IDENTIFICATION AND CHARACTERIZATION OF NESTED ABBREVIATIONS IN SCIENTIFIC DISCOURSE FOR TRANSLATION PURPOSES

A Thesis presented by

NATALIA RIVAS DUQUE, MD.

Submitted to the School of Languages of

University of Antioquia Medellín in partial fulfillment

of the requirements for the degree of

MASTER IN TRANSLATION

June 2016

Master's in Translation

IDENTIFICATION AND CHARACTERIZATION OF NESTED ABBREVIATIONS IN SCIENTIFIC DISCOURSE FOR TRANSLATION PURPOSES

A Thesis Presented

by

NATALIA RIVAS DUQUE

Approved as to style and content by:

Gabriel Quiroz H.

Gabriel Ángel Quiroz, Chair

John Jairo Giraldo

John Jairo Giraldo, Co-Chair

Pedro Patiño

Pedro Nel Patiño, Committee Member

Norman Gómez

Norman Gómez, Committee Member

Paula Echeverri.

Paula Andrea Echeverri, Director
Escuela de Idiomas

To Juanma and Kike, source of unconditional love.

ACKNOWLEDGMENTS

When I think about all the people who made this work possible, there is a risk of producing a document larger than the thesis itself. However, I would try to make it as succinct as possible. In the first place, I want to thank Juanma and Kike for embracing this project as their own and for facing this journey as a family. Because of their love and support I never felt alone in this whole process. And my parents, for having faith in me even when they do not fully understand the reasons for my choices.

My supervisor Gabriel Quiroz, for seeing the potential in me and allowing me to join this Master's program. His guidance gave the change to turn an awful experience into an opportunity to learn and take my knowledge to a different level. And my supervisor John Jairo Giraldo, for his discipline and meticulousness in our meetings. His reviews of every version of this thesis were essential for a successful development and conclusion of this work.

José Luis Rojas, for being a teacher but, mostly, a real friend in my moments of weakness. For making me believe that I will complete this process triumphant. Giovanny Díaz, for his patience and for always being able to help me with a smile on his face. David Vasquez, for his good energy and advice. I really appreciate our "motivational chats" in the University entrance.

Paula Andrea Montoya, Gustavo Lopera, Pedro Patiño and Antonio Tamayo, for making time out of their tight schedules in order to answer my questions and give me recommendations to make my work better.

Alexander MacArthur and Maria Fernanda Rivas, for their valuable opinions and style suggestions, they made this document more readable. Lina Pérez, Diana Sanz, Dorita Cardona, Diana Galvis and the K9 troop, there is a piece of you in this study.

Last but not least, University of Antioquia, especially the School of Languages, all teachers, secretaries, and partners of the Master's program, a big Thank You!

ABSTRACT

IDENTIFICATION AND CHARACTERIZATION OF NESTED ABBREVIATIONS IN SCIENTIFIC DISCOURSE FOR TRANSLATION PURPOSES.

MARCH 2017

M.A NATALIA RIVAS DUQUE, UNIVERSITY OF ANTIOQUIA MEDELLÍN,

COLOMBIA

Directed by: Gabriel Angel Quiroz and John Jairo Giraldo

This Master's Thesis aims to analyze nested abbreviations from a linguistic perspective; describing their morphological, syntactic and semantic features which can be applied for translation purposes. Nested abbreviations are considered *as abbreviated forms, either initialisms or acronyms, which have within their meaning another abbreviation*. 433 nested abbreviations were extracted from two specialized dictionaries in English. These dictionaries include fields such as Military Sciences, Molecular Biology, Medicine, Atmospheric Sciences, among others.

This research was carried out during the second semester of year 2015 and the first semester of 2016 as part of the first cohort of the Master in Translation program of the University of Antioquia. Data processing involved identification of nested abbreviations in each dictionary, tokenization, morphosyntactic tagging and semantic classification of abbreviations and their definitions in order to undertake a linguistic description of the phenomenon.

This study also includes the translation into Spanish of nested abbreviations extracted from corpus, the validation of these translations performed by three professional translators and, the syntactic analysis in both languages, English and Spanish.

Data analysis showed that from the morphological and semantic perspective, nested abbreviations behave like regular abbreviations. Important differences were found from the syntactic perspective where nested abbreviations behave as premodifiers in the noun phrase (NP) 98.93% of cases. They are located as first premodifier 10.96% of times and second or further premodifier in 87.97% of cases. This means that conversion processes are present in over 95% of cases, where abbreviations behave functionally as adjectives regardless they are classified as nouns.

As conclusions, this is the first time nested abbreviations are not only mentioned but also analyzed and defined. Furthermore, it was found that although the percentage of nested abbreviations obtained from the dictionaries is quite low, less than 1% of total abbreviations, it is highly relevant to study this phenomenon because nested abbreviation is a developing process which can continue to grow, as the amount of abbreviations in specialized languages does.

Moreover, it is significant to increase knowledge of this special type of abbreviations in order to improve not only the performance of abbreviations recognition, extraction and disambiguation systems, but also the work of technical and scientific translators.

Keywords: Nested abbreviation, Initialism, Acronym, Minor-word formation.

LIST OF ABBREVIATIONS

Adj. Adjective

Art. Article

Adv. Adverb

Conj. Conjunction

EN. English

HN. Head Noun

N. Noun

NP. Noun Phrase

POS. Part of speech

PP. Past Participle

PPi. Present Participle

ES. Spanish

V. Verb

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

The following Master's Thesis aims to analyze the nested abbreviation phenomenon from a linguistic perspective; describing morphological, syntactic and semantic aspects of these entities for translation purposes. Nested abbreviations are considered *as abbreviated forms, either initialisms or acronyms, which have within their meaning another abbreviation*. They were extracted from two specialized dictionaries in English, which include fields such as Military Sciences, Molecular Biology, Medicine, Atmospheric Sciences, among others.

In an early stage, data processing involved identification of nested abbreviations in each dictionary, tokenization, morphosyntactic tagging and semantic classification of abbreviations and their definitions in order to carry out a linguistic description of the phenomenon.

In a following stage, our study included the translation into Spanish of nested abbreviations extracted from corpus, the validation of these translations performed by three Colombian professional translators and, a syntactic analysis in both languages, English and Spanish. This was made in order to provide some recommendations on the treatment of this type of abbreviations based on regularities found in our data.

This research was developed during the second semester of year 2015 and the first semester of 2016, as part of the first cohort of the Master in Translation program of the University of Antioquia.

This work was motivated for several reasons. First, it is important to remark that our object of study, the nested abbreviation, is an emerging phenomenon. It is becoming widely known among scholars in different fields such as Computer Sciences (Lang & Schreiner, 2002:129; Witten & Bainbridge, 2002:133), Software Engineering (Geriner, Gulledge, & Hutzler, 1994:141), and Biology and Medicine (Eatock, Fay, & Popper, 2006:58), due to the necessity to compress information in scientific discourses for language economy purposes and to facilitate communication among specialists.

However, nesting in minor-word formation processes has not been analyzed neither in English grammars like Quirk, Greenbaum, Leech, & Svartvik (1985); Biber, Johansson, Leech, Conrad, & Finegan (1999); Huddleston & Pullum (2002) and Halliday & Matthiessen (2014) nor in terminology manuals such as Felber (1984); Arntz & Picht (1995); Cabré (1999); Pavel & Nolet (2001) and Lušicky & Wissik (2015).

The same situation is observed when it comes to the analysis of translation manuals. For example, Newmark (1988:148) recommends to translate abbreviations if they are as important in the source language as in the target language and, once again, there is no evidence of recommendations for translation of nested abbreviations.

Second, nested abbreviations have been conceived by authors such as Wren (2000:67) and, Okazaki, Ananiadou, & Tsujii (2010:1249) as an inconvenient for their abbreviation disambiguation and data mining systems. This occurs because of high frequencies of false negatives involving this type of abbreviations. This affects the systems' performance and, we consider that deeper knowledge about morphological aspects, concerning formation patterns of these abbreviations, would improve the systems' work allowing them to provide better results.

Third, as scientific knowledge evolves, the amount of information compressed for its use in specialized discourses becomes larger. Therefore, scientists and translators must deal with more specialized and complex terms every day. And if some of these terms contain more information, as in nested abbreviations, translation tasks become more difficult to perform.

The reasons mentioned above show why this project aims to highlight the existence of nested abbreviations as an emerging linguistic phenomenon, since they have not been described neither in English grammars, nor in terminology and translation manuals yet and their frequency in specialized dictionaries is still low.

By providing a linguistic description, we attempt to present a wider perspective of this new kind of minor-word formation process which can be applied to improve the performance of systems related to abbreviation recognition and disambiguation, data mining or even machine translation. Besides, deeper knowledge of this phenomenon might also help translators to do a more accurate work in the translation of scientific and technical texts, considering that this type of abbreviations might tend to increase as long as specialized texts deal with longer and more complex terms that require further abbreviation.

Thesis Structure

In order to present our work, this thesis is divided in five chapters. The first chapter makes a description of the object of study, exposes the research question and shows general and specific objectives of this work. The second chapter presents the theoretical framework that supports the study. In that section, a review of the definition and main features of specialized languages, minor-word formation processes, including nested abbreviations, and some aspects about translation of abbreviations are shown.

The third chapter describes the general methodological design; the collected corpus, the tagging process and the tools used to process and analyze data.

The fourth chapter exposes a linguistic analysis and a description of the main findings based on our data. The morphologic, syntactic and semantic aspects of nested abbreviations are described in this section. Moreover, an analysis based on the translation of nested abbreviations is performed.

Finally, the fifth chapter contains concluding remarks and lines for future work.

Description of the Object of Study

Minor-word formation is an ancient process, which has been used for centuries. There is some evidence that this activity existed in Sumer, Rome and Ancient Greece where people experienced the need to abbreviate words to save space and paper in their texts (Cannon, 1989:99; Grange & Bloom, 2000:2). Sumerian expressions such as MILH "Mi Iolh Lnv Hshmilh" (Who shall go up for us to heaven?) or Roman expressions like SPQR "Senatus populusque Romanuis" (The Roman Senate and People) and INRI "Jesus Nazarenus Rex Judaeorum" (Jesus of Nazareth, King of the Jews) are good examples of these usages.

This space-economy trend became more popular throughout history in different countries and academic fields. It was in the fifteen century when the first forms of abbreviation dictionaries appeared (Cannon, 1989:99). After that time, a long list of this kind of dictionaries was published and it is an activity that continues to the present day.

According to different authors such as Cannon (1989:101) and Figueroa & Silva (2000:455), one of the global events that contributed to the augmentation in the number of abbreviations was World War II. Cannon (1989) illustrates this fact as follows:

> The real explosion in systematically created initialisms began with World War Two. The many new items, including informal and obscene ones like F.O.[1] and S.O.S[2]., prompted the U.S. War Department's two official lists, particularly Technical Manual 20-205 (Cannon, 1898:101).

[1] "Fuck Off".

[2] "Same Old Shit".

Over the years, minor-word formation processes increased exponentially (Cannon, 1989:104; Fijo, 2003:58) and along with them, the number of people who considered them a solution in communication (Ordóñez, 1992:11). However, other authors believed that the use of abbreviations in texts was excessive, and sometimes annoying (Newmark, 1979:1405; Huth, 1990:138; Goodman & Edwards, 1997:153), or as stated in Fijo (2003:91) a sign of *"mental laziness"*.

On the one hand, there are positive aspects of abbreviation such as language and space economy (Gutiérrez, 1998:137; Figueroa & Silva, 2000:455), mnemonic functions (Fijo, 2003:98), and facilitation of communication thanks to the elimination of the repetition of very long terms that would make the reading process boring and fatiguing. On the other hand, there are negative aspects as well. Some authors such as Fijo (2003:91) affirm that the excessive use of abbreviations in scientific texts may create obscurity and difficulty in communicating original intent, especially when the reader of these texts does not know the meaning of an abbreviation, and its meaning is not found throughout the text (Betancourt, Treto, & Fernández, 2013:97).

Nonetheless, it is important to clarify that in terms of the communicative situation involved in an expert -to- expert communication, the use of such linguistic devices is not obscure or difficult because there is a shared knowledge between the sender and the receiver in an expert channel. And as stated by Arntz & Picht (1995) "un mensaje sólo puede ser claro, si entre los interlocutores existe un acuerdo previo sobre el contenido conceptual de los medios lingüísticos que se van a utilizar" (Arntz & Picht, 1995:146).

Another difficulty with abbreviations is when the author of a scientific paper uses and creates different abbreviations for the same concept, which are not normalized nor explained in the text. This behavior creates an obstacle when the reader tries to discover its meaning for a better understanding of the text. Regarding the translation process, this action also hinders the attempts of the abbreviation disambiguation software to find a solution (Okazaki *et al*., 2010:1249).

Specialized languages have evolved in such a way that the need to compress information is moving to a deeper level. The number of abbreviated forms increases and the amount of information compressed as well. As a consequence, some authors such as Giraldo (2008:255), Quiroz (2008:52) and Dribniuk (2009:208) have identified that some abbreviations have within their definitions other abbreviations. For instance, in the international standard ISO 704 (2000:33) the abbreviation ARC (AIDS Related Complex) has been identified and described, but it was not denominated. This phenomenon has been called by some scholars **nested abbreviation** and is gaining ground extensively.

