Universidad Mayor de San Andrés

Facultad de Humanidades y
Ciencias de la Educación

Carrera de Lingüística e Idiomas

The Individuation, Analysis, and Study of Relevant Syntactic Patterns in Medical Scientific Articles

Postulante:
Univ. Karla Isabel Jimenez Gonzales

Tutor:
Mauro Costantino, Ph.D.

Especialidad: Inglés

December, 2022

Agradezco a la vida por tener a personas tan maravillosas conmigo y que fueron parte esencial de esta tesis. Familia, amigos y profesores por sus enseñanzas, consejos, apoyo y paciencia. Y finalmente agradezco al amor más puro de los animalitos y las plantas.

Abstract

All languages contain linguistic patterns, characteristic that allows us to predict some patterns through corpus linguistics. This work aims to identify significant syntactic patterns, through a computational methodology, in two collections(300 and 600 articles) of medical journal articles. The resulting data shows that both collections present syntactic patterns, and we recommend them as a source for material design in English for Specific Purposes (ESP).

Why English? Regarding Scholarly Publishing and Scientific writing, English is the predominant language. Thus, ESP is a must for professionals that are not native speakers. At the Universidad Mayor de San Andrés of La Paz, in the Medicine Faculty, a study has shown the need for authentic material in writing and reading to share and gain professional expertise.
Keywords: Computational Linguistics, Syntactic Patterns, N-grams, ESP, Corpus Linguistics, Zipf's Law.

Resumen

Todos los idiomas contienen patrones lingüísticos, característica que nos permite predecir algunos patrones a través de la lingüística de corpus. Este trabajo tiene como objetivo identificar patrones sintácticos significativos usando una metodología computacional, en dos colecciones (300 y 600 artículos) de artículos de revistas médicas. Los resultados presentan patrones sintácticos y los recomendamos como fuente para el diseño de material de lectura y escritura en inglés técnico o para propósitos específicos (ESP por sus siglas en inglés).

¿Por qué inglés? Con respecto a las publicaciones académicas y la redacción científica, el inglés es el idioma predominante. Por lo tanto, ESP es una necesidad para los profesionales

que no son hablantes nativos. En la Universidad Mayor de San Andrés de La Paz, en la Facultad de Medicina, un estudio muestra necesidad de material auténtico en escritura y lectura para compartir y adquirir experiencia profesional.

Palabras claves: Lingüística Computacional, Patrones Sintácticos, N-gramas, Inglés Técnico (ESP), Lingüística de Corpus, Ley de Zipf.

Contents

List of Tables

List of Figures

Abbreviations

V	Verb
N	Noun
n	Noun group
v	Verb group
adj	Adjective
adv	Adverb
det	Determiner
num	Number/numeral
prep	Preposition
mod	Modifier
Freq	Frequency
Kwic	Key Word in Context
ESP	English for Specific Purposes
TTR	Type Token Ratio

Introduction

The present study aims to determine if there are relevant syntactic patterns in medical articles in order to help in reading and writing to students and teachers of English for Specific Purposes (ESP). This analysis is carried out using the Concordance Software AntConc in order to identify possible patterns by analyzing the most frequent words. These data lead us to collocations, n-grams, and eventually to syntactic patterns. The analysis will guide us to determine whether these possible patterns could allow teachers to use them as teaching material sources for writing or reading articles in the field of medicine. Our sample is built from/by scientific journal articles in English from many medical researchers from all over the world downloaded randomly from PubMed[1] among the available free articles.

In order for the reader to be able to go easily through the present work, we organized the structures as follows: Chapter 1 (p. 3) presents and develops the need to design ESP teaching material from the language in an actual context, specifically in reading and writing, as well as the need for professionals to produce scientific knowledge through publishing articles about the researches, experiments, or studies. From a didactic point of view, ESP is different from General English, and learning it implies different needs and strategies. By analyzing the syntactic patterns allows us to recommend guidelines when students develop their reading and writing skill. The whole chapter is firmly sustained by many types of research on language patterns and analysis, clearly referred to in the reference section.

The next chapter (2, p. 12) explains the main terms used in this research for a better understanding of what corpus means in linguistics, its characteristics, and concept; moreover, the chapter presents and details computational linguistics and corpus linguistics terms used throughout the project in order for the non-familiar reader to have the needed tool for the

[1]The PubMed database contains more than 33 million citations and abstracts of biomedical literature. It does not include full-text journal articles; however, links to the full text are often present when available from other sources, such as the publisher's website or PubMed Central (PMC)

analysis. We explain the types of patterns, collocations, and n-grams that we may encounter throughout the compilation of articles and the concept of Zipf's law, statistical phenomena, which is a fundamental concept once we broaden our 'corpus' or sample. The chapter aims to offer the basic knowledge of the linguistic procedures in case the reader is familiar with the computational processing of language, computational linguistics, and experienced readers might as well skip the chapter. This description helps us understand the research in general, it also opens the possibilities for further studies or a possible specialized corpus design to identify relevant patterns.

After reviewing the different concepts, 3rd chapter (p. 41) describes step by step the processing followed in the methodology. We describe the use of AntConc, how we built our compilation of the journal articles, the classification of data and analysis, and the broadening of the first collection (300) to 600 new articles in order to compare data and statistical reliability issues. Thus, the present research has two collections to analyze and compare.

Chapter 4 (p. 59) displays supported by data, tables, and examples it's the methodological heart of the work findings of the research. It shows the analysis of frequent words (prepositions, nouns, verbs) also supported by graphics with both collections overlapped to compare, n-grams, statistical analysis of patterns, the analysis of selected sentences of the top 10 nouns. All data compared to the second compilation of texts (600) to prove and validate that we can find a similar pattern. Finally, in chapter 5 (p. 110) we detail the limitations we faced during the research after the recollection and based on the analysis and results. Thus, we present the following subtitles: conclusion, limitations during the research, recommendations, and suggestions for future work. In the Appendix section, you find the additional data that is part of the analysis but not inside the analysis section. For instance, the raw word frequency list(only the top 50 words), data concerning *hapax legomenon*, and the syntactic analysis of some additional examples concerning our N + N structure.

Chapter 1

Objectives and Problem Statement

1.1 Research Problem: Initial Approach

There are different factors to consider when we learn a new language; for instance, our purpose of learning a new language may vary for every student, and it will signify a relevant part of our learning process. Through time teaching another language has evolved to the point that different factors are crucial to determining the teaching content. "The English of research articles and technical specifications is very different from the English of storybooks and street signs, and mastery of this English does not come easily [...]" (Noguchi et al., 2006, p. 155). Therefore, ESP refers to teaching technical English by focusing on the skills (speaking, listening, reading, and writing) or reinforcing one or more of them.

It is not exclusive to any particular language, but English has a relevant role in publishing scientific articles since it is the most used language. Consequently, there might be many authors that are not native speakers. Noguchi et al. (2006) also refers to it stating: "Although English is the language of preference for science in international contexts, the majority of the world's scientists are not native speakers of English and, therefore, they can be considered disadvantaged in professional English communication" (Noguchi et al., 2006, p. 155).

As we know, and Noguchi et al. (2006) clearly points out, there is a disadvantage for a non-native speaker when writing in another language, and consequently, the research to design specialized teaching material is necessary. According to an article (Huttner-Koros, 2015) on The Atlantic about a study on the Scopus website[1] which claims that around 80%

[1] Available on `https://www.scopus.com`(Some sections available only by subscription.)

of the scientific articles are in English; since this database is "the world's largest database for peer-reviewed journals".

The topic is broadly studied due to its relevance in Academia; for example, a recent study about the impact of publishing in English, by Di Bitetti and Ferreras (2017) titled "Publish in English or Perish" directly addresses the problem. The investigation is based on the number of citations that have articles written in English vs. Spanish, and their findings are clear about the audience that reaches out through English. They examine different journals from different countries and arrive at the following conclusion:

> The articles published in English *have a higher number of citations* than those published in other languages, when the effect of the journal, year of publication, and paper length are statistically controlled. This may result because English articles are accessible to a larger audience, but other factors need to be explored. Universities and scientific institutions should be aware of this situation and *improve the teaching of English*, especially in the natural sciences. (Di Bitetti and Ferreras, 2017, p. 1)

Hence, publishing in English is crucial considering the possibility of reaching a wider audience, as proven by several studies (Ramírez-Castañeda (2020); Edwards (2016); Cristina Pabón Escobar and da Costa (2006); Hamel (2007)) in addition to the one cited above. Nevertheless, all things considered, writing scientific articles in English is a process that could result in a challenging task when the writer or author is not a native speaker or deeply familiar with the language. As a consequence, in academia, we often find ourselves in the peculiar situation where a non fully proficient English speaker has great expertise to share research findings, articles, and grant proposals, and at the same time, the desire to learn through reading other professionals' works.

In order to face this problem, as linguists and teachers of ESP, there are many ways in which we can take action toward a solution. For example, one of the most important resources a teacher has is the specific teaching material in ESP[2]. The specific case we address in this study is the access to academic publications, both as a writer or as a reader; we, therefore, focus on those two skills: writing and reading (since they complement each other). These specific skills are crucial when sharing or acquiring professional expertise, and the lack of sufficient work is far from being a local problem.

[2] As a matter of fact, the same is also true for other languages. Nevertheless, the present study only focuses on English due to its relevance.

The work "Corpus linguistics around the world" (Archer et al., 2006), collects some papers presented at the Corpus Linguistics 2003 conference, held at Lancaster University in March 2003. One of the most relevant papers for our study Noguchi et al. (2006) titled What do "we" do in science journal articles?, suggests building a specialized corpus to find or identify features to design better an *adequate material* to fulfill the students training: "[...] the problem of inadequate educational materials for the effective training of non-native speakers in the professional English of the scientific community" and also adds that "one cause of this inadequacy is lack of proper linguistic research based on suitable linguistic research tools" (Noguchi et al., 2006, p. 155).

As a matter of fact, the case also applies to Bolivia, where our professionals and students share this need:

> [...]we realized that the necessity of learning English in the field of medicine is vital for the medical professionals because it is the most used language in international meetings as well as the main medium for medical textbooks, journals and research reports. This necessity is shared by the oncology professionals of the School of Medicine from Mayor de San Andrés University [...]. (Callisaya, 2014, p. 109)

As stated before, some ESP projects were suggested and carried out in the Medicine School of Mayor de San Andres University[3] due to the need of this type of project about teaching technical English. There is still anyway a lack of any in-depth analysis in what respects to writing and reading, as far as we are concerned. One of these projects Callisaya (2014) shows the intention of students to develop their writing skills, above others, because of the need to publish their researches in international specialized academic journals. The findings of Callisaya (2014) perfectly agree with Noguchi et al. (2006) statement: "[...] but it also prevents the world from benefiting from the creative potential of those who cannot disseminate their ideas persuasively because of poor English" (Noguchi et al., 2006, p. 155).

Considering the situation as studied by Noguchi et al. (2006) in general and more closely to our local reality by Callisaya (2014), even though we cannot consider any of the four skills more important than the others due to the intrinsic connection among all of them for the learner to have a complete linguistics competence, for a certain area of interest, namely access to academic journals, the student's needs and requests focus on reading and writing as an immediate needs. This is perfectly understandable since, in the eye of our target

[3]Lic. Soledad Callisaya Callisaya, personal communication

population, reading and writing are the most probable needs skills from the point of view of the researcher. Thus, we limit the present research to reading and writing skills, since due to the real world and professional needs, it is crucial to provide learners with authentic language usage in context as supported by the findings in Callisaya (2014) that analyzing the survey results concluded:

> [...] 90% of oncology professionals answered that they always would use English for reading scientific articles and 80% said for accessing to bibliography from their area. Moreover, 75% of the participants answered they would frequently have conversations and meetings with foreign colleagues and work in groups with them as well as write up scientific reports.(Callisaya, 2014, p. 124)

In order to address the needs and requests of the learners, it is, therefore, important when focusing on an ESP class to provide the students with language in context teaching material. To answer these specific points various works are using computational methods to identify patterns or information to be used as an important source of teaching material. For instance, Peacock (2012) developed a study of high-frequency collocations of nouns in research articles across eight fields through 320 research. The corpus was analyzed using WordSmith Tools, and the conclusions were that: "[...] the differing patterns revealed are disciplinary norms and represent standard terminology within the disciplines". In the same paragraph, Peacock highlights the importance of collocations according to the discipline: "[...] that this evidence of sharp discipline differences underlines the importance of discipline-specific collocation research" (Peacock, 2012, p. 1). The author also refers to another relevant research, Ellis et al. (2008), which emphasizes the importance of collocations, suggesting that they are common in academic discourse and that writers need to know them as a whole. Finally, in the "Implications for Teaching" subsection, the author states:

> The present research provides discipline-specific lists of high-frequency collocations of common nouns [...]. The present findings should improve knowledge of RAs (Research Articles) and have relevance for the teaching of research writing to NNS(Non-Native Speakers) and to NS(Native Speakers), and help teachers prepare discipline-specific materials to teach collocation. (Peacock, 2012, p. 43-44)

Nevertheless, the high frequency of nouns is not the only focus of this project, which illustrates the classification and offer an important material be extracted from the analysis,

therefore suggesting the need to carry out more researches in this field. Moving on from the same needs and assumptions, Peacock states:

> The high-frequency collocations of common nouns may be an important part of academic English including RAs, and worth investigating further. There have been several calls for research into collocation, for example Groom (2005) suggests that disciplines can be differentiated by their favored terminology and that this notion is well worth examining on a larger scale. (Peacock, 2012, p. 32)

In conclusion, as we know from personal experience and from scientific studies, writing can be challenging for non-native English speakers, even more in academic areas. According to various studies cited above there is a need to improve research on this topic in order to provide the ESP teacher with better, more advanced specific materials. Hence through corpus linguistics, we can detect, reveal and analyze relevant syntactic patterns of academic texts in order to present a solid base for newer materials to be developed. As for the present study, although it is not possible to build a corpus in all its specific characteristics, we can nevertheless begin a foundation study to analyze the possibilities that a full corpus study might result. In the present approach we have two separate compilations of articles to establish whether that pattern exists and which they are. Once their existence is established and proved, we will be able to compare and analyze them, since they will be representing the "authentic" language used in this specific context. As Basturkmen (2010) pointed out: "Strategic competence is the link between the context of the situation and language knowledge" (Basturkmen, 2010, p. 139), in (Ahmed, 2014).

1.2 Problem Statement

Considering all of the above mentioned; the problem statement in our research could be summarized as follows: To what extent are syntactic patterns relevant and recurrent in scientific medical articles in order to recommend them as a source for designing teaching ESP material?

1.2.1 Objective

General Objective

To identify relevant syntactic patterns throughout the medical ESP academic journal articles through the analysis of two randomly selected compilations of medical journal articles to recommend them as teaching material for medical ESP learners.

Specific Objectives

- To analyze and select the most frequent collocations and lexical syntactic patterns in scientific medical articles.
- To identify and classify relevant collocations and lexical syntactic patterns.
- To compare the tendency of the most frequent words and syntactic patterns of both collections to prove the correlation.

1.3 Justification

According to the needs established in the above sections, the problem addressed in the present work is a strong justification.

To start there are previous researches based on corpus linguistics done in this field, even though not in this exact area. For example, we can cite Özdemir (2014) whose study explains the concerns about medical students' needs when learning English and the design of a specialized corpus in order to use it as a teaching material: "The corpus comprised 31,731 words. The corpus data were used to teach students collocations and to produce in-house ESP materials [...]" (Özdemir, 2014, p. 1).

Even though it is essential to clarify that both compilations of scientific medical articles are not corpora2.1.2 *per se* because of their characteristics, these represent a great source of information and data in order to establish the presence and the extent of syntactic patterns in scientific medical articles. Once we establish the presence of syntactic patterns in the present study, showing the implications for the ESP area of knowledge, this might lead to applying the same methodology in a specific corpus in order to obtain an improved and broader source of syntactic patterns.

In order for the study to be relevant, we decided to apply this methodology to two different collections so that a comparison could be possible. This step could be a key passage

in the process since, based on Zipf's Law2.2.1 the frequency of the relevant pattern we are looking for should not differ substantially while increasing the number of articles.

In the long run, this will lead us to encourage ESP students to write or/and read them to enhance their future professional visibility. As a matter of fact, according to SCImago Journal Rank and related research ((Di Bitetti and Ferreras, 2017)) if the article is in another language than English, the outreach will be limited in the audience, considering that the top 50 Journals are in English (Huttner-Koros), as explained above.

The mentioned study of Scopus proves that most of these articles were written entirely in English (80%) and the need of achieving international recognition publishing in this language is a must. Under this statement, to keep up and to let our professionals share their studies and research with the academic world, we need to help medical ESP students improve their writing and reading skills.

Adding to the above justification, as the need to answer learners' need for better and more effective ESP material, there are also scientific relevant matters that concern the present work: the theoretical-methodological and practical relevance of the study, which will be addressed in the following sections.

1.3.1 Theoretical Relevance

From the theoretical relevance point of view of the present study, it is important to point out that the study will be able to present relevant and novel data on the structural patterns in Medical ESP. We aim to demonstrate that academic writing in medical English only shows limited variation due to specific topics and the nature of the texts themselves. These limited variations are known as *hapax legomena*[4], and the strong core of repetitive patterns are for the need for clarity in the scientific literature, represent our focus of study. These "core" and syntactic patterns will be evident once the study shows the threshold according to Zipf's Law, which refers to the repetitions of patterns throughout any language and their relation to the frequency and rank (See 2.2.1, p. 28). According to the statistical phenomenon, these occurrences will repeat with no relevant differences in a broader collection of texts, as far as the medical area of study is concerned.

Many researchers focus on patterns in language; for instance, Orasan, associate professor in computational linguistics at the University of Wolverhampton, UK, and deputy head of the Research Group in Computational Linguistics in his work: "Patterns in scientific

[4]Referring to the occurrence of only one word in a whole text see 2.2.2, p. 36.

abstracts" observes:

> The patterns identified in this paper are not only useful for automatic abstracting or computational linguistics in general but they can also be used in order to teach students how to write abstracts. [...] both reading and writing an abstract are not trivial tasks, and many students experience difficulties. Those students who are learning English as a second language have even greater problems with such tasks. The patterns which are identified in this paper could help them to write abstracts. (Orasan, 2001, p. 1)

Unlike Orasan, our focus is not only on the abstract section but to find patterns we decided to investigate the whole article. Further studies like Hanks (2008) mention that language was already discovered to be highly patterned when Hornby was invited[5] to work in an Institute in Tokyo (1930) with English collocations. From this work, researchers formalized some principles, one of them being: "Language in use is highly patterned. Each word is typically associated with only a small number of syntactic patterns" (Hanks, 2008, p. 90).

In the end, the present study situated itself in a very well-studied context nowadays, while still addressing some aspects of the problem that requires further studies and improvement, aiming to allow research on this topic to move a little step further thanks to our small contribution.

1.3.2 Methodological Relevance

From the standpoint of Research Methodology, the relevance we are aiming for is also important. By being able to identify syntactic relevant patterns, we open a whole world of possibilities; it would be possible to follow the same methodology for other areas such as: engineering, economics, law, and so on, and therefore recommend or set other teaching guidelines and tools based on syntactic patterns for the student and the teacher alike. Given the language-independent nature of Zipf's Law, the current methodology has the potential for practical applications in a large number of professional fields. By identifying the syntactic patterns that occur in the vast body of academic literature relating to medical ESP, we are able to demonstrate that such patterns are dependent on neither topic nor language.

Moreover, from the methodological point of view, it would be interesting if, based on the present studies, more local students and researchers would be interested in approaching

[5]Harold E. Palmer, director of the Tokyo Institute for Research into English Teaching (IRET)

computational linguistics, and therefore opening the path for new research lines at the University.

1.3.3 Practical Relevance

Eventually, we would like to underline the practical relevance of the present study. ESP teachers need to provide students with actual content, which means, as far as academic reading and writing are concerned, the language used in real situations and in published journals. As we further explain in (See 2.3, p. 39), ESP matches the students' objectives for learning English by matching topics, skills, situations, etc. The present study aims to present data based on authentic language from genuine scientific articles, published and peer -reviewed, so the results of the syntactic patterns of the present investigation will allow us to recommend guidelines for learners of medical ESP. Its use will let students be more confident in reading, more effective in writing, and more accurate in academic publications.[6].

We also cannot avoid mentioning the practical relevance from the social point of view; the knowledge of English became unavoidable for all professionals not only to acquire information but also to share their professional expertise by working, learning, and publishing. The steep growth for English to suit specific needs for any professional area is even more prominent in medical studies nowadays and might play a critical role for our Universities graduate in order to reach higher publishing rates.

[6]It is important to notice that authors vary from different nationalities, native and non-native speakers and that we didn't classify them under any parameter. In this specific case, we will show that given enough data, differences will be minimal. However, these parameters could reveal other relevant data worth analyzing in future research.

Chapter 2

Theoretical Framework: Corpus and Computational Linguistics

Computer science is constantly growing, and it is hard to imagine a field not related to it. This growth has allowed advances in different areas and became an indispensable tool in research. Linguistics is no exception, and its relationship with computer science is broad. Computational Linguistics is the term used to refer to the overlapping zone between these two sciences, which is defined as:

> A field of linguistics that involves the scientific study of language from a computational perspective. [...] Computational linguistics relates to works such as machine translation, information retrieval, speech recognition, and synthesis, voice response systems, web search engines, text editors, language instruction manuals, and automated content analysis.(Baker et al., 2006, p. 41).

From experimental phonetics, to forensic linguistics, from linguistic atlas in GIS[1] to quantitative syntax the list of interaction could go on and on. As far as the present work is concerned, nevertheless, we focus on the specific aspect of frequency and pattern recurrence as in the broad field of Corpus Linguistics.

Corpus Linguistics is related to computational linguistics, not only to store large bulk of texts but also to provide information about the corpus and its content within seconds. About the scope of corpus linguistics Kennedy (2014) writes: "Corpus linguistics is based on bodies of text as the domain of study and as the source of evidence for linguistic description and

[1] Geographic Information System.

argumentation"(Kennedy, 2014, p. 7). The present research uses computational linguistics to analyze a 'collection'[2] through a corpus-based approach.

In order to clearly circumscribe the field of work and the outreach of the present research, some definitions and specifications are needed. In the following sections, the reader will find a better explanation of Corpus and Computational Linguistics, as well as a brief sketch of some of the term used in the following chapter, such as Zipf's Law or ESP.

2.1 Corpus Linguistics

Corpus, corpora plural, is a Latin word meaning *body*. The definition in linguistics: "[...]a corpus is a collection of texts (a 'body' of language) stored in an electronic database"(Baker et al., 2006, p. 48).

Even though is a relative new field, with only few decades of development, it is already widespread and complex. In the following section, we explore a brief history, underlining the most relevant events from its beginning to the present days of the evolution of corpus linguistics; we will also cover some relevant discussions that have arisen through time, and the criticism Chomsky made in the 1950s towards the reliability of corpus linguistics. Finally, we will review the definitions, discrepancies about whether consider it a methodology or a theory, a branch of linguistics or not, the differences between corpus-based and corpus-driven approaches.

As a manner of warning, this section is not meant to be exhaustive for an introductory course of corpus linguistics; it is written as a tool, for those readers who are not familiar with it, to understand the present work. Plenty of bibliographic reference is offered all through the section in order for the avid reader to know more about it.

2.1.1 Brief History of CL

It is worth mentioning the use of the term "corpus linguistics" will be a point of reference to distinguish two different times in its history. Hence, the term corpus linguistics was used for the first time in 1982 by Aarts and van den Heuvel (1982) in the "Grammars and intuitions in corpus linguistics". However, the idea and concept of corpus linguistics were initiated years before this. Even though different authors may disagree that some of these events are

[2] As stated later, our compilation of texts is not a corpus *per se*

part of corpus linguistics, we consider them as a fundamental part of its development to have an overall understanding.

Nowadays, a corpus is stored electronically, but before computers existed, the very concept was in its beginnings. Alexandre Cruden (1701-1770), for instance, created a corpus to study language and this work was first published in 1736 but had 42 editions in total:

> It included concordances not only for what the author considered to be the major content words in the Bible but also some function words such as how, you, he, once, between (but not on, she or with) and certain collocations such as how long, how many, how much, how much less, how much more, how often, all the nations.(Kennedy, 2014, p. 14)

Edward L. Thorndike, Professor of Educational Psychology, also had an important role, in 1921 due to the migration and the need to learn English in British colonies, wrote the book *The teachers' word book* based on a 4.5 million words corpus conveying the most frequent words in English; the process was carried on manually. This work was mainly to focus on the pedagogical outcome in order to help people to learn English[3] easily.

In the year 1946, an advance in the development of corpus linguistics initiated. The idea of the first concordancer was born with Roberto Busa, a Jewish Priest from Italy, who used to analyze Saint Thomas Aquinas's work and wanted to have a better compilation and save some time. "Digital humanities scholars cite the work of Roberto Busa working with IBM in 1949" (Biber and Reppen, 2015, p. 32). Thus, he asked the president of `IBM`, Thomas Watson, to create a way to have access to other files to redirect you straight to the desired information. The idea of what we know now as hypertext has just started: A total of 10.631.980 words, this project, the Index Thomisticus,which took 34 years in total.

A few years later 1959, the Survey English Usage (SEU) project began (British English). Randolph Quirk developed at the University College London a pre-electronic corpus containing written and spoken language[4]. "The SEU was a precursor of later corpora such as the British National Corpus and the American National Corpus (Ide and Reppen 2004)" (Allan, 2013, p. 8).

[3]For further information there are other works in relation to language teaching through the most frequent words as Biber and Reppen (2015) mentions "[...]common focus was the frequencies of words in lists used in foreign language teaching (see e.g. Thorndike 1921, Palmer 1933, Fries and Traver 1940, Bongers 1947, West 1953)"(Biber and Reppen, 2015)

[4]Jan Svartvik also took part in this project, who will later work on building the London-Lund Corpus.

Two years later, Nelson Francis and Henry Kucera, at the Brown University, started the idea of the Brown Corpus (American English) that was published in 1964 and became the first electronic corpus. Later in 1975, Jan Svartvik and Quirk created a corpus of spontaneous Spoken and written English, the London-Lund Corpus, a printed corpus of English conversation, based on the Survey of English Usage (SEU) and the Brown corpus. It contained the spoken part of the SEU Corpus, but this time electronically available (Facchinetti, 2007).

The Brown Corpus was an important project and crucial for corpus linguistics because, for the first time, the computer was a great advantage to store and analyze the corpus. It decreased the problems such as the possible human error, the time to search for a specific word, obtaining the number of frequencies, the money invested in hiring more people. However, it was also very criticized, in the book "English Corpus Linguistics An Introduction" Meyer (2002) mentions that a leading "generative grammarian" that said it was "a useless and foolhardy enterprise".

Despite the unpopularity of those years, corpus linguistics was growing. There were other projects concerning the creation of a corpus to study the language. For instance, some of the projects were: The British National Corpus project that began in 1991 until 1994, and it was the counterpart of the Brown Corpus for American English.

Some criticism in Corpus Linguistics: Noam Chomsky

The critics towards Corpus Linguistics are also a fundamental part of its evolution. As McEnery and Wilson (2003) remarks: "[...]to cut ourselves off from these discussions is to lose an understanding of how corpus linguistics developed" (McEnery and Wilson, 2003, p. 5). Thus, when corpus linguistics was gaining more space inside the linguistics research, Chomsky criticized the credibility or reliability of it with his famous quote criticizing corpus linguistics cited by various scholars such as Tognini-Bonelli (2001), Facchinetti (2007), Baker.

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still, others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (Chomsky and Hill, 1962, p. 159)

In order to have an overall understanding of the critics made towards corpus linguistics we briefly clarify some terms and concepts. First of all, the dichotomy gave by Chomsky:

Performance - Competence in language refers to the external and the internal language that people have. For instance, competence is the knowledge we have internally; on the contrary, performance refers to the materialization or realization of that knowledge. Then, the main critic was that corpus linguistics focuses on performance[5] rather than in competence[6], which was the main focus for generative grammar[7]. The principal argument was that performance might project or be influenced by external factors in the data such as stress, or alcohol which is not an accurate reflection of competence and that is beyond the control of the person. Related to this point, Baker et al. (2006) adds: "[Chomsky] argues that performance data is therefore degenerate. He also argues that corpus data (which is by nature performance data) is a poor guide to modeling linguistic competence" (Baker et al., 2006, p. 39).

Moreover, we have generative grammar based on rationalism[8] that is mainly focused on internal knowledge, and Chomsky, one of the main advocates. Meanwhile, corpus linguistics, an empirical[9] discipline, which focuses on the observation of the language and therefore, leading to study and analyze just part of the language since language is infinite, a corpus will never contain all language.

Nowadays, things changed, a corpus might be larger than when these critics took place, allowing us to store more texts, therefore balancing the flaws that originated some of the criticism. Besides this, corpus design and corpus organization has been growing enormously, allowing to better understand the whole concept, also thanks to the criticism itself. It will be certainly a matter of designing a corpus considering the purpose of the study and following the parameters or features that we have mentioned.

2.1.2 The Concept of Corpus

There are some characteristics to consider when we talk about corpora in order to set the knowledge base for the understanding of the present research. This section is not meant to be an exhaustive cause in corpus linguistics, just a toolbox for the understanding of

[5]Performance (or E language) is our behavior in real life(Baker et al., 2006, p. 38,39)

[6]Competence (or I-language) consists of our tacit internalized knowledge of the language(Baker et al., 2006, p. 38,39)

[7]A system of linguistic analysis consisting of a limited, unchanging set of rules employing a list of symbols and words to generate or describe every possible sentence in a language.(Collins English, 2014)

[8]The doctrine that knowledge comes from the intellect in itself without aid from the senses; intellectualism (Collins English, 2014)

[9]An empirical approach to knowledge is based on the idea that knowledge comes from our experiences or from observation of the world. In linguistics, empiricism is the idea that the best way to find out about how language works is by analyzing real examples of language as it is used.(Baker et al., 2006, p. 65)

our work relevant references is given for the reader who wants to investigate further. The very concept of corpus has been evolving thorough time, although it's clear there are some common points all scholars agree with. Tognini-Bonelli (2001) in one of the most known works about corpus gathers different concepts. For instance, Francis (1982) describes it as a collection of texts "assumed to be representative"; referring to the lack of certainty when we refer to whether a corpus is representative or not. Novoa (1992), mentions the importance of "natural-occurring" of language. Aarts (1991) highlights the fact that it can be spoken or written.

One definition that condenses many of the characteristics is from Barbera et al.

> Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi.[10] (Barbera et al., 2007, p. 70)

In sum, it is a collection of running texts or electronically stored, authentic language to be studied with a specific purpose. Therefore, different features will also vary according to the type of corpus and the purpose of the research. Diverse authors display other features, but they are very similar, and what is more important each one is closely related and complements each other. Hence, authenticity and representativeness are the most relevant points, Tognini-Bonelli (2001) lists and explains three issues regarding corpus: Authenticity, Representativeness, and Sampling. In the same way, we explain other issues as annotation in the following paragraphs.

Authenticity The language of the texts[11] recollected must fulfill the requirement of 'real language' referring to a spontaneous and natural language which will be biased if the people are aware that the material is for linguistic analysis. Concerning this point, Sinclair maintains: "[. . .] all the material is gathered from the genuine communications of people going about their normal business. Anything which involves the linguist

[10] A collection of texts (written, oral or multimedia) or parts of them in a finite number in electronic format treated in a uniform way (i.e tokenized and added with adequate mark-up) so as to be manageable and searchable by a computer; if (as often) the aims are linguistic (description of natural languages or their varieties), the texts are mostly chosen so as to be authentic and representative.

[11] It refers either to spoken, written, or transcribed language.

beyond the minimum disruption required to acquire the data is reason for declaring a special corpus" Sinclair (1996a) cited by (Stefanowitsch, 2020, p. 23).

Sampling Sampling refers to the selection of texts performed. The established requirements to select certain kinds of texts. Tognini-Bonelli (2001) refers to Biber (1993), who proposes parameters to select the texts, which reflect directly to the corpus evidence. "The decisions involve very clearly a theoretical stance on the part of the corpus builder, and they will have a direct effect on the insights yielded by the corpus" (Tognini-Bonelli, 2001, p. 59).

Representativeness Closely related to authenticity, representativeness concerns the inclusion of all types of texts determined by the corpus type. See (2.1.3 p. 22). Tognini Bonelli (2001) says that: "It is the vexed question of the representativeness of the language included in the corpus" (Tognini-Bonelli, 2001, p. 57). Representativeness is not an easy point to prove and largely discussed. Thus, it is further explained with more details and a close look to our study. (See 2.1.2, p. 20).

Besides these three features, Stefanowitsch (2020) also adds size and annotation.

Size The factors that are closely related to this point are sampling and representativeness. Nevertheless, there is not one specific size for a corpus, because it is available the development of the computer science and the internet to store it, it wouldn't be a problem.

Annotation A corpus may or may not be annotated; annotation will imply an important investment of time and human resources. To this point it is important to distinguish between corpus mark up and corpus annotation. There are three types of annotation:

> 1. Information about paralinguistic features of the text such as font style, size and color, capitalization, special characters, etc. (for written texts), and intonation, overlapping speech, length of pauses, etc. (for spoken text); 2. Information about linguistic features, such as parts of speech, lemmas or grammatical structure; 3. Information about the producers of the text (speaker demographics like age, sex, education) (Stefanowitsch, 2020, p. 39)

Once the terminology about corpora i set clear, there is still some vocabulary that might need clarification. It is in fact relevant to distinguish between the terms mark-up, annotation, tagging and encoding.

> Una prima distinzione possibile è tra "markup esterno", cui sono affidati i riferimenti del testo che di esso non fanno costitutivamente parte (autore, titolo, genere, capitoli, paragrafi, pagine, righe ecc.), e "markup interno e filologico", cui sono affidate le informazioni di carattere filologico (integrazioni, espunzioni, ecc.) e testuale (corsivi, prosa, verso, ecc.)[12].(Barbera et al., 2007, p. 37)

We therefore prefer to distinguish between a corpus and a collection of texts being the features explained above the important distinction. Other features are worth mentioning besides the above analyzed, some of them closely related to each other Dash (2008):

Quantity The corpus size is very important, considering the more words contain it will be more trustful or representative. This feature is conditioned to the availability of technology and to the collecting and processing costs.

Quality The reliability of the texts compounding the corpus; this point is deeply related to authenticity because the source must be genuine and consequently the language must be authentic from "real life". Thus, the texts corpus designer should not interfere in any level to change or alter the data at any point.

Representativeness (See 2.1.2, p. 20). The text data, besides being relevant, must be diverse according to the type of corpus. For instance, if we build a corpus with the editorial section of the newspapers in Bolivia, we will need a proportional sample from each newspaper.

Equality This one is closely related to representativeness; it refers to a proportional recollection of texts, for instance, an equal amount of works from the spoken than the written.

Retrievability Nowadays, corpora are stored in computers, but it should be available for all people who may need it and not only for people with computer training. This point refers to storing the real data for future use.

