47	5
is used *	153	8	the end of *	47	2
* the rib	152	3	* the longwall face	47	7
the * layer	151	20	the * mining area	47	7
the following *	151	30	at the end *	46	2
* when the	150	24	of the deposit *	46	10
* coef cient	150	19	of * mining area	46	2
should be *	150	25	of the upper *	46	8
* model of	150	21	* the analysis of	46	9
the * results	150	19	* large number of	46	2
characteristics * the	149	3	of this study *	45	8
* creative commons	149	4	* are shown in	45	6
* of china	149	12	the mining * is	45	7
thickness of *	149	14	* impact on the	45	10
the * zone	149	14	* in the goaf	45	7
the * rate	149	18	* the amount of	45	6
* a result	149	3	of the * stress	45	6
creative commons *	148	5	* concentration in the	45	3
order to *	148	29	is a * of	45	12
of * working	148	3	the * and lower	45	2
data * the	148	12	* used in the	45	6

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
fully mechanized *	147	6	* roof of the	44	2
there * a	147	3	maps and institutional *	44	2
* process of	147	9	the roof of *	44	2
* caused by	146	24	* the coal sample	44	3
* as follows	146	17	* of coal mining	44	8
* underground mining	146	10	* the surface of	44	5
* mining areas	145	15	the coal * and	44	7
is * that	145	21	not * in the	44	3
* has been	145	23	the same time *	44	5
increase * the	145	4	in the * mining	44	9
as * result	145	2	the hydraulic support *	44	9
china coal *	145	3	* location of the	43	3
mining * is	145	19	* temperature of the	43	5
with an *	145	15	the surface * the	43	2
* possible to	145	7	in the upper *	43	5
more than *	144	12	the surface of *	43	2
* is not	143	9	* sides of the	43	4
* height of	143	12	risk assessment of *	43	6
* degree of	143	5	can be * into	43	3
of * rib	142	2	* carried out in	43	6
* china coal	142	3	as * in table	43	4
the stability *	142	6	is based on *	43	3
* mining institute	142	2	the * for the	42	9
hydraulic support *	142	23	* seam dip angle	42	5
higher than *	142	10	* the fact that	42	4
* stability of	142	4	under * creative commons	42	2
changes in *	142	9	higher than * of	42	2
of water *	142	20	an * of the	42	6
* depth of	141	7	of the total *	42	8
* study of	141	10	higher than that *	42	3
* angle of	141	11	* bottom of the	42	2
than * of	141	4	upper * of the	42	6
results * that	141	9	* of the cable	42	5
* effects of	141	14	is located in *	42	2
will be *	140	26	* upper and lower	42	2
the * mine	140	18	mining * in the	42	10
there * no	139	4	* health risk assessment	42	5
to * and	139	30	* is shown in	42	11
this paper *	139	18	results are * in	42	3
* higher than	139	14	fig * shows the	41	3
are shown *	138	3	of heavy metals *	41	3
* case of	138	2	* of the slope	41	6

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
in * paper	138	2	* binchang mining area	41	3
in * rock	138	12	in * coal seam	41	4
* strength of	137	10	oil and gas *	41	8
coal * is	137	10	it is important *	41	2
up to *	137	11	* soft-rock protective seam	41	3
the stress *	137	19	the range of *	41	7
of * institute	137	2	* shows that the	41	7
* study the	137	4	of the * layer	41	9
angle of *	137	12	is * in figure	41	3
depth of *	136	18	* of the deposit	41	10
* that of	136	9	in the * the	41	7
was * in	135	23	a * amount of	41	3
in * with	135	16	* conditions of the	41	6
a * study	134	7	mining of the *	41	5
properties of *	134	13	used in this *	40	4
the other *	134	11	fully mechanized coal *	40	3
* should be	134	20	on the coal *	40	4
were used *	134	6	* thermal circuit model	40	2
* used for	133	10	lower stope footwall *	40	2
impact of *	133	12	the * drawing stage	40	2
* the use	133	16	* of the roadway	40	8
a result *	133	4	* development of the	40	3
be * that	132	10	* the surrounding rock	40	7
presented in *	132	10	the * characteristics of	40	9
of china *	132	11	of the main *	40	6
* top coal	132	10	* the russian federation	40	3
* which the	130	8	it * found that	40	2
* in mining	130	21	the * waste dump	39	2
* determine the	130	6	* in the return	39	2
* fully mechanized	130	8	in addition * the	39	3
mining industry *	130	19	the * from the	39	7
* into account	129	4	* to determine the	39	8
the process *	129	9	the coal * the	39	7
than that *	129	5	the amount of *	39	6
a case *	129	4	was * to be	39	7
* result of	129	2	health risk assessment *	39	3
* the roadway	128	8	of the * sample	39	3
* the data	128	15	of the model *	39	8
the increase *	127	3	* coal mining area	39	4
is * of	127	10	* away from the	39	4