The nesting process has been described with relation to other linguistic forms such as collocations (Frantzi & Ananiadou, 1996; Vintar, 2004). It was defined as “the occurrence of smaller units within a larger lexical unit” (Vintar, 2004:1). In abbreviation processes, nesting is a phenomenon recognized just by some authors. Dribniuk (2009) defines it as “an acronym where one of the letters represents another acronym” (Dribniuk, 2009:208). The Microsoft Manual of Style for Technical Publications (Microsoft Corporation, 2004) also defines it as “abbreviations and acronyms that include other abbreviations and acronyms” (Microsoft Corporation, 2004:63), and the Guidelines for Creating and Using Abbreviations and Acronyms of the NASA defines it as “an acronym within an acronym (sometimes within yet another acronym, *ad infinitum*)” (Miller, 1995:25).

At this point, we consider important to clarify that in our study an abbreviation is a unit formed with the initial letters of the words of an expression. It is also the superordinate concept which comprises initialism and acronym. The main features and differences between each concept will be explained in more detail in chapter two.

The literature review on nested abbreviations, so far, has been somehow discouraging as there is no evidence of studies that inquire more deeply into this phenomenon. It is evident that researchers in different fields recognize it, to the point of including the term *"nested abbreviation"* in the definition of this type of abbreviations when they are used in a text. An example of this is provided by Williams, Gage, Arvai, Baski, & Cooke (2012) in their article on aeronautics: "the second instrument is the ACIS (nested acronym for Advanced CCD Imaging Spectrometer)" (Williams, Gage, Arvai, Baski, & Cooke, 2012:3). This behavior can also be demonstrated in other academic fields as Computer Sciences (Lang & Schreiner, 2002:129; Witten & Bainbridge, 2002:133), Software Engineering (Geriner, Gulledge, & Hutzler, 1994:141), and Biology and Medicine (Eatock, Fay, & Popper, 2006:58). In his Computer Science Master's Thesis, Nossum (2012) identifies this type of abbreviations when there is a need to spell out large formulas and "the fully expanded formula could potentially be very large indeed" (Nossum, 2012:5).

Furthermore, nested abbreviations have been identified in other studies related to data mining systems as an inconvenience for their performance. Wren (2000) explains the challenge they represented in his research: "for most terms this works well, but some alignments fail because of nested acronyms. This type of nested abbreviation is relatively rare" (Wren, 200:67). Okazaki *et al*., (2010) also identified nested abbreviations as an obstacle to their work in abbreviation disambiguation. In fact, they noted that these abbreviations provide an important percentage of false negatives decreasing their system’s performance: “out of 376 false negatives, 167 instances involved nested abbreviations” (Okazaki *et al*, 2010:1249).

When we reviewed the literature available on nested abbreviations, it was evident that the phenomenon existed and was identified and used by scholars in different fields. We also noticed that neither nesting nor nested abbreviations were described in English grammars (Quirk *et al.*, 1985; Biber *et al*., 1999; Huddleston & Pullum, 2002; Halliday & Matthiessen, 2014). We discovered that nested abbreviations were mentioned in the Microsoft Manual of Style to discourage their use as “these nested abbreviations are often very difficult for worldwide readers to decipher” (Microsoft Corporation, 2004:63).

The above-mentioned features are the reason why it is important to study this phenomenon from an academic perspective, drawing on linguistic features which can be used for translation purposes. In our research, we attempt to make a linguistic description of the phenomenon, showing morphological, syntactic and semantic aspects of these abbreviations in order to provide a better understanding of this new kind of minor-word formation process. As English is the language in which most scientific knowledge is carried, we decided to collect our specialized corpus in that language and compare the aspects mentioned earlier from translations into Spanish.

We consider that if we go deeper in the understanding of the way nested abbreviations are formed and created by authors, in how they are used in texts, and if we manage to understand their linguistic function, we might be able to provide information for improving their use in texts. Besides, it might be possible to identify morphosyntactic patterns that will improve the performance of abbreviation recognition and disambiguation systems. Furthermore, our findings might help scientific and technical translators to do their job more accurately.

Research Question

Taking into consideration the importance of the study of nested abbreviations, the questions that we attempt to answer in this research are:

Which are the linguistic features of abbreviations when nesting phenomena are involved?

What is the behavior of nested abbreviations when translated from English into Spanish?

Objectives

General Objective

To describe from a linguistic perspective the nesting phenomenon in abbreviations extracted from a specialized corpus in English and its implications in their translation.

Specific Objectives

1. To determine the presence and frequency of nested abbreviations in a general and a specialized corpus.

2. To identify different types of nesting in abbreviations extracted from a specialized corpus in English.

3. To classify nested abbreviations based on the linguistic description.

4. To describe morphological, syntactic and semantic aspects of nested abbreviations extracted from a specialized corpus in English.

5. To observe the behavior of nested abbreviations when translated into Spanish from a syntactic perspective.

Theoretical Framework

In order to provide the theoretical foundation that supports our research, this chapter is divided into three sections. The first section reviews the definition and main features of specialized languages. This notion is considered essential taking into account that this study analyzes abbreviations, which are highly used in specialized texts (Arntz & Picht, 1995:153; Cabré, 1999:73). In addition, nested abbreviations studied in this research are extracted from specialized dictionaries, which represent the codification of a specialized terminology, as stated by Cabré (1993): "la terminología se halla en la documentación especializada en su estado natural, y solo cuando haya sido codificada aparece en los diccionarios" (Cabré, 1993:167). The second section describes minor-word formation processes along with each resulting unit. This part of the chapter characterizes different abbreviated forms that might be present in specialized languages, including nested abbreviations. Furthermore, common aspects and differences between all these elements are exposed. The third section of the chapter concludes with a brief description of solutions to translate abbreviations.

Specialized Languages

General and specialized languages are located as subsystems within a wider communicative system, which is called *language* in a global sense and, they are defined by Cabré (1999) as "subcodes that speakers use according to their expressive needs and the nature of the communicative situation" (Cabré, 1999:58).

Considering that these subsystems have different characteristics, it is important to define them separately. On the one hand, general language is defined by Cabré (1999) as "the set of rules, units and restrictions that form part of the knowledge of most speakers of a language" (Cabré, 1999:59) and it is defined by Hoffmann (1998) as "el conjunt de recursos lingüístics de què disposen els membres d'una comunitat lingüística i que fan possible la comprensió entre ells" (Hoffmann, 1998:47). On the other hand, to provide a definition of specialized language is not a simple task because it is a subject studied by several scholars and it has received different denominations and definitions. Some authors such as Hoffmann (1998), Lerat (1995) and Arntz & Picht (1995) denominated it specialized language. Other authors like Gotti (2003) named it specialized discourse, Ciapuscio (2003) called it specialized text and Cabré (1999) preferred the term special language.

When it comes to definitions, it is possible to observe that there are as many as there are denominations on the subject. Hoffmann (1998) limits specialized language to its linguistic aspects when he talks about "el conjunt de tots els recursos lingüístics que s'utilitzen en un àmbit comunicatiu –delimitable pel que fa a l'especialitat– per tal de garantir la comprensió entre les persones que treballen en aquest àmbit" (Hoffmann, 1998:51).

Lerat (1995) remarks the importance of specialized languages on the communication of specialized knowledge as he considers them as "lenguas naturales consideradas como instrumentos de transmisión de conocimientos especializados" (Lerat, 1995:17).

Arntz & Picht (1995) state that specialized languages focus on the communication in a specialized area and, therefore, there is no one specialized language but several. These authors consider specialized languages as "el área de la lengua que aspira a una comunicación unívoca y libre de contradicciones en un área especializada determinada y cuyo funcionamiento encuentra un soporte decisivo en la terminología establecida" (Arntz & Picht , 1995:28).

As Gotti (2003) exposes in the introduction of his book *Specialized Discourse, Linguistic Features and Changing Conventions,* the expression specialized discourse focuses on the context where the language is used and on the community that uses the language as well. This author presents a relevant feature of specialized discourse as follows:

> Specialized discourse reflects more clearly the specialist use of language in contexts which are typical of a specialized community stretching across the academic, the professional, the technical and the occupational areas of knowledge and practice. This perspective stresses the type of user and the domain of use, as well as the special application of language in that setting (Gotti, 2003:24).

Ciapuscio (2003) provides a definition on specialized texts that concentrates more on the register and rhetoric conventions used in texts produced by specialist, she considers them as:

> Productos predominantemente verbales de registros comunicativos específicos, definidos por los usuarios, las finalidades y las temáticas de los textos. Los textos especializados se refieren a temáticas propias de un domino de especialidad y responden a convenciones y tradiciones retóricas específicas. Los factores funcionales, situacionales y temáticos tienen correlato en el nivel de la forma lingüística, tanto en la sintaxis como en el léxico (Ciapuscio, 2003:25).

Cabré (1999) takes into consideration pragmatic aspects in order to provide her own definition of special language as “the subsets of language that are pragmatically characterized by four variables: subject, field, type of user and type of situation in which communication takes place” (Cabré, 1999:65).

Although each author presents its own definition of the concept, it is important to note that they are not as different as one might think. They all have common features like the communication of a specialized knowledge and the definition of the communicative situation, which is made not only by linguistic but also by pragmatic aspects such as the field, the users of language and their communicative intentions.

Since general and specialized languages have been defined, we expose the main differences between them. First, one of the most relevant aspects about the former is that its rules and restrictions are known by all members of a linguistic community (Arntz & Picht, 1995:28; Cabré, 1999:59), while in the latter these subcodes are known only by small groups of people, specialists, that required previous training to learn this type of language.

Second, another important characteristic of general language that differentiates it from special language is that the former is completely autonomous; it can survive without specialized language. However, the situation does not work the other way around, specialized languages need general language to make communication possible (Arntz & Picht, 1995:38).

Third, general language is used in non-specified situations, meanwhile specialized languages are used in specific situations which are determined by “subject field, type of interlocutors, situation, speaker’s intention, the context in which a communicative exchange occurs, the type of exchange” (Cabré, 1999:59).

Fourth, specialized languages derive from general language and act on it with its same lexical resources such as terminologization, compound formation, derivation, conversion, loan and abbreviation. Nevertheless, they are based on different criteria when using these resources (Arntz & Picht, 1995:147).

Regarding the main features of specialized languages, it is important to highlight that there are linguistic and pragmatic characteristics common to all of them, but they differ mainly in frequency. However, each specialized language differs from others based on the terminology it uses. Terminology is defined as “the set of terms of a specific field, which represents the conceptual structure of the area” (Cabré, 1999:81). Likewise, term is considered as “a conventional symbol that represents a concept defined within a particular field of knowledge” (Cabré, 1999:81).

According to Gotti (2003:33), the following are the main features of specialized languages:

a. Monoreferentiality. This means that only one meaning is allowed in a certain context.

b. Lack of emotion, which is the absence of emotive connotations. In texts developed with specialized languages, terms have a purely denotative function. Their main purpose is to inform, leaving other functions aside.

c. Precision. It states that every term must point to its own concept immediately.

d. Transparency. It refers to the possibility to promptly access a term’s meaning through its surface form.

e. Conciseness. It states that concepts are expressed in the shortest possible form. The fulfillment of this characteristic of specialized language causes the appearance of several forms of reduction in texts such as clippings, blends, juxtaposition (omission of prepositions and premodifiers in nominal groups containing two nouns), and abbreviations. As stated by Gotti (2003), “sometimes conciseness in specialized discourse relies on acronyms and abbreviations” (Gotti, 2003:41).

Another aspect that we consider significant to remark is the stratification of specialized languages. As stated by Arntz & Picht (1995:28), it is not possible to talk about one specialized language but several languages. Hoffmann (1998:64) divides them in two types of stratification: horizontal and vertical. In horizontal stratification, the author classifies each specialized language by groups of subjects, subjects, and sub-subjects. An example of this is shown in Table 1.

Table 1. Horizontal stratification of specialized languages.

Chemistry	Theoretical Chemistry			
	Experimental Chemistry	Analytic Chemistry		
		Organic Chemistry		
		Inorganic Chemistry	Nonmetals	
			Noble gases	
			Metal alloys with 6 subgroups	Water
				Heavy Water

Example extracted from Cabré (1999:66).

Conversely, vertical stratification of specialized languages consists of the level of abstraction reached in specialized communication. This means the degree of precision experienced by language in discourses "which allows us to identify several different discourse types which are determined by the degree of abstraction with which the topic is represented or by the style used in a particular communicative situation" (Cabré, 1999:67). The different levels of abstraction range from a very low level to highest level of abstraction as it is presented in Table 2.

Table 2. Vertical abstraction levels.

	Nivell d' abstracció	*Forma lingüística*	*Àmbit*	*Participants en la comunicació*
A	Més elevat	Símbols artificials per a elements i relacions	Ciències fonamentals teòriques	Científic ←→ Científic
B	Molt elevat	Símbols artificials per a elements; llenguatge general per a les relacions (sintaxi)	Ciències experimentals	Científic (tècnic) ←→ Científic (tècnic)
C	Elevat	Llenguatge natural amb terminologia especialitzada i sintaxi molt controlada	Ciències aplicades i tècnica	Científic (tècnic) ←→ Directors cientificotècnics de la producció material
D	Baix	Llenguatge natural amb terminologia especialitzada i sintaxi relativament lliure	Producció material	Directors cientificotècnics de la producció material ←→ mestres ←→ treballadors especialitzats
E	Molt baix	Llenguatge natural amb alguns termes especialitzats I sintaxi lliure	Consum	Representants del comerç ←→ consumidors ←→ consumidors

Extracted from Hoffmann (1998:64).