Verifiability All corpus texts must be reliable to the user and the authenticity should be verified.

[12] A first possible distinction is between "external markup", to which the references of the text that are not constitutively part of it are entrusted (author, title, genre, chapters, paragraphs, pages, lines, etc.), and "internal and philological markup" , which is entrusted with information of a philological nature (additions, expunctions, etc.) and textual (italics, prose, verse, etc.)

Augmentation A language might change, and corpora might show how language has changed through time. Corpora might vary according to their nature; for instance, a monitor corpus is open to see how language is constantly growing and changing through time.

Documentation This feature refers to the information of the texts; sources, publishing dates, editions, pages, authors' details, etc. It could be inside the corpus or as a complement for researchers who need the data.

Management There are four aspects to consider related to management: Maintenance, standardization, augmentation, upgrading. The first one refers to performing maintenance to the corpus; to protect it from viruses or damage. The second one refers to keeping it in a standard manner to contrast to others. Augmentation is about the possibility of adding new text samples. And upgrading refers entirely to the constant computer science evolution, which leads to keeping it updated.

There are different aspects or features to consider and to analyze in further detail corpus design features. Among all features, representativeness has a significant role in our research since it is a specialized collection. In the next subtitle, we develop further and explain this point.

On Representativeness

This feature, as we said, stands for equality towards the texts included in the corpus. Sampling and balance are two crucial factors to reach representativeness while designing a corpus. They are related and to achieve one of them is relevant to consider the other. For instance, a corpus may be representative if the sampling represents the entire field or fields, which means a balance among the samples. Barbera et al. (2007) mentions the question of Kilgarriff - Grefenstette (2003, p. 334): "Representative of what?". For instance, a corpus will contain different sources such as articles, novels, etc. The nature of the corpus is a relevant factor concerning representativeness. Biber (1993 defines it as "[...]the extent to which a sample includes the full range of variability in a population" (Biber, 1993, p. 243)

Barbera et al. (2007) discusses this point considering different authors. He questions the role of sampling when we talk about a collection of texts representative of a language. However, Schütze observes: "[...]there is no easy way of determining whether a corpus is representative" (Manning and Schutze, 1999, p. 120). So, does it mean that a corpus

can't be proved as representative? Tognini-Bonelli (2001) refers to this, quoting Leech (1991) when he says that the findings based on their contents can be generalized to a larger hypothetical corpus. It means that it could be labeled as a representative if the results match a broader corpus.

Nevertheless, as mentioned before, balance is crucial but not always possible; a monitor corpus, for instance, can't keep its balance. Hunston and Francis (2000) mentions, a monitor corpus[13] fulfills representativeness as it is constantly updated; then, balance is not possible in this type of corpus. General and specialized corpora will differ on the concept of representativeness.

In other words, it will be easier to achieve representativeness in a specialized corpus; because it may be easier to accomplish it only focusing on one field, for instance: law, literature, economics or medicine (as this project), the genre; articles, papers, letters, novels. In conclusion, there is not a specific way to prove a corpus as representative despite the guidelines. Nevertheless, the specificity on genre, field, etc., will be an influential factor. In the next section, we discuss representativeness in ESP.

> Rappresentatività (§ 2.3), finalizzazione (§ 2.5) e dimensione (§ 2.7) sono pertanto parametri interrelati: è infatti chiaro però che l'attenzione per la rappresentatività di un corpus significa attenzione per la funzione cui il corpus stesso è destinato; ed ai fini di ricerche specifiche, una raccolta di dimensioni limitate ma in sé equilibrata[14]. (Barbera et al., 2007, p. 50)

Reresentativeness in ESP Representativeness, as we mentioned, is a key feature when building any corpus. Thus, a specialized corpus must as well fulfill this feature. Due to the specificity of texts, nevertheless it might be considered easier to achieve representativeness when focusing in one particular genre or/and field, for instance in our case, scientific articles in medicine. However, how can we be certain that we have achieved representativeness? Even though there is no unique answer to this question, we still can consider some relevant factors, among which possibly the most important is generalization. As long as we can find relevant results in our corpus that allow to be generalized, the results or findings will be

[13] This type of corpus contains general topics and the content constantly grow

[14] Representativeness (§ 2.3), finalization (§ 2.5) and size (§ 2.7) are therefore interrelated parameters: it is in fact clear, however, that attention to the representativeness of a corpus means attention to the function for which the corpus itself is intended; and for the purpose of specific research, a collection of limited dimensions but balanced in itself.

relevant. As mentioned before: "Leech (1991) has suggested that a corpus is 'representative' in the sense that findings based on an analysis of it can be generalized to the language as a whole or a specified part of it"(Kennedy, 2014, p. 62). Nevertheless, even though the generalization of results might be one of the main factors to establish whether one corpus is representative, there is no specific way to ascertain whether corpora are representative or not. The only bright part of this is that specialized corpora might be closer to it.

As for the present work, even though it is not possible to build a corpus *per se* due to time and resources limitation, our two compilations of texts will still show relevant data for our research. As a matter of fact, if we can demonstrate that the same patterns repeat themselves both in the first and in the second, a broader collection of scientific articles, and representativeness may be proved. This assumption is based on Zipf's Law, about the relation between the frequency and rank of the words in a text, as better explained further in the chapter. Due to the relation that Zipf's Law establishes between frequency and rank, it is safe to assume that, once we achieve a sufficiently large sample in which the frequency of the patterns does not change, further increasing the sample would not significantly change the frequency (only the occurrence). In this case, there would be no need to develop a broader corpus since even adding more texts to the analysis would neither affect the significance nor the results.

As clarified above then, in the case we could find syntactic patterns, as our hypothesis states, this preliminary results could be a lead-up to design and build a specialized corpus.

2.1.3 Types of Corpora

As we explained, a corpus design might have different parameters according to the purpose of the linguistic analysis, which will lead us to different types of corpora. In the following paragraphs we explain some of them. Dash (2008) in her book divides them under the following criteria like before and it allow us to understand the diversity that it might have and how much it might vary.

Genre of text There are three types of texts: written, spoken and speech.

Nature of data General corpus, special corpus, controlled language corpus, sub-language corpus, sample corpus, monitor corpus, multimodal corpus.

Genre of text There are three types of texts: written, spoken and speech.

- Written: A corpus exclusively design with written texts. For instance, the Brown Corpus.
- Speech: It consist on verbal interactions, natural, informal.
- Spoken: It is a technical extension of a speech corpus.

Nature of data It refers to the nature of content, such as general, specific, technical, etc.

- General Corpus: As we can deduce this corpus contains different types of texts from different areas. However, it is important to distinguish between formal and informal language, fiction or non-fiction, among other important distinctions.
- Special Corpus: On the contrary, a specialized corpora will characterize from the specific types of language including varieties, dialects, a specific area,
- Controlled Language Corpus: This type of Corpus is focused basically towards a very restricted on the grammar, style, vocabulary. It is still broadly discussed. One example of this Corpus is Caterpillar Fundamental English
- Sublanguage Corpus: It a based on material of an specific language dialect
- Sample Corpus: This corpus it is based on a very specific type of text. Usually based on literary work, like poetry and others.
- Monitor Corpus: This corpus is similar to a General Corpus, but unlike it is open to add more texts under certain criterion. According to Hunston and Francis (2000) this type of corpus will derive into a real representativeness. For instance, the Collins Birmingham University International Language Database is a monitor corpus.

Type of text Monolingual, bilingual, multilingual.

- Monolingual Corpus As the name implies it is a corpus based in one language.
- Bilingual Corpus It contains 2 languages sharing the same text genres,
- Multilingual Corpus The union of monolingual corpora that share the same parameters when designed from different languages

Purpose of design Unannotated corpus, annotated corpus.

Nature of application Parallel corpus, translation corpus, aligned corpus, comparable corpus, reference corpus, learner corpus, opportunistic corpus.

2.1.4 The approach: corpus-based vs. corpus-driven

So far, we mentioned the features required to design a corpus and the different types of corpus based on the research's purpose. Nevertheless, it is important to notice that corpora are nothing more than tools; more or less specific, more or less refines, but still tools. Hence, the corpus-based and corpus-driven approaches are concepts that concern the next stage when working with a corpus, and that eventually will also affect the design parameters. In the following paragraphs, we explain both. First, the prime requirement of a corpus-based approach is a hypothesis or a theory to validate. However, they do not always totally match theory and data. "[...] certain amount of variation that has not been accounted for is not important enough to topple a well-established theoretical position" (Tognini-Bonelli, 2001, p. 67).

In this case, Tognini-Bonelli (2001) explains three stances to face this variation. Insulation, standardization and instantiation, the latter one being crucial to our research.

Insulation (Tognini-Bonelli, 2001, p. 68) As mentioned, theory and data are confronted, founding data as a secondary position. Aarts (1991) contrasts intuition-based and observation-based which refer to, basically equivalent to competence and performance, prioritizing the former. The process starts with a hypothesis about language then "expresses it in a formal grammar". Then is confronted with a corpus as a "test-bed". Aarts also highlights the importance of not only the frequency but also the "normalcy" it means the acceptance of it among the users. The linguist uses the corpus, but also analyses if it is grammatically accepted.

Standardization (Tognini-Bonelli, 2001, p. 71)Leech's is a fundamental part when we talk about standardization. First, the theory tests it out, then the importance of annotation is highlighted to this point. Unlike the previous stance, this one doesn't stick to an 'intuition-based' or competence point of view. Corpus annotation is a plus since the information will be very valuable.

Nevertheless, it could also be consider a disadvantage in means of interference with the linguist work. For instance, an annotated corpus sentence will be displayed with tagging, which will lead up to a hard to read text and as consequence the analyst won't notice other patterns or information. Nowadays, there are tools to remove them so only the plain text will be displayed.

Instantiation (Tognini-Bonelli, 2001, p. 74) The third stance is more related to the work of Halliday and it is an important reference. His work was mostly based in probabilities.

> If a word has appeared ten times per million in a corpus of a hundred million words, there is a good chance that it will do much the same in the next hundred million if there is no major change in the corpus constituency. (Tognini-Bonelli, 2001, p. 75)

Then, corpus-based approach pursues to validate or not a theory, and, if necessary, the mentioned stances are used to adequate the data and the theory. The analogy of Halliday will apply in the same way when frequency as probability.

On the other hand, corpus-driven approach bases the theory on the data. It means, data has priority, first the corpus is analyzed and then the theory may be designed. It also has a special treatment when designing the corpus, for instance balance and representativeness is something that will be accomplish "naturally" Tognini-Bonelli (2001). "[...]the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" (Tognini-Bonelli, 2001, p. 84).

Or as Heine and Narrog states: "[...] attempting to uncover new linguistic constructs through inductive analysis of corpora" (Heine and Narrog, 2015, p. 202).

The advocates of this approach are called Neo- Firthians by the English linguist J. R. Firth. For instance, the Collins Birmingham University International Language Database (COBUILD) project, one of the first corpus-driven lexicography works, founded by J. Sinclair and published in 1987 for the first time.[15]

In general, the difference between both is that corpus-based approach has indeed a theory or hypothesis to be proven, while corpus-driven approach uses the corpus to identify the patterns first. Unlike corpus-based, corpus-driven has a firm objection towards an annotated corpus to avoid interfere in the researcher.

2.1.5 Corpus Linguistics concepts

In the previous sections, we reviewed different concepts about corpora in order for the reader to understand the whole idea of corpus in linguistics. However, there are still some clarifications due to the diverse possibilities and vast amplitude of the area of corpus linguistics there are nonetheless still aspects that need systematization. For instance, not

[15] https://collins.co.uk/pages/elt-cobuild-reference-who-created-cobuild

all scholars agree to whether corpus linguistics is a method or a theory. However, the aim of this work is not to decide among either of them, on the contrary we present here some of the most important positions.

About whether it should be considered a methodology or a theory, Tognini-Bonelli (2001) points out: "[...] corpus linguistics is in a position to define its own sets of rules and pieces of knowledge before they are applied" (Tognini-Bonelli, 2001, p. 1). This position allows the researcher use new parameters which implies a change, and obtaining a theoretical status. About the same subject McEnery et al. (2006) underlines: "As corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory itself" (McEnery et al., 2006, p. 1).

According to both scholars mentioned above, we could define corpus linguistics as a methodology. Even though as we have mentioned before, there are also different points of view about this such as: McEnery et al. (2006), Stefanowitsch (2020). However, as far as the present work is concerned, we are better inclined towards considering it a methodology, since we will use it to carry out linguistics analysis.

Finally, we present here two different concepts of corpus linguistics in order for the reader have a broader understanding. The first one from Stefanowitsch: "[...]the investigation of linguistic research questions that have been framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus" (Stefanowitsch, 2020, p. 56), and from Zufferey (2020) a broader concept that clarifies the general idea about what is corpus linguistics:

> Corpus linguistics is a particularly effective method for establishing the frequent contexts in which a word or an expression is used. But corpus linguistics is also used for conducting research in fundamental areas of linguistics such as the study of syntax, since it makes it possible to identify the types of syntactic structures used in different languages. (Zufferey, 2020, p. 3)

2.2 Computational Linguistics

Computational linguistics might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is just the construction of computer programs to process words and texts in natural language. The processing of natural language should be considered here in a very broad sense that will be discussed

in the present section due to the strict relationship between corpus and computational linguistics.

Nowadays, corpus linguistics is deeply related and linked with computers, which was one of the triggers to its remarkable development. Therefore, it is clear that without computers, it wouldn't have been possible to store big corpora and much less feasible to analyze a bulk of texts at once. Computational linguistics is part of applied linguistics, a term coined by David Hays according to the Oxford Handbook of computational linguistics, (Mitkov, 2004, p. 25) but what is computational linguistics? Besides the understood relation between computer science and linguistics.

For instance, "[. . .] might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is just the construction of computer programs to process words and texts in natural language" (Bolshakov and Gelbukh, 2004, p. 25). Other concept is from McEnery and Hardie (2011) that defines it as "[. . .]the field of computer science that looks at how computer systems can be created that work with language in some way" (McEnery and Hardie, 2011, p. 227).

Hence, the construction and use of computer programs in relation to natural language, but Stanford Encyclopedia of Philosophy adds:

> And since language is our most natural and most versatile means of communication, linguistically competent computers would greatly facilitate our interaction with machines and software of all sorts, and put at our fingertips, in ways that truly meet our needs, the vast textual and other resources of the internet. Schubert (2020)

In other words, Computational Linguistics aims to process Natural Language through the computer: getting computers to handle some aspect of language, to carry out automatically a job that a human analyst would otherwise have to do. Alignment, part-of-speech tagging, named entity recognition and machine translation are examples of applications in NLP. As a matter of fact "any contemporary approaches to NLP make use of corpora: automatic analysis of a corpus is used to build up the base of knowledge about language that the NLP program uses to complete its task." (Baker et al., 2006, p. 121)

Among these programs that allow us to manipulate Natural Language, there is AntConc. See3.2.1, p. 43, which we used in our work. This software does some crucial work for corpus linguistics analysis, since it allows, within seconds, counting words and classifying them by frequency, types, tokens, the range of the word through the corpus, the frequency of a chunk

of different words, collocations, etc. In the next section, we review those terms needed in order to understand our work.

2.2.1 Terminology: Zipf's Law, types, tokens, collocations and colligations

This terminological section set the bases for the understanding of the study developed in the present work. Once again, it is not meant to be exhaustive as to cover the whole universe of computational and corpus linguistics, on the contrary it aims to give the reader the base concepts needed to fully understand the methodology and analysis chapters.

Zipf's Law

In order to verify and support our results, Zipf's law plays a crucial role. As we explain in more detail in this section, this law describes the recurrence in the frequency of words which might signify a repetition in the co-occurrence of words.

Zipf's Law was popularized by George Kingsley Zipf[16] around the year 1935[17]. In his book "Human behavior and the principle of least effort: An introduction to human ecology" there is a section titled "On the economy of words" which explains that people tend to use fewer words to explain something, but the receptor needs more words to understand better. The tendency to use certain words more than others leads to a difference in frequency. It means, there are more words used rarely[18] and, fewer words used more frequently.

Hence, the phenomenon refers to the constant relation between the frequency f and the rank r of words appearing in a text (or collection of texts) $f.r{=}C$ where a pattern repeats in the frequency of words. For instance, a word frequency list of a corpus would show that the second most frequent word will have almost half of the first one; the third, one-third of the first; the fourth, one forth and so on[19] Further studies in Mathematics and Linguistics demonstrated that this pattern appears not only in language; but also the number of times

[16] American linguist and philologist who studied statistical occurrences in different languages

[17] "This relationship between frequency and rank appears first to have been noticed by Estoup 1916 but widely publicized by Zipf and continues to bear his name"(Manning and Schutze, 1999, p. 24). Jean Baptiste Estoup (1868-1950) stenographer. "The first person (to my knowledge) to note the hyperbolic nature of the frequency of word usage was the French stenographer, J. B. Estoup who made statistical studies of cf. his Gammes Stenographiques, Paris, 4 ed.,1916"(Zipf, 2016, p. 546).

[18] Words that appears only once in a text or a corpus is called *Hapax Legomenon* (See subtitle 2.2.2 p. 36

[19] Even though this statistical phenomenon occurs in a close range, around the 100th there is a more great difference in the $f.r{=}C$.

academic papers are cited Redner (1998),or for instance, in the article "A mysterious law that predicts the size of the world's biggest cities" Newitz (2013) discusses about the cities population. Furthermore, there are more fields where Zipf's law presents, however, we would consider only the linguistic relevancy for our research.

The zipfian tendency shows a decrease as in figure 2.2 p. 30 (Stabler, 2003). We find this tendency in various pieces of literature such as: Romeo and Juliet, Moby-Dick, Ulysses, Alice's Adventure of Wonderland etc., but also, and most importantly in huge collections of texts and, of course, in corpora such as Brown Corpus, Lob Corpus, etc. The following figure illustrates the similar zipfian tendency comparing; Brown Corpus, Lob Corpus, Wall Street Journal, the Bible, Shakespeare and Austen (Stabler, 2003). See 2.1 p. 29.

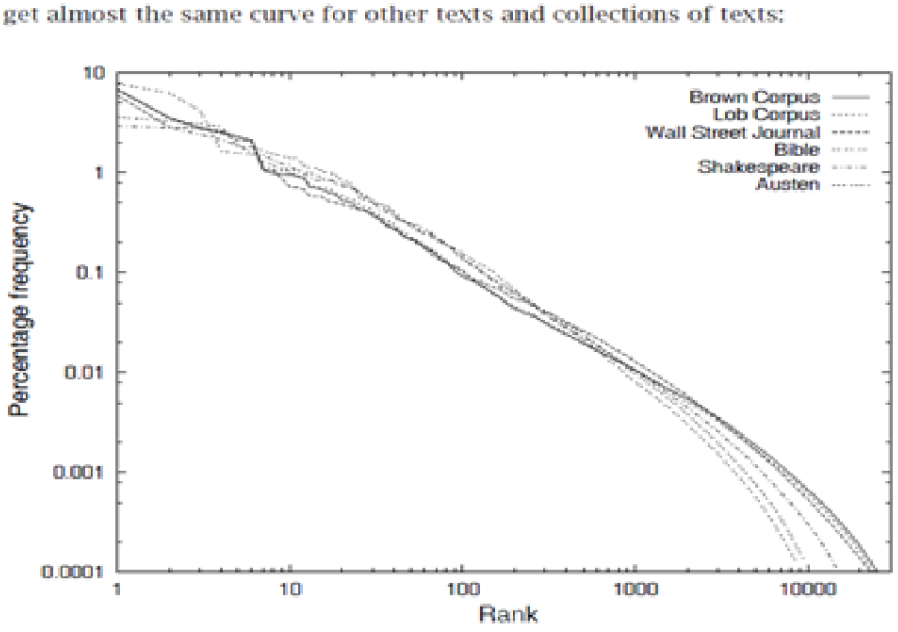

Figure 2.1: Corpora Frequency graph

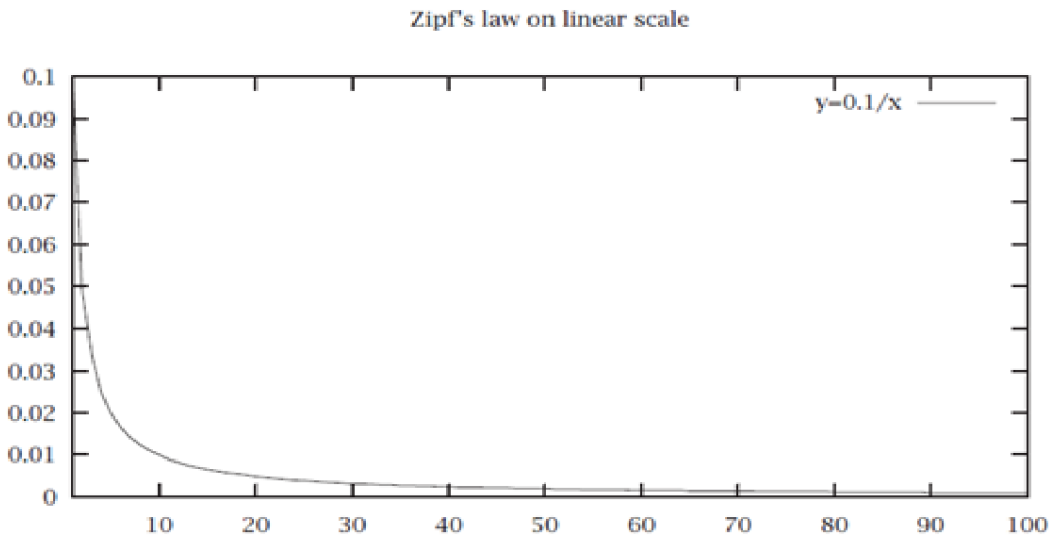

Figure 2.2: Zipfian tendency

In addition, we are able to graphic the word frequency list from the Michigan Corpus of Academic of Spoken English (MICASE) of the list that shares (Schmitt, 2013) of the top 50 words in the corpus. See figure 2.4 p. 32 and the graphic 2.3 p. 31 which also presents a zipfian tendency.

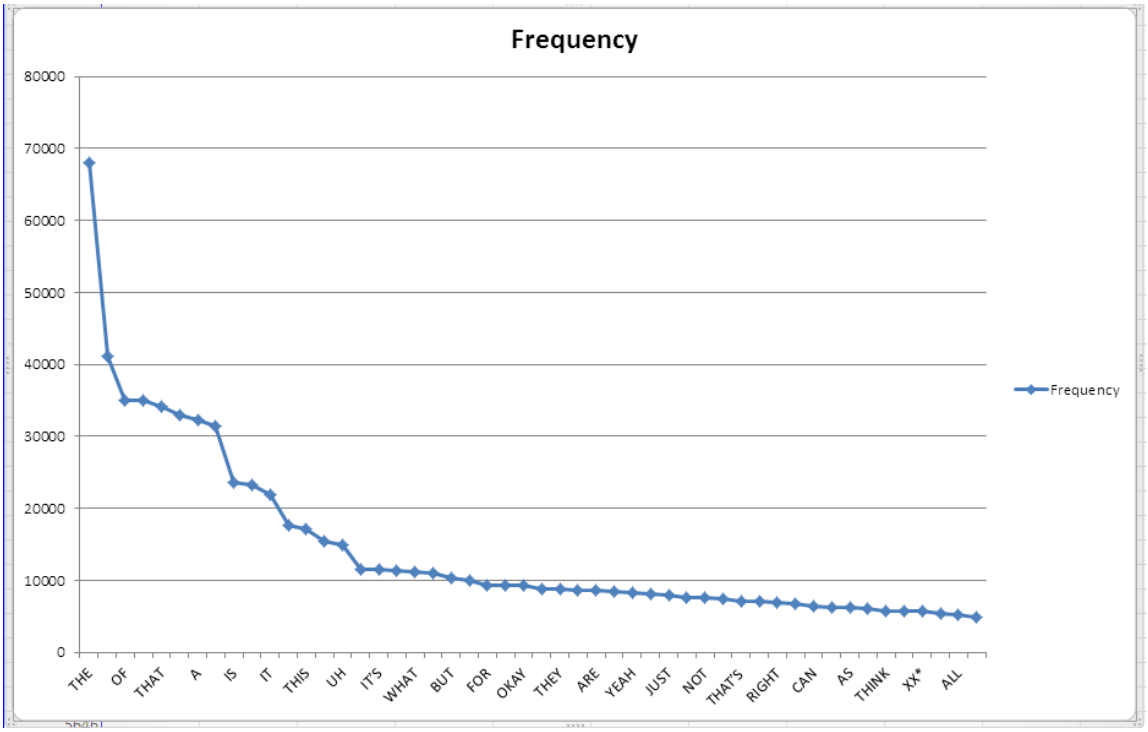

Figure 2.3: Frequency Graph MICASE

The tendency described by Zipf's law is a crucial resource in corpus linguistics, since it allows predicting if a pattern could repeat in a broader compilation of texts or a corpus intended to study the hypothesis.

In addition, Kennedy (2014) presents a comparison showing the most frequent words among corpora, allowing us to state that all collection of texts would present patterns to some extent; in Kennedy (2014) case the word most used in English is *the* and the 20 most frequent words in English are 'the, of, and, to, a, in, is, I, that, it, for, you, was, with, on, as, have, but, be, they'. See figure 2.5 p. 33(Kennedy, 2014, p. 98).

Table 6.1 The 50 most frequent words in the Michigan Corpus of Academic Spoken English (MICASE)

N	*Word*	*Frequency*	*N*	*Word*	*Frequency*
1	THE	68,036	26	BE	8874
2	AND	41,091	27	THEY	8799
3	OF	35,053	28	ON	8650
4	YOU	34,986	29	ARE	8596
5	THAT	34,085	30	IF	8440
6	TO	33,029	31	YEAH	8292
7	A	32,236	32	WAS	8179
8	I	31,483	33	JUST	7970
9	IS	23,535	34	DO	7675
10	IN	23,255	35	NOT	7638
11	IT	21,883	36	OR	7488
12	SO	17,669	37	THAT'S	7042
13	THIS	17,110	38	ABOUT	7014
14	UM	15,346	39	RIGHT	6980
15	UH	14,859	40	WITH	6726
16	HAVE	11,590	41	CAN	6350
17	IT'S	11,560	42	AT	6312
18	WE	11,383	43	AS	6229
19	WHAT	11,236	44	THERE	5991
20	LIKE	11,037	45	THINK	5796
21	BUT	10,402	46	DON'T	5650
22	KNOW	10,000	47	XX*	5646
23	FOR	9282	48	THEN	5443
24	ONE	9267	49	ALL	5289
25	OKAY	9250	50	TWO	4937

Figure 2.4: Frequency list MICASE

Table 3.3 Rank order of the 50 most frequent word types in the *Birmingham Corpus* compared with other corpora

	Birmingham Corpus	*Brown Corpus*	*LOB Corpus*	*Wellington Corpus*	*American Heritage Corpus*	*London–Lund Corpus*
the	1	1	1	1	1	1
of	2	2	2	2	2	5
and	3	3	3	3	3	3
to	4	4	4	4	5	4
a	5	5	5	5	4	6
in	6	6	6	6	6	9
that	7	7	7	10	9	8
I	8	20	17	12	24	2
it	9	12	10	9	10	10
was	10	9	9	8	13	13
is	11	8	8	7	7	11
he	12	10	12	16	11	18
for	13	11	11	11	12	20
you	14	33	32	31	8	7
on	15	16	16	13	14	16
with	16	13	14	14	17	32
as	17	14	13	15	16	29
be	18	17	15	17	21	21
had	19	22	21	23	29	55
but	20	25	24	26	31	15
they	21	30	33	27	19	24
at	22	18	19	18	20	26
his	23	15	18	24	18	85
have	24	28	26	29	25	19
not	25	23	23	25	30	35
this	26	21	22	22	22	14
are	27	24	27	20	15	42
or	28	27	31	32	26	44
by	29	19	20	19	27	65
we	30	41	40	36	36	23
she	31	37	30	28	54	72
from	32	26	25	21	23	53
one	33	32	38	40	28	36
all	34	36	39	41	33	33
there	35	38	36	35	37	38
her	36	35	29	33	64	96
were	37	34	35	30	34	64
which	38	31	28	39	41	43
an	39	29	34	34	39	81
so	40	52	46	48	57	30

Type and Token

These key terms are fundamental in corpus linguistics analysis. The difference between these two is as follows: 'Word types' refers to the distinct words in a text, and 'word tokens' refers to all the words even if it is repeated several times. In other words, the number of tokens in a corpus is the number of the total of words without distinction, but the total of word types is every type of word without considering the repetition of the same word.

Hence, in the sentence (1) there are only 8 types (and 10 tokens) since "the" and "big" are repeated in this sentence.

(1) The big cat from the big purple house is beautiful.

The number of types and tokens is obtained within seconds thanks to the concordance software. This calculations lead us to the one of the corpus most basics, yet important statistics: the Type Token Ratio (TTR). We calculate the TTR by dividing types by tokens, in order to provide us a glance about the lexical variations. For example, in the previous sentence the TTR is 80%. The closer it gets to 100% the greater the vocabulary, the variety and so on. It is evident, as for Zipf's law as well, that a natural text can have a TTR close to 1 only in very small sentences, due to the necessity of repeating functional words, such as articles, prepositions and such.

In order for the machines to calculate TTR, computational linguistics softwares put the texts through a process known as "Tokenization"; this is the process in which the concordancer separates each word and punctuation in order to process the text.

(2) a. He's a student.
b. He ' s a student .

All this processes and counting are possible thanks to the computational approach, since it is evident that, even though it might be possible to pull it out by hand, it would be inefficient and prone to errors.

Collocation and Colligation

In the previous section we analyzed the calculation needed in order to study the occurrence of words in corpora and collections of texts. Another important factor, however, is the co-occurrence of words, that originates sentence patterns. A co-occurrence implies collocation

or colligation, both terms introduced by J. R. Firth, and used mainly in corpus linguistics analysis through statistical analysis; these terms refer to the possibilities that two or more words may occur together or in a range between them. "Specifically, collocation and colligation refer to the likelihood of co-occurrence of (two or more) lexical items and grammatical categories, respectively" (Lehecka, 2015, p. 1). The main difference between them is that collocation refers to the relation of these words in a lexical level, meanwhile, colligation is the relation in a grammatical level. While the grammatical connections of colligations is clear and easy to understand, the concept of collocation might need some more explanation in the following part of this section-

'Collocations' are words connected by a relation that stands even if they are not next to each other, this separation between the node[20] and the collocate[21] is denominated *collocation window,* this separation may be to the left or right, these roles might be exchangeable. "The collocation window is specified as the number of words from the node, for instance 5 words to the left and 5 words to the right of the node (L5–R5)" (Lehecka, 2015, p. 1).

The meaning that a collocation acquires is different from the simple sum of the independent words. This characteristic is non-compositionality which refers to the words are not predictable by the meaning of them. "Either the meaning is completely different from the free combination (...) or there is a connotation or added element of meaning that cannot be predicted from the parts" (Manning and Schutze, 1999, p. 184). Manning also mentions among other characteristics non-substitutability, meaning that the words compounding the collocation cannot be replace by a synonym or other word that implies the same meaning. Non-modifiability, there is not possible to modified by adding other words into the collocation.

In the following example, we can appreciate the seven patterns of collocations Lewis et al. (2000), and therefore it is evident that the expressions acquires a more specific meaning than the simple juxtaposition of terms.

1. Adjective + Noun e.g. an opportune decision
2. Verb + Noun e.g. submit an assignment
3. Noun + Noun e.g. a TV program
4. Verb + Adverb e.g. apologize sarcastically

[20] Word searched highlighted and displayed in the center of the concordancer
[21] Word related to the node

5. Adverb + Adjective e.g. sound asleep
6. Adjective + Preposition e.g. sad of
7. Phrasal Verb e.g. go away

On the other hand, Benson et al. (1986) went even further as to separate collocations into two categories: grammatical and lexical collocations. According to their categorization Grammatical Collocations are those originated by the occurrence of Noun + Preposition. Adjective + Preposition , Preposition + Noun; Meanwhile Lexical Collocations are those patterns that originate by the co-occurrence of Noun + Noun, Verb + Adverb, Verb + Noun.

More recent studies, Boriskina (2009) and also Lewis et al. (2000), prefer to categorize collocations from a different point of view:

1. Unique collocations which refer to fixed and unreplaceable collocations, such as to foot the bill.
2. Strong collocations, which are very strong but not unique, such as, reduced to tears or moved to tears.
3. Weak collocations, which consist of a number of co-occurrences that can be easily inferred, such as a black dress, a red dress, a dark blue dress, etc.
4. Medium-strength collocations, which can be weak, such as, to hold hands or to make a deal, and they are a single item too.

2.2.2 Patterns, Hapax Legomenon, N-grams

The concept of pattern is significant to the analysis section, therefore, we explain what is a pattern and some meaningful combinations to encounter in a corpus study. We also explain the meaning of the terms Hapax Legomenon and N-grams, since are key terms related to these concepts in the following paragraphs in order for readers for definitions background is have the necessary theoretical tools to understand the work.

Patterns, (Hapax)

A pattern is the sequence of words or structures that repeat in a frequent basis. "All the words and structures which are regularly associated with the word and which contribute to

its meaning" (Hunston and Francis, 2000, p. 37). The frequency of these combinations of words through a corpus or texts must represent a relevant information about the word or structure we are analyzing to be considered a pattern.

For instance, when analyzing a noun pattern a left search will show different types of modifiers and to the right its complementation. When there is a verb the complementation, it will be placed to the right. In this sense, we found different structures in texts that are common patterns according to Hornby mentions some of these, but Hunston and Francis (2000) present a complete list (Susan Hunston, 1998, p. 37-341). See table 2.1, p 37.

a N	N in n	it v n N to-inf
the N	N in favour of n	it v-link det N-ing
poss N	N into n	it v-link N -ing
adj N	N of n	v it det N to-inf
nN	N on n	v it as det N to-inf
nN	N over n	there be det N about n
num N	N to n	there be det N in n/-ing
ord N	N towards n	there be det N to n/-ing
Nthat	N with n	at N
Nto-inf	the N be that	by N
v-link adj N to-inf	possN be that	from N
N about n	the N be to-inf	in N
N against n	poss N be to-inf	into N
N among pl-n	the N be-ing	of N
N as n	poss N be -ing	on N
N as to wh	it v-link det N that	out of N
N at n	it v-link a adj N that	to N
N behind n	it v-link poss N that	under N
N between pl-n	it v-link det N to inf	with N
N by n	it v-link N to-inf	within N
N for n	it v-link poss N to-inf	without N
N from n	it v N to-inf	

Table 2.1: Noun Patterns in Grammar Patterns Nouns of Hunston and Francis (2000)

Not only frequency but also the relation between words is a relevant factor when referring to patterns. A pattern could be a collocation or an N-gram2.2.2, it means that it might be next to each other or not, what is important is to determine the relation between them. For instance, the example 2.2.2 *step in* is not the phrasal verb in this case, but the Noun *step* and the preposition *in*, without the analysis this could be mistaken with the phrasal verb.

