



Universidad de Valladolid

Facultad de Filosofía y Letras

Grado en Estudios Ingleses

**A Corpus-Based English/Spanish Glossary of
Pollen Allergy**

Hanane Er-Rouisse Moukrim

Tutora: Isabel Pizarro

Curso: 2023/2024

Departamento de Filología Inglesa

Abstract

In recent decades, terminology has experienced unprecedented significance, beginning to acquire relevance and, to this day, it continues developing, catching the attention of many professionals in the medical field. In this corpus-based study, we elaborated an English-Spanish glossary on pollen allergy due to the lack of availability of an English-Spanish glossary dealing with this topic. This was done by extracting terms from a corpus composed of 50 texts about pollen allergy. The main objective is to compile a glossary to provide access to these terms to experts or specialized translators. As a result, we elaborated a 227 English terms glossary of pollen allergy, with their respective Spanish equivalents, phonological transcriptions, parts of speech, definitions, synonyms, examples of use, fields, subfields and dates.

Keywords: Corpus, specialized language, terminology, pollen allergy, terminological entry, glossary.

Resumen

En las últimas décadas, la terminología ha tenido un auge sin precedentes, empezando a adquirir relevancia que hasta hoy sigue creciendo, atrayendo la atención de muchos profesionales del ámbito médico. En este estudio basado en corpus, hemos elaborado un glosario en inglés-español sobre la alergia al polen. Nuestra elección de abordar este tema es la falta de disponibilidad de un glosario en inglés-español que trate este tema. Esto se ha hecho extrayendo términos de un corpus que hemos elaborado compuesto por 50 textos que tratan sobre la alergia al polen. El objetivo es elaborar un glosario para facilitar el acceso a estos términos a expertos o traductores especializados. Como resultado, hemos elaborado un glosario de 227 términos en inglés sobre la alergia al polen, incluyendo sus respectivos equivalentes en español, transcripciones fonológicas, categorías gramaticales, definiciones, sinónimos, ejemplo de uso, campo, subcampo y fechas.

Palabras clave: corpus, lenguaje especializado, terminología, alergia al polen, entrada terminológica, glosario.

Table of Contents

1. Introduction	4
2. Theoretical Background	5
2.1. Specialized language	5
2.2. Terminology	7
2.3. Corpus	9
2.4. Glossary	10
3. Literature Review	11
4. Materials and Methodology	14
4.1. Corpus	14
4.1.1. Corpus compilation	14
4.1.2. Naming of files.....	15
4.1.3. Corpus Representativeness	16
4.2. Methodology.....	18
5. Results and Discussion	23
6. Conclusion	26
7. Bibliography	28
Annex A	32

1. Introduction

Medicine is considered one of the three oldest sciences of which we have evidence (Fischbach, 1998). Translation is, also, considered an old speciality (Hurtado, 2001). These two disciplines are linked to each other due to the need to make accessible medical knowledge in different languages worldwide. Medical terminology began taking its first steps in the specialized translation field at the end of the 19th century, and from that moment until nowadays, terminology work has become fundamental. Its significance lies in that it allows communication and collaboration across the world of this knowledge: with the translation of medical guidelines, research papers, and clinical studies, experts have access to this knowledge, share it, and facilitate the progression of medical science as well.

However, there is a lack of availability of glossaries in specialized medical subfields, such as pollen allergy, which is the one we are working with in this study, and due to this, we consider it necessary to create a useful tool that offers linguistic information related to these terms. We decided to build an effective reference tool, a bilingual glossary of pollen allergy, which facilitates access to linguistic information related to the terms of subject matter. This was done by using a corpus, from which we extracted the candidate terms. The decision to conduct this study using a corpus is because corpus-based studies present the advantages of containing naturally occurring language texts, which is what makes them more reliable; additionally, they are an efficient means for analysis using corpus tool analyses, thereby helping the task of creating glossaries.

The aim of this work, therefore, is to build an efficient glossary that facilitates access to the terms of pollen allergy. It will provide specific linguistic information, including their Spanish equivalents, phonological transcriptions, parts of speech, definitions, synonyms, examples of use, fields, subfields, and dates. The glossary is designed for adult specialist users, such as medical translators and specialists who require a certain type of linguistic information that we have included. It can also be used for teaching purposes.

This research is divided into two main sections. The first one is the theoretical background, which presents the main concepts of our study, and the second is the

practical part, which involves the compilation of the corpus, the process of extracting candidate terms extraction, and the building of the glossary.

2. Theoretical Background

To lay the foundations of this terminological study, we considered it necessary to define and analyse certain important notions that are part of it. These aspects are the following: specialized language, terminology, corpus, and glossary.

2.1. Specialized language

Specialized language has always attracted the attention of many scholars, as it is often linked to general language due to the two-way contact between them. Therefore, a terminographer needs to account for the differences between these two types of languages. For this, we are going to present a fundamental differentiation between these two language types, after which we will focus on the specialized one, since this is the subject of this study.

On the one hand, general language is the language that we use daily to talk and write about ordinary events in a variety of common situations (Orduña, 2011). General language texts are intended, therefore, to expose the receiver to a topic without requiring any prior knowledge; they are based on every day, non-specialized exchanges and do not use a technical language. On the other hand, specialized language refers to the language used in specific fields to communicate topics and ideas more effectively within their discourse communities. Specialized language differs from general language in its use of specific terminology and specialized expressions (Cabré, 1999). Examples of these types of languages are the language of the experimental sciences, economics, medicine, etc. (Gotti, 2004).

There is debate on whether specialized language is a variant of the general language. Some authors, such as Rondeau (1984), Rey (1995), and Quemada (1990) argue that specialized language is a variant of general language. Meanwhile, other authors, such as Varantola (2006), Picht and Draskau (1985), state that a specialized language acts as a subset of general language and is semiautonomous, meaning that it has characteristics of the general language and characteristics of the specialized language as well. In our case, we favour the second perspective. Our study focuses on

medical terminology, which is a field of specialized language, and specifically, it is about pollen allergy terms, a subfield within the medical one.

When it comes to stating the characteristics of a specialized language, we find no agreement among the different authors that discuss them. The different conceptions given by different authors: Cabré (1999), Gotti (2004), Rondeau (1984), Rey (1995), Quemada (1990), Varantola (2006), Picht and Draskau (1985), and Orduña (2011) influence their characterization. Consequently, we present below a series of characteristics and particularities of this type of language according to the different conceptions given by these authors. The main characteristic of a specialized language is its degree of technicality and specificity. Therefore, it requires knowledge of the subject, which is specific, as advanced technical language is used (Orduña, 2011). The communicative situations in which it is used are another element that characterizes this type of language. It is used in institutional and professional settings, such as hospitals, schools, businesses, universities, courts, etc., and its use takes place within a small discourse community.

To finish this section, we have considered all the characteristics we mentioned and elaborated a table that shows the differences between these two types of languages:

CRITERIA	GENERAL LANGUAGE	SPECIALIZED LANGUAGE
Genres	In mass media	In specialized writings
Function	Conative, emotional...	Mainly referential
User	General	Specialized
Communicative setting	Less formal	More formal
Audience	Lay people	Experts
Use	<ul style="list-style-type: none"> - Everyday use. - Write or speak about ordinary events in various common situations (ordering a meal, writing a 	<ul style="list-style-type: none"> - To develop scientific research papers - To communicate the objective and precise knowledge of an observed phenomenon

	letter to a friend, asking for instructions, etc.)	- Directed to an expert recipient in the content covered, etc.
--	--	--

Table 1. Differences between general language and specialized language

2.2. Terminology

Terminology is another key concept that we consider important to cope with. Terminology and lexicology are often linked to each other, and their differentiation is often a controversial issue. This controversy is because very often general language dictionaries include meanings of words that could be classified as terms, given that the concepts they designate belong to a specialized domain, such as *El Diccionario de la Lengua Española*, which includes terms like “crónico (chronic) or “transplante” (transplant), belonging to the medical specialized field.