As we have mentioned before, abbreviations are frequently used in specialized texts mostly to assure precision and conciseness, and to make the communication of specialized knowledge as clear as possible. Therefore, the next subsection of this chapter explains minor-word formation processes, since abbreviations are one of the resulting elements of these activities.

Minor-Word Formation

The processes of minor-word formation have been studied and defined by several authors in different languages, causing as a consequence a heterogeneity that makes difficult the definition of the process itself and its resulting components. This difficulty has been expressed in English (López, 2004:110), Spanish (Figueroa & Silva, 2000:455; Fijo, 2003:57; Giraldo, 2008:60;) and French (Zolondek, 1991:1), just to name a few languages.

Some authors call initialization the process of combining the initial letters of a sequence of words to form a new abbreviated form, which is available in written and spoken language (Cannon, 1989:116; Huddleston & Pullum, 2002:1632; López, 2004:122). Other authors call the same process acronymy (Quirk *et al*., 1985:1031) or abbreviation (American Medical Association, 2010:1274). However, English is one of the languages that show more uniformity in this regard. Fijo (2003) explains this phenomenon, when comparing the minor-word formation process in Spanish: "Al contrario de lo que sucede en español, en inglés puede observarse una mayor homogeneidad terminológica y conceptual a la hora de analizar los diferentes procedimientos de abreviación" (Fijo, 2003:57).

In order to clarify the concepts that will be used in this study, a theoretical review of English grammars, style, translation, and terminology manuals has been made. The concept of initialism (words formed from the initial letters of words that make up a name and that are pronounced as a sequence of letters) has been named abbreviation by authors such as Cannon (1989:116), Huddleston & Pullum (2002:1632), Plag (2003:163) and The Economist (2005:8). Whereas other authors like Quirk *et al.* (1985:1583) and López (2004:123) call the same entity alphabetism.

Initialism, which is the denomination shared by style manuals (Sabin, 2004:146; Alred, Brusaw, & Oliu, 2006:2; American Medical Association, 2010:1274; Lombard & Kotzé, 2013:26), international standard manuals (ISO, 1999:7; ISO, 2000a:124, 2000b:7, 2000c:33) and terminology manuals (Felber, 1984:178; Pavel &

Nolet, 2001:103; Cardero, 2004b:144; Lušicky & Wissik, 2015:50) is the one that will be used in our research. A summary of the information presented above is summarized in Table 3.

Table 3. Different denominations for Initialism.

Named Abbreviations by	**Named Alphabetisms by**	**Named Initialisms by**
Cannon, 1989	Quirk *et al.*, 1985	Felber, 1984
Huddleston & Pullum, 2002	López, 2004	Pavel & Nolet, 2001
Plag, 2003		Cardero, 2004; Sabin, 2004
The Economist, 2005		Alred, Brusaw, & Oliu, 2006
		American Medical Association, 2010
		Lombard & Kotzé, 2013
		Lušicky & Wissik, 2015
		ISO 12620, 1999; ISO 704, 2000; ISO 10871, 2000; ISO 17241, 2000

Taking into consideration that the present work is located within the translation studies framework, and the nested abbreviations phenomena will be analyzed from a translation point of view, we consider reasonable to use the concept of ***Initialism*** in order to develop our research in the same context. The definition will be explained in more detail later in this chapter.

Acronym is another concept that used in our work and it will be also explained later. Contrary to what happens with initialism, this concept exhibits more homogeneity, and this is clearly seen in English grammars, manuals of style and terminology. The authors tend to agree upon this concept and it is the one that we will use in our work (Quirk *et al.*, 1985:1583; Cannon, 1989:116; Pavel & Nolet, 2001:103; Huddleston & Pullum, 2002:1633; Plag, 2003:164; Sabin, 2004:146; The Economist, 2005:11; Alred et al., 2006:2; American Medical Association, 2010:1274; Oxford, 2014:2; Lušicky & Wissik, 2015:50).

Other processes such as clipping and blending do not have the problems mentioned above. They are also explained in this chapter.

Definitions.

It is important to clarify that the aim of this study is not to provide neither a new taxonomy of minor-word formation nor to correct the existing ones, as each of the analyzed authors have already provided theoretical bases to propose a classification. We just want to work with the one which we consider provides a better understanding of these processes. Nonetheless, it is important to annotate that these categories "are all conceived as overlapping categories with fuzzy boundaries" (López, 2004:124) and it is possible to find some elements located in one category that can be recognized in others as well, *e.g.* acronyms and blends.

The minor-word formation processes are described by López (2004:124). She divides them into *simple* and *complex shortening* of words considering the number of words involved in each process. *Simple shortening* represents the shortening of single lexical units. There are two results of this process, *shortenings* and *clippings.* Shortenings as in Mister (Mr.) or Doctor (Dr.), are only used in the written language and are pronounced in an expanded form, *i.e.* the source word. Clippings as in Telephone (phone) are used in written and oral language and are pronounced in an unexpanded way, this means the newly formed word.

Complex shortening is present when more than one lexical unit is involved in the minor-word formation process. They are phonic and graphic reductions, which means that they can be used in oral and written language. The results of this process are pronounced in an unexpanded way and can be spelled in upper or lower cases. These results are called abbreviations, which can be divided in *initialisms*, for instance DNA (Deoxyribonucleic Acid), and *acronyms* like AIDS (Acquired Immunodeficiency Syndrome), and blending which produces *blends* as in Bit (Binary digit). Figure 1 shows each component of the minor-word formation processes that will be explained in more detail ahead.

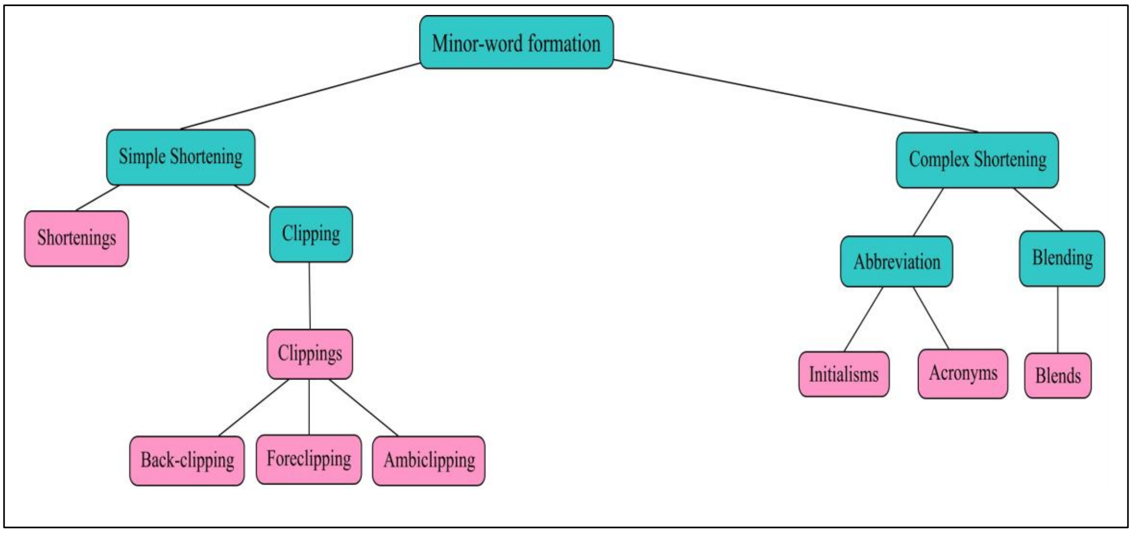


Figure 1. Minor-word formation processes. (Adapted from López, 2004:126).

Simple shortening.

Shortenings.

They are reduced graphic representation of terms, formed by the elimination of some of their graphemes (Gutiérrez, 1998:139). Shortenings are used in the written language in order to save space in texts. Graphically, the most important feature of shortenings is that they are written with a period at the end as stated in ISO 704 (2000:28), but sometimes they can be written with a backslash (Alonso, 2002:161). Phonetically, they are pronounced as the source word. Each shortening represents one term (Gutiérrez, 1998:139). Morphologically, shortenings keep the gender of the abbreviated word. Some examples of shortenings are: Gov. (Governor) and Lit. (Literature).

Clippings.

They are words resulting from the process of “cutting off a part of an existing word or phrase to leave a phonologically shorter sequence” (Huddleston & Pullum, 2002:1634). Most clippings are monosyllabic or disyllabic (Plag, 2003:147) and this clipped form is normally used in informal language to express familiarity with the original word which is also available in regular language (Quirk *et al*., 1985:1583; Plag, 2003:154).

The constituent parts of clipping are:

a. **The original** which is the source of the clipping,

b. **The surplus** which is the phonological material that is cut away, and

c. **The residue** which is the remaining material that forms the new base. Grammatically, clippings normally yield nouns (Huddleston & Pullum, 2002:1634).

There are two operations known in the process of clipping that are described in the Cambridge English Grammar (Huddleston & Pullum, 2002:1635), **plain clipping** where only one operation is made to the original and the result is just the residue of the clipping and, **embellished clipping** where other operations are applied to the residue to produce a longer word; for example adding a suffix (soccer → association football + er).

In the plain clipping, there are three ways to cut of the surplus:

a. The last letters of the original, or the back, which is called back-clipping and is the most common form of clipping (see 1),

b. The first letters from the original, or the front, which is called foreclipping (see 2) and

c. When the surplus is removed from the beginning and the end of the original, which is called ambiclipping (see 3).

Some examples of the different types of plain clipping extracted from Cambridge Grammar (Huddleston & Pullum, 2002:1635) are presented below:

1. Back-clipping: doc (doctor), deb (debutant), lab (laboratory).

2. Foreclipping: bus (omnibus), phone (telephone), cello (violoncello)

3. Ambiclipping: flu (influenza), fridge (refrigerator; BrE), tec (detective; BrE)

Complex shortening.

Abbreviations.

As stated in ISO 704 (2000:33), we consider abbreviation as the superordinate concept that comprises both initialisms and acronyms, taking into account that they are formed by the same process with the initial letters of the words of an expression (or abbreviation). Abbreviations are used both in written and spoken language. This term is used when no distinction between initialism and acronym has been made in the text and no greater specificity is required; as it is used in the Chicago Manual of Style (American Medical Association, 2010:1274).

In view of the fact that abbreviation is the generic form and that the main difference between initialisms and acronyms in English language remains in their pronunciation, all the morphologic, syntactic and semantic aspects that are shared by both units are described in this part of the chapter in order to avoid repetitions in other subsections. Moreover, the characteristics not shared are shown separately.

From the morphological point of view, it is relevant to note that the length of abbreviations ranges from two to five initials (Grange & Bloom, 2000:4; Alcaraz, 2002:42), for instance, ATP (Adenosine Triphosphate) and NATO (North Atlantic Treaty Organization).

There are extreme cases as reported by Giraldo (2006:7) where the length goes up to nine initials, for example in NAMEADSMO (NATO Medium Extended Air Defense System Design and Development, Production and Logistics Management Organization). One more example of these cases is CADASIL (Cerebral Autosomal-Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy).

Another essential aspect of abbreviations is that they are written with upper-case letters, whatever their length, but some have lower case letters (Figueroa & Silva, 2000:457), like abbreviation of Latin phrases such as *i.e* (*id est*) or *e.g* (*exempli gratia*).

Regarding the spelling, Figueroa & Silva (2000:457) state that in English there are some inconsistencies related to the use of periods between the initials which can be present or absent according to the preferences of authors; as opposed to other languages such as French, where periods are used when each letter must be pronounced according to the rules of the French Library Association. Figueroa & Silva (2000:457) also establish that in Spanish the trend is to eliminate the periods. These statements contrast with the arguments of other authors such as Rodríguez (1993:11) who claims that the loss of periods in abbreviations is a consequence of its level of lexicalization and not a trend of a language.

Another morphological aspect that is worth mentioning concerns the plural forms and the gender of abbreviations. In languages such as Spanish, abbreviations adopt the number and the gender according to the head of the noun phrase (NP)

(Gómez, 1992:270; Cuadrado, 1996:262; Alvar, 1996:48). Some examples of this feature in Spanish are: la OTAN, el SIDA and los CD.

Some authors like Cuadrado (1996:262) describe the pluralization of abbreviations in Spanish using the replication of initials in the abbreviations; one example of this is EE. UU. (Estados Unidos, United States in English) where the letters E and U are doubled in order to show the plural character of the term.

Other authors like Gómez (1992:271) state that when the morpheme –S is used to pluralize the abbreviation it is a sign of its level of lexicalization. However, according to Real Academia Española (RAE), this is not an accepted form to pluralize:

> En español, las siglas son invariables en la lengua escrita, es decir, no modifican su forma cuando designan más de un referente. El plural se manifiesta en las palabras que las introducen o que las modifican: varias ONG europeas, unos DVD, los PC. Por eso es recomendable utilizar siempre un determinante para introducir la sigla cuando esta ha de expresar pluralidad.[3]

In English, some of the abbreviations require the article when functioning as head in the NP structure; however, proper name acronyms that stand as full NPs are used without the definite article (Huddleston & Pullum, 2002:1634). One example of this last affirmation is: she works for NATO/ UNESCO, not the NATO/ the UNESCO.

From the syntactic point of view, it is important to note that initialisms belong to the category of noun, and behave grammatically as ordinary nouns as they represent names of political, military and economic organizations, associations, banking institutions, for example IMF (International Monetary Fund) or BYN (Bank of New York) (Rodríguez, 1987:139; Cannon, 1989:103; Gómez, 1992:267; Huddleston & Pullum, 2002:1633; Alcaraz, 2002:48).