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
* an average	127	7	with an * of	39	6
* coal sample	127	7	* the value of	38	8
mechanical * of	127	8	* heavy metals in	38	2
* surrounding rock	127	5	in figure * the	38	2
* there are	127	21	of * fully mechanized	38	2
* one of	127	12	the characteristics of *	38	4
coal seams *	127	18	* be divided into	38	2
* the goaf	126	7	a creative commons *	38	3
to determine *	126	4	* the mining process	38	3
mining in *	126	19	be * in the	38	10
conditions * the	126	8	mining and technology *	38	4
coal mines *	126	19	the maximum * of	38	7
* the soil	125	11	be * as a	37	6
and * mining	125	17	* ecosystem service value	37	6
* coal samples	125	12	and a * of	37	9
amount of *	125	24	* in the case	37	7
* the deposit	124	8	* a thickness of	37	3
of * metal	124	2	in this case *	37	4
in * of	124	4	within the * of	37	4
the presence *	123	2	is greater than *	37	4
be * by	123	18	is * as the	37	7
* have been	123	25	the spatial * of	37	4
top coal *	123	8	the * area is	37	5
within the *	123	21	in the * drawing	37	2
the * river	122	8	carried out in *	37	3
underground mining *	122	15	was used * the	37	3
for * of	122	20	the analysis of *	37	2
* relationship between	122	3	of the cable *	37	6
* the process	121	8	* of the study	37	8
* level of	121	10	rock mass * the	37	7
the * are	121	19	* second longwall panel	37	2
well as *	121	9	are presented in *	36	5
the hydraulic *	120	6	* the coal surface	36	2
greater than *	120	11	impact of * mining	36	5
* length of	120	6	the * distribution of	36	5
* changes in	120	19	in the goaf *	36	5
* application of	120	9	results of the *	36	8
* the highest	120	11	for * of the	36	5
total * of	119	12	during * mining of	36	3
china * of	119	3	was used to *	36	6
* as shown	119	21	* as well as	36	11
of * roof	119	2	in addition to *	36	3
* properties of	119	13	between the * and	36	10

Appendix 2: (Continued)

Mining Corpus Top 500 P-Frames					
Three-word P-frames			**Four-word P-frames**		
P-frame	**Freq**	**Type**	**P-frame**	**Freq**	**Type**
and * is	119	18	* underground coal mining	36	4
* uence of	119	2	the top coal *	36	3
located in *	119	3	side of the *	36	8
in underground *	119	10	in the north *	36	4
length of *	119	8	* increase of the	36	2
* rate of	119	18	for the mining *	36	4
of each *	118	25	* terms of the	36	2
* fly ash	118	9	fully mechanized * mining	36	2
strength of *	118	6	the water-conducting * zone	36	2
out * the	118	11	of this * is	36	5
* dip angle	118	5	* mechanical properties of	36	3
an open *	118	4	area of the *	36	9
* located in	118	9	be seen * the	35	3
the distance *	118	8	of the * mine	35	5
for * mining	117	12	of the * mining	35	8
is * for	117	15	of the proposed *	35	6
* be seen	117	2	* an average of	35	3
related * the	117	2	in underground coal *	35	3
* of heavy	117	16	the presence of *	35	7
* the second	117	11	* water-conducting fractured zone	35	2
* the slope	117	10	when the * of	35	5
* divided into	116	4	of the longwall *	35	3
* concentration of	116	11	* in order to	35	9
* risk assessment	116	11	* working face the	35	2
to a *	115	24	state of the *	35	5
range of *	115	14	the new afton *	35	5
* the longwall	115	12	the dynamic stress *	35	6
* than that	115	7	* affected by the	35	4
the in *	115	5	concentration in the *	35	4
the proposed *	115	11	value of the *	34	8
large * of	115	8	on the * and	34	9

CURRICULUM VITAE

Ayşenur HOŞOĞLU She graduated from Rıdvan Çelikel Anatolian High School in 2016. Then, she graduated from the Department of English Language and Literature at Karadeniz Technical University in 2022 as a high honour student. She started her master's degree in the Applied Linguistics Program at Karadeniz Technical University in 2022. She has been working as a Research Assistant at a private university in Türkiye.

HOŞOĞLU is married and speaks English.