On the one hand, lexicology refers to the study of words that are general, and not specific to any field. On the other hand, terminology refers to the study of terms that are associated with a specific area of specialized knowledge (Cabré, 1999).

For a better understanding of the concept of terminology, we present some definitions given by different authors below. The term is polysemic, and, as previously stated, our study focuses on terminology, thus, we believe it is important to develop its meaning. Sager (1990) and Zanón (2016) give three interpretations of this term. First, it can be understood as a discipline: a set of conceptual principles and bases whose object of study are the terms; second, it can be seen as a vocabulary of a specific subject: a set of terms of a specific field of speciality; and third, it can be interpreted as a methodology: a set of guidelines that are used in terminographic works.

Sager (1990) states that ‘terminology’ is the set of terms belonging to a specialized subject, such as medicine or chemistry, and these terms are codified in the form of vocabularies, glossaries, dictionaries, thesauruses, databases, etc. Zanón (2016) offers two definitions of this concept. The first one refers to a group of terms that belong to a specialized field, for example, chemistry or biology, which may appear in electronic or physical media: databases, dictionaries, glossaries, etc. Meanwhile, in his second definition, he considers ‘terminology’ as an activity, specifically the work of

terminologists. He considers terminology as an activity carried out by terminologists to solve problems that have to do with the use of certain terms. This also includes the creation of terminology for specialized areas through research methods, which are supported by a series of rules and terminological norms.

The work of terminologists is intended to ease and ensure the correct flow of information between specialists and professionals. In this sense, Cabré (1990) exposes the position that terminology borrows some concepts from other subjects. , that is, an interdisciplinary science that must also be considered transdisciplinary since there is no structured discipline that does not use terminology to communicate the specialized knowledge of its area of study (Cabré, 1999).

Regarding the words that are include in a Terminology, they are called terms and are used by experts and professionals in scientific or technological discourse on a specialized topic. These terms are usually presented in a glossary, dictionary or database (Sager, 1990). In our study, we are working with the terminology related to the area of pollen allergy and we will present it in a glossary.

After defining these two concepts, now we discuss whether terminology is an independent discipline or not, because there is debate on this. Although the practice of terminology has developed in parallel with the scientific advances of the last three centuries, the establishment of terminology as an independent science can be considered recent. An author who considers it to be independent is Sager (1990), although he recognizes that there are, and should be, theoretical bases underlying the terminographical practice. By contrast, other authors, such as Cabré (1999) or Santamaría (2006) consider terminology to be a dependent discipline. The last one defines terminology as an interdisciplinary science that supports a body of knowledge related to disciplines such as linguistics, the science of knowledge, communication sciences and information sciences. She considers terminology as an interdisciplinary science that holds across various disciplines such as computer science, linguistics, communication theory, etc. Cabré states that terminology, understood as the compilation, description, and presentation of terms related to a certain field, cannot be considered an activity that stands by itself (Cabré, 1999).

2.3. Corpus

Corpus is a term that is defined by many authors. McEnery and Hardie (2011) define it as a set of texts which are subjected to an analysis and representative of a language. Sinclair (1991, p. 171), among other authors, defines corpus “as a collection of naturally occurring language texts, chosen to characterize a state or a variety of a language”. Another definition we would like to mention is the one presented by Pérez Hernández (2002), who defines it as a collection of more than one text, referring to it as a contextual entity, rather than a body that its only purpose is to be subjected to tool analyses and studies.

In any terminological study of a scientific discipline, it is difficult to achieve exhaustiveness of the investigation. This means it is difficult to conduct a highly detailed analysis due to the considerable number of the terms that belong to the scientific discipline and the difficulties in accessing documentary sources. For this reason, authors like Berber (2019), Biber (2006), Heaps (1978), Sánchez and Cantos (1997), and Gelbukh and Kolesnikova (2022) consider ‘representativeness’ to be a necessary characteristic of any terminological work. In other words, a corpus should be created by selecting a representative sample of the area to be studied, as well as of the subareas. The representativeness of our corpus is explained in section 4.1.3.

Our corpus follows the characteristics that these authors mention. It is a collection of various texts, specifically 50, that are about a single topic, pollen allergy, and they characterize a specialized language, which is the language used in the medical field, specifically, the subfield of pollen allergy subfield. To build our corpus, we had to select texts and sources based on specific and clear design criteria. These criteria can be divided into two types: external criteria and internal criteria (Clear & Ostler, 1992) which we explain in the following paragraphs.

As per the external criteria, they are based on evidence which is external to the body of the texts of the corpus. The source of the texts is one of the elements of these criteria, which in our case is specialist journals, taken on their majority from Google Scholar and medical journals such as Dove. Time span is also part of these criteria, and our texts have been written during the last decade. Another factor we considered is the region of

the texts, being texts that discuss pollen allergy in different countries, such as China or European countries.

The internal criteria cover the internal linguistic features distribution of the corpus, to determine if they are relevant and if they fit into the design and purpose of the study. The language of the texts is part of this, and our corpus is a monolingual English corpus. The topic area is part of these criteria classification, being medical or health in the case of our texts. Finally, the non-verbal elements that are included in the texts, such as graphics or tables, are another element within this type of criterion and, in our case, they were eliminated automatically when we saved them in txt files.

2.4. Glossary

A glossary is a consultation tool intended to provide the terms of a specialized area that is difficult to understand or unknown to a community interested in the topic (Domínguez, 2007, p. 28). It is a repository for organized data that presents a catalogue of terms related to the same field of study, with their definition and/or other pieces of information, which could be, for example, phonological transcriptions, synonyms, or antonyms. A glossary can be created as a dictionary of terms that are specific to a subject for students, as an encyclopaedia, where articles that further explain the concepts are included as entries, or as a database, where users can search for information by author, keyword, date, or other criteria. Its objective is to be consulted by those who need access to this type of terminology or simply want to expand their knowledge on a specific field.

The types of glossaries are many, and they can be classified according to several factors, such as their topic, language or structure. There are glossaries for a wide range of topics, for example, medicine, mathematics, and technology, among others. Their structure can be varied. Some include only the terms with their definition, while others add additional information, such as their phonological transcriptions or examples of use (Lusicky & Wissik, 2023). Finally, according to the number of languages, glossaries can be monolingual, presenting terms in only one language; bilingual, providing equivalents of these terms in another language; or plurilingual, offering equivalents of the glossary terms in more than two languages.

To build a comprehensive glossary, information of a diverse nature needs to be gathered for each entry. Sager (1990) proposes including synonyms in the same language, equivalents in other languages, morphological and grammatical aspects, related terms, definitions, explanations, comments and notes, context and textual type, or its place in a system of concepts. In our case, we decided to include not only the definitions of the terms but also their Spanish equivalents. This allows Spanish users to understand the equivalents of terms in their native language. We also included phonological transcriptions, which help non-native English speakers understand how the words are pronounced. Additionally, we provided parts of speech, to help users understand how a term relates to others so they can construct proper sentences; references, to acknowledge who made the work; synonyms, to ensure clarity of meaning if the definition was not helpful enough; and examples of use, to allow users to see real-world examples of these terms in contexts, which gives them a better idea of how to use these terms.

Regarding the type of users of our glossary, it is intended for adult specialist users, and language learners (L2). This includes specialist translators in the area or any specialist in need of any information that it contains, which could be, for example, non-native English specialists who would need the phonological transcription of a specific term, to be able to pronounce it correctly, translators or experts, who have some knowledge of this type of terms (Svensén, 2009). Furthermore, these users must be familiar with glossary or dictionary conventions, such as phonetic transcriptions, labels, etc., so they can use and understand this glossary, and must possess some knowledge of linguistics and understand how to use dictionaries.