[3] Extracted from: http://www.rae.es/consultas/plural-de-las-siglas-las-ong-unos-dvd on June 10th, 2016.

Although abbreviations behave like nouns, sometimes they can change their grammatical category (Arntz & Picht, 1995:147; ISO 704, 2000:34). These changes are described in Spanish by Rodríguez (1987) when he states that:

> El escaso número de siglas no nominales se ve relativamente incrementado por medio del cambio funcional; es decir, siglas que por su base son sustantivas «cambian» de categoría adoptando funciones adjetivas, verbales, etc. Un cambio muy frecuente en las siglas es el de nombre a adjetivo en función atributiva (Rodríguez, 1987:143).

According to this author, these grammatical changes are called *conversions.* They are functional and made mainly from nouns to adjectives following stylistic and expressive reasons. When these changes take place the structure of the initialism remains without alterations. However, there are changes from nouns to verbs as well, and they are made for rhetorical reasons. In order to accomplish the task, suffixes inherent to the verbal category are added and the structure of the initialism changes.

From the semantic perspective, Figueroa & Silva (2000:461) point out that there are two types of abbreviations. On the one hand, there are nominal abbreviations, which represent proper names of institutions, social organisms, and enterprises among others. One example of this type of abbreviation is CIA (Central Intelligence Agency). On the other hand, we have conceptual abbreviations which abbreviate concepts of the different academic fields and represent the shortened form of noun phrases (NPs). One example of this type of abbreviations is CPK (Creatine Phosphokinase).

Another semantic aspect that is significant to highlight is the proliferation of homonymous terms, as it is shown in Schwager (1991): “GU can mean gastric ulcer to a gastroenterologist, genitourinary to an urologist, or glycogenic unit to other kinds of specialists”. On the contrary, synonymy processes have also been found in abbreviations (Schwager, 1991:165). This phenomenon is specially seen in medical publications where scholars create different abbreviations for the same term; an example of this is also extracted from Fijo (2003:97) where Free Fatty Acids (FFA) can also be found in medical literature as NEFA or NFA (Non Esterified Fatty Acids) and UFA (Unesterified Fatty Acids).

In other languages, such as Spanish, these homonymy and synonymy operations have reached a different level, as specialists often use the Spanish term of a concept and use loans or calques of the same term in English. Fijo (2003) exposes this situation as follows:

> La difusión que alcanzan las denominaciones en inglés, hace que en muchos casos convivan términos originales en español con otros que constituyen calcos de las expresiones inglesas. Esto da origen a multitud de términos sinónimos […] Esto significa que se están usando simultáneamente dos denominaciones para designar un mismo concepto (Fijo, 2003:97).

A separated description of initialisms and acronyms is presented ahead.

Initialisms.

They are abbreviations formed "by combining the initial letter of each word in a multiword term, they are pronounced as separate letters" (Alred *et al*., 2006:2), although, in some cases initialisms might be formed for other letters besides initials (Huddleston & Pullum, 2002:1634; Plag, 2003:161). This process may involve numbers and sounds as it is shown in The Chicago Manual of Style (2010).

> Sometimes a letter in an initialism is formed not, as the term might imply, from an initial letter but rather from an initial sound (as the X in XML, for extensible markup language), or from the application of a number (W3C, for World Wide Web Consortium) (American Medical Association, 2010:1274).

Initialisms are extensively frequent in scientific discourses and they are normally originated from terms that are already established in a certain field. The former serve as substitutes or synonyms for the latter in the text in any context, in order to save space or for stylistic reasons, unlike shortenings that can be used only in certain situations (Gutiérrez, 1998:137).

Some of the aspects of initialisms described by Gutiérrez, (1998) are important to be highlighted. First, as noted earlier, more than one lexical unit must be involved in its formation, the author states that “the initial of an isolated word does not form an initialism” (Gutiérrez, 1998:137), and that this aspect is what differentiates initialisms from shortenings and symbols. Second, the expression that originated the initialism must be frequently used and has to belong to a specific domain. For example, DNA (Deoxyribonucleic Acid) is frequently used in Molecular Biology texts.

According to (Quirk *et al*., 1985:1581), in this form of abbreviation, letters have two possibilities: on the one hand, each letter “represents full words as in EEC (European Economic Community) and FBI (Federal Bureau of Investigation)”. On the other hand, the letters “represent elements in a compound or just parts of a word as in TV (television), GHQ (General Headquarters), ID (Identification Card), and TB (tuberculosis).

One aspect that initialisms have in common with other abbreviated forms such as clipping or blending is that they involve “loss of material but differ from them in that prosodic categories do not play a prominent role. Rather, orthography is of central importance” (Plag, 2003:160).

Acronyms.

The most important difference between acronyms and initialisms is that the former are pronounced like ordinary words, and that the letters have their characteristic phonological value (Huddleston & Pullum, 2002:1633) and “must conform to the phonological patterns of English” (Plag, 2003:163). In other languages, such as Spanish, this difference is not limited to phonological aspects. There are also differences related to the formation of abbreviations separating initialisms from acronyms. If the abbreviation is formed exclusively by the initial letters of the NP (without considering structural words) it is an initialism; if other letters besides initials are used, it is an acronym (Gómez, 1992:269; Figueroa & Silva, 2000:457). This aspect of the formation of acronyms is important in order to facilitate pronunciation and further retrieval of language users (Figueroa & Silva, 2000:460; Alcaraz, 2002:44; Fijo, 2003:105).

As acronyms and initialisms have the same formation process, they share morphologic and syntactic characteristics that were described in the abbreviation subsection of this chapter. Although, we consider relevant to highlight that some acronyms are written with uppercase letters, others with lowercase and others can be written in either way (Huddleston & Pullum, 2002:1633). Some authors such as Lombard & Kotzé (2013:26) recommend that acronyms with less than five letters should be written in uppercase, and those with six letters or more, should be written with the first initial in uppercase and the others in lowercase. According to writing manuals, they are always written without periods (Alred *et al*., 2006:2; Lombard & Kotzé, 2013:26).

An important consideration regarding acronyms has been pointed out by López (2004:119) when she states that all constituents of the acronym must be reduced to some extent, otherwise this will become a blend. The line that divides acronyms and blends is very thin and the difference between them "lies in the degree of shortening of the constituents" (López, 2004:121).

The evolution of an acronym is different in each case. Some acronyms remain as substitutes of terms while others, such as AIDS, become terms themselves and change their written representation: lose the capitals and the periods between them (if it is the case) in the spelling, and are involved in composition and derivation processes, leading to lexicalization of terms. This process of lexicalization does not depend on the structure of the acronym, it depends on its meaning and its social, political, historical relevance (Gutiérrez, 1998:138). As established in ISO 12620 (1999), "an acronym can be so widely accepted that it becomes a term in its own right (*e.g.*, "radar" in the following example)" (ISO 12620, 1999: 6).

Blends.

They are words formed "from a sequence of two bases with reduction of one or both at the boundary between them" (Huddleston & Pullum, 2002:1636). An important aspect of blending is that enough of each constituent part is retained "so that the complex whole remains fairly readily analyzable" (Quirk *et al.*, 1985:1583).

The blending process shares with clipping the loss of phonetic material (Plag, 2003:146), but what distinguishes the first process from the second is that the former always starts "with the first part of the first base and finishes with the final part of the second" (Huddleston & Pullum, 2002:1636) while the latter does not mix with other words to form a new one.

There are 4 types of blending according to Huddleston & Pullum (2002:1636):

a. First part of the first base and the whole of the second base.

Example: telebanking (telephone + banking)

b. The whole of the first base and the final part of the second.

Example: breathalyzer (breath + analyzer)

c. First part of the first base and the final part of the second.

Example: heliport (helicopter + airport)

d. The central part is common to a portion of the two bases, with an overlap between them.

Example: motel (motor + hotel).

Another aspect that is important to note is that the base words that form the blend must be semantically related in order to make the combination of properties possible. Plag (2003) explains in a very detailed way some grammatical aspects of blends that are relevant to consider:

> The structure of blends is constrained by semantic, syntactic and prosodic restrictions. In particular, blends behave semantically and syntactically like copulative compounds and their phonological make-up is characterized by three restrictions. The first is that the initial part of the first word is combined with the final part of the second word. Secondly, blends only combine syllable constituents (onsets, nuclei, codas, rimes, or complete syllables), and thirdly, the size of blends (measured in terms of syllables) is determined by the second element (Plag, 2003:60).

As we have pointed out before, the processes of minor-word formation and the products resulting from each process are complex and sometimes difficult to classify.

As an important part of the scientific discourse, which is always evolving, some of the subjects presented, such as abbreviations, tend to evolve as well. The terms they substitute are more challenging and compress larger amounts of information in smaller lexical units, leading to a new type of minor-word formation, called nested abbreviations.

Definition of Nesting

As presented in chapter one, some authors have provided definitions of nested abbreviations or nested acronyms without any distinction between them. According to Dribniuk (2009) a nested acronym is “an acronym where one of the letters represents another acronym” (Dribniuk, 2009:208). Another definition of nested abbreviation is presented in The Microsoft Manual of Style for technical publications (Microsoft Corporation, 2004) as “abbreviations and acronyms that include other abbreviations and acronyms” (Microsoft Corporation, 2004:63). Finally, the Guidelines for Creating and Using Abbreviations and Acronyms of NASA define it as “an acronym within an acronym (sometimes within yet another acronym, ad infinitum)” (Miller, 1995:25).

The definition of nesting used in our research is *a minor-word formation process in which an abbreviated form, either an initialism, acronym or other, is within the meaning of another abbreviation in order to form a new one*. In Figure 2, nested abbreviations are shown in purple since we consider them a different type of abbreviation. This subject will be fully developed in chapter five; because the features of the phenomenon are provided by the findings of our study and are not cited from the literature review.

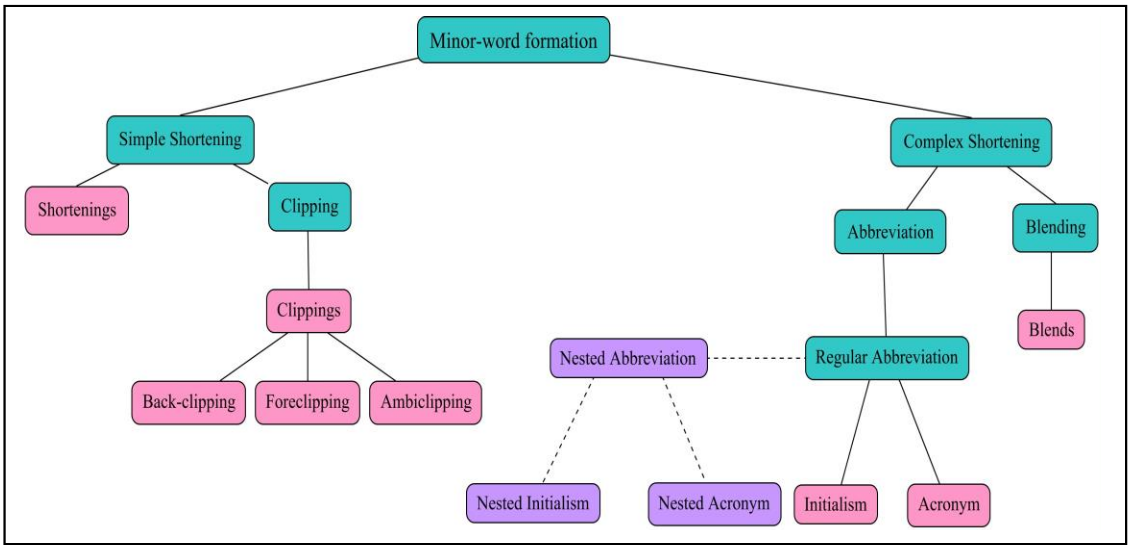


Figure 2. Minor-word Formation with Nested Abbreviation. (Adapted from López, 2004: 126).

Translation of Abbreviations

When translating specialized texts, translators have to face several problems, among them translation of terminology. Authors such as Arntz & Picht (1995) establish that translation of terminology is one the most difficult and time-consuming activity during the translation task, and they explain this fact as follows:

> El traductor se ve obligado a menudo a familiarizarse con la terminología de un texto antes de proceder a su traducción. Este trabajo previo puede robarle mucho tiempo, sobre todo si no está muy versado en la especialidad de la que trata el texto (Arntz & Picht , 1995:18).

This statement is shared by authors like Quiroz & Arroyave (2014) when they claim that “near 40% or 50% of the time employed during process of a specialized translation is dedicated to the resolution of terminological problems” (Quiroz & Arroyave, 2014:138).

As we presented at the beginning of this chapter, abbreviations are highly used in specialized texts and, as shown in subsection 2.2.1.2.1, conceptual abbreviations represent the shortened form of NPs. They abbreviate concepts of the different academic fields and, as stated by Newmark (1988) abbreviations “are frequently created within special topics and designate products, appliances and processes” (Newmark, 1988:148). This is why abbreviations and terminology are closely related and also this is why we attempt to describe some aspects about translation of abbreviations in this section.