2218	The	first	and	crucial	`step`	`in`	disease	data	integration	is
			conj	adj	N	(n)prep	N	N	N	V

On the other hand, *Hapax legomenon*, in Greek 'once said', are words that appear just once in a text or corpus. According to Zipf's law (See 2.2.1, p. 28) a statistic phenomenon about the frequency of words in texts, *hapax legomenon* isn't rare. Thus, it is possible to find a word that repeats only once, but it's more likely to encounter many of these words. For instance, the article "How many words are there" in the Glottometrics Journal[22] mentions that approximately about 50% is *hapax* in a large corpus Kornai(2002) mentions that a corpus of `The America News Paper San Jose Mercury News` (Merc) shows of 56.6% of hapaxes. "[...] consistent with the observation in Baayen (1996) that in large corpora typically more than 50% of the words are hapaxes" (Kornai, 2002, p. 75).

Besides Hapax there are also *dis-legomenon, tris-legomenon* and *tetrakis legomenon* referring to words that appears two, three and four times in a corpus respectively, but these are less common.

N-grams

N-grams[23] are words **series**, words that are next to each other even if they are not related . We can potentially have series of n-words, where N stands for the number of words.

For example, (1) on 34 *The big cat from the big purple house is beautiful.* has ten unigrams, nine bi-grams, five tri-grams, and three 4-grams. Logically there are more bi-grams than tri-grams and 4-grams:

(3) a. The big | big cat | cat from | from the | the big | big purple | purple house | house is | is beautiful.
b. The big cat| cat from the | the big purple | purple house is | house is beautiful.
c. The big cat from | from the big purple | purple house is beautiful.

Bi-grams: Two words next to each other to this point as Crowford and Csomay Crawford and Csomay (2015) mentions all collocates of two words are bi-grams, but not all bi-grams are collocates.

[22] Open access Scientific Journal for the quantitative research of language and text

[23] N-grams is hyperonym for any group of N words, therefore can be articulated into Bi-grams, tri-grams, four-grams, and so on.

Tri-grams: The sequence of three words.

Four-grams: The sequence of four words.

Five-grams or more: The frequency to have a five or more sequence is reduced in probability since it will contain the bi-gram or three-gram.

In order to avoid confusion, it is important do differentiation between collocations and N-grams. Then, the 'collocations' and 'N-grams' difference relies on how words are related to each other. For instance in (4) we have a collocation *spend time* that function without being next to each other.

(4) She *spends* a lot of *time* singing.

2.3 English for specific purposes

As the natural conclusion of the present chapter, dedicated to introduce, explain and organize all terms and references that would allow the reader to follow and understand the analysis developed in the present work, we close this chapter with a brief contextualization about the concept of English for Specific Purposes.

Through time, ESP definition has been evolving. Since the first time this term appeared after World War II due to the need of a world lingua franca[24] more and more scientific texts and articles of different fields were published in English. By the end of the 20th century, English was the lingua franca of science and technology (Dudley-Evans and St. John, 1998). In order to work on the same basic definition, it is important to review some definitions of ESP. The first concept was from (Robinson, 1980, p. 13) who defined it as:

> It is a course that is [...] purposeful and is aimed at the successful performance of occupational or educational roles. It is based on a rigorous analysis of students' needs and should be tailor-made. Any ESP course may differ from another in its selection of skills, topics, situations and functions and also language. It is likely to be of a limited duration. (Robinson, 1980, p. 13)

Later in 1987, Hutchinson and Waters defined ESP as: "An approach to language teaching in which all decisions as content and method are based on the learner's reason for

[24]Cambridge Dictionary defines it as: "a language used for communication between groups of people who speak different languages" (Cambridge English, 2020).

learning" (1987, p. 5). In 1998 (Dudley-Evans and St. John, 1998, p. 4-5) proposed seven main characteristics:

Absolute characteristics:

- ESP is defined to meet specific needs of the learner.
- ESP makes use of the underlying methodology and activities of the discipline it serves.
- ESP is centred on the language (grammar, lexis, and register), skills, discourse and genres appropriate to these activities.

Variable characteristics:

- ESP may be related to or designed for specific disciplines.
- ESP may use, in specific teaching situations, a different methodology from that of General English.
- ESP is generally designed for adult learners, either at a tertiary level institution or in a professional work situation. It could, however, be for learners at secondary school level.
- ESP is generally designed for intermediate or advanced students.
- Most ESP courses assume some basic knowledge of the language systems.

Thus, after reviewing these concepts, we understand the difference from a general English class. For instance, the specificity of the subjects, the learners needs, the importance of the material's authenticity. Even though ESP has been classified in some other more specific branches, for example by purpose, discipline, experience and degree of specificity, in this paper we will, however, refer to the term ESP as the general broad term, therefore considering and including also the different branches.

Chapter 3

Methodology

3.1 Type of Research

This research can be categorized as a descriptive-quantitative study considering the different aspects we developed throughout it. Thus, we will go into further detail to explain the extent of each one of them in this research.

The descriptive component contemplates identifying and categorizing relevant syntactic patterns, through the recollection method of observation and the analysis of both collections. The tool to spot the most frequent words in both collections, which later led to n-grams, was AntConc.

Thus, in order to calculate the frequency and have a complete, detailed description of the syntactic patterns and their frequency in both collections, we rely on a quantitative analysis of our samples. Therefore, the quantitative part of our study develops a simple descriptive statistical analysis to prove the incidence in both collections, allowing the study of their correlation and the assumption of the conclusion section.

Hence, since the research has a strong descriptive base, aiming to the discovery, individuation and identification of the relevant syntactic patterns in the texts, the analysis of their frequency and behavior, it would not be useful or convenient to formulate an hypothesis answering our problem statement. According to Hernández-Sampieri et al. (2018, p.298)

> Cuando el estudio tiene una finalidad puramente exploratoria o descriptiva, debemos interrogarnos: ¿podemos establecer relaciones entre variables? En caso de una respuesta positiva, es factible seguir con la estadística inferencial; pero

> si dudamos o el alcance se limitó a explorar y describir, el trabajo de análisis concluye y debemos comenzar a preparar el reporte de la investigación.[1]

Nevertheless, it is also important to point out that there beyond the descriptive aim of the present work, there is also a fundamental, implicit goal in our problem statement: the fundamental question, as a result of the exploratory study is whether these patterns exist in our compilations and, if they exists whether they are relevant. In our words, if we can demonstrate the relevant occurrence of patterns, we can offer better tools for ESP teachingand learning. Even when we have other studies backing up that language is highly patterned, our research is focused mainly on relevant patterns in relation to ESP and the medical field, specifically focusing on medical scientific journal publications.

As for the methodological structure of the work, we decided to leave the hypothesis out of this section since the main part of our research is essentially descriptive.

3.2 Tools

The present study needed layered and multiple methodological approaches due to the multi-task aspect of the challenge. We, therefore, applied both quantitative and qualitative methods to get a satisfactory answer. A computational approach was employed to obtain data from various scientific journal articles, such as frequency, rank, range, etc. One of the tools that allowed corpus linguistics to grow is the concordance software. Even though there are many available, we use `AntConc`, a freeware concordancer, explained in greater detail in the following subsection.

Another tool we used in order to speed up our work is `AntFileConverter` software to convert `.pdf` or `.doc` files into `.txt`, also by the same designer of `AntConc`. Since `AntConc` specifically requires only `.txt` extension we needed a tool to convert Journal Articles in the correct format. Finally, another useful tool we decided to work with was "The Multi-File downloader", that allows us to download many `.pdf` files at once directly from the web page PubMed and therefore speed up the process.

[1] "When the study has a purely exploratory or descriptive aim, we should ask ourselves whwther we can establish relationships among variables. In case of a positive answer, we can follow with inferential statistics; on the other hand, in case of doubt, or in case of a merely descriptive aim, the analysis work can finish and we can start to prepare the investigation report." author's translation.

3.2.1 AntConc

As mentioned above, `AntConc` is a concordance software, designed and maintained by Anthony Laurence at Waseda University in Tokyo; it is a freeware[2] that allows us to extract different data from a collection of texts, or even a single one. In the following paragraphs, we explain step-by-step how to use it.

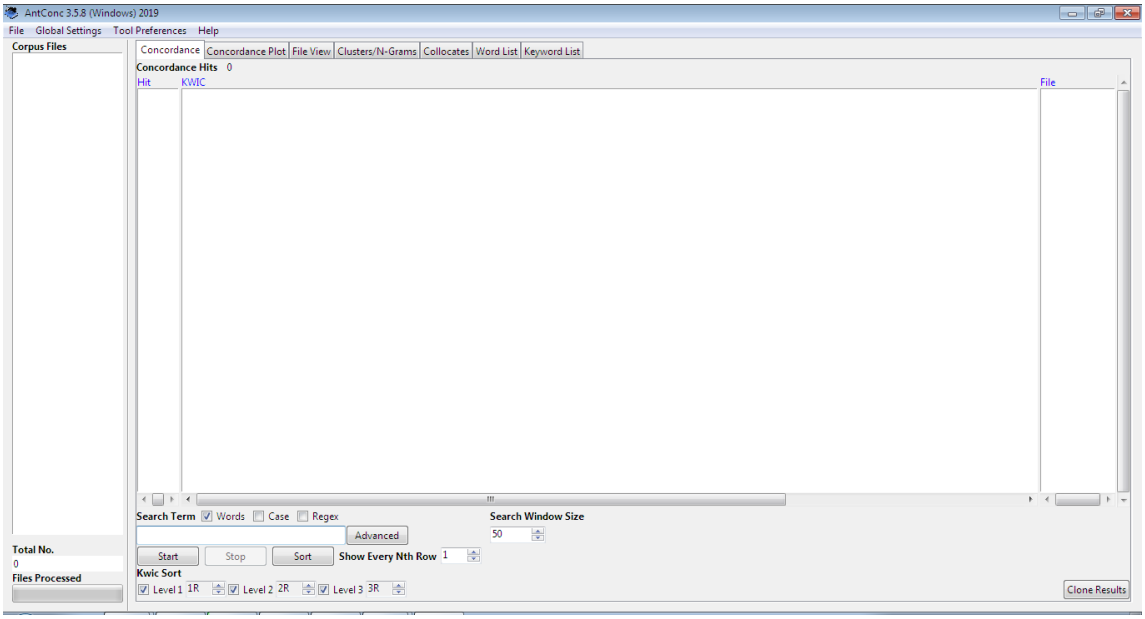

Figure 3.1: Antconc Freeware Concordancer Software

First, there are two options to upload a single document or a directory (many documents)`.txt`, the latter was very convenient for our study since it allows to work with the selected files. The window display shows on the left the names of each of our articles and seven tabs in the center area: Concordance, concordance plot, file view, cluster/N-grams, collocates, word list, and Keyword list. (Fig. 3.1, p. 43)

Initially, the default selected tab is the concordance tab to search for a specific word or words, but we choose the Wordlist tab instead and we click on the start key. The software immediately shows us the most frequent words, the rank,range, total word types and word tokens. So, we could observe that in our first collection (Group A 300) there are: 72,033 Word Types and 2,171,873 Word Tokens (Fig 3.2, p. 44).

[2] `AntoConc` can be downloaded for free at `https://www.laurenceanthony.net/software/antconc/`

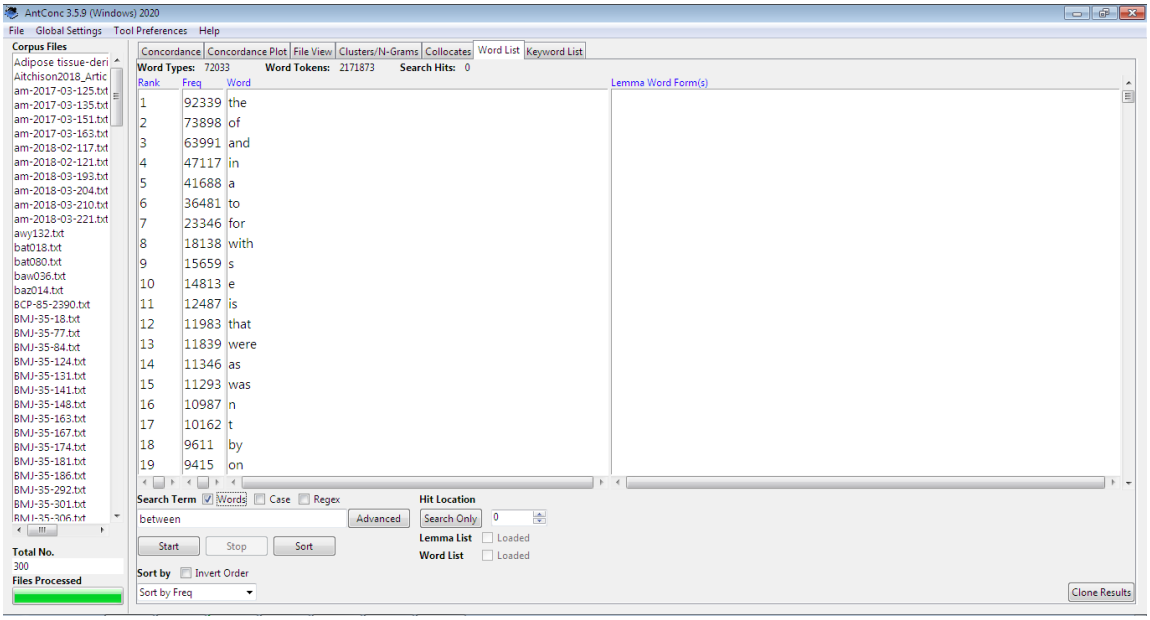

Figure 3.2: AntConc Word List tab

By clicking on one of these words, the software redirect us to the concordance tab displaying the sentences where our selected word appears. On the other hand, the concordance tab displays the information of all the occurrences, and we can also select each one if we need to highlight the words (linguistic context).

The searched word, known as Key Word In Context (hereafter KWCI), appears at the center in blue. There is the option to highlight the surrounded words (up to three) to the right or left. The central column in the KWIC will be in red and sorted alphabetically. For instance, we select the word ***Data***. In figure 3.3, p. 45, we can observe the word ***Data*** in a blue color. The next three words, to the right; in red, green, and purple, respectively. The software gives us the option to highlight the words after and before the selected word. At the bottom area of the window, there are three boxes **Kwic Sort** before and, it can be used to search and sort the results and identify the words either from the left or the right side as we have explained before. (Fig. 3.2, p. 44)

The Cluster/N-Grams tab allows us to search the words next or close to the KWIC. They show the rank of the cluster or N-gram, the frequency and range for the Cluster/N-grams, the total count of Collocates Types and Tokens; on the other hand, the Collocates tab shows the frequency in total (Freq) and the frequency to the left or right of the word (Freq(L) and

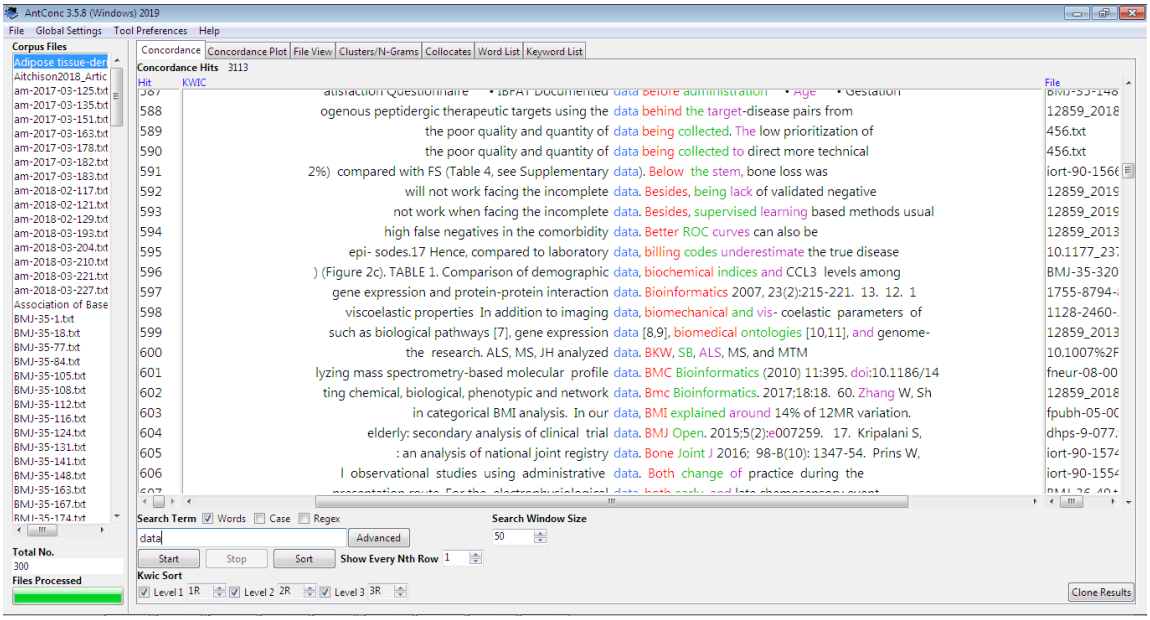

Figure 3.3: The word ***data*** as KeyWord in context

Freq(R), respectively. (Fig. 3.4, p. 46)(Fig. 3.5, p. 47)

The Concordance Plot tab displays diagrams to show the hits, dispersion of that word through the text itself and the total number of characters in the corpus or text. (Fig. 3.6, p. 47)

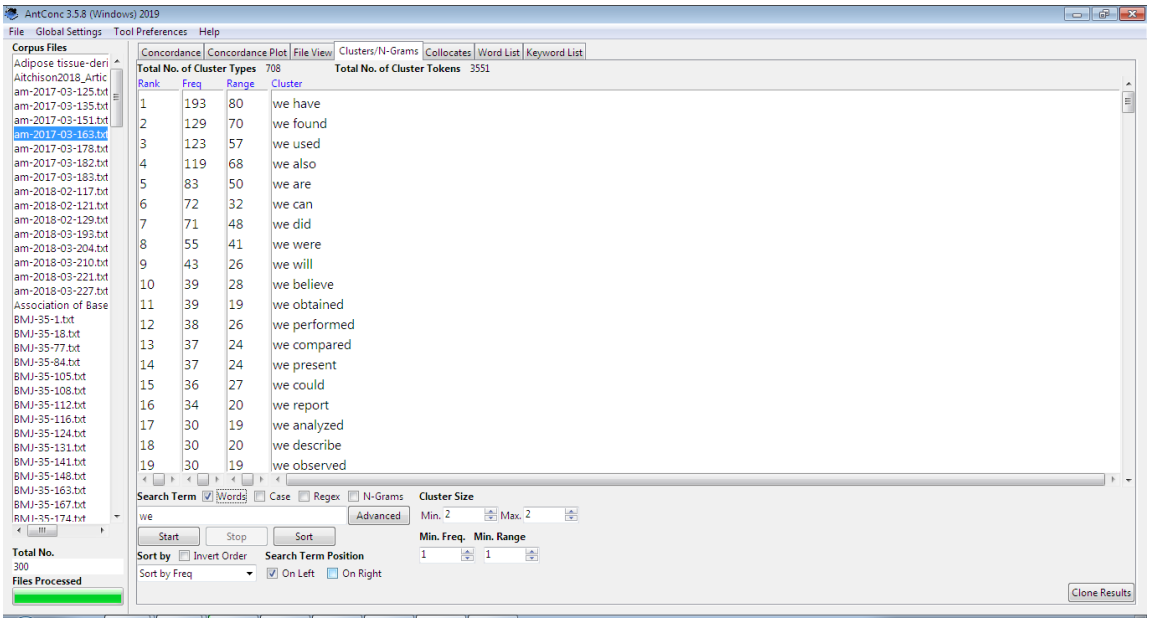

Figure 3.4: The word ***we*** cluster/n-grams. The tool shows automatically the rank, frequency and range for the Cluster/N-grams, the total count of cluster/N-grams Types and Tokens. (Own elaboration).

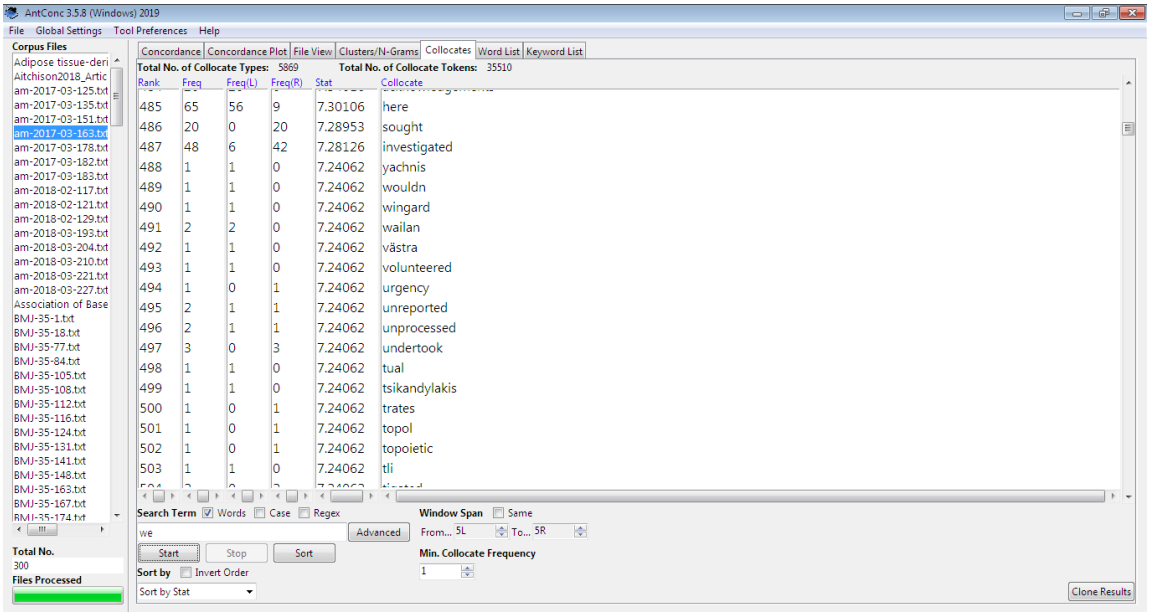

Figure 3.5: The word ***we*** collocates; the software automatically shows the rank, frequency and range for the Collocates, the Left and Right frequency and the total count of Collocates Types and Tokens. (Own elaboration).

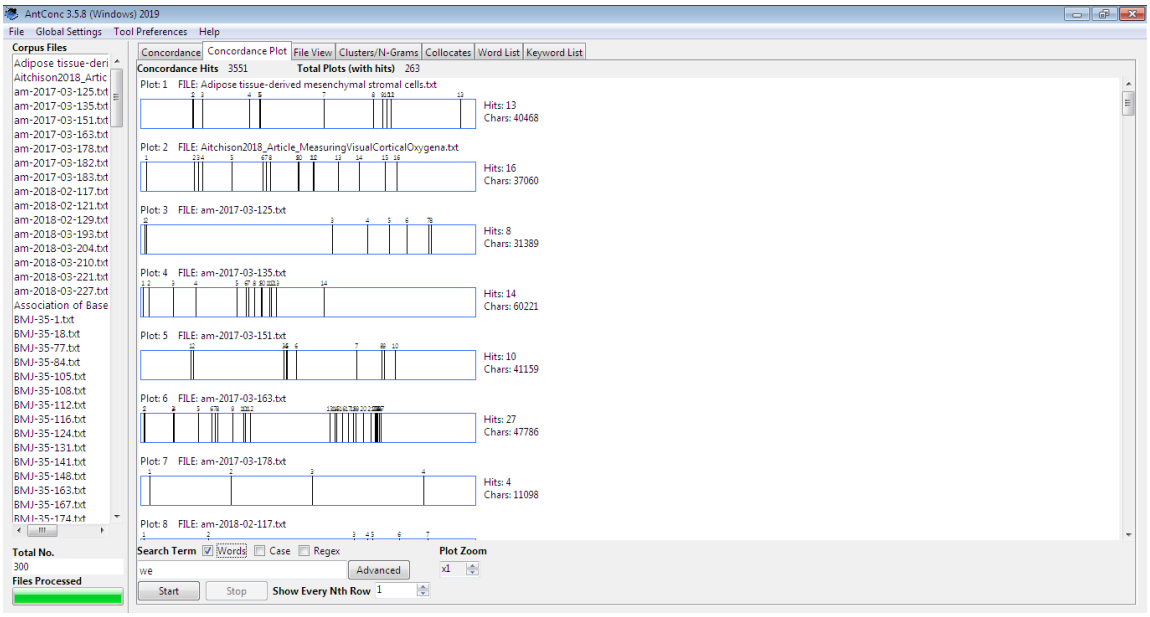

Figure 3.6: Tab Concordance Plot

3.3 Building the collection

The first phase of the research is based on a quantitative method: simple data extraction and analysis. We randomly chose 300 research articles from different journals to individuate, organize the relevant structures and linguistics patterns.

As far as the nationality of the authors is concerned, we found different nationalities which may correspond to native English speakers or not. This should not be considered a bias or a source of mistake since all articles have a revision as editing filter. However, there is often an editor and peer review to check each of the articles, which would give us an authentic language. "Cook's article implies that the model for most learners of English should be not the native speaker but the speaker who operates comfortably in English in international settings" (Tognini-Bonelli, 2001, p. 268).

3.3.1 Downloads

First of all, we downloaded the articles from PubMed Page[3] to build our sample. We selected randomly[4] inside the medical field; without filtering on author or publishing date. All selected journal scientific articles are available to download for free. The web page search tool allows us to filter the results.

Since downloading the articles one by one would be seriously time consuming, a downloader plugin in, that can be installed directly in the browser, gave us the option to get many files at once. However, as we explain later, this interfered in the selection resulting in downloading also other types of documents, therefore calling for a cleaning phase during the preparation of data. (See subsection3.3.2 on p. 51 for cleaning data process). It is nevertheless a great advance and a sensible decision in time, since it allows to download all `.pdf` documents in order to prepare them for the analysis.

3.3.2 Conversion

PubMed only offers articles to download in `.pdf` format, not `.txt`; nevertheless, it is worth considering that `AntConc` Software only accepts a `.txt` extension. We needed to convert all retrieved files in order to proceed with the analysis. Among the great number of on-line conversion tools available, the majority only allows converting one item at a time. Even

[3]PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. It can be accessed here: `https://www.ncbi.nlm.nih.gov/m/pubmed/`

[4]The selection was made through the list of journals that are available on the web page.

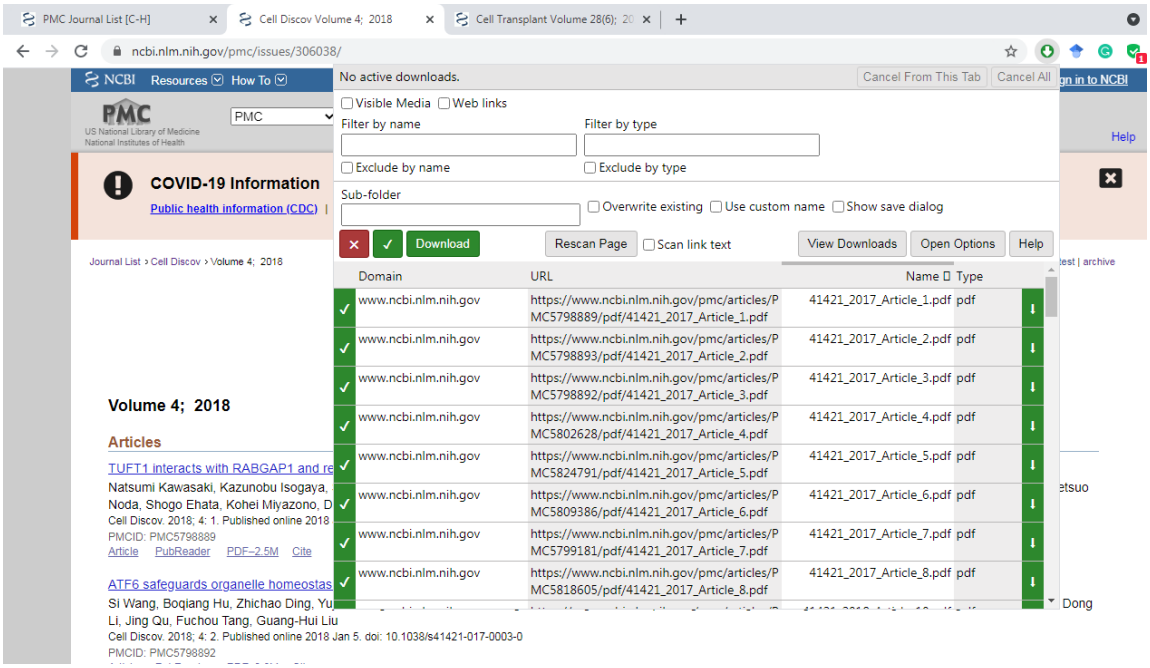

Figure 3.7: Downloader Screeshot. The picture clearly shows the possibility, thanks to the Downloader plugin, the possibility to search through an entire page and filter all possible download by type, by name, by subfolder etc. The filtered results can then be downloaded all together with a simple click on the button.

considering a small sample conversion, this would be a time consuming work. Luckily, the Antconc Software Suite offers another tool (also available for free): AntFile converter. With this tool, we have the option of bulk conversion, in other words to convert a great number of files, accessing an entire directory at once (see figure 3.8, p. 50).

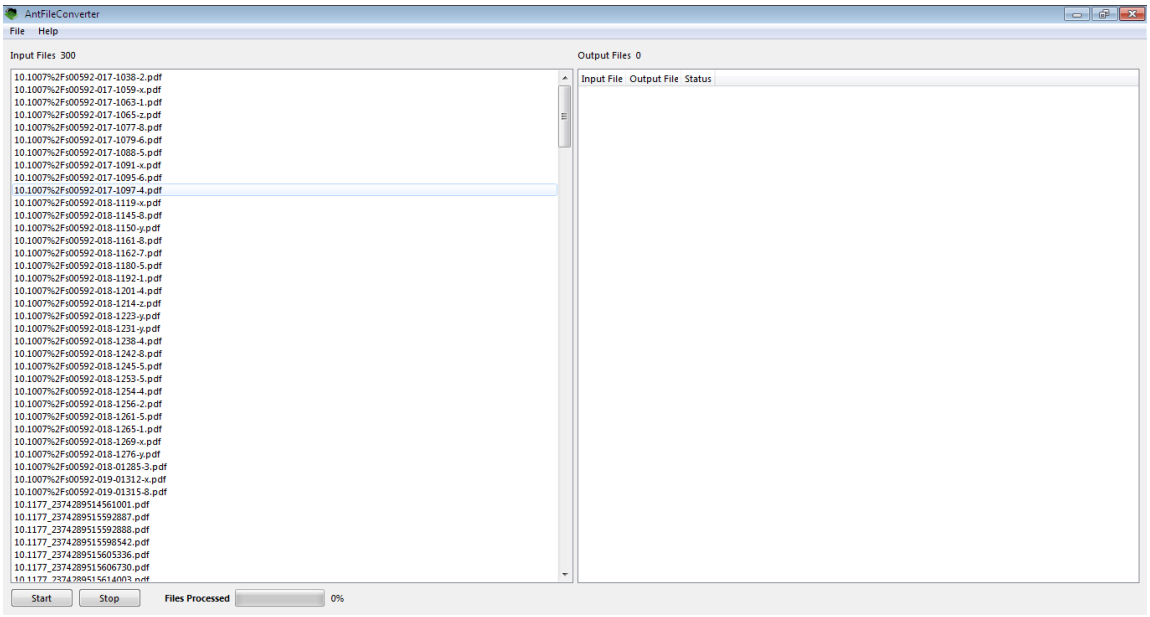

Figure 3.8: AntFile Converter from `.pdf` to `.txt` extension

Data preparation

The data cleansing process is a crucial step to avoid as much as possible potential "noise". There are three factors that might affect our analysis: the text type, the excessive images or tables, and fixed phrases. Since the selection was random and the downloading process was in bulk as explained above, it was inevitable to also retrieve non article part of journals (false positive in the downloading filter selection explained in the previous section), such as:

- Letter to the editor
- Errata/Erratum
- Case study
- Case report
- Educational case
- Brief report
- Abstract
- Obituary
- Commentary
- Editorial
- Correspondence
- Congress program

Along with these, some articles were entirely in Spanish, and even we had cases of abstracts in different languages. Due to some tables, graphics, pictures, format, etc., an article in a `.txt` extension may present, for example, no space between words, characters between letters, vertical lines of letters instead of horizontal. (see figure 3.9, p. 52). The conversion induced "noisy" data in almost all the articles because most of them have images or tables which causes single letters rows but we decided to only remove the articles that had many pages with this "noise".