3. Literature Review

Over the last decades, terminological work has been carried out with the aim of developing glossaries, dictionaries, and databases that today are essential for technical and professional communication; and the Internet has become a necessary tool for disseminating these resources. One of the fields of science in which the most frequent reference works are published on the Internet is medicine. In this section, we discuss some articles about the process of compiling a corpus and a glossary, as well as previously done glossaries, dictionaries, and databases, that are similar to our glossary,

along with a brief description of each one of them, including their author(s), if they are monolingual or bilingual, and their subfields.

Regarding the process of elaborating a glossary, we searched for articles that examine the process of compilation of a corpus and the process of elaboration of a glossary. We found many articles about this, such as Brett's (1997), Losey-León's (2015), and Lukasiks' (2017) articles. We focused on Lareo's (2020) and Sager's articles because we found them better structured, as they describe the process of compiling a corpus by steps, which makes it more visually easy to read. Lareo (2020) explains the different steps of compiling a corpus and the possibilities that corpora offer for linguistic research. Similarly, Sager et al. (1981) analyse the process of building a glossary by using a corpus. We followed the steps explained by these authors, especially Lareo's when we compiled our corpus (explained in section 4.1.1.).

The glossaries, dictionaries and databases that we found lacked quality, in terms of the evaluation criteria stated by Wolf et al.'s (2009). This lack of quality is due to its minimum number of entries. On their coverage, their number of entries is not many, most of them being between 50 to 100 entries. Most of them include only the definition and do not present other relevant information, such as usage, synonyms, equivalents into other languages or phonetical transcriptions.

First, we are going to discuss the medical monolingual glossaries we found and after that, we will analyse the bilingual ones. An important aspect to point out is the extensive availability of medical monolingual glossaries, dictionaries and databases. Most of them, whether English monolingual or Spanish monolingual, deal with medical terms in general and not with specific medical subfields.

Regarding the monolingual glossaries *MedTerms* (2021) is an English monolingual medical glossary of medicine containing about 16,000 terms of medical concepts and diseases explained in an easy and uncomplicated way. *Saludalia* is a Spanish monolingual medical glossary. It lists 1,015 medical terms that can be searched by selecting the letter of the alphabet that you want to review. These two glossaries provide only one type of linguistic information for the terms they include, which are the definitions. In the case of monolingual dictionaries, Navarro's (2000) dictionary offers answers to doubts related to medical words and expressions (400,000 in total) that are

difficult or misleading to translate. Finally, regarding monolingual databases, an example of a monolingual medical one is Collen's database (1950-2011). It includes the description of problematic medical terms that often lead to confusion.

Concerning the bilingual glossaries, dictionaries, and databases, we have found *The English-Spanish Dictionary of Health-Related Terms*, which includes 14,000 terms. The difference between this dictionary and others is that it only focuses on the most frequently used terms. It presents the equivalents in both languages (English and Spanish).

With respect to bilingual glossaries, we have found a specialized one: Klosa-Kückelhaus and Kernerman's glossary of coronavirus, published in 2022, and including neologisms related to coronavirus in three languages, which are English, Korean and German.

Regarding a medical database we found is TriMED (2018-2024), which contains 1135 terms in total, of which 436 are English, 410 are French and 289, Italian ones. This source is intended for both experts and non-experts. Non-experts could be patients who want to consult the definition of a specific term they do not understand. And experts could include translators seeking equivalents of terms in different languages that provided by the resource, and physicians who might need this information as well. IATE (Interactive Terminology for Europe, 2020) is another database worth mentioning. It was launched in 1999 and contains 6,958,766 terms in 24 languages and more than 100 different domains of the EU legislation, including medicine. We found an Association as well: IMIA: the International Medical Interpreters Association (2015-2024). It lists dictionaries, glossaries databases, encyclopaedias, manuals and other documents in the field of medicine, available in 70 languages.

After having carried out research, we can summarize it by stating that most of the medical glossaries, dictionaries, and databases are monolingual, whether in English or Spanish, and general, concerning medical terms in general, including key medical terms about different illnesses and allergies. Very few of them focus on a specific illness, and none of them exhibit the characteristics of our glossary, which is English-Spanish bilingual and specifically related to pollen allergy. We did not find any monolingual or bilingual glossary about this topic. This lack of a bilingual glossary on pollen allergy is the reason we decided to create a glossary on this topic, as it is an area that has not been explored before.

4. Materials and Methodology

Once the study has been contextualized, we continue with its practical part, which is about the process of compilation of the specialized corpus from which we extracted the information that we included in our glossary. This section is divided into many steps that we explain below.

The first step for the creation of the corpus was to look on the Internet for reliable web pages where we could take the texts from, such as the case of a medical journal, *Dove Medical Press* (Tóth-Czifra, 2022) or Google Scholar. We have chosen these websites because they are reliable sources and due to their simplicity in their use, as they follow a common encoding scheme, which makes them simple and easy to use. Once we found the texts, we downloaded and saved them, in total 50 texts. Later, we gave a name for each text with the labelling that is explained in section 4.1.2.

Once we had all our texts named and saved in txt format, we used the Lextutor tool to combine all the files in one single document, since we found it the easier and simpler tool to do this task. After that, we used ReCor to prove both the quantitative and qualitative representativeness of our corpus, which is explained in section 4.1.3. Once we had this done, we looked for tools that could help us identify the most frequent terms of our corpus, choosing TermoStat (Drouin, 2010) to obtain a list of candidate terms. Then, we used AntConc (Anthony, 2011) concordance tool to extract examples of the use of the terms, and finally, we used Excel as a terminology management tool for our terms (entries) and worked on adding a series of linguistic information that we will discuss in the following paragraphs.

4.1. Corpus

4.1.1. Corpus compilation

The first phase was the choice of a series of criteria and characteristics to decide the composition of our corpus. We will first explain the type of our corpus and following that, we will detail the criteria and the process of its compilation. We decided to classify it according to Bowker and Pearson's (2002), Biel's (2009), Breyer's (2011), and Sinclair's (1991) classifications.

Our English corpus is specialized, since it is designed to study a particular topic, which is pollen allergy, therefore, it is smaller in size than other corpora (Bowker and Pearson, 2002, p. 12). Moreover, our corpus is monolingual, since the texts that compose it are examples written in one language, English (Biel 2009, p. 3). According to the written vs. Spoken criterion, our corpus is in written form (Bowker and Pearson, 2002, p. 12). It is a monitor corpus, not having a final extent since it can be regularly updated by adding new texts about this topic (Sinclair, 1991, p.25). Our corpus is specialized because it has been compiled with the specific purpose of building a glossary of pollen allergy. The corpus has been compiled according to a synchronic criterion, and this is due to our will to analyse pollen allergy texts written in the last decade, due to the many changes that constantly take place in the medical world, therefore, it is better to look up for texts within a recent period. It is not a learner corpus because the texts are not written by language learners, but rather by experts (Breyer, 2011, p. 29). Finally, it is not a parallel corpus, as it is not made up of texts that are written in one language and translated into another language.

The first step of the process of compiling the corpus is the documentation search and access to the information available on the Internet

Two types of searches are reliable: the institutional search, which is the one that is carried out on specific websites of national and international institutions, associations or organizations; and the keyword search, through search engines. In our case, we have decided to use the second option. Firstly, we looked for reliable websites, such as Google Scholar, PubMed Central or ResearchGate and found different texts dealing with pollen allergy from which we intended to take 50 complete texts.