It is important to analyze the amount of abbreviations that are adapted from English and translated into other languages, and the number of abbreviations that remain untranslated. This can be explained because translators and specialists in different fields, which have to perform as translators in their daily work, do not know how to proceed when it comes to abbreviations (Gutiérrez, 1998:251). This situation is seen in fields like Medicine as exposed by Navarro (2000):

> La mayor parte de los médicos españoles e hispanoamericanos ejercen [la traducción] con frecuencia de manera informal durante sus estudios universitarios y a lo largo de su carrera profesional. Las publicaciones médicas en lengua española, nadie puede negarlo, son hoy en gran medida el resultado de un proceso de traducción a partir del inglés (…) Debemos aceptar, pues, que en países como los nuestros, de ciencia secundaria y dependiente, todo autor médico es en buena medida un buen traductor (Navarro, 2000:XIII).

According to Navarro (2000), each author proceeds as he considers best leading to the development of all kinds of solutions, which are not always accurate. Gutiérrez (1998) agrees with this author when she states:

> Y es que nadie les ha aclarado a nuestros investigadores si deben traducir o no las siglas inglesas. Ante la falta de una postura clara y unitaria, cada profesional de la ciencia o de la traducción opta por la solución que cree más conveniente y mientras unos mantienen las siglas originales, otros se dedican a traducirlas, cada uno de la mejor forma que puede (Gutiérrez, 1998:278).

Solutions proposed by translators/specialists are presented by Fijo (2003:115) and grouped in three categories, which are described ahead:

1. Translation of the term into a second language and creation of an abbreviation from the resulting term. An example of this is Discoid Lupus Erythematous (DLE) and Lupus Eritematoso Discoide (LED).

According to the author, this solution is the most frequent form of translation of abbreviations. Nonetheless, is the one that leads to more homonymy of all, because one term can have several translations into the second language, originating more than one abbreviation for the same term.

2. Translation of the meaning of abbreviations, leaving abbreviations untranslated. An example is Blood Urea Nitrogen (BUN) and Nitrógeno Ureico en Sangre (BUN).

3. Loans of abbreviations and meanings. This means that abbreviation and its meaning remain untranslated. An example that shows this is *laser* (Light Amplification by Stimulated Emission of Radiation).

Another possible solution of translation involves proper names of institutions such as NACOSA (NATO CIS Operating and Support Agency), where abbreviations and their meanings remain untranslated. Regarding this fact the Academia Real Española (RAE) recommends:

> Solo en casos de difusión general de la sigla extranjera y dificultad para hispanizarla, o cuando se trate de nombres comerciales, se mantendrá la forma original: Unesco, sigla de United Nations Educational, Scientific and Cultural Organization; CD-ROM, sigla de Compact Disc Read-Only Memory; IBM, sigla de International Business Machines.[4]

[4] Extracted from: http://lema.rae.es/dpd/?key=sigla on June 10th, 2016.

There is an evident lack of criteria concerning translation of abbreviations, and it is described by authors like Alonso (2002:163) who explains how foreign abbreviations like Irish Republican Army (IRA) rest untranslated in Spanish newspapers and books, while the meaning is translated. This is consisting with the second solution presented by Fijo (2003).

Another example is presented by Figueroa & Silva (2000:465). In their work, the authors aim to create an abbreviation dictionary. They recommend several translation strategies such as calques in an attempt to put in a second language the meaning of a term and using the words of the target language to create a specific neologism. This is similar to the first solution presented above. They also recommend loans, when the abbreviation remains untranslated, as in the third solution. Moreover, the authors have a strategy when an equivalent in the second language does not exist, and it is described as follows:

> En el caso de que no exista una sigla equivalente en la lengua de llegada, se procede a su explicación o traducción literal, de manera que el usuario pueda completar su información en esa lengua sobre el texto con el que trabaje (Figueroa & Silva, 2000:465).

This lack of criteria on translation of abbreviations might cause problems not only to understand specialized texts, but also in the use of abbreviation identification systems. Dannélls (2005) presents a work that aims to recognize Swedish abbreviations and their definitions in biomedical texts using a system called SARP (Swedish Acronym Recognition Program). The author shows how translation of abbreviations without specific criteria can lower system's performance. This fact is explained by the author because initials in abbreviation do not match letters in the definition:

> The variety of acronym pairs is large and involves different structures which are hard to detect, for example < "VF", "kammarflimmer" >, < "CT", "datortomografi" >, where the acronym is in English while the extension is written in Swedish […]. It becomes more difficult because there are also counter examples in the Swedish text where both the acronym and the definition are in English (Dannélls, 2005:10).

Translation of abbreviations, and therefore nested abbreviations, is a delicate subject and, as mentioned above, there is no evidence of a consistent theory that supports or makes recommendations on how to proceed in these cases. Terminology manuals such as Felber (1984); Arntz & Picht (1995); Cabré (1999) and Pavel & Nolet (2001) define abbreviations but they do not make any suggestions about their translation. Gutiérrez (1998) attempts to provide a recommendation on this matter when she states that:

> Existe una serie de siglas utilizadas en el lenguaje científico que gozan de una implantación internacional, son universalmente aceptadas; esas siglas, lógicamente, no se traducirán. Sin embargo, hay muchísimas otras que, procedentes generalmente del inglés, si se pueden traducir. Como no existe una norma clara al respecto, el traductor deberá consultar los glosarios mono o multilingües de siglas existentes y tratar de seguir siempre la costumbre que haya entre los especialistas en ese tema (Gutiérrez, 1998:257).

Likewise, translation manuals as Newmark (1988) only recommend to translate abbreviations if they are "as important in the SL as in the TL" (Newmark, 1988:148). And again, the decision is left to translators/specialists, who have to decide if the translation of an abbreviation should be made and, what kind of solution should be offered to arrive at the best version they can think of.

Methodology

In order to accomplish the objectives of the present study, a quantitative and qualitative methodological strategy with an exploratory scope is proposed. The general methodological design of this work is shown in Figure 3. The collected corpus, tokenization and tagging processes, and tools used to process and analyze data are presented in this chapter.

This process started with literature review on minor-word formation, in order to define minor-word forms, including nested abbreviations, and linguistic criteria to analyze them.

Afterwards, criteria for dictionary selection was established and collection of corpus started. Data extraction included: identification of nested abbreviations in selected dictionaries, tokenization, syntactic tagging, validation of terms in general and specialized corpora. Translation of abbreviations and validation of translations by three Colombian professional translators is an important step before examining our data.

Data analysis involved a linguistic description of nested abbreviations extracted from dictionaries. This description comprised morphological, syntactic and semantic aspects of abbreviations. Morphological aspects covered information such as formation, type, length, spelling, number and use of articles. Aspects like lexical category, morphosyntactic patterns and syntactic relations are included in the syntactic description. Semantic aspects like type of abbreviation (nominal or conceptual), field, and synonymy and homonymy process were also analyzed.

Translation analysis involved a comparison of syntactic aspects between abbreviations and meanings extracted from dictionaries in English and their translation into Spanish. Finally, we present the concluding remarks and recommendations.

General methodology of this study is shown in Figure 3, and it is presented ahead.

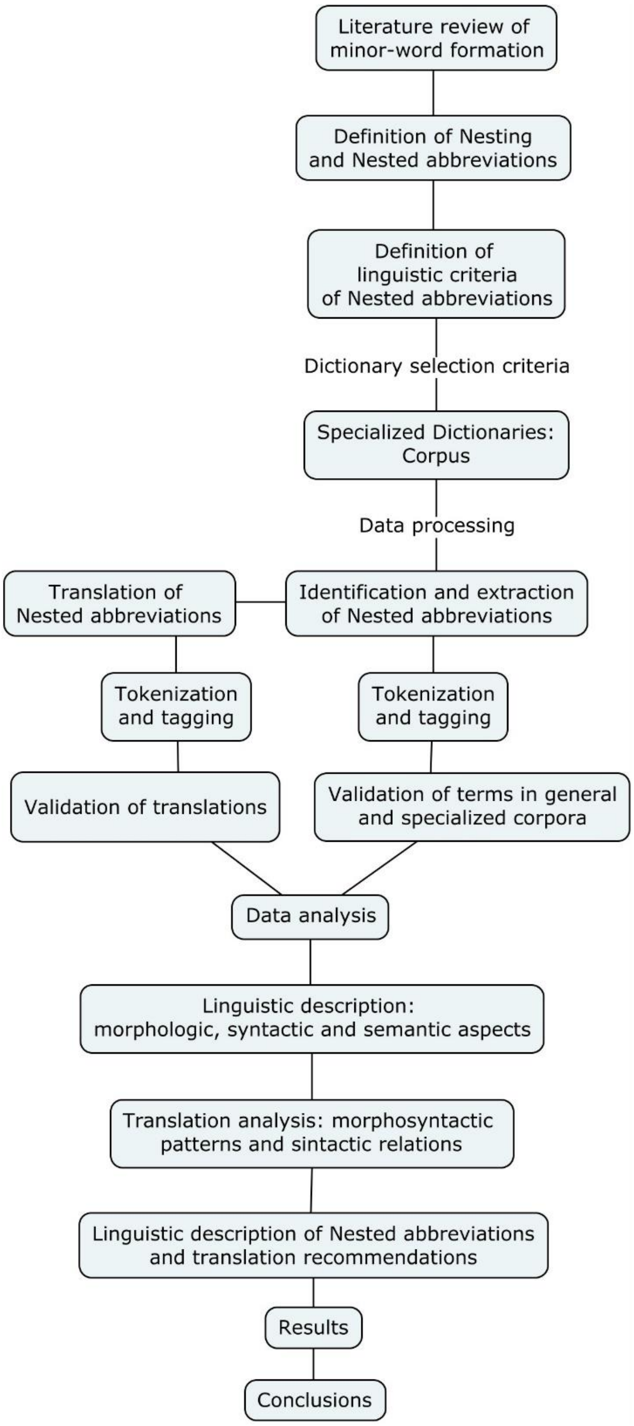

Literature review of minor-word formation
Definition of Nesting and Nested abbreviations
Definition of linguistic criteria of Nested abbreviations
Dictionary selection criteria
Specialized Dictionaries: Corpus
Data processing
Translation of Nested abbreviations
Identification and extraction of Nested abbreviations
Tokenization and tagging
Tokenization and tagging
Validation of translations
Validation of terms in general and specialized corpora
Data analysis
Linguistic description: morphologic, syntactic and semantic aspects
Translation analysis: morphosyntactic patterns and sintactic relations
Linguistic description of Nested abbreviations and translation recommendations
Results
Conclusions

Figure 3. General work methodology.

Description and Collection of Corpus

The corpus used in this work was compiled from two dictionaries of acronyms and abbreviations: Jablonski's Dictionary of Medical Acronyms and Abbreviations, 6th edition (Jablonski, 2009) and Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols (Mattia, 2003). As shown in Table 4, from these dictionaries, 62068 abbreviations and 433 nested abbreviations were obtained.

Table 4. General description of the dictionaries.

Dictionary	**Field**	**Entries**	**Nested Abbreviations**	**% From the total entries**
Jablonski's	Medicine	28068	85	0.30%
Elsevier's	3000 fields and subfields	34000	348	1.02%
Total		62068	433	0.69%

To collect the corpus, one of the criteria used was the selection of dictionaries in related areas to control the results, this is because "El tener varios ámbitos, abriría otras puertas pero también variables que no podríamos controlar de manera fiable" (Quiroz, 2008:75). The dictionaries were selected according to the following criteria:

a. Availability in electronic format.

b. Update: The year of publication must be between 2003 and 2013, in order to have abbreviations that are still in use and not obsolete terms.

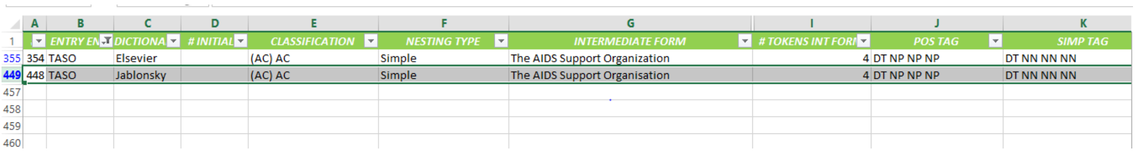
c. The author of the dictionary must be an expert in the field and the publishing house must be recognized and be considered to be prestigious by the scientific society.

d. The dictionary must be cited in other scientific papers as a sign of quality (Sánchez-Gijón, 2004:33).

1. Jablonski's Dictionary of Medical Acronyms and Abbreviations, 6th edition (2009), was written by Stanley Jablonski, a Polish indexer of medical literature, who worked an important part of his life in the production of medical indices and dictionaries. The dictionary was published by Saunders Elsevier, which has been an authority in scientific books since 1880 and manages subjects like Medicine, Chemistry, Technology, among others. The dictionary is available in electronic format and consists of 544 pages with 28068 abbreviations related to Medicine. It has been cited in 35 articles according to the information provided by Google Scholar®.

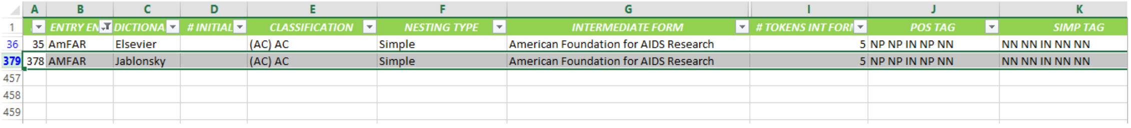
2. Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols (2003). The author of this dictionary, Fioretta Benedetto Mattia, has a strong academic background, she studied Commercial Sciences and Languages, afterwards she studied Philosophy, Psychology, Journalism, Law and Economics of the European Communities, Criminal Law and International Business, besides, she is a certified translator. The dictionary published by Elsevier and covers over 3000 different fields and subfields in sciences. The dictionary is available in electronic format and consists of 721 pages with 34000 acronyms, according to the author. It has been cited in three academic papers according to Google Scholar®.