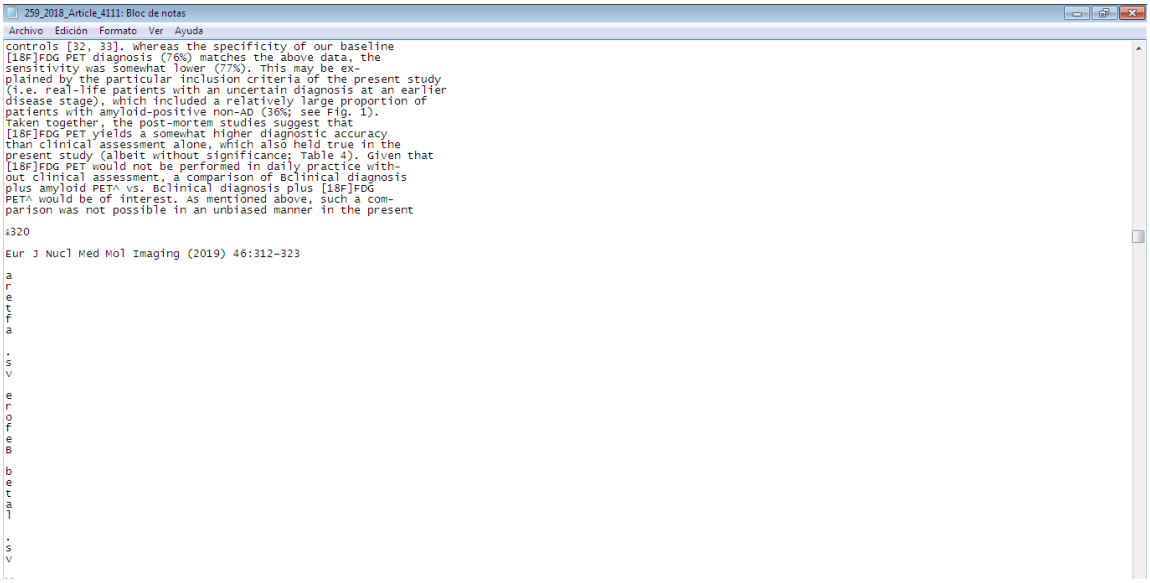

Figure 3.9: Vertical lines of letters found in `.txt` converted articles

As a matter of fact, other factors weren't possible to clean due to time and resources, such as the bibliography, reference section, formulas, codes, tables, etc. whose nature as text is somehow different and its study is beyond. All these "uninteresting data" created more `Hapax legomenon` and could produce a lot of difference in a study that focus on single appearance of words. Anyways, since our research is specifically focused on the most frequent patterns, the presence of (real or not) Hapax Legomenon will not be a relevant issue in our analysis. Other similar cases are represented by things such as last names (See figure 3.10, 53) or names of institutions. We didn't carry on the cleansing process because it will take more time and the conclusive data wouldn't be changed significantly, as stated before. As a matter of completeness, might it be the case for the need of cleaning "dirty data", we would suggest to work with Python and its Natural Language Project

False Positives

A false positive occurs when the information retrieved is unreliable or does not fulfill the requirements. As we have mentioned in the previous section; at first glance, we can immediately notice that the software has also identified single letters and other parts of text such as:

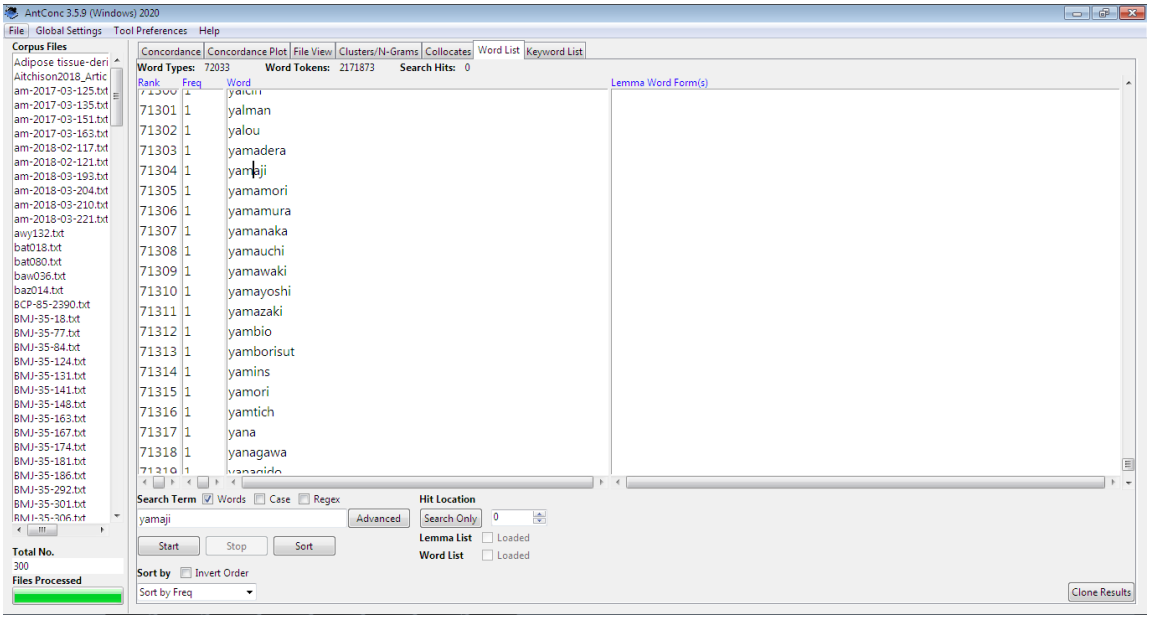

Figure 3.10: Uncleaned data detail screenshot. The picture shows the highest presence of researchers that, being listed by their family name.

- Acronyms
- Formulas
- Words split by a hyphen

It also refers to the different grammatical categories that a word may have; for instance, 'study' may be a Noun or a verb according to the context of the sentence. Because our compilation is not tagged, it shows one word without distinction among them in our frequency list. Punctuation is another factor that shows false positives, punctuation marks, etc. As our compilation is no tagged, and the punctuation is not differentiated, allowing us to focus only on the words that meet our requirements, we encounter false positives.

For instance, in the following examples we can see that at the right or left of our KWIC there is punctuation mark.

1416 the time for change is 24. `Health` Employment in the United States. National
1417 Solli O, Stavem K, Kristiansen IS (2010) `Health`-related quality of life in diabetes:

1418 definition of chronic disease, which is " `health` 1 3 Acta Diabetologica (2018)
1419 health system: Framework and research issues. `Health` Policy. 2012;105(2):185–91
28 Srivastava A, Thomson SB. Framework `analysis`: a qualitative methodology for applied
29 tal knee arthroplasty using radiostereometric `analysis`: a randomised controlled trial.

The collections are not tagged 'corpora' which doesn't allow us to differentiate the verb from the noun, for instance *study*. In this case, we can't determine the occurrences of each one.

120	localisation	between	pairs	of	genetic	association	`studies`	using
	N	prep	adj	prep	adj	N	N	v
31		damage.	In order	`to`	`study`	heterochromatin	maintenance	in
			prep N	prep	V	N	N	prep

28	gave	us	the	unique	opportunity	`to study`	an	
30	pro- vide	the	unique	opportunity	`to study`	the	pathogenesis	
122	test	menu	offered;	quality	assurance	`studies`;	engaging	
123	Since	currently	available	compar- ative	`studies`	between	CyA	

3.4 Data Analysis

The analyses developed in 4 (p. 59) are both qualitative and quantitative in order to grasp the whole scenario. A qualitative analysis since the patterns and structures needed a careful semantic, morpho-syntactic, and linguistics study to individuate the relevant ones for the reading and writing skills of medical ESP learners. And a quantitative analysis in order to complement the qualitative analysis and back up the implementation of the methodology in specialized corpora and consequently recommend the results for ESP teaching material.

3.4.1 Frequency word list

The main task of this phase is to individuate the threshold that separates the most common and repeated words according to Zipf's Law (*The, is, a...*, etc.) and the non-repeated ones (*hapax legomena*), namely the specific and unique words of each article. However, throughout this stage, we spot other documents besides articles (case studies, editor letters, etc.) when identifying questions more frequently than usual for articles. Consequently, it was pertinent to filter these documents. After this second cleansing, the data were finally ready to work

with: we started by classifying the frequency of nouns, prepositions, adjectives, adverbs, etc. In order to find N-grams, we check these frequent words to determine which co-occur.

Thus, the frequency word list table is organized by rank (from the most frequent to the less frequent), the occurrence frequency in the middle column, and the right column with the list of words. (See table 3.1, (p. 55). Then, we select the most frequent words related to the top rank and we gather relevant patterns and structures common among our analyzed sample and therefore valuable to teach potential learners to cover the majority of their needs both in reading and writing.

Once we have classified them, we can check the surrounding of the most frequent words for n-grams for the most frequent syntactic patterns. Therefore, we have a clear and structured data organization built around the patterns obtained from the first phase.

Rank	**Freq**	**Words**
1	92339	the
2	73898	of
3	63991	and
4	47117	in
5	41688	a
6	36481	to
7	23346	for
8	18138	with
9	15659	s
10	14813	e
11	12487	is
12	11983	that
13	11839	were
14	11346	as

Table 3.1: Sample of Frequent words without cleansing

3.4.2 Statistic Data

In this stage, we found the need to run one by one the articles in Antconc to obtain tokens and types of each of them and divide the results by range of tokens in order to reinforce our statistic data. In addition, and to obtain the maximum, minimum, mode, mean and medium we extract the n-grams occurrences in each article. Since, the n-gram tables displayed in Antcon illustrates the range, frequency but in our analysis we also add the syntactic

structure. However, and as we have mentioned one missing data is the individual frequency which offers a glance of how the frequency is distributed through each article.

AntConc has the Concordance Plot Key that presents these data but are not exportable to excel so to obtain the number of repetitions in each article we use the Concordance Key that also displays the files name and sort on the right side `File`(See image 3.3, 45) which will give us the data to export in `.txt` and then to excel to filter the times the n-gram appears in each file to finally obtain the number of repetitions.

3.4.3 Nouns selection

Furthermore, after filtering the nouns from both collections, we select the top 10 Nouns of each list plus, their plurals or singular will be deeply analyzed. Thus, we focus on the words that surround them as we encounter the N + N structure. The structure N+N is usual in English but not in Spanish. Consequently, it might result in more mistakes when writing or reading. To find these patterns, we check the right and left context of the word since this is not a tagged "corpus" that would also show the grammatical category.

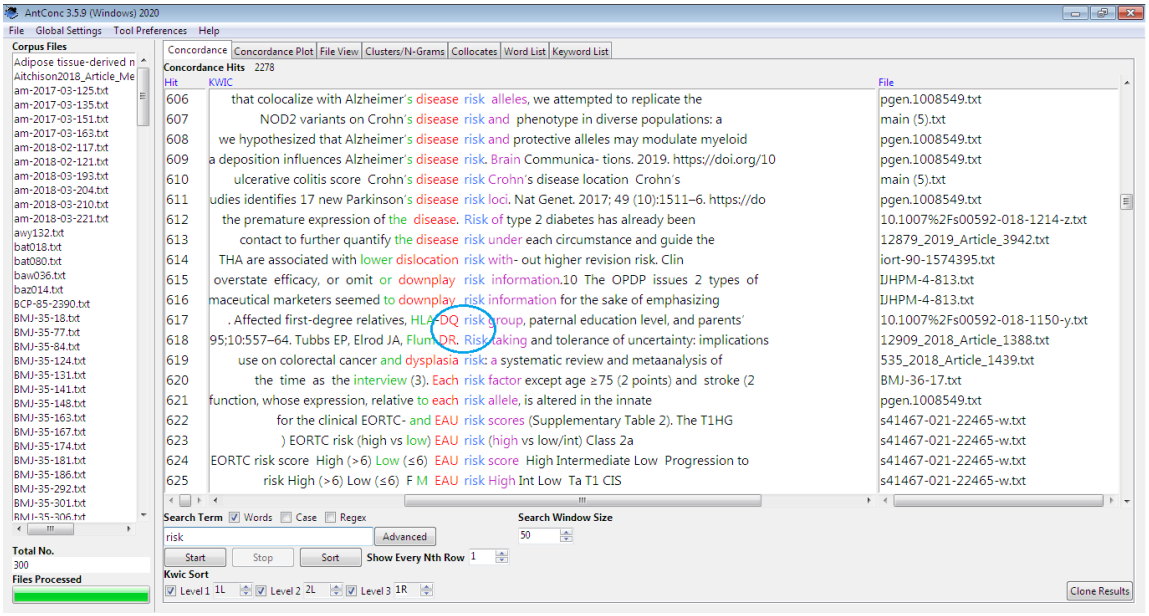

Figure 3.11: Analysis: Punctuation Marks

3.5 Broadening

Even if there is a slight deviation from the initial sample, the broadening will show the non-relevance of the variation and the resulting effectiveness of the founded patterns. The third, and last phase, consists of processing a new compilation with a broader number of scientific articles to analyze the relevant patterns that we already found to prove the existence of relevant syntactic patterns. This sample contains 600 journal articles, so the total number of scientific journal articles is 900. Since the patterns are mainly dependent on the topics (in this case, medical research), the variation encounter is limited between the Zipf's Law repetitions and the *hapax legomena* (See subsection 2.2.2, p. 36).

However, the goal is to demonstrate that this analysis, applied to a larger sample, shows similar results, therefore supporting reliable data, to recommend some guidelines, suggestions based on the analysis. For this reason, the same process of data treatment and preparation is repeated during the third phase; to broaden the data sample.

3.6 Statistical Reliability

Throughout this phase, we replicate the process of the first stage on a broader sample, 600 new articles. Due to the repetition and already proven application of Zipf's Law, the middle section of *any* sample big enough to be representative will not show relevant difference even after a great incrementation of the collection itself. In other words, the relevant part of syntactic and lexical patterns, i.e. the middle part of the frequency list in the collections is our focus of interest. As per Zipf's Law (see 2.2.1, p. 28), once the sample is big enough, the rates and proportion of the frequency ot the text words would not change. Since we find the same patterns with very similar frequency in both collections of 300 and 600 articles, it is safe to assume that, since the frequency and relevance didn't change from 300 to 600 articles sample, there would be no relevant change in patterns even if we keep increasing the sample. It is important to notice that the relevance of the patterns is to be calculates among the patterns itself, not among the whole number of words. As per Zipf's law, as we saw, the most frequent words in any text are, and always will be, the same words: verb to be, prepositions, articles. The relevance of medical patterns is therefore to be calculates, we stress it once again, as relevant among the patterns in the collections, not among the simple number of repeated words. This allows us to say which relevant patterns to teach to our medical ESP students. Hence, during this phase, we compare both data of frequency to

back up the existence of relevant syntactic patterns in both collections.

Chapter 4

Patterns in Scholarly Publications

4.1 General Analysis

In order to explain the data analysis, we divided this section into different subtitles. First, we present the information in detail and the word frequency lists of both collections and then the statistical and linguistic analysis of the bi-grams, tri-grams, and 4-grams. Finally, the last section concerns the filtered frequency list of nouns and the top ten nouns analysis from both collections.

4.1.1 Collection Details

The collections are the 300 (hereafter Collection A) and the 600 (hereafter Collection B). The former has 3,329 pages in total, an average of 11 pages per article. The minimum is three pages in two articles and, the maximum is 83 pages in just one of them. The TTR is about 3.3%; hapax legomena 33,229 (almost 50%)[1] of 72,033 word types. The latter has 5,867 pages in total, an average of 9.7 pages per article. The minimum is two pages in one article, and the maximum is 26 pages in three. The TTR is about 2.85%; hapax legomena 53,248 (almost 50%) of 110,612. (See table 4.1, p. 60)

While all these data are important, the number of tokens in each article provides a complete panorama of our collections. As mentioned in the methodology, it is not available in AntConc, so we upload the articles one by one to gather the data. (See chapter 3, 41). This data is a reference point to distinguish the articles by length and the frequency of

[1] As explained in the subtitle 2.2.2 (p. 36) the percentage in a corpus is between 40% to 60%

Collection	A	B
Number of articles	300	600
Number of Pages	3,329	5,867
Word Tokens	2,171,873	3,873,219
Word Types	72,033	110,612
Hapax Legomena	33,229	53,248
TTR	3.3%	2.85%

Table 4.1: Collections A and B descriptive information

N-grams, as explained later in section 4.2.1, 74. To illustrate it, in the following bar charts, we observe the range of tokens in each article. For instance, over half of the articles in Collection A has between 1,588 to 6,500 tokens (See figure 4.1, 61)and Collection B has 396 articles from 2,501 to 7,502 tokens. (See figure 4.2, 62).

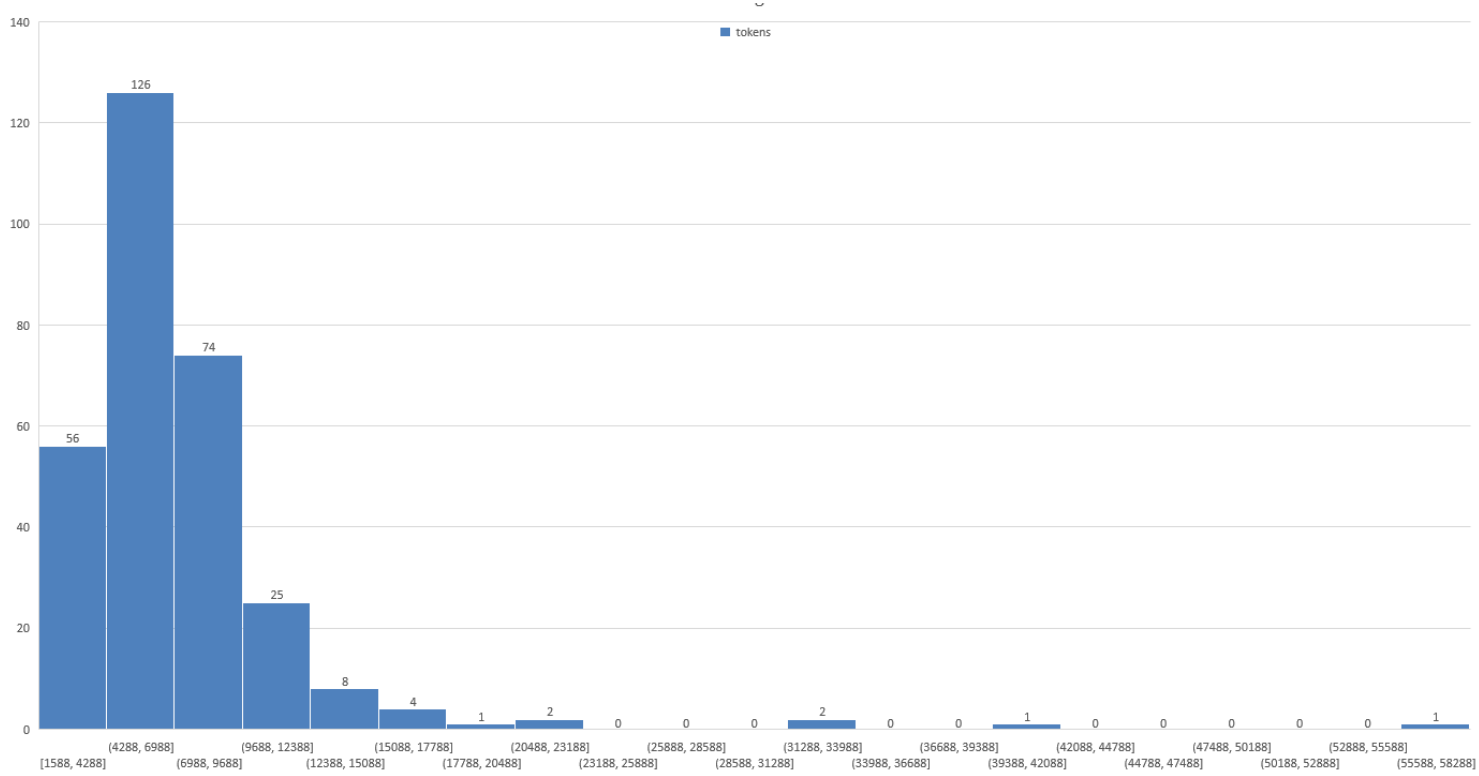

Figure 4.1: Tokens per article Collection A

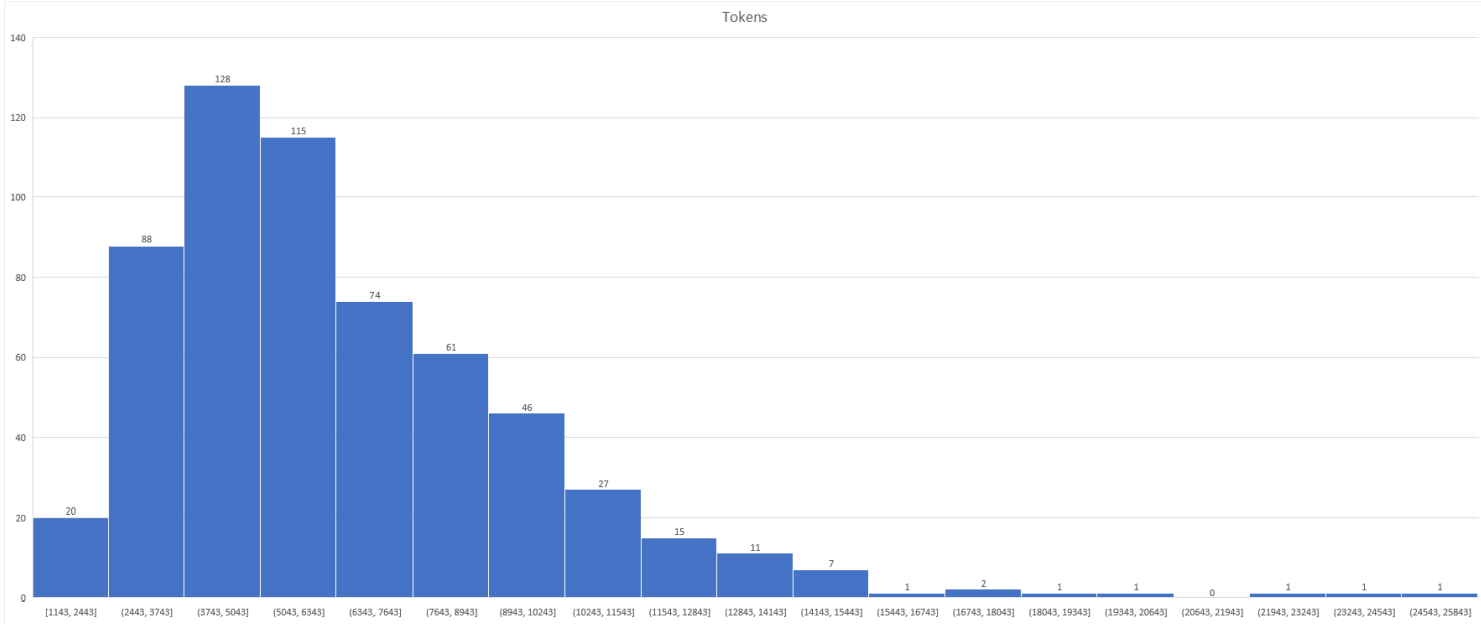

Figure 4.2: Tokens per article Collection B

4.1.2 Words Frequency List

After running our collections in AntConc, we have the following clean frequency lists[2] of the top 50 words. (See table 4.37, p. 104 and 4.38, p. 105). As explained in the methodology section (see 3, p. 41), we excluded acronyms or single letters such as et, al, doi, vol, e, n, etc. Another important exclusion is regarding the letter I we did not include it in the list since the most frequent use was not of the pronoun but the Roman number, single letter, etc. We expected this data since the pronoun I barely appears in a journal article or/and academic English; on the contrary, the pronoun WE might be more frequent in academic writing.

According to the frequency lists, the top five words are identical in both collections and we confirm the most frequent words in English with a slight variation in the rank order.(See subsection 2.2.1, p. 28). There are prepositions, conjunctions, determiners, and linking verbs in the top rank of the lists. Nevertheless, the frequency doesn't follow Zipf's law when referring to $f.r=C$ (See section 2.2.1, p. 28)both collections have a Zipfian distribution tendency. For instance, in Collection A the first-word *the* repeats 92,339 times, the second-word *of* repeats 73,898 when it should be around half of the first but if we observe figure 4.4 p. 66, it illustrates the frequency overlapped to compare Collection A and B and figure 4.3, p. 65 in a percentage scale. Both figures present a Zipfian distribution although the tendency frequency/rank doesn't apply.

[2]However, both raw lists of the top 50 words are available in the appendix section for further consultation.

Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word
1	92,339	the	32	7,674	this	54	3,934	which
2	73,898	of	33	7,532	disease	55	3,661	between
3	63,991	and	34	6,996	be	57	3,583	these
4	47,117	in	35	6,726	patients	58	3,472	no
5	41,688	a	37	6,333	at	59	3,424	their
6	36,481	to	40	5,928	an	60	3,311	it
7	23,346	for	41	5,823	medical	61	3,267	analysis
8	18,138	with	42	5,554	we	63	3,070	more
11	12,487	is	43	5,498	study	64	3,070	using
12	11,983	that	44	5,305	not	66	2,970	also
13	11,839	were	47	4,929	data	67	2,952	one
14	11,346	as	48	4,672	pathology	68	2,923	based
15	11,293	was	49	4,314	have	69	2,923	our
18	9,611	by	51	4,050	health	70	2,890	other
19	9,415	on	52	4,032	clinical	71	2,811	table
24	8,505	or	53	4,014	all	74	2,728	may
29	8,169	are						

Table 4.2: Cleaned Word frequency Collection A

Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word
1	157,259	the	35	12,663	patients	60	5,755	analysis
2	138,652	of	36	11,829	are	61	5,545	these
3	118,783	and	37	11,683	this	62	5,537	all
4	91,598	in	38	11,576	no	65	5,214	treatment
5	73,638	a	39	11,098	cells	66	5,157	it
6	55,798	to	41	10,766	study	67	5,137	also
7	38,306	with	42	10,396	be	68	5,125	group
8	33,927	for	44	9,811	an	69	5,017	used
11	25,592	was	45	9,385	cell	70	4,630	clinical
12	24,645	were	46	8,329	not	71	4,623	results
13	21,370	is	47	7,741	we	72	4,473	has
14	20,940	by	49	6,868	disease	73	4,456	may
17	19,166	as	51	6,757	which	74	4,453	been
20	18,746	that	53	6,553	between	75	4,436	protein
28	14,524	from	54	6,514	using	76	4,428	figure
29	14,479	on	55	6,377	have	77	4,303	studies
30	14,171	or	58	6,117	data	78	4,301	control
32	13,395	at	59	5,808	after	79	4,281	can

Table 4.3: Cleaned Word frequency Collection B

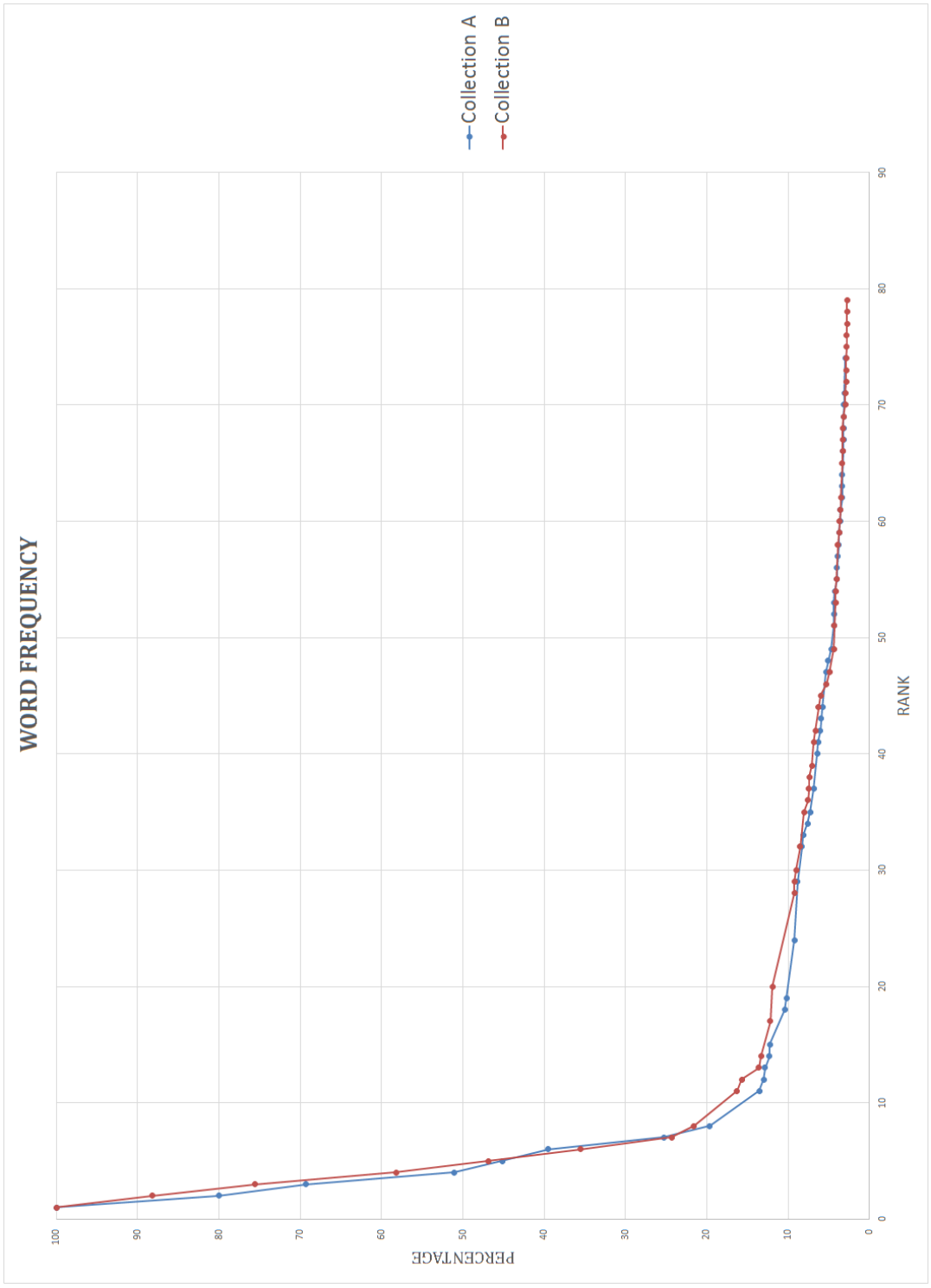

Figure 4.3: Word frequency both collections in percentage

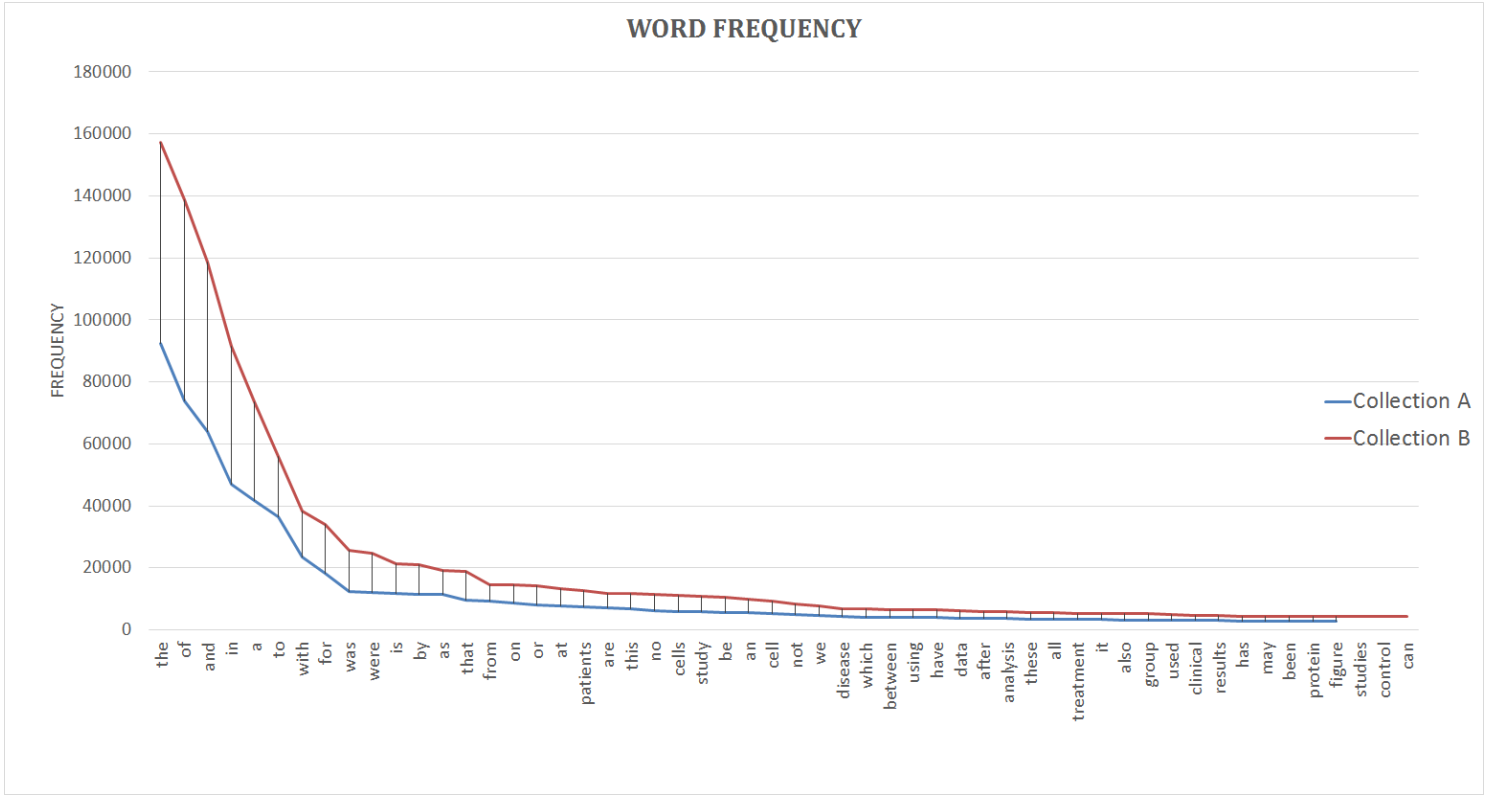

Figure 4.4: Word frequency both collections

4.2 N-grams

The search of the most frequent n-grams is based on the top words of the lists (See section 3, p. 41. Firstly, we have the most frequent bi-grams (prepositions and determiners) at the top of the list. For instance, *of the*, *for the*, and *in the*. (See tables 4.4, p. 68 for the group A most frequent bi-grams and 4.5, p. 69 for group B). The most frequent structures are N + PREP, ADJ + PREP, PREP + N, PREP + DET and V + PREP the frequency decreases when combining words since they have different ranks. (See tables 4.10, p. 72; 4.11, p. 72).

Furthermore, as expected tri-grams and 4-grams aren't many and as frequent as bi-grams. In addition and as explained in the subsection N-grams 2.2.2, p. 38, bi-grams are also part of the structures in tri-grams and 4-grams. For instance the most frequent patterns are N + PREP, PREP + N + PREP and PREP + DET + N + PREP in bi-grams, tri-grams and 4-grams respectively. (See table 4.6, p. 70; table 4.8, p. 71; 4.7, p. 71; table4.9, p. 72).