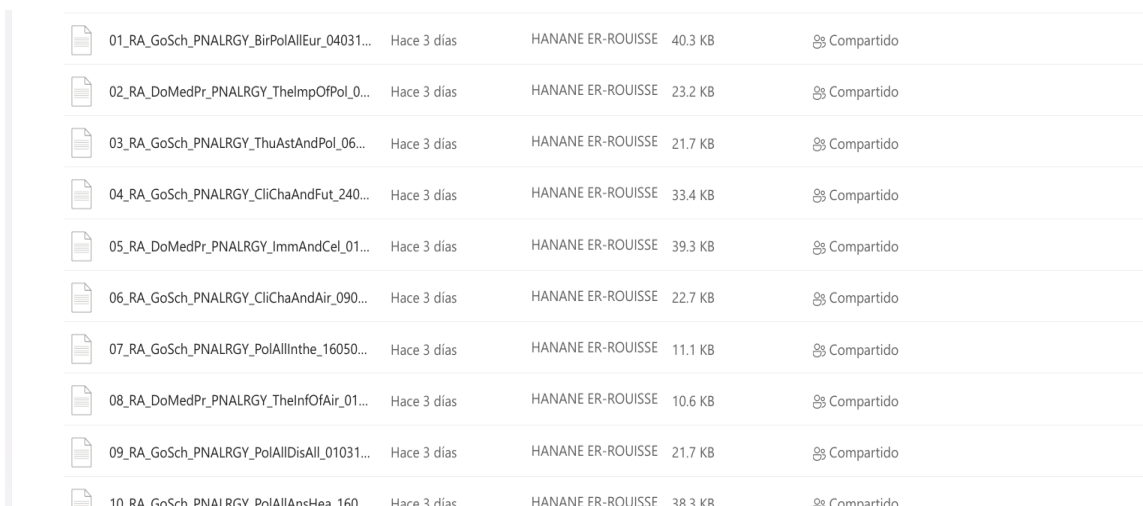
The following step is the “Data Downloading” process. We copied and pasted each text in txt file, cleaned them, removed images, tables, etc., and saved them giving them file names that are explained in the labelling section below.

4.1.2. Naming of files

We have chosen the names of the files considering key elements, such as the topic, the date in which they were written and the language they were written in. Below, we give an example of a specific file of our corpus to explain and justify the labelling we have chosen for the names of files:

02_RA_DoMedPr_PNALRGY_TheImpOfPol_061206_En

The first two characters refer to the number of the text. We named all the files from 01 to 50 since our corpus is composed of 50 texts. The next two characters correspond to the type of genre of this project, which is a research article, abbreviated, as seen above, as *RA*. This was followed by the source from which we took the texts: Google Scholar, abbreviated as *GoSch*, or Dove Medical Journal, presented with *DoMedJo* abbreviation. The next six characters refer to the topic of the text, which is pollen allergy, abbreviated as *PNALGY*. The following characters correspond to the first letters of each word that form part of the title of each text. The next six characters refer to the date on which each text was written, including the day, month and year. Finally, the last characters correspond to the language in which the texts are written, which is English, abbreviated as *En*.



01_RA_GoSch_PNALRGY_BirPolAllEur_04031...	Hace 3 días	HANANE ER-ROUISSE	40.3 KB	Compartido
02_RA_DoMedPr_PNALRGY_TheImpOfPol_0...	Hace 3 días	HANANE ER-ROUISSE	23.2 KB	Compartido
03_RA_GoSch_PNALRGY_ThuAstAndPol_06...	Hace 3 días	HANANE ER-ROUISSE	21.7 KB	Compartido
04_RA_GoSch_PNALRGY_CliChaAndFut_240...	Hace 3 días	HANANE ER-ROUISSE	33.4 KB	Compartido
05_RA_DoMedPr_PNALRGY_ImmAndCel_01...	Hace 3 días	HANANE ER-ROUISSE	39.3 KB	Compartido
06_RA_GoSch_PNALRGY_CliChaAndAir_090...	Hace 3 días	HANANE ER-ROUISSE	22.7 KB	Compartido
07_RA_GoSch_PNALRGY_PolAllInthe_16050...	Hace 3 días	HANANE ER-ROUISSE	11.1 KB	Compartido
08_RA_DoMedPr_PNALRGY_TheInfOfAir_01...	Hace 3 días	HANANE ER-ROUISSE	10.6 KB	Compartido
09_RA_GoSch_PNALRGY_PolAllDisAll_01031...	Hace 3 días	HANANE ER-ROUISSE	21.7 KB	Compartido
10_RA_GoSch_PNALRGY_PolAllAncHea_160...	Hace 3 días	HANANE ER-ROUISSE	38.3 KB	Compartido

Figure 1. Names of the files of our corpus

4.1.3. Corpus Representativeness

The determination of the minimum size that a corpus must present is a controversial aspect (Berber, 2019). Concerning this representativeness, the size of the corpus is a key element when it comes to considering if a corpus is representative of the search study (Biber, 2006). There are different proposals given by different authors. Many authors discuss this, such as Heaps (1978), Sánchez and Cantos (1997). According to Sánchez and Cantos Gómez, all the proposals given have deficiencies.

Biber (2006) argues that a corpus aims to present a part of a language or a language, and it is not simply a collection of texts about a certain topic. The quantitative representativeness of a corpus depends on its quality or density, as well as the relationship between the number of units that are part of the corpus, which are called tokens, and the variety of the type of words, called types.

In our case, we decided to use a tool called ReCor, because it gives an effective and quick solution to figure out the minimum size that a corpus has to be representative, thus, determining the smallest threshold of its representativeness through the analysis of its lexical density. This program does so by providing the level of representativeness of a given corpus in a graph form, showing in statistics whether the corpus under analysis presents an appropriate size.

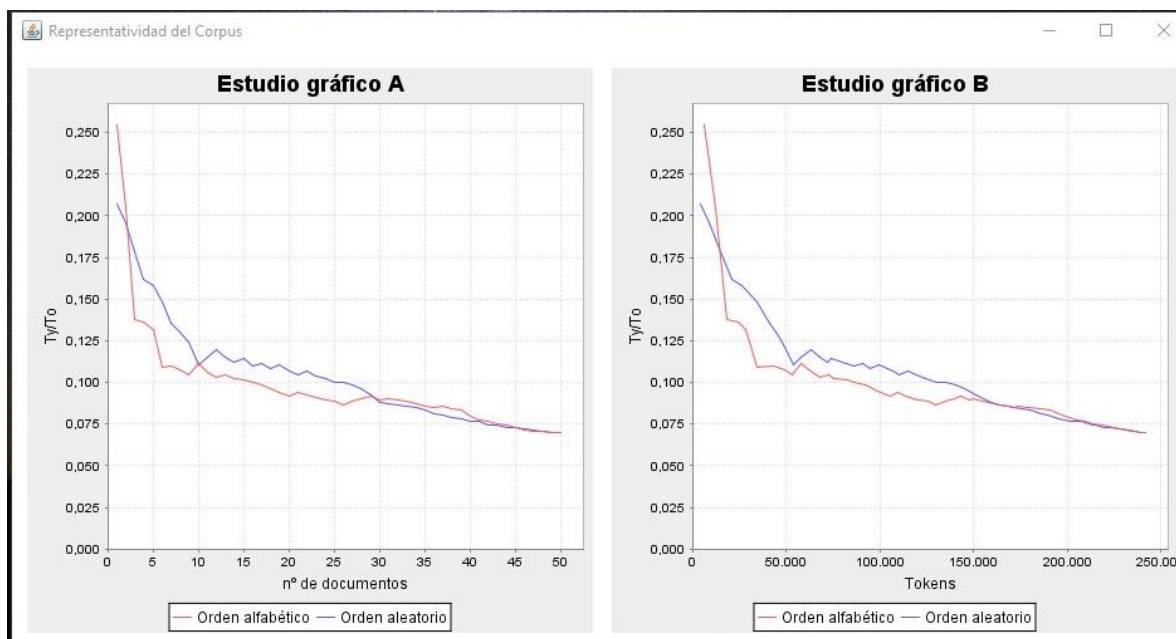


Figure 2. Quantitative criteria of our corpus using ReCor

As we can see in Figure 2, the tokens ratio is shown on the vertical axis, while the number of files is in the horizontal one. We can see, therefore, that the corpus starts to be representative at the point where both the blue line, which represents the files introduced, and the red line, which stands for the files ordered alphabetically, coincide. Our corpus begins to be representative with the first 30 documents and 170,000 words in total.