From Jablonski's (2009) dictionary we extracted 85 nested abbreviations, after excluding 2 abbreviations that were also present in the other dictionary used (Figures 4 and 5) and from Elsevier's (2003) dictionary we extracted 349 nested acronyms.



	A	B	C	D	E	F	G	I	J	K
1		ENTRY EN	DICTIONA	# INITIAL	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM	# TOKENS INT FORI	POS TAG	SIMP TAG
355	354	TASO	Elsevier		(AC) AC	Simple	The AIDS Support Organization	4	DT NP NP NP	DT NN NN NN
449	448	TASO	Jablonsky		(AC) AC	Simple	The AIDS Support Organisation	4	DT NP NP NP	DT NN NN NN
457										
458										
459										
460										

Figure 4. Acronym TASO excluded from the corpus.



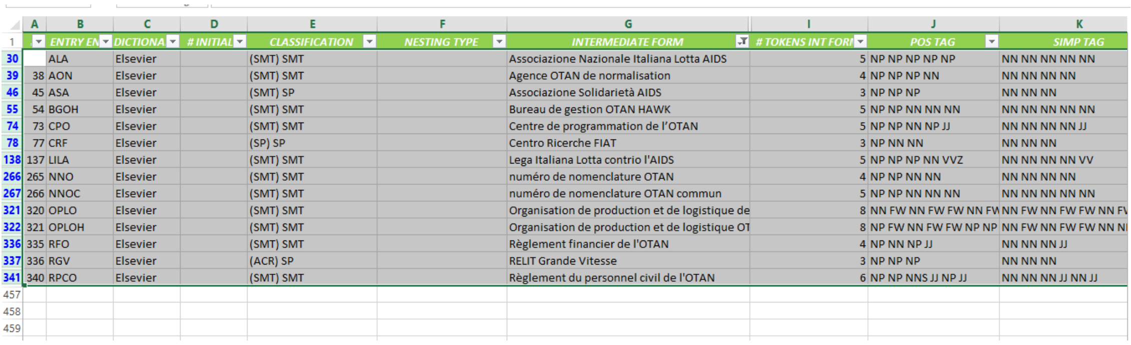
	A	B	C	D	E	F	G	I	J	K
1		ENTRY EN	DICTIONA	# INITIAL	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM	# TOKENS INT FORI	POS TAG	SIMP TAG
36	35	AmFAR	Elsevier		(AC) AC	Simple	American Foundation for AIDS Research	5	NP NP IN NP NN	NN NN IN NN NN
379	378	AMFAR	Jablonsky		(AC) AC	Simple	American Foundation for AIDS Research	5	NP NP IN NP NN	NN NN IN NN NN
457										
458										
459										

Figure 5. Acronym AMFAR excluded from the corpus.

During the collection of data several difficulties arose. To begin with, a bilingual dictionary of abbreviations that complied all the criteria mentioned above was not available online. To overcome this difficulty, a Spanish abbreviation dictionary was consulted to collect information in a second language. So, the Dictionary of Medical Acronyms / Diccionario de Acrónimos Médicos written by Yetano & Alberola (2004) was consulted and the data was processed. Unfortunately, only 3 nested abbreviations were obtained from over 20000 present in the dictionary.

A similar situation happened with the Routledge Spanish Dictionary of Business, Commerce and Finance / Diccionario Inglés de Negocios, Comercio y Finanzas (1997), a bilingual specialized dictionary with more than 38000 entries of business. Although this dictionary was published in 1997 and did not fulfill the update criteria, we decided to process this data, because it is a bilingual specialized dictionary. However, this dictionary also provided just 3 nested abbreviations. This is the reason why we decided to exclude these two dictionaries from the present study.

We also decided to work only with English abbreviations and exclude 14 abbreviations found in Elsevier's Dictionary written in French (9) and Italian (5). They represented 3.5% of the initial number of abbreviations which was 455 (Figure 6).



	A	B	C	D	E	F	G	I	J	K
1		ENTRY EN	DICTIONA	# INITIAL	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM	# TOKENS INT FORI	POS TAG	SIMP TAG
30		ALA	Elsevier		(SMT) SMT		Associazione Nazionale Italiana Lotta AIDS	5	NP NP NP NP NP	NN NN NN NN NN
39	38	AON	Elsevier		(SMT) SMT		Agence OTAN de normalisation	4	NP NP NP NN	NN NN NN NN
46	45	ASA	Elsevier		(SMT) SP		Associazione Solidarietà AIDS	3	NP NP NP	NN NN NN
55	54	BGOH	Elsevier		(SMT) SMT		Bureau de gestion OTAN HAWK	5	NP NP NN NN NN	NN NN NN NN NN
74	73	CPO	Elsevier		(SMT) SMT		Centre de programmation de l'OTAN	5	NP NP NN NP JJ	NN NN NN NN JJ
78	77	CRF	Elsevier		(SP) SP		Centro Ricerche FIAT	3	NP NN NN	NN NN NN
138	137	LILA	Elsevier		(SMT) SMT		Lega Italiana Lotta contrio l'AIDS	5	NP NP NP NN VVZ	NN NN NN NN VV
266	265	NNO	Elsevier		(SMT) SMT		numéro de nomenclature OTAN	4	NP NP NN NN	NN NN NN NN
267	266	NNOC	Elsevier		(SMT) SMT		numéro de nomenclature OTAN commun	5	NP NP NN NN NN	NN NN NN NN NN
321	320	OPLO	Elsevier		(SMT) SMT		Organisation de production et de logistique de	8	NN FW NN FW FW NN FW	NN FW NN FW FW NN F
322	321	OPLOH	Elsevier		(SMT) SMT		Organisation de production et de logistique OT	8	NP FW NN FW FW NP NP	NN FW NN FW FW NN N
336	335	RFO	Elsevier		(SMT) SMT		Règlement financier de l'OTAN	4	NP NN NP JJ	NN NN NN JJ
337	336	RGV	Elsevier		(ACR) SP		RELIT Grande Vitesse	3	NP NP NP	NN NN NN
341	340	RPCO	Elsevier		(SMT) SMT		Règlement du personnel civil de l'OTAN	6	NP NP NNS JJ NP JJ	NN NN NN JJ NN JJ
457										
458										
459										

Figure 6. French and Italian abbreviations excluded from corpus.

In order to analyze the nested abbreviations phenomenon and its relation to translation, 433 English abbreviations obtained from our corpus were translated. Terminology online databases such as IATE[5], UNTerm[6], HonSelect[7], Snomed[8], Termium Plus[9] and NATO bilingual glossaries (NATO, 2000, 2005, 2010, 2013) were used. Spanish translations of nested abbreviations and their meanings (or intermediate form) were validated by three Colombian professional translators, graduated from the University of Antioquia. They work as freelance translators.

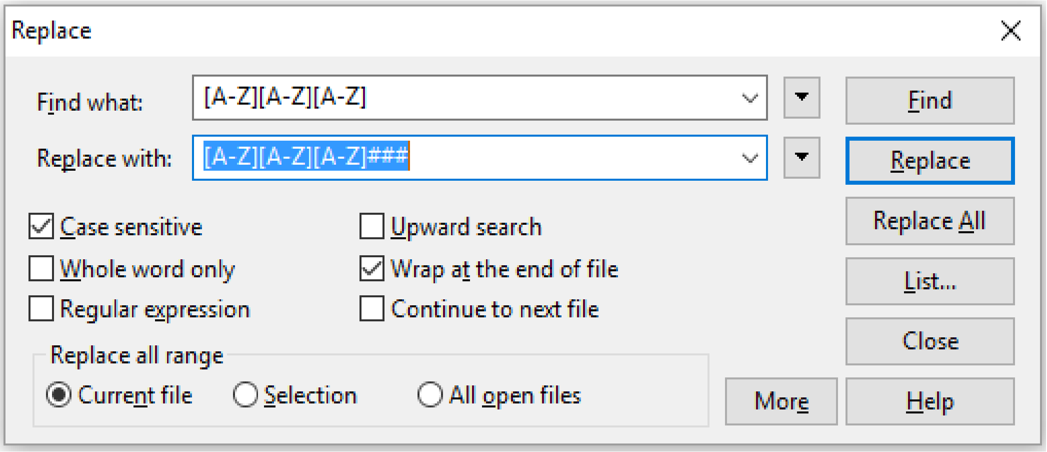
To collect data from the dictionaries mentioned above, the following process was carried out. First, the dictionaries were converted from PDF file to a .docx file using the Solid Converter® computer program in order to handle the information contained in the dictionaries.

[5] Available in http://iate.europa.eu/switchLang.do?success=mainPage&lang=en. Consulted on May, 2016.
[6] Available in http://unterm.un.org/. Consulted on May, 2016.
[7] Available in https://www.hon.ch/MeSH/. Consulted on May, 2016.
[8] Available in http://www.ihtsdo.org/snomed-ct. Consulted on May, 2016.
[9] Available in http://www.btb.termiumplus.gc.ca/. Consulted on May, 2016.

Second, the regular expression **^&^t** was used in the resulting text to separate abbreviations, which were written in bold, from their meaning, and then the resulting information was taken to an Excel® spreadsheet.

Third, to avoid noise in corpus, information such as explanations and further reading recommended by the author of the dictionary, was suppressed from the intermediate form.

Fourth, the column containing the meaning of each abbreviation was copied into a .txt file using the text editor EditPlus®, in order to tag the presence of abbreviations with the symbol ###. The program searched the regular expression [A-Z] [A-Z] [A-Z] and replaced it with [A-Z] [A-Z] [A-Z] ### (see Figure 7).



Find what: [A-Z][A-Z][A-Z]

Replace with: [A-Z][A-Z][A-Z]###

☑ Case sensitive
□ Whole word only
□ Regular expression
□ Upward search
☑ Wrap at the end of file
□ Continue to next file

Replace all range
● Current file ○ Selection ○ All open files

Figure 7. Use of regular expressions to label with ### three consecutive upper case letters.

Finally, intermediate forms that have 3 consecutive uppercase letters were manually reviewed to make sure that they were, in fact, abbreviations.

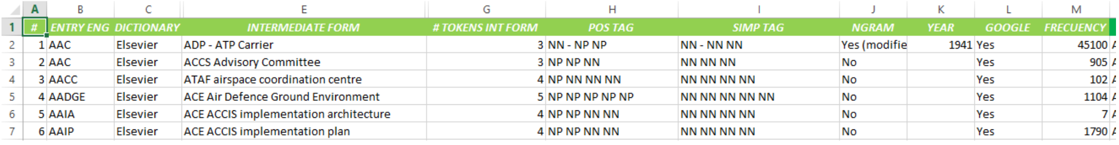
Once the corpus was collected, data was organized in an Excel® spreadsheet; and separated into two main groups: the intermediate form, which was provided by the dictionary and where the abbreviation is undeveloped, and the extended form, which is our creation and contains the abbreviation within the meaning fully developed.

Each group was processed separately to get the number of tokens in the same way used by Rojas (2014:76) using the Excel formula:

```
=SI(LARGO(ESPACIOS(H2))=0;0;LARGO(ESPACIOS(H2))-
   LARGO(SUSTITUIR(H2;" ";""))+1)
```

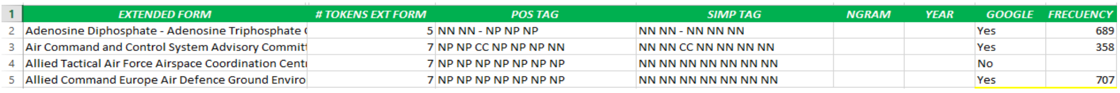
Besides the number of tokens, the part of speech (POS) of each component of the NP was also analyzed and tagged using TreeTagger[10].

In order to validate each term, the frequency of appearance in Google, which represents the general corpus, and Ngram viewer, which is the specialized corpus (Figure 8 and 9) was another aspect considered.



	A	B	C	E	G	H	I	J	K	L	M
1	#	ENTRY ENG	DICTIONARY	INTERMEDIATE FORM	# TOKENS INT FORM	POS TAG	SIMP TAG	NGRAM	YEAR	GOOGLE	FRECUENCY
2	1	AAC	Elsevier	ADP - ATP Carrier	3	NN - NP NP	NN - NN NN	Yes (modifie	1941	Yes	45100
3	2	AAC	Elsevier	ACCS Advisory Committee	3	NP NP NN	NN NN NN	No		Yes	905
4	3	AACC	Elsevier	ATAF airspace coordination centre	4	NP NN NN NN	NN NN NN NN	No		Yes	102
5	4	AADGE	Elsevier	ACE Air Defence Ground Environment	5	NP NP NP NP NP	NN NN NN NN NN	No		Yes	1104
6	5	AAIA	Elsevier	ACE ACCIS implementation architecture	4	NP NP NN NN	NN NN NN NN	No		Yes	7
7	6	AAIP	Elsevier	ACE ACCIS implementation plan	4	NP NP NN NN	NN NN NN NN	No		Yes	1790

Figure 8. Intermediate form.