Freq	Range	Bi-grams	Tag	Freq	Range	Bi-grams	Tag
12632	300	of the	prep det	347	146	differences in	N prep
9944	299	in the	prep det	329	155	in table	prep N
3190	292	for the	prep det	328	122	prior to	adj prep
1786	280	in a	prep det	318	95	of human	prep N
1768	279	as a	prep det	302	108	out of	prep prep
1751	260	number of	N prep	279	139	a significant	det N
1480	227	associated with	adj prep	269	118	obtained from	V prep
1114	259	department of	N prep	257	125	defined as	V prep
1062	254	as the	prep det	256	115	data from	N prep
949	238	due to	adj prep	250	79	association between	N prep
917	128	in patients	prep N	244	119	over the	prep det
799	213	analysis of	N prep	243	133	in all	prep adj
670	131	risk of	N prep	240	97	relationship between	N prep
636	129	follow-up	V prep	238	141	supported by	V prep
609	210	in addition	prep N	214	120	up to	prep prep
550	183	at least	prep adj	210	109	used as	V prep
545	136	changes in	N prep	203	118	with respect	prep N
509	125	compared with	V prep	191	77	no significant	adv N
435	174	of all	prep adj	173	98	needed to	V prep
421	169	the following	det prep	169	62	in human	prep N
412	165	development of	N prep	160	67	in total	prep N
406	135	levels of	N prep	153	69	across the	prep det
376	142	in order	prep N	136	109	available at	adj prep
372	212	of interest	prep N	100	61	carried out	V prep
370	127	likely to	adj prep	100	61	of total	prep N
370	148	in both	prep adj	313	112	presence of	N prep
368	144	level of	N prep	793	203	use of	N prep
359	162	review of	N prep	419	142	role of	N prep

Table 4.4: Bi-grams (Collection A)

Freq	Range	Bi-grams	Tag	Freq	Range	Bi-grams	Tag
22527	600	of the	prep det	632	228	of human	prep N
18408	600	in the	prep det	598	270	in all	prep adj
4561	576	for the	prep det	592	244	no significant	adv N
3385	553	in a	prep det	586	242	in order	prep N
2856	541	as a	prep det	560	282	the following	det prep
2733	476	associated with	adj prep	546	207	prior to	adj prep
2064	459	number of	N prep	518	257	obtained from	V prep
2063	506	due to	adj prep	516	262	in table	prep N
2049	330	in patients	prep N	500	221	defined as	V prep
1706	480	as the	prep det	495	197	relationship between	N prep
1630	487	based on	V prep	468	251	of all	prep adj
1523	463	analysis of	N prep	464	218	used as	V prep
1458	471	department of	N prep	452	181	association between	N prep
1286	283	levels of	N prep	447	216	over the	prep det
1218	411	in addition	prep N	430	212	up to	prep prep
1198	276	risk of	N prep	411	293	supported by	V prep
1079	296	changes in	N prep	399	207	carried out	V prep
1021	291	treatment of	N prep	331	148	likely to	adj prep
1141	207	follow-up	V prep	308	163	data from	N prep
968	275	compared with	V prep	288	126	out of	prep prep
810	297	level of	N prep	249	122	across the	prep det
796	399	of interest	prep N	231	166	needed to	V prep
774	308	development of	N prep	190	115	available at	adj prep
708	284	a significant	det N	155	98	with respect	prep N
691	293	at least	prep adj	145	84	of total	prep N
675	192	in human	prep N	125	83	in total	prep N
662	277	differences in	N prep	1248	349	presence of	N prep
642	262	in both	prep adj	1143	390	use of	N prep
				1118	343	role of	N prep

Table 4.5: Bi-grams (Collection B)

Freq	Range	3-Grams	Tag
763	202	as well as	prep adv prep
686	177	the number of	det N prep
551	99	in patients with	prep N prep
396	149	in this study	prep det N
376	142	the use of	det N prep
358	136	in order to	prep N prep
313	120	a total of	det N prep
246	114	the development of	det N prep
243	93	in terms of	prep N prep
231	107	in the study	prep det N
217	115	*of this article	prep det N
204	120	in addition to	prep N prep
204	69	the present study	det adj N
203	121	was used to	V part prep
254	67	of patients with	prep N prep
200	118	with respect to	prep N prep
200	76	the risk of	det N prep
183	83	in our study	prep poss N
172	69	more likely to	adj adv prep
160	79	at the time	prep det N
156	46	in response to	prep N prep
126	72	a number of	det N prep
109	78	review of the	N prep det
108	51	the association between	det N prep
100	65	as part of	prep N prep
86	55	data from the	N prep det
83	81	for the research	prep det N

Table 4.6: Tri-grams (Collection A)

Freq	Range	tri-grams	Tag
1510	290	in patients with	prep N prep
1167	392	as well as	prep adv prep
1042	316	The presence of	det N prep
960	345	in this study	prep det N
794	273	the number of	det N prep
627	273	the use of	det N prep
571	238	in order to	prep N prep
468	261	of this study	prep det N
447	177	the risk of	det N prep
352	131	in the presence	prep det N
341	198	in addition to	prep N prep
323	161	in terms of	prep N prep
295	145	in the present	prep det N
205	111	at the time	prep det N
193	124	as a result	prep det N
188	143	in accordance with	prep N prep
187	154	with regard to	prep v prep
174	128	of the most	prep det adv
585	240	the effect of	det N prep
459	220	the development of	det N prep
190	100	in relation to	prep N prep

Table 4.7: Tri-grams (Collection B)

Freq	Range	4-grams	Tag
137	68	at the time of	prep det N prep
105	60	on the other hand	prep det adj N
102	88	with respect to the	prep N prep det
82	40	in the present study	prep det adj N
56	42	as a result of	prep det N prep
49	24	in the presence of	prep det N prep
40	31	in the form of	prep det N prep
31	25	of the number of	prep det N prep
22	22	the development of a	det N prep det
16	12	in patients with a	prep N prep det

Table 4.8: 4-grams (Collection A)

Freq	Range	4-grams	Tag
317	125	in the presence of	prep det N prep
274	153	on the other hand	prep det adj N
259	96	in the absence of	prep det N prep
204	113	for the treatment of	prep det N prep
160	133	as well as the	prep adv prep det
121	86	as a result of	prep det N prep
97	97	as long as you	prep adj prep pron

Table 4.9: 4-grams (Collection B)

Bi-grams	Freq
N prep	13
prep N	10
adj prep	5
V prep	8
prep det	8
det prep	1
adv N	1
det N	1
prep prep	2
prep adj	4

Table 4.10: Syntactic structures Bi-grams (Collection A)

Bi-grams	Freq
N prep	13
prep N	10
adj prep	5
V prep	9
det prep	1
adv N	1
det N	1
prep prep	2
prep det	8
prep adj	4

Table 4.11: Syntactic structures Bi-grams Collection B

Tri-grams	Freq
det N prep	7
prep det N	5
det adj N	1
N prep det	2
prep N prep	8
prep adv prep	1
prep poss N	1
adj adv prep	1
v prep	1

Table 4.12: Syntactic structures Tri-grams Collection A

Tri-grams	Freq
det N prep	6
prep det N	6
prep N prep	6
prep adv prep	1
prep det adv	1
prep v prep	1

Table 4.13: Syntactic structures Tri-grams Collection B

Four-grams	Freq
prep det N prep	5
prep det adj N	2
det N prep det	1
prep N prep det	2

Table 4.14: Syntactic structures 4-grams Collection A

Four-grams	Freq
prep det N prep	4
prep adj prep pron	1
prep adv prep det	1
prep det adj N	1

Table 4.15: Syntactic structures 4-grams Collection B

4.2.1 Statistical Analysis

We decided to organize data by the number of tokens since the average obtained with the general frequency and rank wouldn't show a precise frequency distribution. We export the n-grams data from AntConc and filter it by the article's name to later organize them to obtain the maximum, minimum, mode, median, and mean to include in the analysis.

Hence, the data is distributed as follows: Collection A is divide into two tables composed by 165 and 118 articles, while Collection B has three tables with 214, 182, and 109 articles respectively. Thus, the following tables in collection A 4.16, p. 76; 4.18, p. 78; 4.25, p. 93; 4.29, p. 97; 4.30, p. 98; 4.31, p. 98).

Also in Collection B 4.19, p. 87 and 4.20, p. 88; 4.22, p. 90; 4.24, p. 92; 4.26, p. 94; 4.27, p. 95; 4.28, p. 96; 4.32, p. 98; 4.33, p. 99; 4.34, p. 99.

In addition to tables we illustrate the data about bi-grams and tri-grams with plot box graphics, we divided bi-grams in two sections. See figures 4.2.1 p. 79, 4.5 p. 79, 4.2.1 p. 81, 4.2.1 80, 4.6 p. 80, 4.7 p. 81, 4.2.1 p. 82, 4.2.1 p. 84, 4.2.1 p. 83 , 4.9 p. 83, 4.10 p. 84, 4.11 p. 85, 4.12 p. 85, 4.13 p. 86.

At first glance, the data obtained doesn't show statistic relevancy *per se* of the structures, if considered among the whole collection data; nevertheless, this should not fool ourselves into thinking the retrieved patterns do not have relevance from a semantic point of view. As a matter of fact both from a semantic and a syntactic structures point of view our data are relevant as we are about to explain.

Our data show that mean is not as high as the top bi-grams composed by prepositions and determiners, but it shows consistency in both collections. We anticipated the results since prepositions are more frequent than other words and this will decrease when combined with less frequent words. It is important to notice again that almost half of our collection types are `hapax legomenon` and the most frequent words are prepositions, conjunctions and determiners which means that other type of words will be less frequent. In case of the tri-grams and four-grams the mean, mode and median is zero because the range is less than half of the total and they are not as frequent as bi-grams as mentioned.

The relevance of the data is not therefore to be calculated among the whole collection, including *hapax legomena* (the bottom part of the data) and preposition alike (the top part of the data); the most important, key part of the study is that those retrieved patter are relevant *among semantic and syntactic patterns*, so they have to be relevant among the middle "slice" of the ranking, not among the whole collection. Due to what we know about

Zipf's Law and distribution of frequency in the natural languages (see theoretical framework at the beginning of the present work, 2, p. 12) the most frequent words in *any* language and *any subcollection* of language are, and always will be, prepositions and functional words. This is due to the nature and characteristics of natural languages. This does not mean that there are not relevant pattern in the languages; it just mean that in order to discover the relevant pattern of a language (or a subcollection of it) we have to investigate that part of the text that refer to the content, namely the middle part of the frequency distribution, therefore setting apart the top and bottom of our lists.

4.2.2 Linguistic Analysis

As explained in the previous section, even though the occurrence of syntactic patterns might not appear to be statistically relevant when comparing the absolute numbers of occurrence against the total number of Token in the collection, it does have relevancy as syntactic structure because it follows a pattern in both collections such as N + PREP, PREP + N, PREP + DET bi-grams, DET + N + PREP, PREP + N + PREP in tri-grams, PREP + DET + N + PREP in 4-grams. As we observe N + PREP repeats in all the n-grams. Due to the similarity in both word lists we find similar lexical items in the n-grams except for some n-grams.

In addition, representativeness (as discussed in subtitle 2.1.2 on p. 20) is a key part of our collection, since our collection is a specialized 'corpus' it is more possible to reach it in a specific field and gender such as our collections as mentioned before in 4.2.1 (p. 74) and in the design of the research (3, p. 41).

165 A		Max	Min	Mode	Med	Mean			Max	Min	Mode	Med	Mean
of the	prep det	65	5	18	27	29	in table	prep N	5	0	0	1	1
in the	prep det	60	0	24	22	24	prior to	adj prep	16	0	0	0	1
for the	prep det	28	0	3	6	7	of human	prep N	5	0	0	0	0
in a	prep det	16	0	1	4	4	out of	prep prep	16	0	0	0	1
as a	prep det	17	0	2	3	4	a significant	det N	8	0	0	0	1
number of	N prep	27	0	2	2	4	obtained from	V prep	8	0	0	0	1
associated with	adj prep	22	0	0	2	3	defined as	V prep	6	0	0	0	1
department of	N prep	12	0	2	3	3	data from	N prep	12	0	0	0	1
as the	prep det	36	0	0	2	3	association between	N prep	17	0	0	0	1
due to	adj prep	28	0	1	2	2	over the	prep det	9	0	0	0	1
in patients	prep N	22	0	0	0	3	in all	prep adj	6	0	0	0	1
analysis of	N prep	14	0	0	1	1	relationship between	N prep	18	0	0	0	1
risk of	N prep	32	0	0	0	2	supported by	V prep	2	0	0	0	0
follow up	V prep	36	0	0	0	2	up to	prep prep	7	0	0	0	1
in addition	prep N	8	0	0	1	1	used as	V prep	9	0	0	0	0
at least	prep adj	13	0	0	1	1	with respect	prep N	5	0	0	0	0
changes in	N prep	16	0	0	0	1	no significant	adv N	10	0	0	0	1
compared with	V prep	20	0	0	0	2	needed to	V prep	7	0	0	0	0
the following	det prep	7	0	0	0	1	in human	prep N	8	0	0	0	0
development of	N prep	15	0	0	1	1	in total	prep N	7	0	0	0	0
levels of	N prep	9	0	0	0	1	across the	prep det	8	0	0	0	0
in order	prep N	5	0	0	0	1	available at	adj prep	3	0	0	0	0
of interest	prep N	6	0	2	1	1	carried out	V prep	5	0	0	0	0
likely to	adj prep	16	0	0	0	1	of total	prep N	6	0	0	0	0
in both	prep adj	8	0	0	0	1	preference of	N prep	7	0	0	0	1
level of	N prep	15	0	0	0	1	use of	N prep	39	0	0	1	2
review of	N prep	10	0	0	1	1	role of	N prep	9	0	0	0	1
differences in	N prep	16	0	0	1	1	of all	prep adj	12	0	0	0	1

Table 4.16: Bi-grams statistics / Collection A / 165 articles

118 A		**Max**	**Min**	**Mode**	**Med**	**Mean**
of the	prep det	130	9	32	43	50
in the	prep det	106	10	29	34	38
for the	prep det	54	1	5	11	13
in a	prep det	23	0	8	6	7
as a	prep det	34	0	6	7	8
number of	N prep	31	0	2	6	8
associated with	adj prep	36	0	0	3	6
department of	N prep	21	0	2	2	4
as the	prep det	25	0	1	3,5	4
due to	adj prep	28	0	0	3	4
in patients	prep N	28	0	0	0	2
analysis of	N prep	18	0	0	2	3
risk of	N prep	37	0	0	0	2
follow up	V prep	28	0	0	0	2
in addition	prep N	14	0	1	2	3
at least	prep adj	17	0	0	1	2
changes in	N prep	32	0	0	0	2
compared with	V prep	25	0	0	0	2
the following	det prep	13	0	0	1	2
development of	N prep	12	0	0	1	2
levels of	N prep	12	0	0	0	1
in order	prep N	17	0	0	1	2
of interest	prep N	8	0	0	1	1
likely to	adj prep	16	0	0	1	2
in both	prep adj	29	0	0	1	2
level of	N prep	7	0	0	1	1
review of	N prep	6	0	0	0	1
differences in	N prep	9	0	0	0	1

Table 4.17: Bi-grams statistics / Collection A / 118 articles / First Part

118 A		**Max**	**Min**	**Mode**	**Med**	**Mean**
in table	prep N	12	0	0	1	1
prior to	adj prep	14	0	0	0	1
of human	prep N	16	0	0	0	2
out of	prep prep	20	0	0	0	1
a significant	det N	14	0	0	0	1
obtained from	V prep	9	0	0	0	1
defined as	V prep	9	0	0	0	1
data from	N prep	7	0	0	0	1
association between	N prep	14	0	0	0	1
over the	prep det	6	0	0	0	1
in all	prep adj	8	0	0	0	1
relationship between	N prep	12	0	0	0	1
supported by	V prep	15	0	0	1	1
up to	prep prep	7	0	0	0	1
used as	V prep	10	0	0	0	1
with respect	prep N	7	0	0	0	1
no significant	adv N	12	0	0	0	1
needed to	V prep	9	0	0	0	1
in human	prep N	12	0	0	0	1
in total	prep N	8	0	0	0	1
across the	prep det	7	0	0	0	1
available at	adj prep	5	0	0	1	1
carried out	V prep	4	0	0	0	0
of total	prep N	6	0	0	0	0
preference of	N prep	12	0	0	0	1
use of	N prep	18	0	0	2	3
role of	N prep	10	0	0	1	1
of all	prep adj	14	0	1	1	2

Table 4.18: Bi-grams statistics / Collection A / 118 articles / Second Part

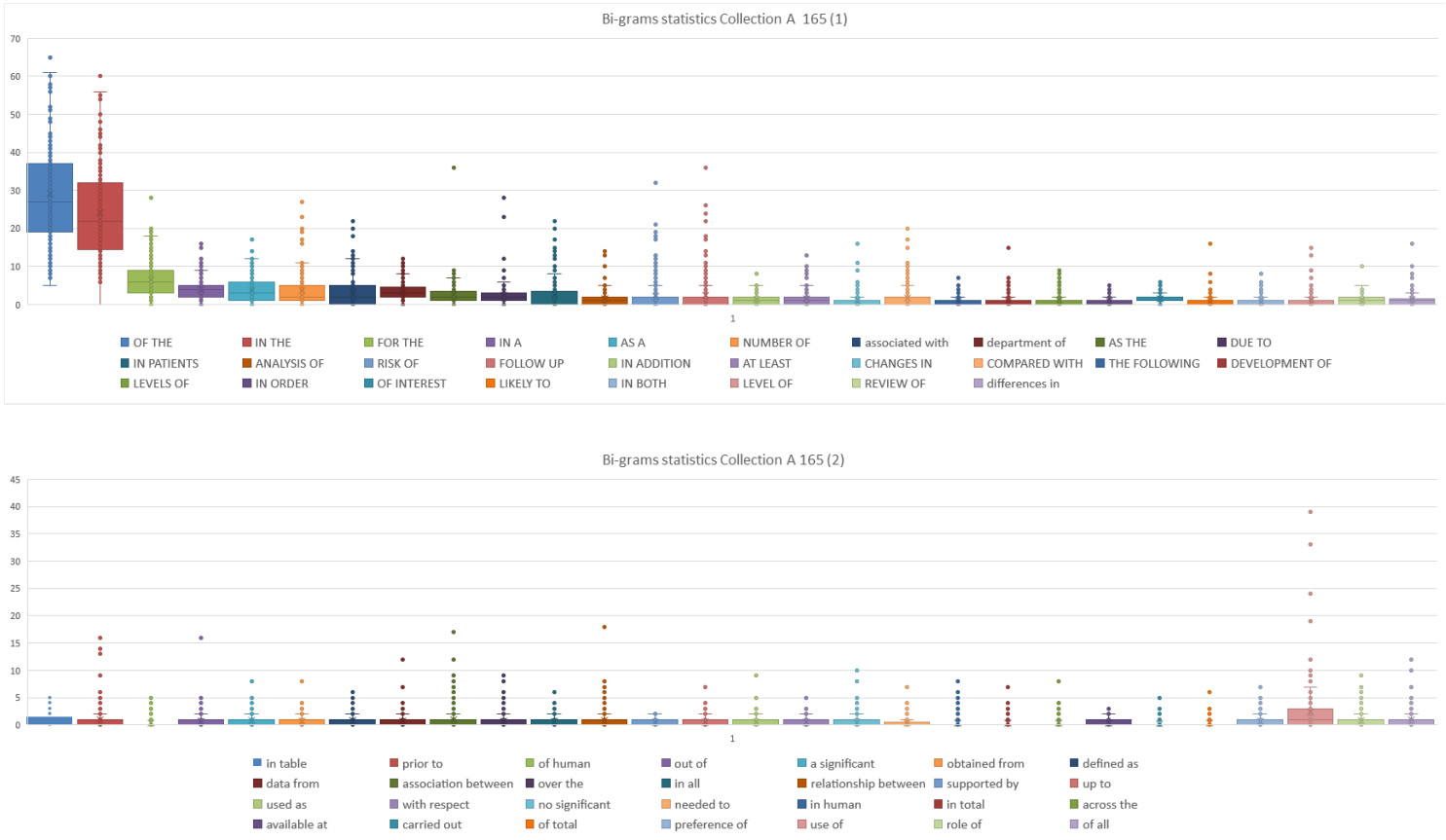

Figure 4.5: Box Plot Graphic Bi-grams Collection A (165) First Part (above). Box Plot Graphic Bi-grams Collection A (165)Second Part (below)

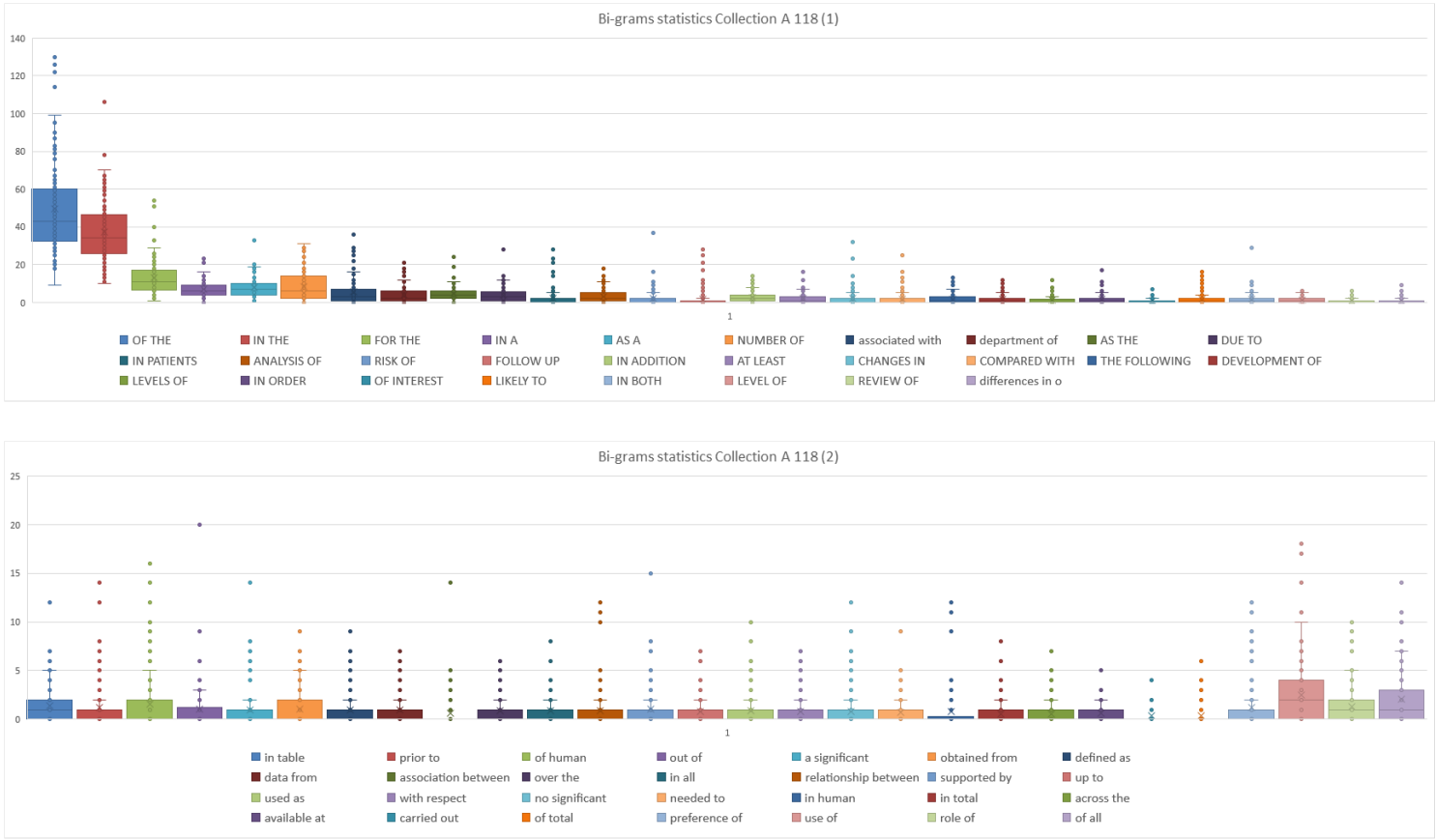

Figure 4.6: Box Plot Graphic Bi-grams Collection A (118)First Part (above). Box Plot Graphic Bi-grams Collection A (118) Second Part (below)

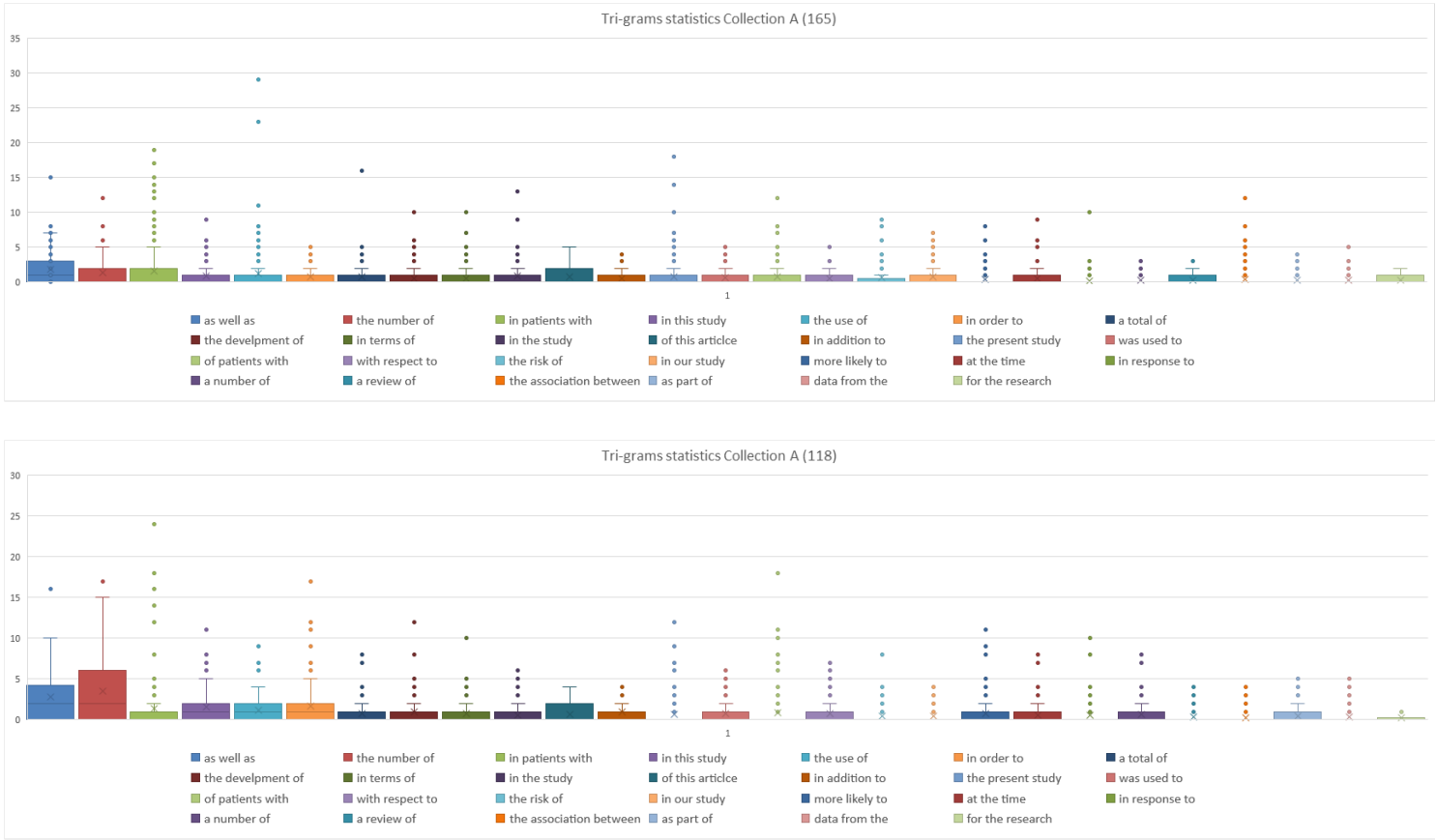

Figure 4.7: Box Plot Graphic Tri-grams Collection A (165) (above). Box Plot Graphic Tri-grams Collection A (118) (Below)

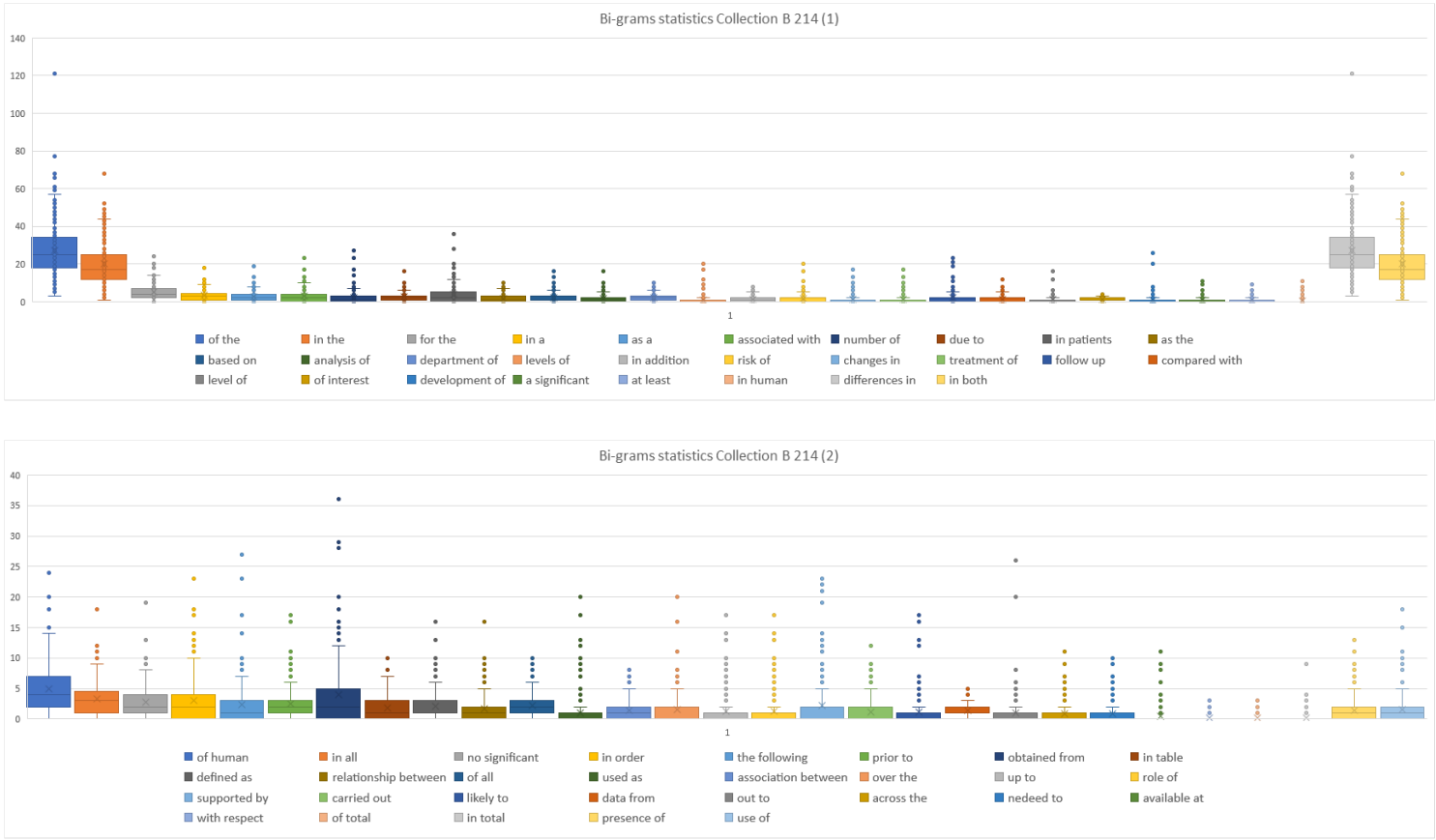

Figure 4.8: Box Plot Graphic Bi-grams Collection B (214)First part (above). Box Plot Graphic Bi-grams Collection B (214) Second part (below).

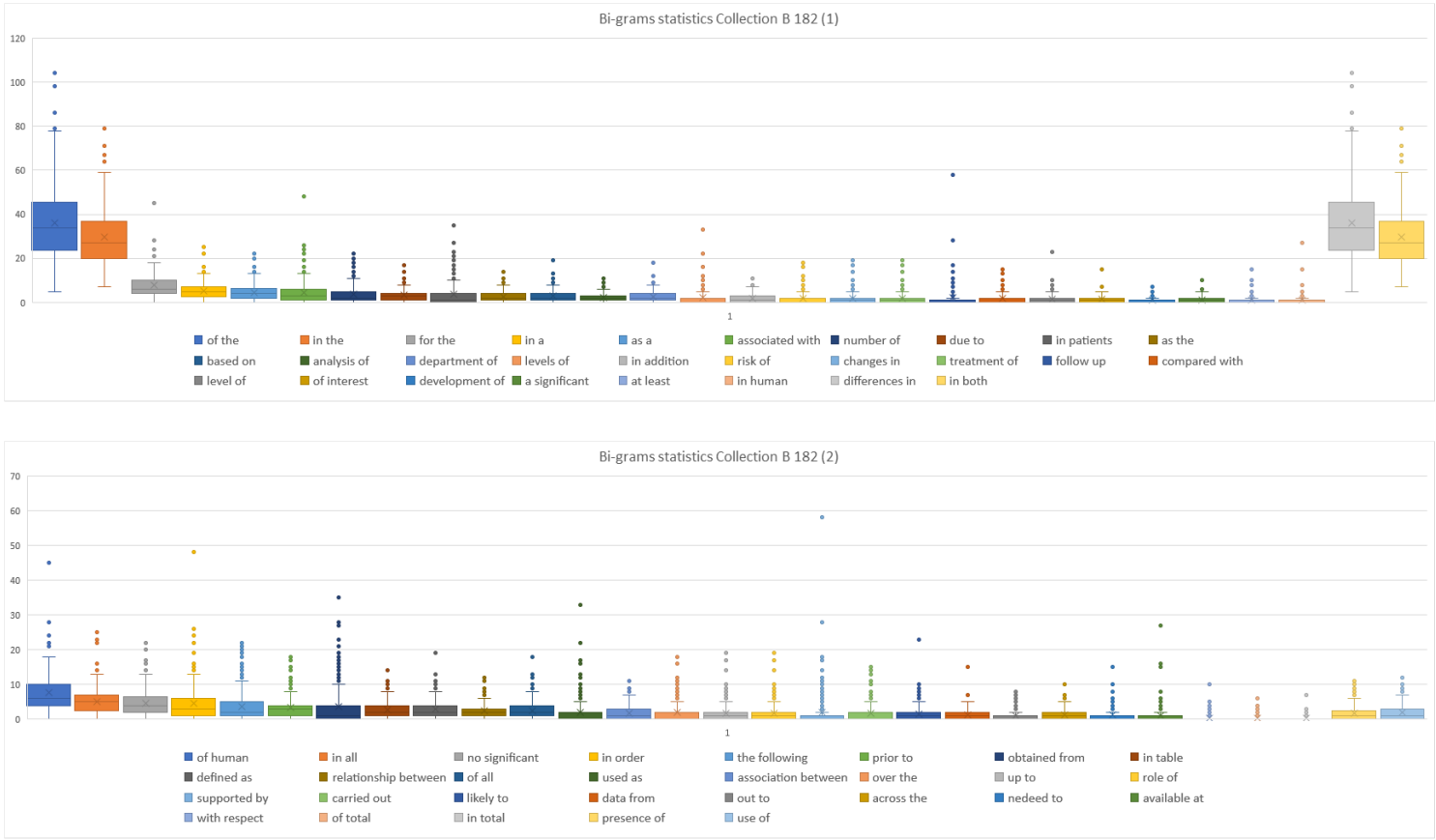

Figure 4.9: Box Plot Graphic Bi-grams Collection B (182)First part (above). Box Plot Graphic Bi-grams Collection B (182)Second part (below)

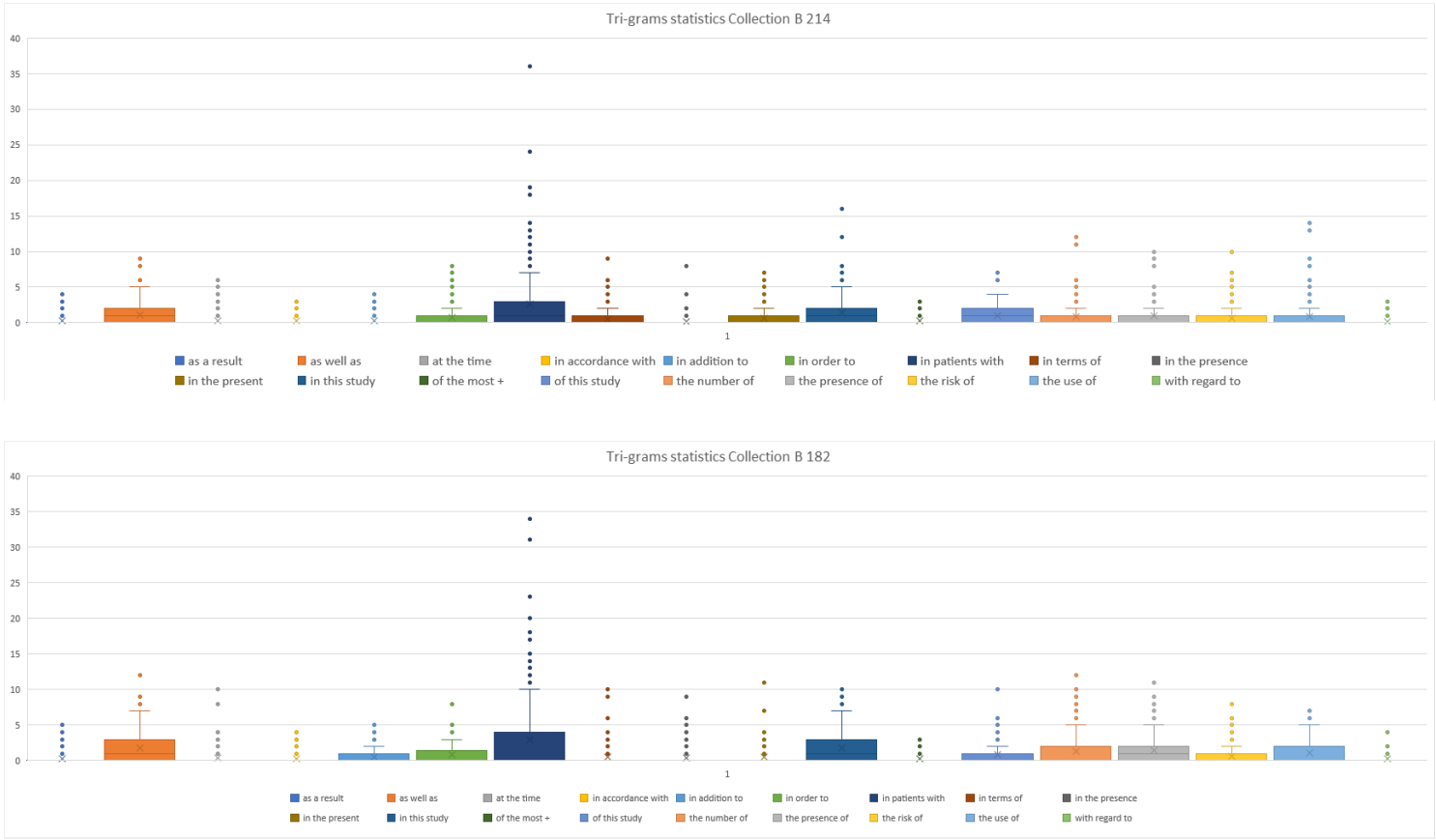

Figure 4.10: Box Plot Graphic Tri-grams Collection B (214) (above). Box Plot Graphic Tri-grams Collection B (182) (below).