Regarding the qualitative representativity of our corpus, it has to do with the quality of the texts that make up the corpus (Gelbukh & Kolesnikova, 2022). This type of representativity is determined, therefore, by the quality of the linguistic material that makes up the corpus. The data was taken from reliable sources such as Google Scholar and Dove Medical Press. Google Scholar is reliable since the texts it contains are from publishers, libraries, scientific journals, books, dissertations, conferences or professional societies. Dove Medical Press's texts are journals of science, technology and medicine, which means that they are taken from sources that are reliable as well.

4.2. Methodology

Before addressing the terminological extraction process from our corpus, we want to proceed to explain the definition of the term “terminological extraction”. Taljard (2012) defines it and explains it as a process carried out by a computer program to detect and extract candidate terms automatically. Carrying out this operation is not a straightforward task, since some people in charge of doing it continue to have doubts when deciding and reviewing whether the terms suggested by the program have the term status (De Schryver & Taljard, 2011). Due to this, a manual review is required after the automatic extraction of the candidate terms to guarantee the quality of the obtained results. The choice and identification of the specialized terms is a key phase in determining what the terms are that we should include as entries in a terminological database, and what is noise. However, as Cabré points out (1999), the choice of such terms will depend on the objectives and users in question, therefore, not all the terms identified as terms will be included in a glossary, specialized dictionary or terminological database.

The first step that we followed for the compilation of a list of candidate terms was to choose a computer program to extract such terms. In our case, we have chosen TermoStat tool to get the list of candidate terms, since this tool allows us to save them in an Excel file, clean the list and choose the terms to include in our glossary. Then, we have used AntConc to search for the KWIC (Key Words In Context), since it identifies in which texts of the corpus each specific term has been used, unlike TermoStat, which does not identify in which text a term appears. Finally, we used Excel to compile our

glossary, which is composed of the selected terms, and included their respective linguistic information.

Termostat

As stated before, the first tool we used was TermoStat. It only allows uploading a single file, this is why we first combined all the files using the Lextutor tool, which combines multiple files into one single file.

Then, we uploaded the file to TermoStat and chose the search option for single-word terms, including adjectives, adverbs, nouns and verbs, and for multi-word terms (see Figure 3).

hananeer | Help | Log out

TermoStat Web 3.0

New corpus

File Pollen Allergy Corpus (2).txt

Language ▾

Extraction single-word terms multi-word terms (no

Categories (4) ▾
 adjectives
 adverbs
 nouns
 verbs

Figure 3. Selection of Terms in TermoStat

Nombre de termes: 5219

Matrices

- Adjectif Nom= 1433 (27 %)
- Nom= 1311 (25 %)
- Nom Nom= 1078 (21 %)
- Nom Préposition Nom= 260 (5 %)
- Adjectif Nom Nom= 239 (5 %)
- Adjectif Adjectif Nom= 151 (3 %)
- Nom Nom Nom= 146 (3 %)
- Nom Préposition Adjectif Nom= 93 (2 %)
- Nom Préposition Nom Nom= 67 (1 %)
- Adjectif Nom Préposition Nom= 56 (1 %)
- Adjectif Nom Nom Nom= 45 (1 %)
- Adjectif Coord_Conjunction Adjectif Nom= 36 (1 %)
- Adjectif Nom Préposition Adjectif Nom= 30 (1 %)
- Adjectif Adjectif Nom Nom= 29 (1 %)
- Nom Nom Nom Nom= 26 (0.5 %)
- Nom Adjectif Nom= 22 (0.4 %)
- Nom Préposition Adjectif Nom Nom= 18 (0.3 %)
- Adjectif Nom Préposition Nom Nom= 12 (0.2 %)
- Nom Préposition Adjectif Adjectif Nom= 10 (0.2 %)
- Adjectif Adjectif Nom Préposition Nom= 8 (0.2 %)
- Nom Nom Préposition Nom= 8 (0.2 %)
- Nom Nom Préposition Adjectif Nom= 8 (0.2 %)
- Adjectif Nom Nom Nom Nom= 7 (0.1 %)
- Nom Préposition Nom Nom Nom= 7 (0.1 %)
- Nom Adjectif Nom Nom= 6 (0.1 %)

Figure 4. Patterns of the terms of our corpus

In total, this tool identified 5,219 candidate terms. In terms of the terms chosen to be part of our glossary, their selection was guided by different criteria that are essential to ensure the complete reliability of our glossary. These criteria include the suitability of the topic, representativeness of the field, and frequency of use. The suitability and representativeness will be guaranteed by the varied, sufficient and representative texts selected for the corpus from which we extracted the candidate terms for our glossary. The frequency of use is also an important factor, especially if a term occurs often. We downloaded the candidate terms from TermoStat and copied and pasted them into Excel, as we can be seen in Figure 5 below. Then, the terms that adhere to these criteria were labelled as terms (T), those that met only one or two of the three mentioned criteria were labelled as semi-specialized (S), and finally, the ones that did not meet any of the criteria were identified as noise (N).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1749	S	fructofuranosidi	4	9.27	fructofura	Common_Noun												
1750	T	bioaerosols	4	9.27	bioaerosol	Common_Noun												
1751	N	childrenand adc	4	9.27	childrenar	Common_Noun Common_Noun												
1752	S	whole ragweed	4	9.27	whole rag	Adjective Common_Noun												
1753	T	pollen wall	4	9.27	pollen wal	Common_Noun Common_Noun												
1754	N	development of	4	9.27	developmi	Common_Noun Preposition Common_Noun												
1755	T	wheat pollen	4	9.27	wheat pol	Adjective Common_Noun												
1756	T	sensitiza-tion	4	9.27	sensitiza-t	Common_Noun												
1757	N	full-size image fi	4	9.27	full-size in	Adjective Common_Noun Common_Noun												
1758	S	marker protein	4	9.27	marker pri	Common_Noun Common_Noun												
1759	S	symp-toms	4	9.27	symp-tom	Common_Noun												
1760	S	daily clinical pra	4	9.27	daily clinic	Adjective Adjective Common_Noun												
1761	S	poaceae family	4	9.27	poaceae fi	Common_Noun Common_Noun												
1762	S	airway mucosa	4	9.27	airway mu	Common_Noun Common_Noun												
1763	T	bromelain glyco	4	9.27	bromelain	Common_Noun Common_Noun												
1764	N	n c	4	9.27	n c	Common_Noun Common_Noun												
1765	N	t o	4	9.27	t o	Common_Noun Common_Noun												
1766	T	control of ragwe	4	9.27	control of	Common_Noun Preposition Common_Noun												

Figure 5. Selection of the terms of our glossary

In the end, we selected 227 terms, since I considered that these terms are useful and followed the criteria above explained. After having the list of terms that we chose to include in our glossary, we moved into the last part of the creation of the glossary, which was to choose the information we wanted to include about the selected terms. We made a search about what we could include, and we consulted a wide range of sources, with monolingual, bilingual and multilingual printed, online dictionaries, parallel texts in English, and supporting texts in Spanish. We reached the conclusion that creating a specialized terminological glossary is something that goes further than simply collecting a set of terms with their definitions, but it is a complex work that also consists of collecting various relevant information about such terms for its users. Sager states that information of a diverse nature needs to be gathered for each entry (1990). He proposes including synonyms in the same language; equivalents in other languages; morphological and grammatical aspects; related terms, for example: antonyms; definitions and explanations, comments and notes, context and textual type, or its place in a system of concepts. In our case, we have chosen to include phonological transcriptions, part of speech, definitions, synonyms, equivalents in Spanish and examples of use extracted from our corpus.

Regarding the sources of the definitions and phonological transcriptions, it is important to mention that these two elements were built up by consulting various resources and specialized dictionaries, to finally create my own definitions and transcriptions.