1	EXTENDED FORM	# TOKENS EXT FORM	POS TAG	SIMP TAG	NGRAM	YEAR	GOOGLE	FRECUENCY
2	Adenosine Diphosphate - Adenosine Triphosphate (	5	NN NN - NP NP NP	NN NN - NN NN NN			Yes	689
3	Air Command and Control System Advisory Commit	7	NP NP CC NP NP NP NN	NN NN CC NN NN NN NN			Yes	358
4	Allied Tactical Air Force Airspace Coordination Cent	7	NP NP NP NP NP NP NP	NN NN NN NN NN NN NN			No	
5	Allied Command Europe Air Defence Ground Enviro	7	NP NP NP NP NP NP NP	NN NN NN NN NN NN NN			Yes	707

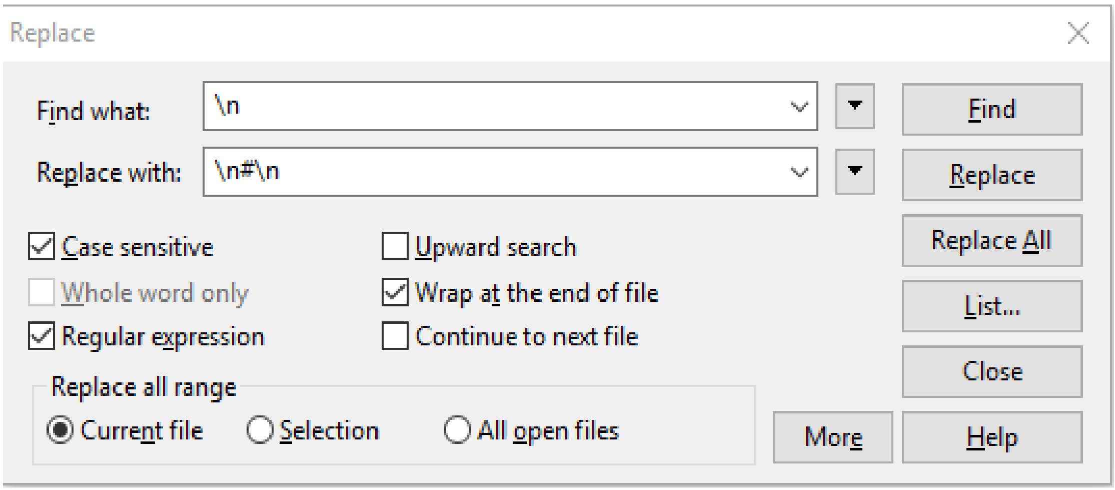
Figure 9. Extended form.

Corpus Tagging

In order to obtain the part of speech of each word of the intermediate and extended form, the TreeTagger tool was used. A brief description of the tool will be provided in the next subsection. For now, the process involved to tag each syntagm is described.

[10] Available in http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/. Consulted on April and May 2016

First, the column containing intermediate and extended forms were pasted in a new .txt document in EditPlus. Second, each New Line (\n) was replaced with regular expression **\n # \n** (Figure 10), and later each (\n) was replaced by (space). Third, the document mentioned above was opened in TreeTagger (Figure 11) and a second file called file_tag.txt, which had all intermediate and extended forms tagged, was created. The information contained in file_tag.txt was copied later in a new document in EditPlus. Through regular expressions, each New Line (\n) was replaced by (space), and then all (space) # (space) were replaced by New Line (\n) and, finally, this column was copied in an Excel document.



Replace

Find what: \n

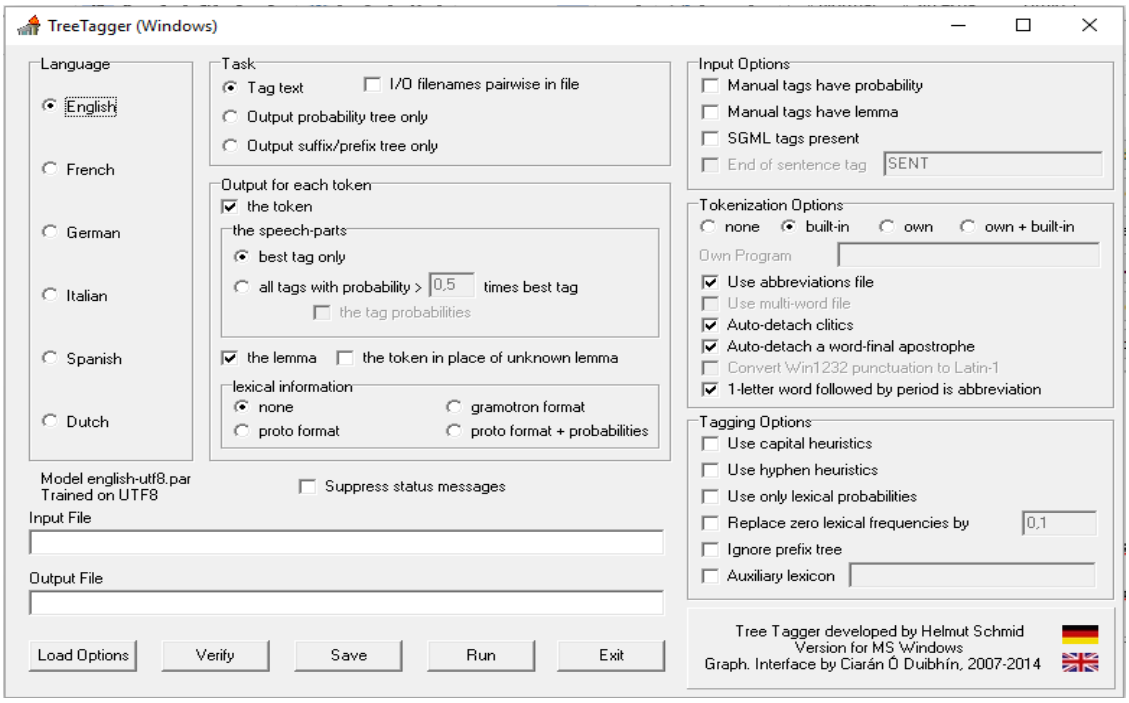
Replace with: \n#\n

☑ Case sensitive
□ Whole word only
☑ Regular expression
□ Upward search
☑ Wrap at the end of file
□ Continue to next file

Replace all range
● Current file ○ Selection ○ All open files

Figure 10. Use of regular expression in EditPlus.

In order to avoid tagging errors with the TreeTagger tool, all NPs of the intermediate and extended forms were written in lowercase letters, except the abbreviations. The reason for that is because when any word is written with its initial in uppercase, the tagging tool recognizes it as a proper noun.



Language
● English
○ French
○ German
○ Italian
○ Spanish
○ Dutch

Model english-utf8.par
Trained on UTF8

Task
● Tag text
☐ I/O filenames pairwise in file
○ Output probability tree only
○ Output suffix/prefix tree only

Output for each token
☑ the token
the speech-parts
● best tag only
○ all tags with probability > 0,5 times best tag
☐ the tag probabilities
☑ the lemma
☐ the token in place of unknown lemma
lexical information
● none
○ gramotron format
○ proto format
○ proto format + probabilities

☐ Suppress status messages

Input File

Output File

Input Options
☐ Manual tags have probability
☐ Manual tags have lemma
☐ SGML tags present
☐ End of sentence tag SENT

Tokenization Options
○ none ● built-in ○ own ○ own + built-in
Own Program
☑ Use abbreviations file
☐ Use multi-word file
☑ Auto-detach clitics
☑ Auto-detach a word-final apostrophe
☐ Convert Win1232 punctuation to Latin-1
☑ 1-letter word followed by period is abbreviation

Tagging Options
☐ Use capital heuristics
☐ Use hyphen heuristics
☐ Use only lexical probabilities
☐ Replace zero lexical frequencies by 0,1
☐ Ignore prefix tree
☐ Auxiliary lexicon

Tree Tagger developed by Helmut Schmid
Version for MS Windows
Graph. Interface by Ciarán Ó Duibhín, 2007-2014

Figure 11. TreeTagger interface.

Tools

Office automation tools.

For the purpose of processing the information that was initially not in an editable format, Solid Converter® program was used. It is an efficient, fast and user-friendly tool to convert PDF files into fully editable Windows documents[11]. To organize the information, a Microsoft Office Excel® spreadsheet was used. This program allowed to structure data in columns and rows to separate each feature of nested abbreviations to be analyzed. EditPlus®, a text editor, was also implemented in our work due to its efficiency at the time to use regular expressions and process large amounts of information.

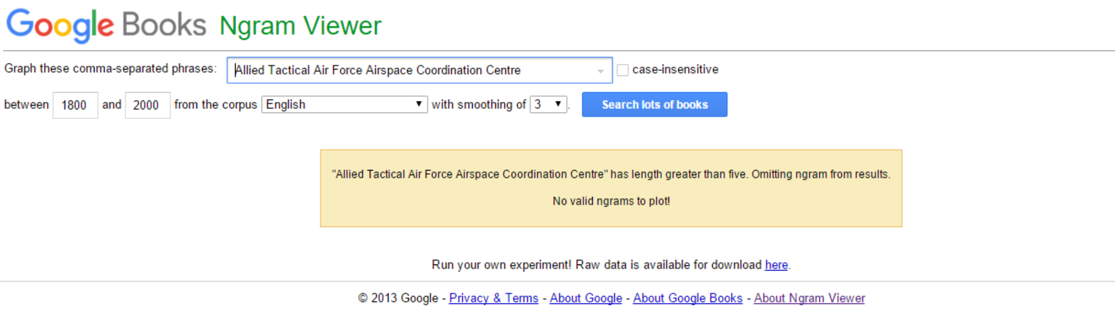
[11] http://www.soliddocuments.com/features.htm?product=SolidConverterPDF. Consulted on October 2015.

Linguistic tools.

Ngram viewer[12].

The Ngram viewer is an online tool that allows the graphic representation of the trends in different ngrams. It shows their behavior in a corpus of specialized books in a selected range of years, different languages and areas of interest. The tool also provides other services such as wildcard (top ten substitutions), inflections (modification of a word in different grammatical categories), case insensitive searches, part of speech tags and ngram compositions.

One limitation that has been noted when using this tool is that it only works with ngrams smaller than five tokens, after five tokens the tool cannot find them, as shown in Figure 12. This is the reason why all intermediate and extended forms were searched in Google, the general corpus.



Google Books Ngram Viewer

Graph these comma-separated phrases: Allied Tactical Air Force Airspace Coordination Centre □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 . Search lots of books

"Allied Tactical Air Force Airspace Coordination Centre" has length greater than five. Omitting ngram from results.

No valid ngrams to plot!

Run your own experiment! Raw data is available for download here.

© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

Figure 12. Ngram viewer[13].

TreeTagger.

TreeTagger is a tool designed for morphosyntactic tagging and lemma information using a binary decision tree to “obtain reliable estimates of transition probabilities” (Schmid, 1994:1).

[12] https://books.google.com/ngrams. Consulted on December 2015 and January 2016.
[13] Obtained from http://books.google.com/ngrams on May 15th 2016.

It is possible to tag texts in several languages such as German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Portuguese, Galician, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Coptic and old French. In this study, this tool was used in order to obtain the morphosyntactic category of each element constituting the intermediate and the extended form of the all abbreviations. The set of steps followed to obtain the parts of speech was described in the corpus tagging section.

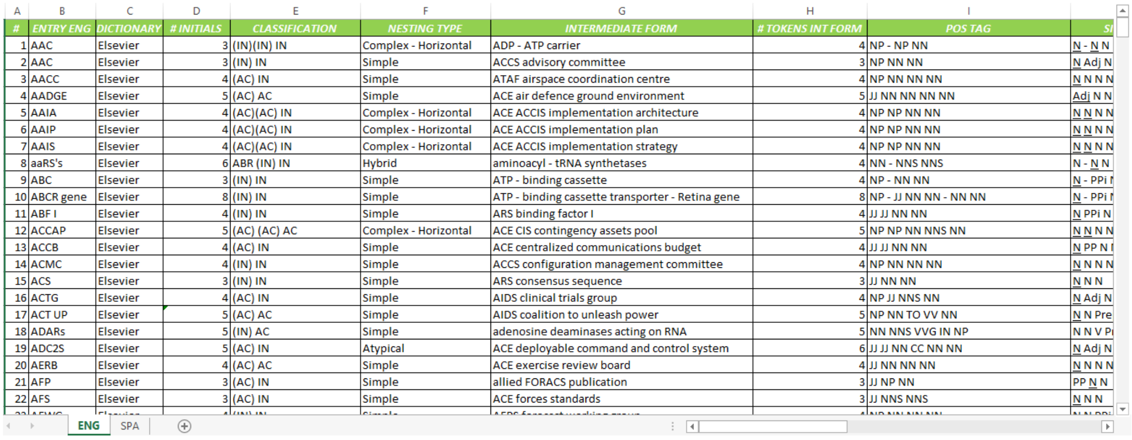
Statistics tools.

Statgraphics Centurion®.

Since our study has a quantitative approach with an exploratory scope, the version 16.1 of this statistic tool was used in order to perform a descriptive analysis of our data. This program allows the description of numerical and categorical data and tabulate information. The tool provides frequency tables, bar diagrams and sector diagrams of requested variables that are interpreted and analyzed in chapter four.

Description of the Database

The Database used in our work is constituted by 2 spreadsheets of 24 columns for English data and 12 columns for Spanish data, since this last spreadsheet only has information of intermediate forms. Information is distributed as follows: *Number, Entry EN / Entry ES, Dictionary, # Initials, Classification, Nesting type, Intermediate form, # Tokens int form, Pos tag, Simp tag, Ngram, Year, Google, Frequency, Semantic, Field, Extended form, # Tokens ext form, Pos tag, Simp tag, Ngram, Year, Google, Frequency,* (Figure 13).