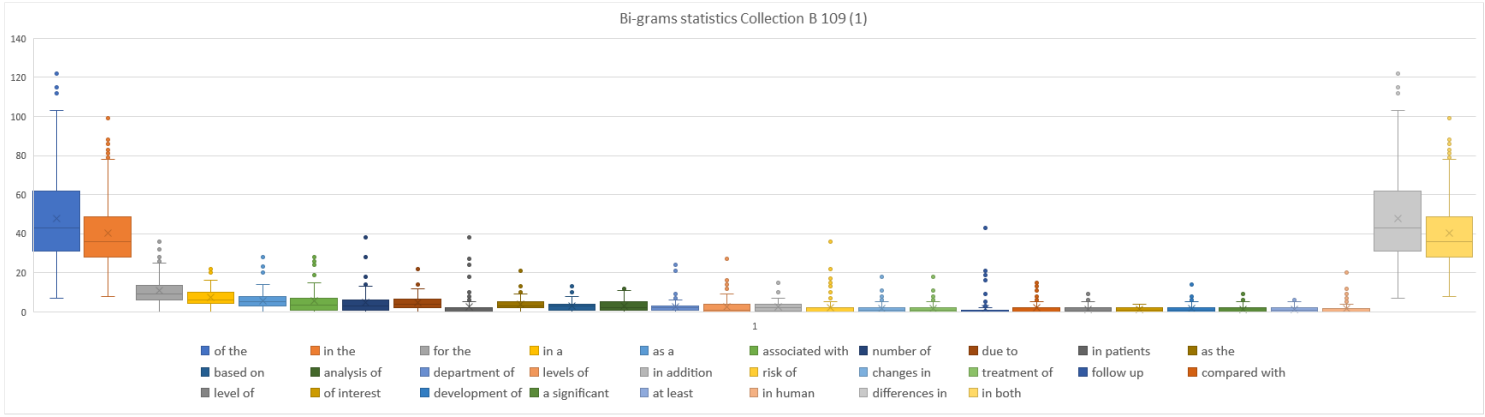

Figure 4.11: Box Plot Graphic Bi-grams Collection B (109)First Part

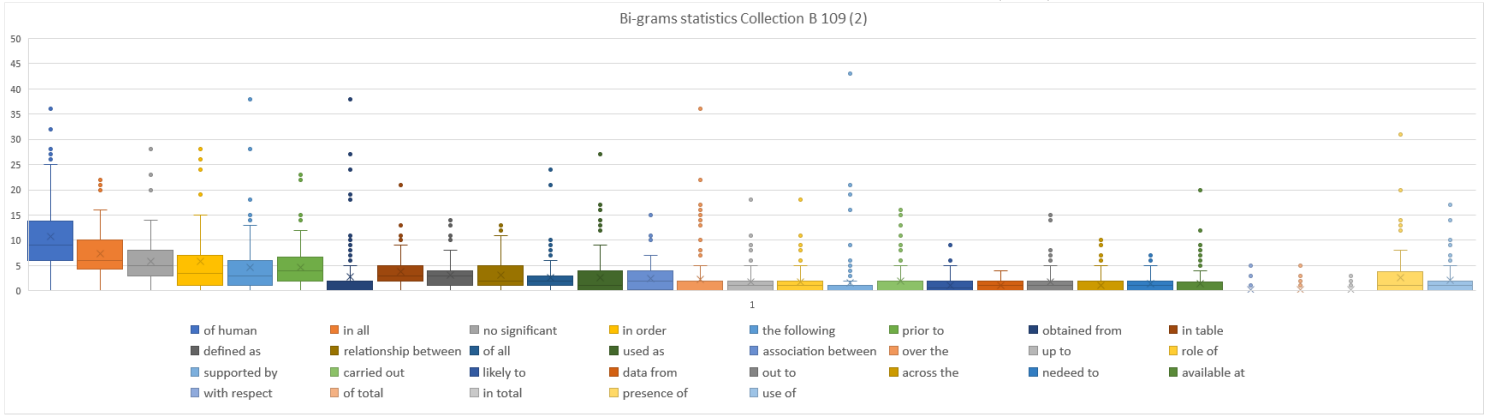

Figure 4.12: Box Plot Graphic Bi-grams Collection B (109)First Part (above). Box Plot Graphic Bi-grams Collection B (109) Second Part (below)

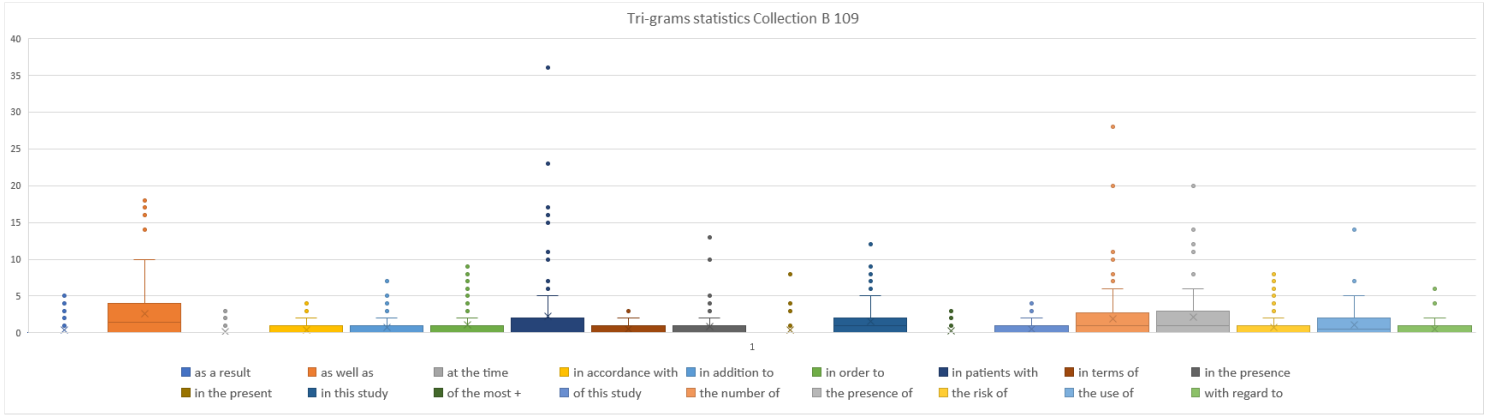

Figure 4.13: Box Plot Graphic Tri-grams Collection B (109)

214 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
of the	prep det	121	3	26	25	27
in the	prep det	68	1	16	17	20
for the	prep det	24	0	4	4	5
in a	prep det	18	0	2	3	3
as a	prep det	19	0	1	2	3
associated with	adj prep	23	0	0	2	3
number of	N prep	27	0	0	1	2
due to	adj prep	17	0	1	2	3
in patients	prep N	36	0	0	2	4
as the	prep det	10	0	0	1	2
based on	V prep	16	0	1	1	2
analysis of	N prep	16	0	1	1	2
department of	N prep	10	0	2	2	2
levels of	N prep	20	0	0	0	1
in addition	prep N	8	0	0	1	1
risk of	N prep	20	0	0	0	2
changes in	N prep	17	0	0	0	1
treatment of	N prep	17	0	0	0	1
follow up	V prep	23	0	0	0	2
compared with	V prep	12	0	0	0	1
level of	N prep	17	0	0	0	1
of interest	prep N	5	0	2	2	1
development of	N prep	26	0	0	0	1
a significant	det N	11	0	0	0	1
at least	prep adj	10	0	0	0	1
in human	prep N	11	0	0	0	0
differences in	N prep	121	3	26	25	27
in both	prep adj	68	1	16	17	20
of human	prep N	24	0	4	4	5
in all	prep adj	18	0	2	3	3
no significant	adv N	19	0	1	2	3
in order	prep N	23	0	0	2	3
the following	det prep	27	0	0	1	2
prior to	adj prep	17	0	1	2	3
obtained from	V prep	36	0	0	2	4
in table	prep N	10	0	0	1	2

Table 4.19: Bi-grams statistics / Collection B / 214 articles

214 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
defined as	V prep	16	0	1	1	2
relationship between	N prep	16	0	1	1	2
of all	prep adj	10	0	2	2	2
used as	V prep	20	0	0	0	1
association between	N prep	8	0	0	1	1
over the	prep det	20	0	0	0	2
up to	prep prep	17	0	0	0	1
role of	N prep	17	0	0	0	1
supported by	V prep	23	0	0	0	2
carried out	V prep	12	0	0	0	1
likely to	adj prep	17	0	0	0	1
data from	N prep	5	0	2	2	1
out to	prep prep	26	0	0	0	1
across the	prep det	11	0	0	0	1
nedeed to	V prep	10	0	0	0	1
available at	adj prep	11	0	0	0	0
with respect	prep N	3	0	0	0	0
of total	prep N	3	0	0	0	0
in total	prep N	9	0	0	0	0
presence of	N prep	13	0	0	1	1
use of	N prep	18	0	0	1	2

Table 4.20: Bi-grams statistics / Collection B / 214 articles

182 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
of the	prep det	104	5	35	34	36
in the	prep det	79	7	24	27	30
for the	prep det	45	0	3	6	8
in a	prep det	25	0	2	5	5
as a	prep det	22	0	2	4	5
associated with	adj prep	48	0	1	3	5
number of	N prep	22	0	0	2	4
due to	adj prep	18	0	1	3	3
in patients	prep N	35	0	0	1	4
as the	prep det	14	0	1	2	3
based on	V prep	19	0	1	2	3
analysis of	N prep	12	0	1	2	2
department of	N prep	18	0	1	2	3
levels of	N prep	33	0	0	0	2
in addition	prep N	11	0	0	1	2
risk of	N prep	18	0	0	0	2
changes in	N prep	19	0	0	1	2
treatment of	N prep	19	0	0	1	2
follow up	V prep	58	0	0	0	2
compared with	V prep	15	0	0	0	2
level of	N prep	23	0	0	1	2
of interest	prep N	15	0	0	1	1
development of	N prep	8	0	0	0	1
a significant	det N	10	0	0	1	1
at least	prep adj	15	0	0	0	1
in human	prep N	27	0	0	0	1
differences in	N prep	104	5	35	34	36
in both	prep adj	79	7	24	27	30
of human	prep N	45	0	3	6	8
in all	prep adj	25	0	2	5	5
no significant	adv N	22	0	2	4	5
in order	prep N	48	0	1	3	5
the following	det prep	22	0	0	2	4
prior to	adj prep	18	0	1	3	3
obtained from	V prep	35	0	0	1	4
in table	prep N	14	0	1	2	3

Table 4.21: Bi-grams statistics / Collection B / 182 articles

182 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
defined as	V prep	19	0	1	2	3
relationship between	N prep	12	0	1	2	2
of all	prep adj	18	0	1	2	3
used as	V prep	33	0	0	0	2
association between	N prep	11	0	0	1	2
over the	prep det	18	0	0	0	2
up to	prep prep	19	0	0	1	2
role of	N prep	19	0	0	1	2
supported by	V prep	58	0	0	0	2
carried out	V prep	15	0	0	0	2
likely to	adj prep	23	0	0	1	2
data from	N prep	15	0	0	1	1
out to	prep prep	8	0	0	0	1
across the	prep det	10	0	0	1	1
nedeed to	V prep	15	0	0	0	1
available at	adj prep	27	0	0	0	1
with respect	prep N	10	0	0	0	0
of total	prep N	6	0	0	0	0
in total	prep N	7	0	0	0	0
presence of	N prep	11	0	0	1	2
use of	N prep	12	0	0	1	2

Table 4.22: Bi-grams statistics / Collection B / 182 articles

109 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
of the	prep det	122	7	46	43	48
in the	prep det	99	8	29	36	40
for the	prep det	36	0	8	9	11
in a	prep det	22	0	5	6	7
as a	prep det	28	0	4	5	6
associated with	adj prep	28	0	0	3,5	6
number of	N prep	38	0	2	3	5
due to	adj prep	23	0	4	4	5
in patients	prep N	38	0	0	0	3
as the	prep det	21	0	2	3	4
based on	V prep	14	0	2	3	3
analysis of	N prep	13	0	1	2	3
department of	N prep	24	0	1	2	3
levels of	N prep	27	0	0	1	3
in addition	prep N	15	0	0	2	2
risk of	N prep	36	0	0	0	2
changes in	N prep	18	0	0	1	2
treatment of	N prep	18	0	0	1	2
follow up	V prep	43	0	0	0	2
compared with	V prep	16	0	0	0	2
level of	N prep	9	0	0	0,5	1
of interest	prep N	4	0	0	1	1
development of	N prep	15	0	0	1	2
a significant	det N	10	0	0	0	1
at least	prep adj	7	0	0	1	1
in human	prep N	20	0	0	0	2
differences in	N prep	122	7	46	43	48
in both	prep adj	99	8	29	36	40
of human	prep N	36	0	8	9	11
in all	prep adj	22	0	5	6	7
no significant	adv N	28	0	4	5	6
in order	prep N	28	0	0	3,5	6
the following	det prep	38	0	2	3	5
prior to	adj prep	23	0	4	4	5
obtained from	V prep	38	0	0	0	3
in table	prep N	21	0	2	3	4

Table 4.23: Bi-grams statistics / Collection B / 109 articles

109 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
defined as	V prep	14	0	2	3	3
relationship between	N prep	13	0	1	2	3
of all	prep adj	24	0	1	2	3
used as	V prep	27	0	0	1	3
association between	N prep	15	0	0	2	2
over the	prep det	36	0	0	0	2
up to	prep prep	18	0	0	1	2
role of	N prep	18	0	0	1	2
supported by	V prep	43	0	0	0	2
carried out	V prep	16	0	0	0	2
likely to	adj prep	9	0	0	0,5	1
data from	N prep	4	0	0	1	1
out to	prep prep	15	0	0	1	2
across the	prep det	10	0	0	0	1
nedeed to	V prep	7	0	0	1	1
available at	adj prep	20	0	0	0	2
with respect	prep N	5	0	0	0	0
of total	prep N	5	0	0	0	0
in total	prep N	3	0	0	0	0
presence of	N prep	31	0	0	1	3
use of	N prep	17	0	1	1	2

Table 4.24: Bi-grams statistics / Collection B / 109 articles

165 A		**Max**	**Min**	**Mode**	**Med**	**Mean**
as well as	prep adv prep	15	0	0	1	2
the number of	det N prep	12	0	0	0	1
in patients with	prep N prep	19	0	0	0	2
in this study	prep det N	9	0	0	0	1
the use of	det N prep	29	0	0	0	1
in order to	prep N prep	5	0	0	0	1
a total of	det N prep	16	0	0	0	1
the develpment of	det N prep	10	0	0	0	1
in terms of	prep N prep	10	0	0	0	1
in the study	prep det N	13	0	0	0	1
of this articlce	prep det N	5	0	0	0	1
in addition to	prep N prep	4	0	0	0	0
the present study	det adj N	18	0	0	0	1
was used to	V part prep	5	0	0	0	1
of patients with	prep N prep	12	0	0	0	1
with respect to	prep N prep	5	0	0	0	0
the risk of	det N prep	9	0	0	0	1
in our study	prep poss N	7	0	0	0	1
more likely to	adj adv prep	8	0	0	0	0
at the time	prep det N	9	0	0	0	1
in response to	prep N prep	10	0	0	0	0
a number of	det N prep	3	0	0	0	0
a review of	N prep det	3	0	0	0	0
the association between	det N prep	12	0	0	0	0
as part of	prep N prep	4	0	0	0	0
data from the	N prep det	5	0	0	0	0
for the research	prep det N	2	0	0	0	0

Table 4.25: Tri-grams statistics /Collection A/165 articles

214 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
as a result	prep det N	4	0	0	0	0
as well as	prep adv prep	9	0	0	1	1
at the time	prep det N	6	0	0	0	0
in accordance with	prep N prep	3	0	0	0	0
in addition to	prep N prep	4	0	0	0	0
in order to	prep N prep	8	0	0	0	1
in patients with	prep N prep	36	0	0	1	3
in terms of	prep N prep	9	0	0	0	1
in the presence	prep det N	8	0	0	0	0
in the present	prep det N	7	0	0	0	1
in this study	prep det N	16	0	0	1	1
of the most +	prep det adv	3	0	0	0	0
of this study	prep det N	7	0	0	1	1
the number of	det N prep	12	0	0	0	1
the presence of	det N prep	10	0	0	0	1
the risk of	det N prep	10	0	0	0	1
the use of	det N prep	14	0	0	0	1
with regard to	prep v prep	3	0	0	0	0

Table 4.26: Tri-grams statistics /Collection B/214 articles

182 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
as a result	prep det N	5	0	0	0	0
as well as	prep adv prep	12	0	1	1	2
at the time	prep det N	10	0	0	0	0
in accordance with	prep N prep	4	0	0	0	0
in addition to	prep N prep	5	0	0	0	1
in order to	prep N prep	8	0	0	0	1
in patients with	prep N prep	34	0	0	0	3
in terms of	prep N prep	10	0	0	0	0
in the presence	prep det N	9	0	0	0	0
in the present	prep det N	11	0	0	0	0
in this study	prep det N	10	0	0	1	2
of the most +	prep det adv	3	0	0	0	0
of this study	prep det N	10	0	0	0	1
the number of	det N prep	12	0	0	0	1
the presence of	det N prep	11	0	0	1	1
the risk of	det N prep	8	0	0	0	1
the use of	det N prep	7	0	0	0	1
with regard to	prep v prep	4	0	0	0	0

Table 4.27: Tri-grams statistics /Collection B/182 articles

109 B		**Max**	**Min**	**Mode**	**Median**	**Mean**
as a result	prep det N	5	0	0	0	0
as well as	prep adv prep	18	0	0	1,5	3
at the time	prep det N	3	0	0	0	0
in accordance with	prep N prep	4	0	0	0	0
in addition to	prep N prep	7	0	0	0	1
in order to	prep N prep	9	0	0	0	1
in patients with	prep N prep	36	0	0	0	2
in terms of	prep N prep	3	0	0	0	1
in the presence	prep det N	13	0	0	0	1
in the present	prep det N	8	0	0	0	0
in this study	prep det N	12	0	0	1	2
of the most +	prep det adv	3	0	0	0	0
of this study	prep det N	4	0	0	0	1
the number of	det N prep	28	0	0	1	2
the presence of	det N prep	20	0	0	1	2
the risk of	det N prep	8	0	0	0	1
the use of	det N prep	14	0	0	1	1
with regard to	prep v prep	6	0	0	0	1

Table 4.28: Tri-grams statistics /Collection B/109 articles

118 A		**Max**	**Min**	**Mode**	**Median**	**Mean**
as well as	prep adv prep	16	0	0	2	3
the number of	det N prep	17	0	0	2	3
in patients with	prep N prep	24	0	0	0	1
in this study	prep det N	11	0	0	1	2
the use of	det N prep	9	0	0	1	1
in order to	prep N prep	17	0	0	1	2
a total of	det N prep	8	0	0	0	1
the develpment of	det N prep	12	0	0	0	1
in terms of	prep N prep	10	0	0	0	1
in the study	prep det N	6	0	0	0	1
of this articlce	prep det N	4	0	0	0	1
in addition to	prep N prep	4	0	0	1	1
the present study	det adj N	12	0	0	0	1
was used to	V part prep	6	0	0	0	1
of patients with	prep N prep	18	0	0	0	1
with respect to	prep N prep	7	0	0	0	1
the risk of	det N prep	8	0	0	0	0
in our study	prep poss N	4	0	0	0	0
more likely to	adj adv prep	11	0	0	0	1
at the time	prep det N	8	0	0	0	1
in response to	prep N prep	10	0	0	0	1
a number of	det N prep	8	0	0	0	1
a review of	N prep det	4	0	0	0	0
the association between	det N prep	4	0	0	0	0
as part of	prep N prep	5	0	0	0	0
data from the	N prep det	5	0	0	0	0
for the research	prep det N	1	0	0	0	0

Table 4.29: Tri-grams statistics /Collection A/118 articles

165 A		Max	Min	Mode	Med	Mean
as a result of	prep det N prep	3	0	0	0	0
at the time of	prep det N prep	9	0	0	0	0
in patients with a	prep N prep det	3	0	0	0	0
in the form of	prep det N prep	2	0	0	0	0
in the presence of	prep det N prep	2	0	0	0	0
in the present study	prep det adj N	7	0	0	0	0
of the number of	prep det N prep	2	0	0	0	0
on the other hand	prep det adj N	5	0	0	0	0
the development of a	det N prep det	1	0	0	0	0
with respect to the	prep N prep det	2	0	0	0	0

Table 4.30: 4-grams statistics /Collection A/165 articles

118 A		Max	Min	Mode	Med	Mean
as a result of	prep det N prep	3	0	0	0	0
at the time of	prep det N prep	7	0	0	0	0
in patients with a	prep N prep det	2	0	0	0	0
in the form of	prep det N prep	4	0	0	0	0
in the presence of	prep det N prep	4	0	0	0	0
in the present study	prep det adj N	6	0	0	0	0
of the number of	prep det N prep	2	0	0	0	0
on the other hand	prep det adj N	4	0	0	0	0
the development of a	det N prep det	1	0	0	0	0
with respect to the	prep N prep det	3	0	0	0	0

Table 4.31: Four-grams statistics /Collection A/118 articles

214 B		Max	Min	Mode	Med	Mean
as a result of	prep det N prep	2	0	0	0	0
as long as you	prep adj prep pron	1	0	0	0	0
as well as the	prep adv prep det	3	0	0	0	0
for the treatment of	prep det N prep	3	0	0	0	0
in the absence of	prep det N prep	4	0	0	0	0
in the presence of	prep det N prep	4	0	0	0	0
on the other hand	prep det adj N	4	0	0	0	0

Table 4.32: Four-grams statistics /Collection B/214 articles

182 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
as a result of	prep det N prep	4	0	0	0	0
as long as you	prep adj prep pron	1	0	0	0	0
as well as the	prep adv prep det	3	0	0	0	0
for the treatment of	prep det N prep	9	0	0	0	0
in the absence of	prep det N prep	5	0	0	0	0
in the presence of	prep det N prep	9	0	0	0	0
on the other hand	prep det adj N	6	0	0	0	0

Table 4.33: Four-grams statistics /Collection B/182 articles

109 B		**Max**	**Min**	**Mode**	**Med**	**Mean**
as a result of	prep det N prep	2	0	0	0	0
as long as you	prep adj prep pron	1	0	0	0	0
as well as the	prep adv prep det	4	0	0	0	0
for the treatment of	prep det N prep	18	0	0	0	1
in the absence of	prep det N prep	12	0	0	0	1
in the presence of	prep det N prep	13	0	0	0	1
on the other hand	prep det adj N	8	0	0	0	0

Table 4.34: Four-grams statistics /Collection B/109 articles

4.3 Nouns

By filtering the nouns of our frequency lists, we select the top 143, some nouns repeat without much difference in the rank order between collection A and B, and almost two-thirds are the same words. However, we may encounter some words that do not function as nouns since our collections are not tagged. See tables 4.35, p. 101 and table4.36, p. 102.

Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word
33	7532	disease	151	1500	system	249	941	potential
35	6726	patients	155	1481	education	252	933	blood
43	5498	study	157	1474	model	253	930	tissue
47	4929	data	158	1471	program	254	926	response
48	4672	pathology	159	1461	article	255	918	participants
51	4050	health	160	1457	diabetes	258	911	network
61	3267	analysis	163	1451	academic	259	902	tumor
71	2811	table	164	1447	practice	260	902	values
75	2728	research	168	1403	virus	261	895	author
80	2615	time	171	1345	value	263	886	mortality
83	2523	results	173	1306	hospital	265	880	learning
84	2499	patient	174	1287	expression	267	873	weight
85	2484	cancer	177	1279	levels	268	872	glucose
89	2375	cells	179	1271	samples	269	871	report
90	2356	cell	184	1235	methods	270	866	approach
91	2332	group	185	1226	protein	274	852	terms
92	2285	university	186	1226	residents	275	851	resident
93	2278	risk	187	1204	work	276	850	months
94	2265	number	190	1191	change	280	839	process
97	2246	years	192	1176	drugs	281	836	support
99	2152	students	194	1168	population	284	826	public
100	2134	laboratory	196	1166	changes	285	824	present
101	2114	test	200	1156	association	288	819	center
103	2097	gene	201	1156	mean	302	798	part
105	2075	diseases	203	1140	control	305	788	pathologists
106	2069	care	205	1123	role	306	783	days
111	1996	drug	206	1115	primary	307	780	loss
112	1986	studies	207	1100	function	308	779	increase
116	1967	medicine	208	1098	quality	309	779	outcomes
117	1958	diagnosis	209	1081	development	310	777	surgery
118	1956	total	210	1079	women	311	775	autopsy
119	1947	figure	212	1076	rate	312	774	family
121	1940	case	213	1074	factors	314	769	access
122	1938	human	214	1073	score	315	769	infection
125	1923	department	219	1068	groups	316	767	tests
128	1862	year	227	1044	activity	317	765	service
130	1845	age	231	1024	journal	322	750	science
133	1782	treatment	234	1008	survey	324	745	findings
134	1779	training	235	998	effect	325	742	nature
136	1757	cases	236	991	performance	326	741	distribution
137	1692	faculty	238	989	programs	328	738	evaluation
138	1669	genes	239	988	testing	329	730	cohort
140	1645	information	240	984	therapy	333	725	impact
142	1615	school	242	978	evidence	336	723	authors
143	1592	level	244	972	management	337	721	scores
145	1578	type	245	970	knowledge	338	717	addition
150	1510	review	246	961	effects			

Table 4.35: Frequent Nouns Collection A

Rank	Freq	Word	Rank	Freq	Word	Rank	Freq	Word
39	11098	cells	171	2162	model	252	1559	visual
41	10766	study	172	2157	proteins	253	1549	authors
45	9385	cell	174	2140	dna	254	1546	day
49	6868	disease	175	2119	including	255	1540	drug
58	6117	data	176	2103	due	258	1530	primary
60	5755	analysis	177	2103	surgery	259	1525	antibody
65	5214	treatment	183	2051	system	260	1524	post
68	5125	group	184	2050	effect	261	1522	range
71	4623	results	185	2032	diseases	264	1515	ophthalmology
75	4436	protein	188	1987	serum	266	1507	work
76	4428	figure	189	1979	review	267	1492	case
77	4303	studies	190	1976	cases	268	1492	stress
78	4301	control	194	1952	effects	270	1476	symptoms
85	3978	cancer	195	1941	children	272	1472	nature
87	3883	table	196	1934	value	273	1471	association
88	3874	human	197	1902	factor	275	1456	center
89	3805	expression	198	1875	therapy	282	1432	participants
92	3695	health	200	1862	growth	283	1431	rna
95	3678	journal	202	1857	brain	286	1423	china
96	3654	time	208	1829	tissue	287	1416	concentration
98	3574	age	209	1828	values	290	1398	months
99	3493	research	210	1807	methods	291	1386	presence
101	3448	test	212	1797	rate	292	1367	activation
104	3396	levels	215	1772	tumor	293	1367	antibiotics
107	3298	risk	221	1744	usa	294	1367	days
115	2996	number	222	1735	population	295	1366	area
118	2932	years	223	1734	acid	298	1354	diagnosis
122	2874	activity	225	1713	medicine	300	1351	loss
125	2800	university	227	1691	parkinson	301	1347	antibodies
128	2743	infection	228	1690	published	302	1343	responses
131	2703	article	229	1689	genes	303	1343	year
133	2604	groups	231	1660	score	306	1331	liver
139	2516	blood	233	1649	increase	307	1331	mitochondrial
141	2489	total	234	1644	muscle	310	1324	receptor
143	2451	samples	235	1644	virus	311	1315	molecular
144	2431	patient	236	1640	subjects	312	1309	phase
148	2406	type	237	1635	department	315	1306	assay
149	2390	response	238	1634	resistance	318	1301	membrane
155	2357	role	239	1618	prevalence	328	1261	mouse
157	2342	factors	241	1611	baseline	329	1256	death
162	2248	vaccine	242	1604	potential	330	1254	information
163	2241	gene	243	1600	present	336	1229	conditions
164	2236	level	245	1578	addition	339	1223	ratio
165	2229	hiv	246	1578	hospital	340	1223	science
166	2210	changes	248	1574	signaling	341	1221	change
168	2184	function	250	1569	sample	342	1220	surface
169	2175	medical	251	1568	bone	343	1212	distribution
170	2162	development						

Table 4.36: Frequent Nouns Collection B

4.3.1 Linguistic and statistic analysis

Thus, due to lack of resources and time we carry out the analysis on the top ten most frequent nouns, plus the plural or the singular word (as required) to have a complete analysis of the word. For instance, we analyze disease and diseases as shown in table 4.37, p. 104 and4.38, p. 105. Consequently, some of these words (plural or singular) might be in a lower rank than the top ten nouns. In the data displayed we present the number of times the N + N pattern occurs.

Both collections share words such as patients, study, disease, data, and analysis. The tendency in these words doesn't change in collection B and it shows an increase according to the number of occurrences. For instance, the word *patients* appears 6,726 times in collection A, 12,663 in collection B and the pattern N+N has a similar and proportional occurrence in both. However, *disease* is less frequent in collection B, which might be due to the type of subject inside medicine[3] but, the frequency of the syntactic structure and the occurrences are relevant.

The rest of the nouns in our list follow the same pattern N + N structure in both collections. Even if certain words vary there is a correlation in the syntactic structure. For instance, `protein`, `group`, `treatment`, `pathology` have a relevant number of occurrences of N + N.

The particular case of the word `table` show zero cases of N + N structure since it is often used to introduce a table number. For example:

555 | 20:755 Page 8 of 15 `Table` 2 Correlation between medical school
558 | clinical severity (Table 4). `TABLE` 3. Correlation of serum chemokine

4.3.2 Specific Cases

We chose some examples for a deeper analysis in the following examples we have some specific analysis, we observe the number of the concordance hit[4] and the KWIC the main word highlighted[5]. Sorting to the right to see the complementation and left to find a

[3] As mentioned before, articles were chosen randomly from different journals
[4] Number of line
[5] Some of the sentences are not complete due to the concordancer software

300							
Rank	**Freq**	**Words**	**Noun+Noun**	**Rank**	**Freq**	**Words**	**Noun + Noun**
33	7532	disease right	3226	105	2075	diseases right	10
		disease left	2564			diseases left	194
35	6726	patients right	104	84	2499	patient right	672
		patients left	618			patient left	30
43	5498	study right	669	112	1986	studies right	2
		study left	480			studies left	154
47	4929	data right	845				
		data left	264				
48	4672	pathology right	1531	3055	85	pathologies right	0
		pathology left	105			pathologies left	0
51	4050	health right	2270				
		health left	113				
61	3267	analysis right	32	419	609	analyses right	0
		analysis left	1003			analyses left	68
71	2811	table right	0				
		table left	0				
75	2728	research	710	22587	4	researches right	0
		research	208			researches left	0
80	2615	time right	2235				
		time left	75				

Table 4.37: Frequency Noun + Noun Collection A

600							
Rank	**Freq**	**Words**	**Noun+ Noun**	**Rank**	**Freq**	**Words**	**Noun + Noun**
35	12663	patients right	156	144	2431	patient right	638
		patients left	2219			patient left	86
39	11098	cells right	18				
		cells left	1665				
41	10766	study right	1234	77	4303	studies right	0
		study left	527			studies left	162
45	9385	cell right	4746				
		cell left	908				
49	6868	disease right	1526	185	2032	diseases right	0
		disease left	3065			diseases left	147
58	6117	data right	529				
		data left	104				
60	5755	analysis right	64	419	1039	analyses right	0
		analysis left	1476			analyses left	169
65	5214	treatment right	691	1202	402	treatments right	0
		treatment left	358			treatments left	9
68	5125	group right	77	133	2604	groups right	0
		group left	1378			groups left	258
71	4623	results right	4	498	922	result right	4
		results left	163			result left	15
75	4436	protein right	1090	172	2157	proteins right	2
		protein left	266			proteins left	161

Table 4.38: Frequency Noun + Noun Collection B

modifier. First, the selected examples[6] of both collections and but separated according to the collection (A-B) in the appendix.