AntConc

The second tool we used is AntConc. This software offers many options, but the most important are the Word List, which allows users to create a list of words from your corpus, and the Concordance tool, which is the one we used in our study to search for the context of the words that compose our glossary, to include them as examples of use later. This tool shows search results of words in context (KWIC). The steps we followed to get the concordances from our corpus are the following: first, we selected our corpus using the “Open File(s)” option in the “File” menu; and then, we entered every term we have collected and cleaned from the candidate list obtained using TermoStat, and

clicked on the “Start” bottom to start the concordance generation. Figure 6 shows the results for the term “pollen allergy”.

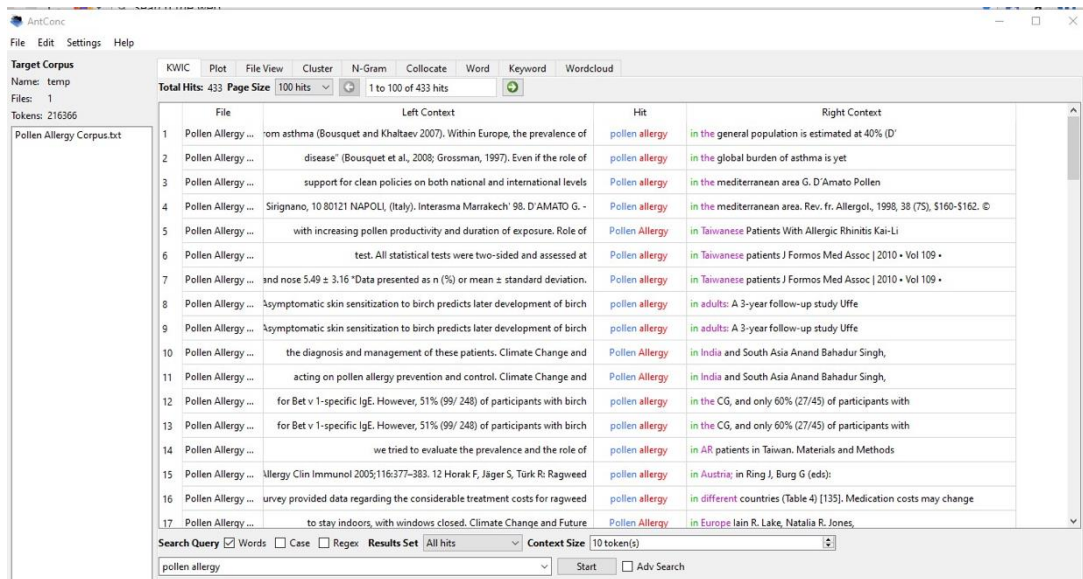


Figure 6. Search of the terms in context using AntConc

Finding synonyms was the main challenge we encountered when creating the glossary because some of them did not convey the meaning of the terms in their literal sense, making them untrustworthy. As a result, we have decided to not include the synonyms of some terms.

Excel

The tool that we chose to store our glossary entries was Excel. We selected it over any other options due to the many advantages it offers, even though it has some disadvantages that we explain below.

The main advantage of Excel is its versatility, as it offers many features and options, among which we have data analysis, calculations, or charting. It is also a suitable option for simple terminology projects, such as the one we are undertaking, as it offers effective tools for managing and storing data. Its customizability allows the adaptation of our sheets to the needs of our terminological project. Another positive point is its cost-effectiveness, since it comes together with Microsoft, which means that there is no additional cost required, unlike other tools, which are more expensive or not available.

Although the advantages of this program are many, it does not mean that it presents some disadvantages as well, such as its limited data storage, which makes it not suitable for projects with much data to store; its slowness, when there is much data included; and the inability for viewing all information on the screen at once (definitions, examples of use...). However, this did not affect our study, since our glossary, due to its high degree of specialization, was not large enough to reach the data limit of the program.

5. Results and Discussion

As shown in Figure 4 we obtained a total of 5,219 candidate terms and, as indicated in Annex A, our corpus includes 13,748 types, among which we have chosen 227 terms to include in our glossary: 217 nouns, 5 verbs and 5 adjectives.

We decided to include the following information for each term: Spanish equivalent, phonological transcriptions part of speech, definition, sources synonym, example of use taken from the corpus, field, subfield and date.

In relation to the Spanish equivalents, most of them were straightforward to identify without the help of any resources, this is the case of terms like ‘pollen allergy’, ‘pollen grain’, ‘pollination’, ‘surgical mask’ or ‘vaccinate’”. As for the phonological transcriptions, we reviewed some notes from the course “Fonética Inglesa” (Grado en Estudios Ingleses). With respect to the definitions, we crafted some of them based on the information present in the corpus, and we remaining definitions come from specialized articles in this field, which we read and adapted to finally create our own.

In the following tables, 2 and , we present two examples of terminological sheets from our glossary.

Term	Allergoid
Spanish Equivalent	Alergoide
Phonological Transcription	/ə'ɪz:(r)dʒɔɪd/
POS	n.

Definition	Plant that presents pollination through the wind (anemophilous), being able to transport the grain of pollen over a long distance.
Ref.	HE
Synonym	
Example of Use	The efficacy of sublingual immunotherapy with the allergoid ragweed sublingual tablets will be tested.
Field	Medicine
Subfield	Pollen Allergy
Date	May 2024

Table 2. Terminological Sheet Example 1 from our Glossary.

Term	Polcalcín
Spanish Equivalent	polcalcina
Phonological Transcription	/pól'kælsíns/
POS	n.
Definition	Calcium-binding proteins found in the pollen of weed, tree and grass. Many allergy sufferers are sensitized to these birch pollen-like proteins.
Ref.	HE
Synonym	Allergist
Example of Use	Protein families, such as profilins, polcalcins, and non-specific lipid transfer proteins.
Field	Medicine
Subfield	Pollen Allergy
Date	May 2024

Table 3. Terminological Sheet Example 2 from our Glossary.

Below we present a screenshot representing part of our glossary. The terms are listed in the left column, followed by the previously mentioned information, each in its respective column:

A	B	C	D	E	F	G	H	I	J
Term	Spanish Equivalent	Phonological Transcription			POS	Definition	Ref.	Synonym	Sample of Use
acid regurgitation	regurgitación ácida	/ˈæsið rɪ ɡɜː(r)dʒɪf(ə)n/			n.	Sensation that gastric fluids rise through the chest and may reach the mouth. Other less common symptoms associated with this include chest pain, sore throat, wheezing, and cough.	HE		Acid regurgitation occurs when it presents a burning sensation in the chest.
aeroallergen	aeroalérgeno	/eəɹəʊ ˈælə(r)dʒən/			n.	come into contact with the human respiratory tract or other mucous membranes, and produce an allergy.	HE	antigen	It was demonstrated that intranasal exposure induced marked oesophagus eosinophil.
aeroallergen sensitization	sensibilización a aeroalérgenos	/eəɹəʊælə(r) dʒən sensɪtəˈzeɪʃən/			n.	Pollen or spores that trigger pollen allergy.	HE	asthma development	Southern regions on aeroallergen showed the highest sensitization to Dermatophagoides pteronyssinus.
aeroallergens	aeroalérgeno	/eəɹə ˈælə(r) dʒəns/			n.	production of specific antibodies in predisposed individuals. These antigens become allergens depending on chemical, environmental or physical factors.	HE		Climate change affects aeroallergy particular plant allergens.
aerobiology	aerobiología	/eəɹəʊbɪˈɒlədʒi/			n.	such as bacteria, fungal spores, small insects, and pollen, which are passively transported through the air.	HE	bacteriology	It is disappointing that research on aerobiology in a developed country like Australia is lagging.
aeropalynology	aeropalínología	/eəɹəʊpæliˈnɒlədʒi/			n.	Part of aerobiology that exclusively studies pollens and spores, which can be produced by various types of plants.	HE		There have been a few attempts to study aeropalynology.
air filtration	purificación de aire	/ˈeər fɪl treɪʃən/			n.	Treatment against pollen allergy that rids the air of the substances that cause them using special filters.	HE	air cleaner	Air filtration can help with this type of allergy by trapping the small particles that

Figure 7. Graphic example of a part of the glossary

As we previously stated in the literature review regarding the lack of a monolingual or bilingual glossary on pollen allergy, we can now state that we have contributed to the creation of a bilingual glossary that has never been done before.