Compound words

Through the analysis we found compound nouns such as: *ill-health*, *meta-analysis*, *out-patient*, and *face-time*, which are in the dictionary as a sole word. For instance, *ill-health* means "a condition of inferior health in which some disease or impairment of function is present but is usually not as serious in terms of curtailing activity as an illness" (Merriam-Webster, 2021). However, they are considered as N + N pattern.

1252	period	of	functional	decline	and	`ill health`	at	
	N	prep	adj	N	conj	N N	prep	
1541	studies	included	in	the	main	`meta-analysis`,	the	
308	a published	model	to	mitigate	`face-time`	bias	associated	with
1359	cholesterol;	MOPD:	Medical	`out-patient`	department;			
5049	transfections	Human	liver	hepatocellular	`cell`	`line`	HepG2	

Hyphen and dash

We found a dash to establish a relationship between two subjects N + N such as number 325 `patient-caregiver` but in number 891 `chemical-disease` ADJ + N to modify another noun.

891	database	provides	curated	and	inferred	`chemical-disease`	associa-tions.	
	N	V	adj	conj	adj	adj+N	N	
893	important	resource	for	predicting	unobserved	`drug-disease`	associa- tions.	
	adj	N	prep	V	adj	N N	N	
325	support,	all	of	which	may	impair	`patient-caregiver`	relational

[6]Both compilations of occurrences of the mentioned structure are available in the appendix section for your further consultation

Proper Names

The N + N structure appears in names of diseases, institutions, associations, hospitals, book titles, etc. With respect to diseases, authors may use or not 'the possessive' Merriam Webster Dictionary accepts both versions. For instance, Alzheimer's disease in line 2877 and 195.

2875	changes	and	arrhythmia in	Fabry	disease,	
2877	between	spirochetal	infections	and	ALZHEIMER'S	disease.
195	Correlation	of	Alzheimer	disease	neuropathologic	
2878	Mitochondria	in	Parkinson's	Disease.		
2876	Rapid	Resolution	of	Lyme	Disease.	
2879	injury;	CKD,	chronic	kidney	disease;	

906	1,361	patients	using	the	National	Trauma	Data	Bank.
	num	N	V	det	N	N	N	N

1071 SUS foundations, Kristianstad Central Hospital Research and Development Fund,
1072 the Helsinki University Central Hospital Research Foundation (Grant No.
1073 Imaging, Polish Mother's Memorial Hospital Research Institute, Lodz,
1074 was received from Skaraborgs Hospital research founda- tion. There is
1075 Medicine, King Faisal Specialist Hospital & Research Centre, Riyadh, Saudi Arabia;

Noun as modifier

The following sentences are chosen to exemplify the N + N structure functioning as a modifier. For instance, the number 1777 drug-disease is modifying associations or number 4260 Alzheimer's disease modifies patients.

1777	repositioning	to	express	drug-	disease	as-sociations		
	v	prep	V	N	N	N		
4260	cohort	of	moderate-	severe	Alzheimer's	disease	patients	
	N	prep	adj	N	N	N	N	

However, it may work along and adjective inside an adjective phrase. In line 244 the adjective phrase optic nerve disease modifies the nounactivity.

244	optic	nerve	disease	activity	and			
	adj	N	N	N	conj			
4904	microbial	profiles	vary	with	inflammatory	bowel	disease	phenotypes.
	adj	N	V	prep	adj	N	N	N
5766	the	disease-	severity	subgroups,	although	no	significant	
	det	N	N	N	conj	adv	adj	

66	profiling	of	lung	adenocarcinoma	patients	reveals
	N	(n)prep	N	N	N	V
6323	Symptom	-disease	network	reasoning	In	the
	N	N	N	N		
138	rank	correlation	was used	for	correlation	analyses.
		N	V	prep	N	N

6106	and	cell	migration	assays	were		
	conj	N	N	N			
2218	and	crucial	step	in	disease	data	integration
	conj	adj	N	(n) prep	N	N	N
631	Coronary	artery	disease	in	police	officers	
	N	N	N	(n) prep	N	N	

1129	The	prioritization	candidate	disease	gene	function	is
	det	N	N	N	N	N	v
1691	(analogous	to	a	postsophomore	pathology	fellowship	in
		prep	det	N	N	N	prep
1808	performance	of	only	some	health	indicators	was
	N	prep	adj	det	N	N	V

2172	important	to	solve	the	health	manpower	crisis
	adj	prep	V	det	N	N	N
3218	personal	or	a	family	member's	health	conditions.
		conj	det	N	N	N	N
130	increased	breast	cancer	cell	adhesion	to	the
	adj	N	N	N	N	prep	det

On the one hand and on the other hand

The data of the connector ON THE ONE HAND and ON THE OTHER HAND is also worth mentioning. First, these connectors are not always together. It means that we may use only the latter one but not only the former one. We found two versions ON ONE HAND and *on the one hand* the latter one found in the Cambridge Dictionary.

	A	B
In the other hand	1	0
on one hand	6	12
on the one hand	4	15
on the other hand	104	274

Table 4.39: Connector: On the one hand

Chapter 5

Conclusions

In conclusion, scientific medical articles do present relevant syntactic patterns as expected. Even though n-grams might not present statistical reliability if calculated among the whole number of words in the collections, there are relevant syntactic structure patterns in both collections which are *the most frequent and relevant* among the content part of the texts (excluding functional words, as per Zipf's Law). It is important to notice, moreover, that due to the very nature of the academic writing, as we saw in our data sets, the number of *hapax legomena* represent almost 50% of the tokens. This means, together with the Zipf's law top frequency occurrence, that the core of a text is, and will be, the central "slice" of the collection. If we focus on that core part, the retrieved pattern are relevant and play a key role in the behavior of the articles.

Therefore, it is safe to state that if we apply the same methodology in a specialized corpus we will gather proportional and reliable data for generalization to recommend ESP material, gathering those syntactic (and semantic) patters that the specific area of language shows.

As per our findings, in the present work the following syntactic patterns are shown to be relevant in medical scientific articles. First, the syntactic patterns in the analysis of n-grams are as follows: N + PREP, PREP + N, PREP + DET bi-grams, DET + N + PREP, PREP + N + PREP in tri-grams, and PREP + DET + N + PREP in 4-grams, and N + N.

According to our analysis, therefore, this specific set of syntactic patterns is fundamental for understanding in reading and for production (writing) of scientific articles, since are recurrent in and above all other structures. Students of English for Specific Purposes in the medical area who do not achieve understanding of these patterns are likely to not understand

the majority of medical scientific journal articles. It is important therefore to elaborate both materials and evaluation focused on these patterns, in order to guarantee a minimum base level of comprehension.

It is also important to say that, since we are working on medical scientific content, these patterns and the relevant lexical content derived from them, are a part and above any competence level according to the Common European Framework of Reference. This is due to the need of medical students to read all through their courses about recent publication, a need that goes beyond their need to communicate in English. It is likely to assume that at the and of their university studies, medical student would acquire a 360 degrees competence in English, as desired, nevertheless, during the first years the competence will necessarily be skewed, as for the need driven aims.

The achievements in the present research allow to work on both materials and evaluation specifically thought for that part of the medical student development stage, before the full competence, when the students struggle with early steps and need support in focusing on effective learning. The syntactic patterns (and derived lexical items) found in the present work come to help and support both students and teachers alike on the task.

5.1 Limitation of the current study

There were limitations in different stages of the research, from the recollection of scientific articles to data analysis. For instance, in the recollection stage, the downloader allowed us to obtain numerous articles but it was inevitable to also retrieve non-article parts of journals resulting in noisy data inside our collection and, therefore, more time invested in the cleaning process. This limitation was overcome, but at the cost of working hours.

In addition, about the process of cleansing data, when converting articles from *.pdf* to *.txt* extension page headers, tables, figures, bibliography caused noise in our data, the possibility to access the articles without these would have saved some time to focus on the analysis itself or to clean other possible 'noise' in the corpus. As a matter of fact, the above limitations could be put aside, but at a relatively high cost in working hours and manual analysis.

The solution for further and more efficient studies in the area is the possibility to build a tagged corpus, which would have resulted in more detailed data concerning word categories and their frequency and co-occurrence. However, it would have implied a major organization, and investment in both human and economic resources.

5.1.1 Recommendations

About the teaching material design, we recommend the use of N + PREP, PREP + N, PREP + DET, DET + N + PREP, PREP + N + PREP, PREP + DET + N + PREP and the N + N syntactic structures to design medical ESP teaching material in order to help no native speakers avoid mistakes and achieve native proficiency specially the syntactic structure N + N is not usual in romance languages and would represent a challenge when writing or reading.

In addition, the design of a specialized tagged corpus of scientific articles in order to have more specific data in the frequency of syntactic structures. Not only for the present research but also for linguistic research in general and in such a manner to encourage linguistic students and professionals to use corpus linguistics as a teaching and research resource.

Finally, as we explain above, tables, figure, bibliography, etc. caused noise in our collections resulting in the need of an increased manual work, therefore, we recommend other tools such as R, Phyton and specific corpus building process. Further investigation should be based on tagged and formal corpus.

Nevertheless, the important statement of the present work is that, since the presence of the relevant pattern could be proved, the study can be considered as a strong and sufficient preliminary base in order to justify the submission for research proposal and funding. Due to the findings presented here, we can safely state that the investment in corpus building for ESP will be giving positive results at a larger scale.

5.1.2 Implications

Scientific articles are crucial for new insights, professional growth by sharing findings and expertise. The data found allow us to use these patterns to provide the teacher and student with reliable teaching material data. Furthermore, the confirmation of the existence of relevant syntactic patterns means the possibility to follow the same methodology for other areas such as: engineering, economics, law, and so on, and therefore recommend or set other teaching guidelines and tools based on syntactic patterns for the student and the teacher alike. Thus, as mentioned in the methodological relevance, given the language-independent nature of Zipf's Law we are able to demonstrate that such patterns are dependent on neither topic nor language.

In addition, the structure N + N is not standard in other languages such as; French or Italian, which means that it can apply in other languages.

5.2 Suggestions for Future Work

There are many aspects to be studied in more detail that might benefit not only in relation to ESP students and teachers, but Corpus linguistics and computational linguistics might be a new source of data for Bolivian linguists.

In this specific case, there are more patterns related to the syntactic structures. For instance, the pattern N + N, as mentioned, is not a usual structure in romance languages, and building two different corpora to compare native and non-native speakers in the target language would be a source of data of possible mistakes. Thus, open the discussion between grammatically correct and near-native proficiency when writing.

In relation to the authors' background (nationality or mother tongue) would reveal to us the use of the N + N syntactic structure. It means that since English is an important language in scientific articles, there will be authors from all over the world.

Concerning the structure of an article, it would be relevant to identify syntactic structure or lexicography according to the structure. For instance, when explaining tables, graphs, results, comparisons, etc. The field is still open to much more investigation and studies, offering a long line of possible research towards a deeper exploration of patterns in ESP.

Bibliography

Aarts, J. (1991), *Intuition-based and observation-based grammars*, Taylor & Francis, chapter 4, pp. 44–62.

Aarts, J. and van den Heuvel, T. (1982), 'Grammars and intuitions in corpus linguistics', *Computer corpora in English language research* pp. 66–84.

Ahmed, M. K. (2014), 'The esp teacher: Issues, tasks and challenges', *English for specific purposes world* **42**(15), 1–33.

Allan, K. (2013), *The Oxford handbook of the history of linguistics*, OUP Oxford.

Archer, D., Rayson, P. et al. (2006), *Corpus linguistics around the world*, BRILL.

Baker, P. (2012), *Contemporary corpus linguistics*, Bloomsbury Publishing.

Baker, P., Hardie, A. and McEnery, T. (2006), 'A glossary of corpus linguistics.'.

Barbera, M., Corino, E. and Onesti, C. (2007), *Corpora e linguistica in rete*, Guerra.

Basturkmen, H. (2010), *Developing courses in English for specific purposes*, Springer.

Benson, M., Benson, E. and Ilson, R. (1986), *Lexicographic description of English*, Vol. 14, J. Benjamins Publishing Company.

Biber, D. (1993), 'Representativeness in corpus design', *Literary and linguistic computing* **8**(4), 243–257.

Biber, D. and Reppen, R. (2015), *The Cambridge handbook of English corpus linguistics*, Cambridge University Press.

Bolshakov, I. A. and Gelbukh, A. (2004), *Computational linguistics models, resources, applications*, Ciencia de la computación.

Boriskina, O. (2009), 'cryptotype approach to the study of metaphorical collocations in english', *Corpus-Based Approaches to Figurative Language. A Corpus Linguistics* .

Callisaya, S. C. (2014), 'Oncology esp course design addressed to oncology professionals from the school of medicine of mayor de san andres university'.

Cambridge English, D. (2020), 'Cambridge english dictionary'.
URL: *https://dictionary.cambridge.org/*

Chomsky, N. and Hill, A. (1962), 'Third texas conference on problems of linguistic analysis in english'.

Collins English, D. (2014), 'Collins english dictionary'.
URL: *https://collins.co.uk/pages/reference-collins-english-dictionary*

Crawford, W. and Csomay, E. (2015), *Doing corpus linguistics*, Routledge.

Cristina Pabón Escobar, S. and da Costa, M. C. (2006), 'Visibility of latin american scientific publications: the example of bolivia', *Journal of Science Communication* **5**(2), A01.

Dash, N. S. (2008), *Corpus linguistics: An introduction*, Pearson Education India.

Di Bitetti, M. S. and Ferreras, J. A. (2017), 'Publish (in english) or perish: The effect on citation rate of using languages other than english in scientific publications', *Ambio* **46**(1), 121–127.

Dudley-Evans, T. and St. John, M. J. (1998), *Developments in ESP, A multi-disciplinary approach*, Cambridge University Press.

Edwards, P. N. (2016), 'Michael d. gordin. scientific babel: How science was done before and after global english.'.

Ellis, N. C., Simpson-Vlach, R. and Maynard, C. (2008), 'Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol', *TESOL quarterly* **42**(3), 375–396.

Facchinetti, R. (2007), *Corpus linguistics 25 years on*, number 62, Rodopi.

Francis, W. N. (1982), *Problems of assembling and computerizing large corpora*, na.

Groom, N. (2005), 'Pattern and meaning across genres and disciplines: An exploratory study', *Journal of English for Academic Purposes* **4**(3), 257–277.

Hamel, R. E. (2007), 'The dominance of english in the international scientific periodical literature and the future of language use in science', *Aila Review* **20**(1), 53–71.

Hanks, P. (2008), Lexical patterns: from hornby to hunston and beyond, *in* 'Proceedings of the XIII EURALEX International Congress', Vol. 1, pp. 89–129.

Heine, B. and Narrog, H. (2015), *The Oxford handbook of linguistic analysis*, Oxford Handbooks in Linguistic.

Hernández-Sampieri, R., Fernández Collado, C., Baptista Lucio, P. et al. (2018), *Metodología de la investigación*, Vol. 4, McGraw-Hill Interamericana México.

Hunston, S. and Francis, G. (2000), *Pattern grammar: A corpus-driven approach to the lexical grammar of English*, number 4, John Benjamins Publishing.

Hutchinson, T. and Waters, A. (1987), *English for Specific Purposes*, Cambridge University Press.

Huttner-Koros, A. (2015), 'The hidden bias of science's universal language'.
URL: *https://www.theatlantic.com/science/archive/2015/08/english-universal-language-science-research/400919/*

Kennedy, G. (2014), *An introduction to corpus linguistics*, Routledge.

Kornai, A. (2002), 'How many words are there?', *Glottometrics* **4**, 61–86.

Leech, G. (1991), 'The state of the art in corpus linguistics, english corpus linguistics aijmer k., altenberg b.(eds) pp. 8-29'.

Lehecka, T. (2015), Collocation and colligation, *in* 'Handbook of pragmatics online', Benjamins.

Lewis, M., Conzett, J., Hargreaves, P. H., Hill, J., Lewis, M. and Woolard, G. C. (2000), *Teaching collocation: Further developments in the lexical approach*, Vol. 244, Language Teaching Publications Hove.

Manning, C. and Schutze, H. (1999), *Foundations of statistical natural language processing*, MIT press.

McEnery, T. and Hardie, A. (2011), *Corpus linguistics: Method, theory and practice*, Cambridge University Press.

McEnery, T. and Wilson, A. (2003), 'Corpus linguistics', *The Oxford handbook of computational linguistics* pp. 448–463.

McEnery, T., Xiao, R. and Tono, Y. (2006), *Corpus-based language studies: An advanced resource book*, Taylor & Francis.

Merriam-Webster (2021), 'An encyclopædia britannica company'.
URL: *https://www.merriam-webster.com/*

Meyer, C. F. (2002), *English corpus linguistics: An introduction*, Cambridge University Press.

Mitkov, R. (2004), *The Oxford handbook of computational linguistics*, Oxford University Press.

Newitz, A. (2013), 'A mysterious law that predicts the size of the world's biggest cities'.
URL: *https://gizmodo.com/the-mysterious-law-that-governs-the-size-of-your-city-14792441*
utm_expid=66866090-48.Ej9760cOTJCPS_Bq4mjoww.0&
/the-mysterious-law-that-governs-the-size-of-your-city-1479244159

Noguchi, J., Orr, T. and Tono, Y. (2006), Using a dedicated corpus to identify features of professional english usage: what do "we" do in science journal articles?, *in* 'Corpus Linguistics around the world', Brill, pp. 155–166.

Novoa, P. (1992), 'John sinclair: Corpus, concordance, collocation oxford: Oxford university press. 1991', *Lenguas Modernas* (19), 167–171.

Orasan, C. (2001), Patterns in scientific abstracts, *in* 'Proceedings of Corpus Linguistics 2001 Conference', Citeseer, pp. 433–443.

Özdemir, N. Ö. (2014), 'Using corpus data to teach collocations in medical english', *Journal of Second Language Teaching & Research* **3**(1), 37–52.

Peacock, M. (2012), 'High-frequency collocations of nouns in research articles across eight disciplines', *Iberica: Revista de la Asociacion Europea de Lenguas para Fines Especificos (AELFE)* (23), 29–46.

Ramírez-Castañeda, V. (2020), 'Disadvantages in preparing and publishing scientific papers caused by the dominance of the english language in science: The case of colombian researchers in biological sciences', *PloS one* **15**(9), e0238372.

Redner, S. (1998), 'How popular is your paper? an empirical study of the citation distribution', *The European Physical Journal B-Condensed Matter and Complex Systems* **4**(2), 131–134.

Robinson, P. C. (1980), *ESP (English for Specific Purposes): The Present Position*, Prentice Hall.

Schmitt, N. (2013), *An introduction to applied linguistics*, Routledge.

Schubert, L. (2020), Computational Linguistics, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Spring 2020 edn, Metaphysics Research Lab, Stanford University.

Stabler, E. (2003), 'Notes on computational linguistics', *Course notes (under revision), Department of Linguistics, UCLA. Downloadable at: http://www. linguistics. ucla. edu/people/stabler/185-03. pdf* .

Stefanowitsch, A. (2020), *Corpus linguistics: A guide to the methodology*, Language Science Press.

Susan Hunston, Elizabeth Manning, G. F. (1998), *Grammar Patterns 2: Nouns and Adjectives*, Collins Cobuild.

Tognini-Bonelli, E. (2001), *Corpus linguistics at work*, Vol. 6, John Benjamins Publishing.

Zipf, G. K. (2016), *Human behavior and the principle of least effort: An introduction to human ecology*, Ravenio Books.

Zufferey, S. (2020), *Introduction to Corpus Linguistics*, John Wiley & Sons.

Chapter 6

Appendix

We present here the supplementary data concerning our research divided in three sections. First, the raw frequency list of the top 50 words of both collections. See 6.1 p. II. Then, there is a table about the dis-legomenon, tris-legomenon, and tetrakis-legomenon in both collections as additional data. See 6.2 and 6.3 p. III. Finally, some additional examples for our N + N structure.

6.1 Raw Word Frequency Lists

Collection A			Collection B		
Rank	**Freq**	**Word**	**Rank**	**Freq**	**Word**
1	92339	the	1	157259	the
2	73898	of	2	138652	of
3	63991	and	3	118783	and
4	47117	in	4	91598	in
5	41688	a	5	73638	a
6	36481	to	6	55798	to
7	23346	for	7	38306	with
8	18138	with	8	33927	for
9	15659	s	9	29877	s
10	14813	e	10	27426	e
11	12487	is	11	25592	was
12	11983	that	12	24645	were
13	11839	were	13	21370	is
14	11346	as	14	20940	by
15	11293	was	15	20242	p
16	10987	n	16	20181	c
17	10162	t	17	19166	as
18	9611	by	18	19125	n
19	9415	on	19	18873	m
20	9162	c	20	18746	that
21	9110	i	21	18005	j
22	8704	from	22	17992	t
23	8540	al	23	17126	r
24	8505	or	24	15413	d
25	8465	p	25	15160	al
26	8463	r	26	15147	i
27	8286	et	27	14996	et
28	8265	j	28	14524	from
29	8169	are	29	14479	on
30	8088	d	30	14171	or
31	8041	m	31	13615	h
32	7674	this	32	13395	at
33	7532	disease	33	12847	b
34	6996	be	34	12781	l
35	6726	patients	35	12663	patients
36	6589	o	36	11829	are
37	6333	at	37	11683	this
38	6227	h	38	11576	no
39	6198	l	39	11098	cells
40	5928	an	40	10970	o
41	5823	medical	41	10766	study
42	5554	we	42	10396	be
43	5498	study	43	9856	g
44	5305	not	44	9811	an
45	5202	b	45	9385	cell
46	5023	g	46	8329	not
47	4929	data	47	7741	we
48	4672	pathology	48	7615	f
49	4314	have	49	6868	disease
50	4136	f	50	6772	vol

Table 6.1: Raw Word Frequency Lists Top 50 (Collection A-B)

6.2 Hapax Legomenon

300	
Word Types	72,033
Hapax	33229
Dis legomenon	10388
Tris legomenon	5047
Tetrakis legomenon	3192

Table 6.2: Hapax Legomenon Collection A

600	
Word Types	110,612
Hapax legomenon	53248
Dis legomenon	15736
Tris legomenon	7510
Tetrakis legomenon	4640

Table 6.3: Hapax Legomenon Collection B

6.3 Linguistic analysis Nouns

6.3.1 DISEASE/RIGHT (Collection A) 2376

153	a	useful	indicator	of	disease	activity.			
	det	adj	N	(n) prep	N	N			
156	a	negative	effect	on	disease	activity	in	UC	
	det	adj	N	(n) prep	N	N	prep		
158	less	weight	loss	and	disease	activity	index	(DAI)	
	adj	N	N	conj	N	N	N	N	
746	by	PSDs	were validated using	independent	disease	annota- tions			
	prep	N	V	adj	N	N			

752	improving	and	unifying	disease	annotations	across	species.	
	V	conj	V	N	N	prep	N	
753	The	strain-associated	disease	annotations	are viewed	under	the	
	det	N adj	N	N	v	prep	det	
2217	disease-gene	data.	RGD	regularly	imports	disease	data	
	N N	N	N	adv	v	N	N	
2218	and	crucial	step	in	disease	data	integration	is
	conj	adj	N	(n) prep	N	N	N	v

2220	for	linking	biomedical	knowledge	through	disease	data.
	prep	V	adv	N	prep	N	N
2470	observation	of	heterogene- ous	disease	distribution	in	axial
	N	prep	adj	N	N	prep	adj
2506	Then,	disease-drug	relationships	are obtained	from		
	adv	N N	N	V	prep		

154	effective	in	reducing	disease	activity	in	active	CD,
	adj	prep	V	N	N	prep	adj	N
157	stricturing	type.	The	Crohn's	disease	activity	index	(CDAI),
			det	N	N	N	N	
2219	Disease	data	The	disease	data	is downloaded	from	
			det	N	N	v	prep	

6.3.2 DISEASE/LEFT (Collection A) 2564

630	management	of	peripheral	artery	`disease`	in	patients
	N	prep	adj	N	N	(n) prep	N
631	Coronary	artery	`disease`	in	police	officers	participating
	N	N	N	(n) prep	N	N	V
816	outstanding	questions	in	the	biodiversity–	`disease`	literature
	adj	N	prep	det	N	N	N

819	different	biodiversity–	`disease`	patterns	to	emerge	at
	adj	N	N	N	prep	v	prep
1060	calculated	different	inflammatory	bowel	`disease`	risk	scores
	adj	adj	adj	N	N	N	N
1129	The	prioritization	candidate	`disease`	gene	function	is denoted
	det	N	N mod	N	N	N	v

1777	drug	repositioning	to	express	drug-	`disease`	as-sociations
	N	v	prep	V	N	N	N
1778	known	drug-	`disease`	as- sociations,	we	adopted	
	adj	N	N	N	pron	V	
3435	hip	dysplasia	as	solely	a joint	`disease`	should be
	N	N	prep	adv	det N	N	v

3436	hip	dysplasia	is considered	a	joint	`disease`	with	
	N	N	v	det	N	N	prep	
5423	man	with	typical	Parkinson's	`disease`	and	mild	mental
	N	prep	adj	N	N	conj	adj	adj
5424	in	mild	and	moderate	Alzheimer's	`disease`	and	mixed
	prep	adj	conj	adj	N	N	conj	adj

5425	for	autoimmune	diseases,	including	Crohn's	`disease`	and		
	prep	adj	N	(n) prep	N	N	conj		
5752	the	management	of	Alzheimer's	`disease`	is based	on	the	
	det	N	(n) prep	N	N	v	prep	det	
5753	The	diagnostic	concept	of	Parkinson's	`disease`	is changing	and	
	det	adj	N	(n) prep	N	N	V	conj	

5754	in	which	colonic	Crohn's	disease	is	intermediate
	prep	pron	adj	N	N	v	adj
6322	The	bipartite	network	of	symptom-	disease	interactions
	det	adj	N	(n) prep	N	N	N
6323	so	far.	Symptom-	disease	network	reasoning	In
			N	N	N	N	N

6.3.3 DISEASE /LEFT (Collection B) 3065

195	Correlation	of	Alzheimer	disease	neuropathologic	changes	with	
	N	(n) prep	N	N	adj	N	prep	

1634	rubbing,	or	atopic	eye	disease)	[20, 21].	A	
		conj	adj	N	N			
2098	correlations	were found	with	specific	HIV disease	measures.		
	N	v	prep	adj	N N	N		
2421	fibrosis	markers	and	chronic	kidney	disease	among	adults
	N	N	conj	adj	N	N	prep	N

2575	mouse	model	for	chronic	alcoholic	liver	disease.	
	N	N	prep	adj		N	N	
2577	lung	and	liver	disease	like	chronic	obstructive	
	N	conj	N	N	prep	adj	adj	
3560	American	Parkinson	Disease	Association	and	salary	support	
	adj	N	N	N	conj	N	N	

3701	Caregiver	burden	in	patients	with	Parkinson	disease	
	N	N	(n) prep	N	prep	N	N	
4068	the	detection	of	mild	Alzheimer's disease	and	mild	
	prep	N	prep	adj	N	N	conj	adj

4260	cohort	of	moderate-severe	Alzheimer's	disease	patients	and	
	N	prep	adj	N	N	N	conj	
4261	Secondary	outcomes	in	mild	Alzheimer's	disease	patients.	
	adj	N	prep	adj	N	N	N	
4367	for	ulcerative	colitis	or	Crohn's	disease	also	
	prep	adj	N	conj	N	N	adv	

6.3.4 DISEASE/ RIGHT (Collection B) 1526

243	Moreover,	disease	activity	and	severity	are		
	adv	N	N	conj	adj	v		
245	'Mild-moderate'	pattern	[≤5	quarters	of	disease	activity]	
	adj	N		N	(n) prep	N	N	

287		allowed	for	determination	of	disease	activity	in
		V	prep	N	prep	N	N	prep
288	The	overall	disease	activity	in	uSpA	is	
	det	adj	N	N	prep	N	V	
359	reflects	only	subjective	disease	activity	rather	than	
	V	adv	adj	N	N	adv	pron	

360	The	remaining	patients	have	persistent	disease	activity	
	det	adj	N	V	adj	N	N	
1258	a	direct	consequence	of	a	complicated	disease	behaviour.
	det	adj	N	prep	det	adj	N	N
1303	source	of	samples	for	detection	of	disease	biomarkers.
	N	(n)prep	N	(n)prep	N	(n)prep	N	N

1405	impact	by	preventing	disease	burden	quantification	and	
	N	prep	V	N	N	N	conj	
1968	and	PSP	patients	were chosen	as	disease	controls.	
	conj	N	N	v	prep	N	N	
2176	of	C-reactive	protein	for	cardiovascular	disease	detection	
	prep		N	prep	adj	N	N	

4791	middle-	income	countries	is altering	disease	patterns	and	vector
	adj	N	N	v	N	N	conj	N
4903	associated	with	multiple	disease	phenotypes	cor- relate	with	
	adj	prep	adj	N	N	V	prep	
4904	profiles	vary	with	inflammatory	bowel	disease	phenotypes.	
	N	V	prep	adj	N	N	N	

5110	as	having	strong	associa- tions	with	`disease`	progression
	conj	v	adj	N	(n)prep	N	N
5181	at	distinct	levels	of	the	`disease`	progression,
	prep	adj	N	prep	det	N	N
5182	in	Ab42	concen- trations	with	`disease`	progression,	mimics
	prep	N	N	prep	N	N	N

5710	with	radiographic	`disease`	severity [32, 33],	as	measured	
	prep	adj	N	N			
5766	the	`disease`	-severity	subgroups,	although	no	significant
	det	N	N	N	conj	adv	adj

5767	across	the	spectrum	of	diabetic	`disease`	severity	that
	prep	det	N	prep	adj	N	N	
5936	oxytocin	was	possible	during	early	`disease`	stages	and
	N	v	adj	prep	adj	N	N	conj
6358		and	reducing	the	rate	of	`disease`	transmission is
	conj	v	det	N	prep	N	N	v

6.3.5 DISEASES/LEFT (Collection B) 194

615	Though	he	remembered	urological	and	heart	`diseases`,
	conj	pron	V	adj	conj	N	N
1484	For	example,	Graves'	and	Addison's	`diseases`	are correlated
	prep	N	N	conj	N	N	v
1941	Distribution	of	the	reported	virus	`diseases`	in
	N	prep	det	adj	N	N	prep

6.3.6 DISEASES RIGHT (Collection A) 10

1632	command	system	to	get	`diseases`	surveillance	information
	N	N	prep	V	N	N	N

6.3.7 DISEASES LEFT (Collection B) 147

201	Inflammatory	bowel	diseases	and	human	reproduction:		
	adj	N	N	conj	adj	N		
202	emerging	links	with	inflammatory	bowel	diseases	and	
	v	N	prep	adj	N	N	conj	

214	imbal- ances	in	human	inflammatory	bowel	diseases.	
	adj	prep	adj	adj	N	N	
215		in	patients	with	inflammatory	bowel	diseases:
		prep	N	prep	adj	N	N

451	criteria	included;	evidence	of	other	eye	diseases,	
	N	adj	N	prep	adj	N	N	
452	diseases,	lens	diseases,	and	other	eye	diseases	that
	N	N	N	conj	det	N	N	

517	The	study,	however,	mixes	heart	diseases	acquired	in	
	det	N	adv	adj	N	N	V		
1222	in	patients	who	had	malignant	liver	diseases.		
	(n) prep	N	pron	V	adj	N	N		
1318	safe	and	beneficial	for	several	muscle	diseases,	but,	
	adj	conj	adj	prep	det	N	N	conj	

6.3.8 PATIENTS/ RIGHT (Collection A) 104

1128	During	routine	clinical	follow-up,	the	patients'	data
	prep	adj	adv	V prep	det	N	N
1132	out	based on	the	patients'	declaration,	and	therefore
	prep	V prep	det	N	N		

6.3.9 PATIENTS /LEFT (Collection A) 608

34	Many	ACHD	patients	have altered	cardiovascular	physiology	due	
	det	N	N	v				
66	Genomic	profiling	of	lung	adenocarcinoma	patients	reveals	
	adj	N	(n)prep	N	N	N	V	
585	not	all	resveratrol	arm	patients	showed	an	increase
	adv	det	N	N	N	V	det	N

644	chronic	hepatitis B	patients	with	any	hepatitis B	virus DNA
	adj	N	N	prep	det	N	N
765	medicine	use	among	cancer	patients	at	the end
	N	v	prep	N	N	prep	det N
818	two thirds	of	lung	cancer	patients	were seen	in
	num num	prep	N	N	N	v	prep

846	virus	was	successfully	isolated	from 34	case	patients
	N	v	adv	adv	prep num	N	N
878	induction	of	remission	in active	CD	patients	is
	N	prep	N	prep adj	N	N	v
936	and	reduces	protein	excretion	in CKD	patients.	
	conj	v	N	N	(n) N	N	

1113	of	cardiac	abnormalities	in	neonatal	diabetes	patients,
	prep	adj	N	(n)prep	N	N	N
1228	treating	complex	chronic	Lyme	disease	patients	that
	v	adj	adj	N	N	N	pron
1794	quality	of	care	in	hip	fracture	patients:
	N	(n) prep	N	prep	N	N	N

1882	796	young	femoral	neck	fracture	patients	were treated
	num	adj	adj	N	N	N	V
3311	1	mutations	have been reported	in	LGMD	patients	from
	num	N	V	prep	N	N	prep
4850	cluster	analysis	in	untreated	PD	patients,	although
	N	N	(n)prep	adj	N	N	conj

4972	blood	cells	from	rheumatoid	arthritis (RA)	patients	represents
	N	N	prep	N	N	N	v
5083	metabolism	in	high-risk	patients	using	HPLC	analysis
	N	prep	adj N	N	v	N	N
5427	In	Patna,	61%	of	TB	patients	first
	prep	N	num	prep	N	N	num

5542	cause	of	death	in	elective	THA	patients	in Finland.
	N	(n)prep	N	prep	adj	N	N	(n)prep N
6160	In	summary,	we	found	that	TKA	patients	were
	prep	N	pron	v	det	N	N	v

6.3.10 PATIENTS/RIGHT (Collection B) 156

4655	should consider	and	ask	about	patients'	need	for	spiritual
	V	conj	v	prep	N	N	prep	adj
5194	relationships	exemplify	how	MSM	patients'	perceptions	about	
	N	V	adv	N	N	N	prep	
5238	parameters	between	the	patients'	populations	with		
	N	prep	det	N	N	prep		