Our glossary presents two main advantages compared to those discussed in the literature review section. The first one is concerned with the specific topic (pollen allergy) that we did not find in any English/Spanish glossary. The second one is related to the comprehensive information that we provide for each entry, containing more information than what is typically provided, which is the Spanish equivalent, phonological transcription, part of speech, definition, reference, synonym, example of use, field, subfield, and date. Most of the glossaries, dictionaries and databases we found that address any medical subtopic only present the part of speech and definition of the terms. With this glossary, therefore, we facilitate and reduce the documentation time that specialized translators invest in collecting information on a topic. This is possible because it contains comprehensive linguistic information, such as the definition of the term or the context of use, among other informative elements. The glossary and the corpus are saved in One Drive, an online cloud storage option where all the information related to this study is uploaded.

6. Conclusion

To conclude, our study aimed to elaborate an English/Spanish glossary dealing with pollen allergy. To do this, in the theoretical part, we discussed some key concepts related to this topic. Then in the literature review framework, we discussed some of the previous works related to the compilation of a glossary. We also describe some glossaries, dictionaries and databases that are similar in nature to our topic. This is followed by the materials and methodology section, in which we explained the process, step by step, of the compilation of our corpus, as well as the tools we used to extract the terminology and create the specialized glossary.

After having finished the process of building the glossary, we ended up with a total of 5,219 candidate terms that were identified by TermoStat, of which we have chosen to include 227 terms of them in our glossary: 217 are nouns, 5 adjectives and the other 5, verbs.

Regarding the contributions of our final dissertation, we offer a bilingual glossary dealing with a topic that has never been part of any study. As mentioned in the literature review section, we identified that there is a great need for a bilingual terminological resource focused on pollen allergy, given the lack of such glossaries.

Concerning its final utility, the glossary was built to make accessible information on the terms related to pollen allergy. It is addressed to diverse types of users, all of them specialized users, such could be the case of specialized translators or non-native English specialists.

Finally, we want to motivate other researchers to compile more glossaries, since there is a need to develop new glossaries related to various types of allergies or diseases, particularly bilingual ones. It is worth mentioning as well that our glossary is a Spanish/English glossary, meaning that other researchers from different nationalities could work on a similar project with other language pairs. They could also include additional information that we did not include, such as images and illustrations, which can provide valuable support in understanding the terms.

7. Bibliography

- Anthony, L. (2011). *AntConc (Version 4.0.0)*. [Computer Software]. Tokyo, Japan: Waseda University. <https://doi.org/10.4135/9781529774009>.
- Berber, T. (2019). *Multi-Dimensional Analysis. Research Methods and Current Issue*. Bloomsbury Publishing.
- Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamin Publishing Company.
- Biel, L. (2009). Meng Ji: Phraseology in corpus-based translation studies. *Yearbook of Phraseology*, 2 (1), 186–191. <https://doi.org/10.1515/9783110236200.186>.
- Bowker, L., & Pearson, J. (2002). *Working with Specialized Languages: A Practical Guide to Using Corpora*. Taylor & Francis.
- Brett, P. (1997). *An Illustrated Dictionary of Building: A Reference Guide for Practitioners and Students*. Butterworth-Heinemann.
- Breyer, Y. (2011). *Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges*. Peter Long.
- Brown, C. (n.d.). *Google Scholar [dataset]*. In CC Advisor. The Charleston Co. <https://doi.org/10.5260/cca.199423>.
- Cabré, M. T. (1999). *Terminology. Theory, Methods, and Applications*. Antártida/Empúries.
- Sánchez, A., & Cantos P. (1997). Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus. *International Journal of Corpus Linguistics*, vol. 2, (2), 1-16. <https://doi.org/10.1075/ijcl.2.2.06san>.
- Collen, M. F. (2011). Medical Knowledge Databases. *Computer Medical Databases*, 217–232. https://doi.org/10.1007/978-0-85729-962-8_8
- Clear, J., & Ostler N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, vol. 7, (1), 181 –190. <https://doi.org/10.1093/llc/7.1.1>.

- De Schryver, G.-M., & Taljard, E. (2011). Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho. *Lexikos*, 17(0).
<https://doi.org/10.5788/17-0-1163>
- Domínguez, P. (2007). *Inglés para Hispanohablantes. Una Opción Real para Mejorar la Enseñanza del Inglés en Nuestro País*. Universidad Autónoma de Baja California.
- Drouin, P. (2010). *TermoStat* (Version 3.0). [Computer Software]. Montreal, Canada. Université de Montréal. <https://doi.org/10.1109/icacea.2015.7164706>.
- Fischbach, H. (1998). *Translation and Medicine*. John Benjamins.
- Gelbukh, A., & Kolesnikova, O. (2022). *Métodos de Clasificación Automática de Textos Para El Español. Lingüística de Corpus En Español*. Routledge.
- Gotti, M. (2004). Specialised Discourse in Multilingual and Multicultural Contexts, 45–46, 5–20. <https://doi.org/10.4000/asp.839>.
- Heaps, H. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press.
- Hurtado, A. (2011). *Traducción y Traductología: Introducción a la Traductología*. Cátedra.
- IATE. (2020). *Interactive Terminology for Europe*. Libsearch. <https://iate.europa.eu/>.
- IMIA. (2015-2024). *International Medical Interpreters Association*.
<https://thetranslationcompany.com/translation-directory/translation-associations/international-medical-interpreters-association/>.
- Klosa-Kückelhaus A., & Kernerman, I. (2022). Lexicography of Coronavirus-Related Neologisms: An Introduction. *Lexicography of Coronavirus-Related Neologisms, De Gruyter*, 1-10. <https://doi.org/10.1515/9783110798081-001>.
- Lareo, I., (2020). A Corpus of English Life Sciences Texts (CELIST). *Universidade da Coruña. Servizo de Publicacións*, 12–15,
<https://doi.org/10.17979/spudc.9788497497848>.

- Losey-León, M. (2015). Corpus Design and Compilation Process for the Preparation of a Bilingual Glossary (English-Spanish) in the logistics and Maritime Transport Field: LogisTRANS. *Procedia - Social and Behavioural Sciences*, 173. <https://doi.org/10.1016/j.sbspro.2015.02.068>.
- Lukasik, P. (2017). Antecedents of Consumer-Based Store Brand Equity – Conceptual Model. *Annales Universitatis Mariae Curie-Skłodowska, Sectio H, Oeconomia*, 51, (2). <https://doi.org/10.17951/h.2017.51.2.169>.
- Lusick, V., & Wissik, T. (2023). “Procedural Manual on Terminology”. *Handbook of Terminology*. 58–84. <https://doi.org/10.1075/hot.3.mull1>.
- MacWhinney, B. (2000). *The Children Project: The Database*. Lawrence Erlbaum.
- McEnery, T., & Hardie, A. (2011). *Corpus Linguistics. Method, Theory and Practice*. Cambridge University Press. Cambridge, UK., 2011.
- Maggi, R., & Pérez T. (2006). *ReCor* (Version 1.0). [Computer Software]. Lexicografía y Traducción. <https://www.lexytrad.es/en/>.
- MedTerms*. (2021). *Medical Dictionary, Medical Definitions & Medical Terminology*. <https://doi.org/10.32388/hytlc3>.
- Navarro, F. (2000). *Diccionario Crítico de Dudas Inglés/Español de Medicina*. McGraw-Hill/Interamericana de España S.L.
- Orduña, J. L. (2011). Estudio Gramatical de las Locuciones Verbales con Doble Pronombre Clínico Estudio. *Revista de Lingüística Teórica y Aplicada*, 49 (2), 87–110. <https://doi.org/10.4067/s0718-48832011000200005>.
- Pérez Hernández, C. (2002). Estudios Basados en Corpus: La Necesidad de Estudiar la Lengua en Uso. *Estudios de Lingüística del Español*, vol. 18, 20-21. <https://elies.rediris.es/elies18/>.
- Picht, H., & Draskau, J. (1985). *Terminology. An Introduction*. University of Surrey. Department of Linguistic and International Studies.
- Quemada, B. (1990). Lexikographie. *Französisch*, 869–96. <https://doi.org/10.1515/9783110966091.869>.