6.3.11 PATIENTS/ LEFT (Collection B) 2219

85	concentration	was	significantly	reduced	in	AD	patients
	N	v	adv	V	prep	N	N
630	prognosis	and	management	of	peripheral	artery	disease
	N	conj	N	prep	adj	N	N
631	Coronary	artery	disease	in	police	officers	participating
	N	N	N	(n) prep	N	N	V

816	outstanding	questions	in	the	biodiversity–	disease	literature
	adj	N	prep	det	N	N	N
1060	calculated	different	inflammatory	bowel	disease	risk	scores
	adj	adj	adj	N	N	N	N

1129	The	prioritization	candidate	disease	gene	function	is denoted	
	det	N	N	N	N	N	v	
1777	for	drug	repositioning	to	express	drug-	disease	as-sociations
	prep	N	v		V	N	N	N

3436	hip	dysplasia	is considered	a	joint	disease	with	insufficient
	N	N	v	det	N	N	prep	adj
5423	man	with	typical	Parkinson's	disease	and	mild	mental
	N	prep	adj	N	N	conj	adj	adj

5424	in	mild	and	moderate	Alzheimer's	`disease`	and	
	prep	adj	conj	adj	N	N	conj	
5425	for	autoimmune	diseases,	including	Crohn's	`disease`	and	
	prep	adj	N	(n) prep	N	N	conj	

5753	The	diagnostic	concept	of	Parkinson's	`disease`	is changing	
	det	adj	N	(n) prep	N	N	V	
5754	colonic	Crohn's	`disease`	is	intermediate	between	ileal	Crohn's
	adj	N	N	v	adj	prep	N	N
6322	The	bipartite	network	of	symptom-	`disease`	interactions	
	det	adj	N	(n) prep	N	N	N	

6323	so	far.	Symptom-	`disease`	network	reasoning	
			N	N	N		
6751	Tumors	from	melanoma	`patients`	in	the	vemurafenib
	N	prep	N	N	prep	det	N
10309	systemic	lupus	erythematosus (SLE)	`patients`	is	relatively	high,
	adj	N	N	N	V	adv	adj

10480	Among	ischemic	stroke	`patients`	who	were admitted	within	
	prep	adj	N	N	pron	v	prep	
12166	problems	faced	by	tuberculosis	`patients`	in	the	
	N	adj	prep	N	N	prep	det	
12293	We	also	excluded	UC	`patients`	who	had undergone	
	pron	adv	V	N	N	pron	v	

6.3.12 PATIENT /RIGHT (Collection B) 638

282	on	modern	medicine	and	patient	care	outcomes.	
	prep	adj	N	conj	N	N	N	
331	The	virtual	patient	case	comprised	15–30	pages	in
	det	adj	N	N	V	num	N	prep
432	Diarrhoeal	patient	charts	that	are diagnosed	and	treated	
	adj	N	N	pron	v	conj	V	

560	Patient	data	were gathered	from	published	data,	
	N	N	v	prep	adj	N	
656	fluctuations,	patient	diaries	have	problems	with	
	N	N	N	V	N	prep	
749	patients	correlate	with	lower	patient	expectations	regarding
	N	V	prep	adj	N	N	V

854	quantitative	analysis	(AngioTool)	in	two	patient	groups,	
	adj	N		prep	num	N	N	
949	consisted	of	patient	history,	results	from	a	clinical
	V	(n) prep	N	N	V	prep	det	adj
1034	enrollment	using	structured	patient	interviews	and	chart	review.
	N	V	adj	N	N	conj	N	N

1320	may shed	viral	particles	into	the	patient	plasma
	v	adj	N	prep	det	N	N
1340	disease	activity	in	two	kinds	of	patient	population,
	N	N	prep	num	adj	prep	N	N

1568	in	accordance	with	the	patient's	baseline	characteristics,
	prep	det		N	N	N	
1576	prosthesis	and	the	patient's	body	surface	area (BSA).
	N	conj	det	N	N	N	
1634	travel	to	the	patient's	home	and	collect
	V	prep	det	N	N	conj	V

1649	fatigue	and	EDS	on	XVIII	patient's	life	is	different.
	N	conj	N	prep	N	N	V	adj	
1792	Patient	satisfaction	improved	significantly	in	the	active		
	N	N	V	adv	prep	det			
1831	There	were	no	patient	selection	or	matching		
	det	V	adv	N	N	conj	N		

6.3.13 PATIENT /LEFT (Collection B) 86

446	expression	and	gastric	cancer	`patient`	relapse-free			.
	N	conj	adj	N	N	N adj			
1021	single	melanoma	`patient`	with a	negative	family	cancer	history,	
	adj	N	N	prep det	adj	N	N	N	
1505	prosthesis-	`patient`	mismatch	and	prosthetic	stenosis			
	N	N	adj	conj	adj	N			

6.3.14 PATIENT / RIGHT (Collection A) 672

220	data	in order	to	improve	patient	care	and	care	
	N	prep N	prep	V	N	N	conj	N	
593	and	mortality	are based	on	patient	cohorts	from	North America.	
	conj	N	v	prep	N	N	prep	N	

816	retaining	the	patient	–doctor	interface		
	N	det	N	N	N		
955	profiles	of	the	four	patient	groups	were compared
	N	prep	det	num	N	N	v

1052	Comply	with	positive	patient	identification	processes/	protocols.	
	V	prep	adj	N	N	N		
1126	the	rich	data	collected	from	the	patient	interviews,
	det	adj	N	V	prep	det	N	N

1189	another	layer	of	com- plexity	to	patient	management.	
	det	N	(n)prep	N	prep	N	N	
1261	retrospective	in	nature,	and	the	patient	numbers	
	adj	prep	N	conj	det	N	N	

1435	pathology;	regional	diversity	of	patient	population;	
	N	adj	N	(n) prep	N	N	
1595	Activation	rates	sorted	by	self-declared	patient	race.
	N	N	V	prep	adj	N	N

1648	in	patient	recruitment	and	assembling	phenotypic	data.
	(n)prep	N	N	conj	V	adj	N
1701	learning	Meaningful	physician–	patient	relationships	and	
		adj	N	N	N		

1819	Students	act	patient	roles	based	on			
	N	V	N	N	V	prep			
1886	as	a	patient	or	a	patient's	family	member	[10].
	prep	det	N	conj	det	N	N	N	

1911	newly	diagnosed	during	the	patient's	hospital	stay.
	adv	V	prep	det	N	N	N
2095	surgery	group	showed	better	patient	satisfaction	
	N	N	V	adj	N	N	

6.3.15 PATIENT /LEFT (Collection A)30

1052	The	index	patient	was diagnosed due	to	idiopathic		
	det	N	N	v	prep	adj		
1053	muscle	biopsy	of	the	index	patient	with	LGMD2J
	N	N	prep	det	N	N	prep	N
1588	important	components	of	the	physician–	patient	relationship,	
	adj	N	prep	det	N	N	N	

6.3.16 STUDY / RIGHT (Collection A) 669

828	The	study	characteristics	and	methodological	quality		
	det	N	N	conj	adj	N		
871	included	in	this	study	cohort	attended	child	welfare
	V	prep	det	N	N	V	N	N

1111	role	in	study	design,	data	collection		
	N	(n) prep	N	N	N	N		
1197	The	study	design	as	a	retrospective	cohort	study
	det	N	N	prep	det	adj	N	N

1361	A	study	nurse	prepared	the	`study`	drug	and	
	det	N	N	V	det	N	N	conj	
1800	patients	from	the	`study`	group	received	sodium		
	N	prep	det	N	N	V	N		

3377	The	mean	follow-up	for	`study`	participants	was	15.1	
	det	N	V prep	prep	N	N	V	num	
3515	During	the	`study`	period,	the	province	of		
	prep	det	N	N					
3610	Our	`study`	population	consisted	of	patients	aged	60	
	det	N	N	V	prep	N	adj	num	

3732	All	participants	were informed	of	the	`study`	process	and	
	det	N	v	prep	det	N	N	conj	
3758	in	accordance	with	our	primary	`study`	protocol	we	
	prep	N	prep	det	adj	N	N	pron	
4017	The	`study`	sample	size	was	small;	however,	to	
	det	N	N	N	V	adj			

6.3.17 STUDY / LEFT (Collection A) 480

341	A	genome-wide	association	`study`	identifies	four	
	det	N adj	N	N	V	num	
499	From	the	case	`study`	it	can	also
	prep	det	N	N	pron	v	adv

581	In	summary,	this	nationwide	cohort	`study`	based	on
	prep	N	det	adj	N	N	V	prep
671	A	prospective	cohort	`study`	of			
	det	adj	N	N	(n)prep			

777	Additionally,	the	case–control	`study`	including	60	patients	
	adv	det	N	N	V	num	N	
2391	joint	hypermobility:	a	population	`study`	of	female	twins.
	N	N	det	N	N	prep	adj	N

6.3.18 STUDIES / LEFT (Collection B) 162

180	is	consistent	with	results	from	animal	`studies`.
	v	prep	N	prep	N	N	
251	in	genome-wide	association	`studies`	late-onset	AD14–19,	
	prep	adj	N	N	adj	N	

504	results	in	large	cohort	`studies`	or	randomized .	
	N	prep	adj	N	N	conj	adj	
517	for	supporting	the	high-risk	cohort	`studies`	in	Kilifi,
	prep	V	det	adj	N	N	(n) prep	N

1447	subcutaneous	immunization	`studies`	described	below			
	adj	N	N	V	prep			
1634	The	authors	rec- ommended	intervention	`studies`	in	future	
	det	N	v	N	N	prep	adj	
3235	in	timely	dissemination	of	research	`studies`	during	the
	prep	adj	N	(n) prep	N	N	prep	

4195	However,	in	validation	`studies`	from	independent	cohorts,
	adv	(n)prep	N	N	prep	adj	N

6.3.19 STUDIES / RIGHT (Collection A) 2

364	for	inclusion	in	general	`studies`	curriculums	offered
	(n) prep	N	prep	adj	N	N	V

6.3.20 STUDIES / LEFT (Collection A) 154

136		Autopsy	studies	should be able	to	confirm	or
		N	N	v	prep	V	
137	autopsy	studies	which	greatly	influ- enced		
	N	N	pron	adv	adj		

204	The	results	from	the	case	studies	presented
	det	N	prep	det	N	N	V
205	The	following	case	studies	provide	evidence	as
	det	prep	N	N	V	N	prep
280	including	population-	based	cohort	studies,	the	standardized
	V	N	adj	N	N	det	adj

281	the	qual- ity	of	nonrandomized	cohort	studies	
	det	N	prep	adj	N	N	
370	observational	Alzheimer's	disease	studies	offer	similar	
	adj	N	N	N	V	adj	
371	new	class	of	approaches	to	disease	studies
	adj	N	prep	N	prep	N	N

821	Careful	laboratory	studies	on	various	fishes	have
	adj	N	N	prep	adj	N	v
822	in	methodology	between	the	two	laboratory	studies,
	prep	N	prep	det	num	N	N
1925	experiments	and	some	clinical	trial	studies.	
	N	conj	det	adj	N	N	

6.3.21 STUDY / RIGHT (Collection B) 669

750	Plasma	samples	from	all	study	animals	in
	N	N	prep	det	N	N	prep
1034	vulnificus	isolates	in	the	study	area	indicated
		V	prep	det	N	N	V

2156	and	during	study	days	4 to 7	.			
	conj	prep	N	N	num prep num				
2318	this	study	include	the	randomized	study	design	and	
	det	N	V	det	adj	N	N	conj	
2591	The	study	drug	was provided	by	Gilead Sciences.			
	det	N	N	v	prep	N			

3036	were included.	The	study	flow	chart ,				
		det	N	N	N				
3433	The	baseline	characteris- tics	of	the	study	groups	are shown	
	det	N	N	prep	det	N	N	v	
5167	took	study	medication	for	<3	days	and		
	V	N	N	prep	num	N	conj		

6394	The	PD	study	participants	had	an	asymmetric	bradykinesia
	det	N	N	N	v	det	adj	N

7542	common	na- tive	language	in	the	study	settings,
	adj	adj	N	prep	det	N	N
7938	Recruitment	of	study	subjects			
	N	prep	N	N			

6.3.22 STUDY / LEFT (Collection B) 480

582	of	epilepsy	in	the	health	areas	`study`,	
	prep	N	prep	det	N	N	N	
821	cognitive	assessment	in	the	Bruneck	`Study`	cohort	
	adj	N	prep	det	N	N	N	
899	A	case	`study`	reported	the	effects	of a	6-week
	det	N	N	v	det	n	prep	det num

900	the	baseline	images	of	the	case	`study`	representing
	det	adj	N	prep	det	N	N	V
1062	lupus	erythematosus:	a	prospective	cohort	`study`,		
			det	adj	N	N		

1063	nationwide	cohort	`study`	based	on	national	administrative
	adj	N	N	V	prep	adj	adj
2148	this	multi-system	feasibility	`study`	to	exclusively	focus
	det	adj	N	N	prep	adv	V

4519	those	found	in	another	general	population	`study`	in
	pron	V	prep	det	adj	N	N	prep
5684	Clinical	trial	and	research	`study`	participants		
	adj	N	conj	N	N	N		
5685	research	`study`	participants	should be				
	N	N	N	v				

6.3.23 ANALYSES / LEFT (Collection A) 68

96	The	data	`analyses`	were conducted	using SAS	version.
	det	N	N	v		N

6.3.24 ANALYSIS / LEFT (Collection A) 100

599	For	‘decision	`analysis’`,	candidates	were presented	
	prep	N	N			
715	influenza	vaccine	effectiveness	`analysis`	in	2014-2015
	N	N	N	N	prep	Num
818	biol- ogy	obtained	via	gene	expression	`analysis`.
	N	v	prep	N	N	N

6.3.25 ANALYSIS / RIGHT (Collection B) 64

3957	De-identified	data	and	`analysis`	pipeline	may be shared	
	adj	N	conj	N	N	v	
4495	ATLAS-Ti v7	qualitative	data	`analysis`	software		
		adj	N	N	N		

6.3.26 ANALYSIS / LEFT (Collection B) 1476

458	STAT3	was measured	by	Western	blot	`analysis`.
	N	v	prep	adj	N	N
797	Table 5:	Pearson	correlation	`analysis`	of	relationship
		N	N	N	prep	N

839	variables	from	the	univariate	Cox	`analysis`	were
	N	prep	det	adj	N	N	v
881	The	ROC	curve	`analysis`	was performed	only	for
	det	N	N	N	v	adv	prep

1117	the	ATLAS-Ti v7	qualitative	data	`analysis`	software	
	det	N	adj	N	N	N	
1171	Preparation,	sequencing	and	data	`analysis`	were performed	as
	N	N	conj	N	N	v	prep

1301	diagnosing	late-onset	diseases	by	DNA	`analysis`,	
	V	adj	N	(n) prep	N	N	
1342	NOTE:	Weights	are	from	random	effects	`analysis`
		N	V	prep	adj	N	N

1486	exploratory	factor	`analysis`	was conducted	to	extract	the
	N	N		v	prep	v	det
1917	The	multiple	logistic	re- gression	`analysis`	with	age
	det	adj	adj	N	N	prep	N
2299	for	systematic	literature	`analysis`	have been	strictly	
	prep	adj	N	N	v	adv	

3068	“Genome-wide	promoter	methylation	`analysis`	identifies	epigenetic	
	adj	N	N	N	V	adj	
3125	of	cell	migration	`analysis`	on	laser	machined
	prep	N	N	N	prep	N	adj

3612	criteria,	through	photomicrographs	`analysis` XXX	under:	
	N	prep	N	N	prep	
4403	and	chi-square	`analysis`	was used	to	determine
	conj	N	N	v	prep	V

6.3.27 ANALYSES / LEFT (Collection B) 169

137	Canonical	correlation	`analyses`	showing	fungal	and
	adj	N	N	V	adj	conj
138	rank	correlation	was used	for	correlation	`analyses`.
	N	N	V	prep	N	N

212	Gene	expression	`analyses`	of	cells	with	high	brain
	N	N	N	(n)prep	N	prep	adj	N
632	Finally,	the	multiple	regression	`analyses`	indicated	that	
	adv	det	adj	N	N	V	det	

6.3.28 DATA / RIGHT (Collection A)845

70	`Data`	acquisition	was performed	with	a	data	
	N	N	v	prep	det	N	
155	The	`data`	analyses	were conducted	using	SAS	version
	det	N	N	v			

906	patients	using	the	National	Trauma	`Data`	Bank.	
	N	V	det	N	N	N	N	
1090	data	sets	for	its	latent	`data`	characteristics	extraction
	N	V	prep	det	adj	N	N	N

1438	Normality	of	the	`data`	distributions	was tested using			
	N	prep	det	N	N	v			
1545	extraction	sheet	(based	on	the	Cochrane	`data`	extraction	
			V	prep	det	N	N	N	

2552	were	some	`data`	limitations	and	necessary	assumptions		
	V	det	N	N	conj	adj	N		
2619	data	manager	in a	web-based	`data`	management	system	in order	
	N	N	prep det	adj	N	N	N	prep N	

2700	literature	search- ing,	`data`	mining	from	the	EHR	and
			N	N	prep	det	N	conj
3221	`Data`	points	collected	(deidentified)	were	age,		
	N	N	adj		V	N		

6.3.29 DATA LEFT (Collection A) 227

361	Disease- gene	association	`data`	were downloaded	via			
	N	N	N	V	prep			
1204	the	PEER	model	to	gene	expression	`data`	with
	det	N	N	prep	N	N	N	prep

1659	independent	of	input	`data`	label	distributions	
	adj	prep	N	N	N	N	
1789	for	billing	and	laboratory	`data`	were calculated	using
			conj	N	N	v	

2666	performance	outcome	`data`	to improve	the	program's	quality.	
	N	N	N	prep V	det	N	N	
1022	WMH	and	other	brain	`pathology`	using	positron	emission
		conj	det	N	N	V		

6.3.30 DATA / RIGHT (Collection B) 529

333	Data	analysis	showed	a	statistically	significant			
	N	N	V	det	adj	N			
1673	purpose	of	study	and	procedures	of	data	collection	
	N	prep	N	conj	N	prep	N	N	

1707	Data	collection	of	soft	tissue	and	microbiological		
	N	N	prep	adj	N	conj			
3357	were	reviewed	by	the	data	management	team	and	
	V	prep	det	N	N	N	conj		

6.3.31 DATA / LEFT (Collection B) 104

604	Ocular	baseline	data	including	visual	acuity,	intraocular	
	adj	N	N	V	N	N		
703	All	case	data	used	in	the	analysis	are
	det	N	N	V	prep	det	N	v
1117	(based	on	our	initial	growth	curve	data)	and
	V	prep	pron	adj	N	N	N	conj

1901	for the	supplement	of	the	health	data	in	Cambodia.	
	prep det	N	prep	det	N	N	prep	N	
2008	knee	imaging	data	set	from	the	Osteoarthritis Initiative (OAI),		
	N	N	N	V	prep	det	N N		

2243	hydration	requires	the	application	of	laboratory	data
	N	V	det	N	(n)prep	N	N
2549	forms	were fitted	to	the	mouse	data	(for
	N	v	prep	det	N	N	

6.3.32 RESEARCH / RIGHT (Collection A) 710

43	primarily	involved	in	clinical	or	`research`	activity.
	adv	V	prep	adj	conj	N	N

2308	the	critical	evaluation	of	`research`	results.	
	det	adj	N	(n) prep	N	N	
2712	to	the	`research`	world	of	mechanisms	of
	prep	det	N	N	prep	N	prep

6.3.33 RESEARCH /LEFT (Collection A) 208

975	The	disease	gene	`research`	will open	a	new	era
	Det	N	N	N	v	det	adj	N
1952	increased	as	the	faculty's	`research`	and	clinical	abilities
	adj	prep	det	N	N	conj	adj	N
2158	stu- dents	for	a	10-week	summer	`research`	fellowship.	
	N	prep	det	num N	N	N	N	

6.3.34 TIME / RIGHT (Collection A) 2235

1059	and	differences	in	definitions	of the	`time`	intervals	used
	conj	N	(n) prep	N	prep det	N	N	v
1524	compared	to	the	pilot	`time`	period	analyzed,	
	V	prep	det	N	N	N	V	
1669	the	knee	joint	at	dif- ferent	`time`	-points,	comparable
	det	N	N	prep	adj	N	N	adj

6.3.35 TIME /LEFT (Collection B) 75

531	as an	opportunity	to	minimize	face-	`time`	bias	
	prep det	N	prep	v	N	N		
1415	give	patients	`time`	to	adjust	to	their	new
	v	N	N	prep	V	prep	pron	adj

6.3.36 PATHOLOGY / LEFT (Collection A) 105

1431	to	investigate	patterns	of	disease	`pathology`	within	
	prep	V	N	(n)prep	N	N		
1432	Infectious	Diseases	`Pathology`	and	pathogenesis	of	pulmonary	tuberculosis.
	adj	N	N	conj	N	(n) prep	N	N

6.3.37 PATHOLOGY / RIGHT (Collection A) 1531

1081	management	section,	this	`pathology`	consultation	requires		
	N	N	det	N	N	V		
1224	introducing	such	`pathology`	coursework	at	grade	levels	
	V	det	N	N	prep	N	N	
1400	Chairs	of	academic	`pathology`	departments	include	basic	
	N	prep	adj	N	N	V		

1875	have been	intertwined	in	`pathology`	graduate	medical			
		v	prep	N	N	adj			
2193		the	access	to	and	safety	of	`pathology`	information.
		det	N	prep	conj	N	prep	N	N

2445	publishable	in the	general	medical	or	`pathology`	literature.
	adj	prep det	adj	N	conj	N	N
2719	Other	`pathology`	organizations	followed	with	programs	in
	det	N	N	V	prep	N	prep
2876	Exposure	to	`pathology`	practice	can counter	negative	stereo- types
	N	prep	N	N	v	adj	N

2932	Telemedicine	Program	and	the	Department	of	`Pathology`	programs
	N	N	conj	det	N (n)	prep	N	N
3019	explain	the	results	of	a	`pathology`	report	to
	V	det	N	prep	det	N	V	prep

3227	improving	forensic	pathology	education	in	`pathology`	residency.	
	V	adj	N	N	(n) prep	N	N	
3729	restructured	to	include	clinical	exposure	to	`pathology`	services,
		prep	V	adj	N	prep	N	N

6.3.38 HEALTH / RIGHT (Collection A) 2270

912	economic	problems	due	to	patient'	**health**	care.	
	adj	N	V	prep	N	N		
1150	increased	stu- dent	interest	in	professional	**health**	careers.	
	adj	N	N	(n) prep	N	N	N	
1490	medical	teams	respond	to	international	**health**	emergencies.	
	adj	N	V	prep	adj	N	N	
180	However	performance	of	only	some	**health**	indicators	was
	conj	N	prep	adj	det	N	N	V
2172	be	important	to	solve	the	**health**	manpower	crisis
	V	adj	prep	V	det	N	N	N

2732	and	are regarded	as	a	serious	**health**	problem.	
	conj	v	prep	det	adj	N	N	
2983	In	the	analysis	of	effect	modification,	**health**	regions
	prep	det	N	prep	N	N	N	N

3134	grade	4 which	is	an	"unacceptable	**health**	risk."
		num pron	V	det	adj	N	N
3583	a	situational	approach	to	**health**	status	deviations.
	det	adv	V	prep	N	N	N

3654	pathologist	can have	on	patients,	a	**health**	system	as	
	N	v	prep	N	det	N	N		
4031	The	shortage	of	trained	**health**	workforce	in	smaller	
	det	N	prep	adj	N	N	prep	adj	

6.3.39 HEALTH / LEFT (Collection A) 113

2433	less	likely	to	utilize	their	patient	`health`	portals,
	adj	adj	prep	v	det	N	N	N
2493	new	roles	involving	population	`health`	management,	increased	
	adj	N	V	N	N	N		

3217	Ebola	epidemic	in	Sierra Leone's	`health`	system:	a	
	N	adj	prep	N	N	N		
3218	personal	or	a	family	member's	`health`	conditions.	
		conj	det	N	N	N	N	

6.3.40 TABLE (Collection A) 0

555	20:755	Page	8 of 15	`Table` 2	Correlation	between	medical	
558	both	BMD	and	clinical	severity	(Table 4).	`TABLE` 3.	

CELLS / RIGHT (Collection B) 18

6697	potential	link	between	this	particular	T `cells`	population	
	adj	N	prep	det	adj	N	N	
6698	expansion	of	BMSCs	yields	homogenous	`cells`	population	for
	N	prep	N	V	adj	N	N	prep

6.3.41 CELLS / LEFT (Collection B) 1665

439	the	genetic	loss	of	B `cells`	alone	or	
	det	adj	N	prep	N N	adj	conj	
766	DNA	was extracted	from	peripheral	blood	`cells`	following	standard
	N	v	prep	adj	N	N	v	adj

2965	cells	and	their	conversion	to	foam	`cells`	is
	N	conj	det	N	prep	N	N	v
3137	of	cells	with	functions	resembling	germ	`cells`.	
	prep	N	prep	N	v	N	N	
3430	survival	signals	to	HCC	`cells`	that	limit	
	N	N	prep	N	N			

6740	unbalanced	when	comparing	control	and	patient	`cells`.	
	adj	pron	V	N	conj	N	N	
10662	and	cytotoxic	effects	in	human	tumour	`cells`	independently
	conj	adj	N	prep	adj	N	N	adj

6.3.42 CELL / LEFT (Collection B) 908

1714	Establishment	of a	human	lung	cancer	`cell`	line	
	N	prep det	adj	N	N	N		
3347	stimulation	to apoptotic	heart	`cell`	death	through	protein	kinase
	N	prep adj	N	N	N	conj	N	N

3418	B virus	X protein	promotes	hepatoma	`cell`	proliferation		
	N	N	V	N	N	N		
3434	Analysis	of hES	`cell`	phosphorylation	dynamics	during		
	N	prep N	N	N	N	prep		
3472	of	positive	controls	of	known	host-	`cell`	DNA
	prep	adj	N	prep	adj	N	N	N

4961	PBMCs	with	or	without	CD33+	myeloid	`cell`	depletion.
	N	prep	conj	prep	N	N	N	N
6197	B cell	activation,	expansion,	and	plasma	`cell`	generation	
	N	N	N	conj	N	N N	N	

6732	In d,	Schwann	`cell`	diameters	were measured.			
	prep	N	N	N	v			
7103	There	was	one	case	of spindle	`cell`	sarcoma,	diagnosed
	det	V	det	N	prep N	N	N	adj
9144	we	developed	M¨uller	`cell`	lines	in	vitro	and
	pron	v	N	N	N	prep	N	conj

6.3.43 CELL / RIGHT (Collection B) 4746

14	Photo-inducible	cell	ablation	in	Caenorhabditis	elegans		
	adj	N	N	prep	N			
251	cytometric	methods	to	a	single	cell	analysis	of
	adj	N	prep	det	adj	N	N	prep

566	A	master	cell	bank	was established	and	stored	
	det	adj	N	N	v	conj		
1480		permanent	sections	showed	malignant	squamous	cell	carcinoma.
		adj	N	V	adj	adj	N	N

1666	"Utility	of	ganglion	cell	complex	analysis	in
	N	(n) prep	N	N	adj	N	prep
2864	of	pyroptotic	cell	death	in	activated	macrophages
	prep	adj	N	N	prep	adj	N

3110	Toward	col- orimetric	cancer	cell	detection	and	targeted
	prep	adj	N	N	N	conj	
3188	"Vascular	smooth	muscle	cell	differentiation	from	human
	adj	adj	N	N	N	prep	

4310	After	incubation,	the	bacterial	cell	growth	on	the
	prep	N	det	adj	N	N	prep	det
5812	Expression	of	stem cell	and	differentiated	cell	marker	genes
	N	prep	N	conj	adj	N	N	N

5944	Identification	of	a	lipase-linked	cell	membrane	receptor
	N	prep	det	adj	N	N	N
6289	To	determine	whether	the	increased	cell	motility
	prep	V	conj	det	adj	N	N
7539	normal	functioning	by	using	the	cell's	machinery
	adj	N	prep	V	det	N	N

6.3.44 TREATMENT / LEFT (Collection B)358

1098	as	a	molecular	target	of	cancer	`treatment`,		
	prep	det	adj	N	prep	N	N		
1658	EVD	infection	were reported	in	Ebola	`treatment`	centers (ETCs)		
	N	N	V	prep	N	N	N		
2501	efficacy	of	YAG	laser	`treatment`	was achieved	depending		
	N	prep	N	N	N	V			
2515	Responders	to	LCIG	`treatment`	were observed	across	the		
	N	prep	N	N	V	prep	det		
2711	irradiation	power	and	time	of	microwave	`treatment`	should	
	N	N	conj	N	prep	N	N	V	
2748	(not	including	the	6-month	`treatment`	gap	period)		
		V	det	num N	N	N	N		
2783	after	2	years	of	natalizumab	`treatment`,	neurologists	in	
		num	N	(n) prep	N	N	N		

6.3.45 TREATMENT / RIGHT (Collection B) 691

64	role	of	self-efficacy	in	HIV	`treatment`	adherence:
	N	(n)prep	N	(n)prep	N	N	N
468	models	were used	to	fit	each	`treatment`	arm
	N	v	prep	V	det	N	N
595	The	`treatment`	beliefs	examined	in	the	literature
	det	N	N	V	prep	det	N

596	Exactly	how	the	antibiotic	`treatment`	benefits	this	
	adv	adv	det	adj	N	N	det	
727	HIV	infection	at a	support	and	`treatment`	centre	in
	N	N	prep det	N	conj	N	N	(n) prep
917	to	advance	clinical	pathology	and	inform	`treatment`	decisions,
	prep	V	adj	N	conj	V	N	N

1194	improve	the	ability	to	identify	`treatment`	effects
	V	det	N	prep	V	N	N
1573	of	the	montelukast	`treatment`	group	and	45.1%
	prep	det	N	N	N	conj	num
1670	“Adherence	to	`treatment`	guidelines	for	acute	diarrhoea
	N	prep	N	N	prep	adj	N

3745	In	brief,	at	the	ending	of	`treatment`	period,
	prep	N	prep	det	N	prep	N	N
3811		indicate	significant	differences	between	`treatment`	populations:	
		V	adj	N	prep	N	N	

3951	for	nearly	all	patients	despite	aggressive	`treatment`	regimens.
	prep	adv	det	N	prep	adj	N	N
4288	with	the	combination	strategy	of	the	`treatment`	strategy.
	prep	det	N	N	prep	det	N	N

6.3.46 TREATMENTS / LEFT (Collection B)9

157	Drug	`treatments`	were	applied	next	day	after	transfection,
	N	N	v		adj	N	adj	N
158	and	other	drug	`treatments`,	G	allele	carriers	are
	conj	adj	N	N				

6.3.47 GROUP / LEFT (Collection B) 1378

422	a	positively	charged	amine	`group`	or	other	polar
	det	adv	V	N	N	conj	det	adj
695	evaluate	the	adjusted	effect	of	blood	`group`	on
		det	adj	N	prep	N	N	prep
863	that	of	chemotherapy	`group`,	without	significant	differ- ence.	
		prep	N	N				

948	the	cases	in	the	combination	`group`	received
	det	N	prep	det	N	N	V
2117	the	single	fraction	`group`,	but	this	was
	det	adj	N	N			
2603	age	of	mothers	in	the	intervention	`group`
		prep	N	prep	det	N	N

2788	In	the	active	treat- ment	`group`,	rTMS	was
	prep	det	adj	N	N		
3291	in	the	hard	pancreas	`group`	were	higher
	prep	det	adj	N	N	V	adj
3331	in	the	patient	`group`	with	audiological	impairments
	prep	det	N	N	prep	adj	N

3815	Lauritano's	`group`	found	two	species	belonging
	N	N	V	num	N	V

6.3.48 GROUP / RIGHT (Collection B) 77

912	Toll- like	receptor	4	and	high-mobility	`group`	box-1	
	adj	N	num	conj	adj	N	N num	
1469	literature	review	and	several	`group`	discussions	held	by
	N	N	conj	det	N	N	V	prep

6.3.49 GROUPS / LEFT (Collection B) 258

223	bonds	was due	to	the	amine	`groups`	of	proteins
	N	v	prep	det	N	N	(n)prep	N
427	The	distribution	of	blood	`groups`	were	as	
	det	N	(n) prep	N	N	v	prep	

659	be- tween	the	diabetes	and	control	`groups`	were not
	prep	det	N	conj	N	N	V
719	number	of	participants	in the	9 control	`groups`	ranged
	N	(n)prep	N	prep det	num N	N	V
1541	percentage	(26–30)	reported	by	other	research	`groups`.
	N		V	prep	adj	N	N

1648	lead	us	to	design	our	study	`groups`	with
	V	pron	prep	V	pron	N	N	prep
2075	were	broadly	similar	between	treatment	`groups`	at	
	V	adv	adj	prep	N	N	(n) prep	

6.3.50 RESULTS / LEFT (Collection B)163

129	present	analysis	`results`	obtained	from	the	ExonScan	
	adj	N	N	V	prep	det	N	
1217	We	gathered	laboratory	`results`	regarding	patients'	nadir	
	pron	V	N	N	V	N	N	
2313		The	research	`results`	were treated	by	methods	of
	det	N	N	v	prep	N	(n)prep	
2900	described	the	photoscreening	-based	vision	test	`results`	
	V	det	adj		N	N	N	

6.3.51 PROTEIN / RIGHT (Collection B) 1090

151	-promoting	is involved	in	protein	aggregation,	inclusion		
		v	(n) prep	N	N			
921	In	addition	to variations	in Htt	protein	concentrations		
	prep	N	prep N	(n)prep N	N	N		
1124	sample	to stop	further	protein	degradation,	and	samples	
		prep V	adj	N	N			

1554	The	subcellular	protein	frac- tion	that	enabled		
	det	adj	N	N				
1681	The	purpose	of	viral	envelope	protein	glycosylation	remains
	det	N	prep	adj	N	N	N	V
1861	Since	the	confidence	of	protein	identification	depends	
		det	N	(n)prep	N	N		
2810	since	a	purity-	adjusted	total	protein	measurement	
	conj	det	adj	adj		N	N	
2812	an	explanation	for	failed	protein	mechanisms	rather	
	det	N	prep	adj	N	N	adv	
3305	or	a	determinant	for	specific	protein-	protein.	
	conj	det	N	prep	adj	N	N	

6.3.52 PROTEIN / LEFT (Collection B) 266

909	phenotypes	determined	by	immune	cell	protein	expression.
	N	V	prep	adj	N	N	N
2196	Conserved	outer	membrane	protein	of	neisseria	meningitidis
	adj	adj	N	N	(n) prep	N	N
2346	distur- bance	of	muscle	protein	synthesis	and	reduced
	N	(n)prep	N	N	N	conj	adj

2924	protein	-protein	interaction	(PPI)	network	analysis		
	N	N	N	N	N	N		
3456	Water	absorption	capacity	of	pumpkin	seed	protein	
	N	N	N	(n) prep	N	N	N	
3723	role	of	oligomeric	forms	of	tau	protein	as
	N	prep	adj	N	(n)prep	N	N	prep