- Rey, A. (1995). *Essays on Terminology*. John Benjamins Publishing.
- Rondeau, G. (1984). *Introduction à la Terminologie*. Centre Educatif et Culturel.
- Sager, J. (1990). A Practical Course in Terminology Processing. *International Journal of Lexicography*, 6, 5-7. <https://doi.org/10.1075/z.44>.
- Sager, J. (1981). Thesaurus integration in the social sciences. Comparison of thesauri. *Knowledge Organization*, 8(3), 133–138. <https://doi.org/10.5771/0943-7444-1981-3-133>.
- Santamaría, M. I. (2006). *La Terminología: Definición, Funciones y Aplicaciones*. Liceus, Servicios de Gestión y Comunicación.
- Sinclair, J. (1991). *Corpus Concordance and Collocation*. Oxford University Press.
- Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge University Press.
- Taljard, E. (2012). Corpus-Based Language Teaching: An African Language Perspective. *Southern African Linguistics and Applied Language Studies*, 42 (3), 37–93. <https://doi.org/10.2989/16073614.2012.739318>.
- Cobb, T. *Lextutor* (2024). (Version 3.0). [Computer Software]. Thornbury, UK. <https://www.lexutor.ca/>.
- Tóth-Czifra, E. (2022). *Dove Medical Press Super Collection*. <https://doi.org/10.14293/s2199-1006.1.sor-uncat.clkpwfs.v1>.
- TriMED*. (2018-2024). *A Multilingual Terminological Database*. European Language Resources Association. <https://doi.org/10.32388/hytlc3>.
- Varantola, K. (2006). Finnish Lexicography. *Encyclopaedia of Language & Linguistics*, 80–81. <https://doi.org/10.1016/b0-08-044854-2/04357-1>.
- Wolf, M., Mikyung K., Kao, J.C., & Rivera, N. (2009). Examining the Effectiveness and Validity of Glossary and Read-Aloud Accommodations for English Language Learners in a Math Assessment. *Applied Measurement in Education*, 25, (4), 34–74. <https://doi.org/10.1080/08957347.2012.714693>.

Zanón, N. (2016). *A University Handbook on Terminology and Specialized Translation*. UNED.

Annex A

Quantitative data of the corpus

File name	Number of words	Number of types
01_RA_GoSch_PNALRGY_BirPolAllEur_040314_En	5,705	1,351
02_RA_DoMedPr_PNALRGY_TheImpOfPol_061206_En	3,530	812
03_RA_GoSch_PNALRGY_ThuAstAndPol_061206_En	3,221	830
04_RA_GoSch_PNALRGY_CliChaAndFut_240816_En	4,555	1,020
05_RA_DoMedPr_PNALRGY_ImmAndCel_010217_En	5,840	989
06_RA_GoSch_PNALRGY_CliChaAndAir_090424_En	3,341	968
07_RA_GoSch_PNALRGY_PolAllInthe_160505_En	2,230	569
08_RA_DoMedPr_PNALRGY_TheInfOfAir_011221_En	1,603	1,603
09_RA_GoSch_PNALRGY_PolAllDisAll_010314_En	3,180	945
10_RA_GoSch_PNALRGY_PolAllAnsHea_160116_En	5,555	1,176
11_RA_GoSch_PNALRGY_AshPolAllAns_300919_En	5,565	1,184
12_RA_GoSch_PNALRGY_RisAssOfPoll_061222_En	5,506	1,325
13_RA_DoMedPr_PNALRGY_SymPatAndCom_220322_En	4,286	1,126
14_RA_GoSch_PNALRGY_PollAllAndHea_170215_En	5,565	1,184
15_RA_GoSch_PNALRGY_RagPolAllBur_210518_En	7,878	1,797
16_RA_DoMedPr_PNALRGY_TrePolAllRus_050221_En	5,033	1,069
17_RA_DoMedPr_PNALRGY_SeaIntInfInp_090903_En	3,804	869
18_RA_GoSch_PNALRGY_OralImmAgaApo_230605_En	4,706	1,366
19_RA_DoMedPr_PNALRGY_EosInfOfThe_091105_En	5,493	1,056
20_RA_GoSch_PNALRGY_AirPolEnhRhi_121001_En	4,001	1,157
21_RA_GoSch_PNALRGY_ASurOnTheMam_281003_En	2,414	676
22_RA_DoMedPr_PNALRGY_CiChaAndPo_070218_En	7,329	1,796
23_RA_GoSch_PNALRGY_RoIOfPolAll_131210_En	2,562	2,562
24_RA_DoMedPr_PNALRGY_RepGlyPot_280120_En	5,767	1,371
25_RA_GoSch_PNALRGY_TheRelBetAir_020622_En	2,968	567
26_RA_GoSch_PNALRGY_AsySkiSen_060103_En	3,917	950
27_RA_DoMedPr_PNALRGY_TomalycEscAll_050314_En	4,036	928
28_RA_GoSch_PNALRGY_ArtAllInChi_290818_En	3,756	1,121

29_RA_GoSch_PNALRGY_TheEffOfFac_080222_En	2,538	664
30_RA_DoMedPr_PNALRGY_RecAllForAll_260211_En	3,986	758
31_RA_GoSch_PNALRGY_RisOfPolAll_260607_En	4,194	937
32_RA_GoSch_PNALRGY_ChaOfPolRel_230520_En	3,598	709
33_RA_DoMedPr_PNALRGY_TrePolAllAn_170715_En	5,031	1,608
34_RA_GoSch_PNALRGY_TrePolAllEfg_280608_En	3,943	1,119
35_RA_GoSch_PNALRGY_AllToCypPol_280105_En	3,424	1,269
36_RA_DoMedPr_PNALRGY_ChaOfPolRel_240520_En	3,790	767
37_RA_GoSch_GoSch_PNALRGY_SofOfUltTit_211210_En	1,766	727
38_RA_GoSch_PNALRGY_CitAllFroPoll_040113_En	1,766	727
39_RA_DoMedPr_PNALRGY_AllandPolAll_310318_En	3,553	1,129
40_RA_DoMedPr_PNALRGY_CitAllFroPol_040113_En	6,923	1,411
41_RA_GoSch_PNALRGY_ChaOfPolAll_100203_En	3,361	923
42_RA_GoSch_PNALRGY_AllSymCauBy_010909_En	1,314	502
43_RA_GoSch_PNALRGY_MonTecForPol_100421_En	11,404	2,510
44_RA_GoSch_PNALRGY_PolAllDevMul_220322_En	5,205	1,487
45_RA_DoMedPr_PNALRGY_CliChaAndPol_010221_En	7,330	1,797
46_RA_GoSch_PNALRGY_PepGlyAsPot_020320_En	5,653	1,332
47_RA_GoSch_PNALRGY_PolAllForMol_160416_En	4,309	1,033
48_RA_DoMedPr_PNALRGY_ChaForAllDia_150215_En	3,732	905
49_RA_GoSch_PNALRGY_MicWheSeeAns_070709_En	4,324	735
50_RA_GoSch_PNALRGY_PolAllansHea_030619_En	5,565	1,184
Total	216,366	13,748