#	ENTRY ENG	DICTIONARY	# INITIALS	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM	# TOKENS INT FORM	POS TAG	SI
1	AAC	Elsevier	3	(IN)(IN) IN	Complex - Horizontal	ADP - ATP carrier	4	NP - NP NN	N - N N
2	AAC	Elsevier	3	(IN) IN	Simple	ACCS advisory committee	3	NP NN NN	N Adj N
3	AACC	Elsevier	4	(AC) IN	Simple	ATAF airspace coordination centre	4	NP NN NN NN	N N N N
4	AADGE	Elsevier	5	(AC) AC	Simple	ACE air defence ground environment	5	JJ NN NN NN NN	Adj N N
5	AAIA	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation architecture	4	NP NP NN NN	N N N N
6	AAIP	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation plan	4	NP NP NN NN	N N N N
7	AAIS	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation strategy	4	NP NP NN NN	N N N N
8	aaRS's	Elsevier	6	ABR (IN) IN	Hybrid	aminoacyl - tRNA synthetases	4	NN - NNS NNS	N - N N
9	ABC	Elsevier	3	(IN) IN	Simple	ATP - binding cassette	4	NP - NN NN	N - PPi N
10	ABCR gene	Elsevier	8	(IN) IN	Simple	ATP - binding cassette transporter - Retina gene	8	NP - JJ NN NN - NN NN	N - PPi N
11	ABF I	Elsevier	4	(IN) IN	Simple	ARS binding factor I	4	JJ JJ NN NN	N PPi N
12	ACCAP	Elsevier	5	(AC) (AC) AC	Complex - Horizontal	ACE CIS contingency assets pool	5	NP NP NN NNS NN	N N N N
13	ACCB	Elsevier	4	(AC) IN	Simple	ACE centralized communications budget	4	JJ JJ NN NN	N PP N I
14	ACMC	Elsevier	4	(IN) IN	Simple	ACCS configuration management committee	4	NP NN NN NN	N N N N
15	ACS	Elsevier	3	(IN) IN	Simple	ARS consensus sequence	3	JJ NN NN	N N N
16	ACTG	Elsevier	4	(AC) IN	Simple	AIDS clinical trials group	4	NP JJ NNS NN	N Adj N
17	ACT UP	Elsevier	5	(AC) AC	Simple	AIDS coalition to unleash power	5	NP NN TO VV NN	N N Pre
18	ADARs	Elsevier	5	(IN) AC	Simple	adenosine deaminases acting on RNA	5	NN NNS VVG IN NP	N N V P
19	ADC2S	Elsevier	5	(AC) IN	Atypical	ACE deployable command and control system	6	JJ JJ NN CC NN NN	N Adj N
20	AERB	Elsevier	4	(AC) AC	Simple	ACE exercise review board	4	JJ NN NN NN	N N N N
21	AFP	Elsevier	3	(AC) IN	Simple	allied FORACS publication	3	JJ NP NN	PP N N
22	AFS	Elsevier	3	(AC) IN	Simple	ACE forces standards	3	JJ NNS NNS	N N N
23	AFWG	Elsevier	4	(IN) IN	Simple	AFPS forecast working group	4	NP NN NN NN	N N PPi

Figure 13. Database.

Each component of the database is explained below in more detail:

a. Number: the number that each abbreviation received in the database.

b. Entry: the alphabetical organization of the acronyms.

c. Dictionary: the source from where the abbreviation was extracted.

d. # Initials: the number of letters used in the abbreviations.

e. Classification: indication of whether the abbreviation is an initialism, acronym or other.

f. Type of nesting: indication of Simple or Complex, Vertical or Horizontal nesting.

g. Intermediate form: the definition of the abbreviation provided by the dictionary.

h. # Tokens int form: the number of words of the syntagm in the intermediate form.

i. TreeTagger: the part of speech of the intermediate form provided by TreeTagger.

j. Simp tag: the simplified tag.

k. Ngram: indication of yes (1) or no (0) the intermediate form was found in Ngram viewer (NA if the syntagm had more than five tokens).

l. Year: the first appearance of the intermediate form in Ngram viewer (NF if the intermediate form was not found).

m. Google: indication of Yes (1) or No (0) the intermediate form was found in Google.

n. Frequency: the number of occurrences of the intermediate form in Google.

o. Semantic: indication of Nominal or Conceptual according to the semantic classification.

p. Field: the discipline related to abbreviation.

q. Extended form: the abbreviation fully developed, provided by the author of this study.

r. # Tokens ext form: the number of words of the syntagm in the extended form.

s. TreeTagger*:* the part of speech of the extended form provided by TreeTagger.

t. Simp tag: the simplified tag.

u. Ngram: same as above.

v. Year: same as above.

w. Google: same as above.

x. Frequency: the number of occurrences in Google of the extended form.

Data Analysis

In order to analyze data obtained from our corpus, the statistic tool Statgraphics Centurion® was used. This program allows to analyze statistical values and their graphic representation through tabulation of numeric and categorical data. Excel ® was used in this phase as well.

As the first moment of this research involved a linguistic description of nested abbreviations, all nested abbreviations extracted from two specialized dictionaries in English were analyzed. Using Excel® filters and Statgraphics Centurion® tool, a descriptive statistical analysis, including quantification of frequencies and percentages, is performed. Linguistic description involved the following features:

1. Morphological features such as formation, type, length, spelling, number and use of articles. They are examined and presented along with examples to improve understanding of the phenomenon.

2. Syntactic aspects like lexical category, morphosyntactic patterns and syntactic relations. They were observed and compared with other studies like Cannon (1989); Grange & Bloom (2000); Alcaraz (2002) and Giraldo (2006), in order to validate our results.

3. Semantic features, for instance, semantic type, field, homonymy and synonymy. They were also analyzed and tabulated.

Apart from the linguistic description, the development of abbreviations within each meaning of the nested abbreviations, extracted from the corpus, allowed the identification and classification of different types of nesting. Their frequencies and percentages were also tabulated and examples of each type of nesting are shown.

In the second moment of this study the qualitative part of our analysis was conducted, since we attempted to examine and compare translations from a syntactic perspective. Although, we present our results with numerical data, the underlying processes have qualitative implications.

Due to a high number of morphosyntactic patterns in English and Spanish (149 vs 165), a sample of 173 noun phrases (NPs) out of 13 most common morphosyntactic patterns of our corpus was manually selected. This sample is denominated *analysis corpus* and represented 86.30% of the most common patterns found in corpus.

The size of the sample was determined using a sample size calculator available online[14], setting a confidence level of 98% and a confidence interval of 5%. To accomplish the task, 173 NPs were selected out of 252 NPs corresponding to the 13 patterns mentioned above, accounting for 68.65% of NPs.

Moreover, NPs were proportionally distributed according to the frequency of the morphosyntactic pattern. This means that the most prevalent patterns have more NPs assigned, and the least frequent the lowest number. For example, pattern N N N N is the most frequent and has 47 NPs assigned to analyze and pattern N – PPi N is the least frequent and has 4.

Each morphosyntactic pattern was related to its syntactic relation and compared with their equivalents in Spanish. This is made in order to provide some recommendations associated to the translation of these entities.

[14] Calculator.net, Available at http://www.calculator.net/sample-size-calculator.html. Consulted on May 15th, 2016

Results and Analysis

In this chapter, which is divided in three sections, we present the main findings taken from the data collected from our corpus. The first section shows a description of morphological, syntactic and semantic aspects of nested abbreviations. The second section introduces findings related to types of nesting with their corresponding examples. The final section presents a comparison between nested abbreviations extracted from our corpus and the proposed translations, from a syntactic perspective.

It is important to mention that most of the characteristics used to explain our results are given as basic criteria to identify and analyze abbreviations. These characteristics were extracted from theories of several authors exposed in chapter two.

Linguistic Analysis

From the two dictionaries used in this research, we obtained 62068 abbreviations. 28068 abbreviations were extracted from Jablonski's Dictionary of Medical Acronyms and Abbreviations, 6th edition. 85 were nested abbreviations, and they represent 0.30% from the total entries featured in this dictionary. From Elsevier's Dictionary of Acronyms, Initialisms, Abbreviations and Symbols, 34000 abbreviations were processed and 348 nested abbreviations were obtained, which represents 1.02% of all entries. This information is summarized in Table 5:

Table 5. General Information of Dictionaries

Dictionary	**Field**	**Entries**	**Nested Abbreviations**	**% From Total Entries**
Jablonski's	Medicine	28068	85	0.30%
Elsevier's	3000 fields and subfields	34000	348	1.02%
Total		62068	433	0.69%

Morphological Aspects of Nested Abbreviations.

Formation.

Regarding the formation of the abbreviations from our corpus, the percentage of the abbreviations formed with the initial letters is shown in Figure 14, along with those abbreviations formed with other letters besides initials or with the use of structural words.

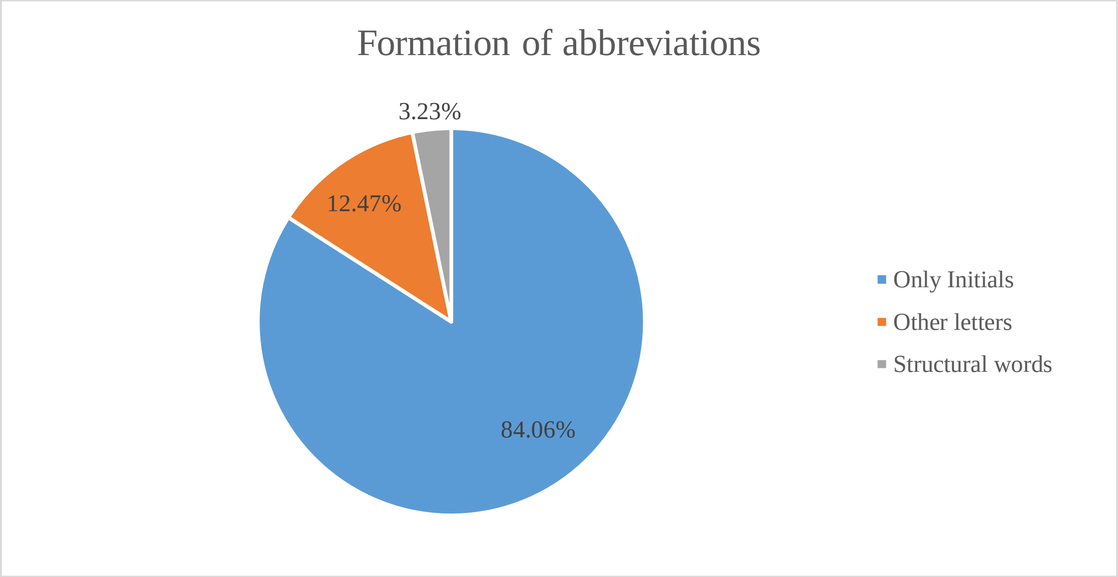


Figure 14. Formation of abbreviations.

As shown in Figure 14, there are three groups of abbreviations. The first one is represented by the abbreviations formed only by the initial letters of the words of the NP. The second one includes the abbreviations formed with initials and other letters. The third group is constituted by abbreviations which include structural words in their formation. The first group represents 84.06% of the abbreviations analyzed in our corpus. Out of 433 nested abbreviations, 364 were formed only by the initial letters of the words from the NP. A case showing this type of abbreviations is presented in Example 1.

(Example 1). ABC = **A**TP-**B**inding **C**assette (Adenosine Triphosphate-Binding Cassette).

The second group of abbreviations corresponds to 54 cases that had other letters besides initials; this is equivalent to 12.47% of all the abbreviations of our corpus. Of this second collection, it is worth mentioning that in 12 cases, 22.22% of the 54 abbreviations, the letter was a “s” (10 written in lowercases and 2 in uppercases) that indicated plural (see Example 2). In 6 cases, which represent 11.11% of this group, the additional letters corresponded to clippings that were inside the definition of the abbreviation, (see Example 3). In 50% of the cases, 27 abbreviations, the additional letter was employed to form acronyms (see Example 4).

(Example 2). ADARs = **A**denosine **D**eaminases [plural] **A**cting on **R**NA (Adenosine Deaminases Acting on Ribonucleic Acid).

(Example 3). LysRs = **Lys**yl [clipping] – t**R**NA **S**ynthetase (Lysyl – transfer Ribonucleic Acid Synthetase).

(Example 4). DIABLO = **D**irect **IA**P - **B**inding protein with **LO**w pI (Direct Inhibitor of Apoptosis Protein - Binding Protein with low pI).

The 15 cases of the third group of abbreviations constituted 3.23% of the nested abbreviations of our corpus, and used “structural words” in their formation process. From this group, it is relevant to mention that in 7 cases, which represents 50% of this group, these words were used to form acronyms (see Example 5), in 4 cases out of 15 (28.58%) the symbol “&” was introduced to replace the word “and” (see Example 6).

(Example 5). ENCODE = **Enc**yclopedia **o**f **D**NA **E**lements (Encyclopedia of Deoxyribonucleic Acid elements).

(Example 6). NAEW**&**C FC = NATO Airborne Early Warning **and** Control Force Command (North Atlantic Treaty Organization Airborne Early Warning and Control Force Command).

If we go back to the first group, 335 abbreviations (92% of the abbreviations formed with initial letters only) used each word of the NP in order to constitute the abbreviation (see Example 7). In 29 cases, 8% of abbreviations, there was an omission of one or more components of the NP (see Example 8). This information is presented in Figure 15:

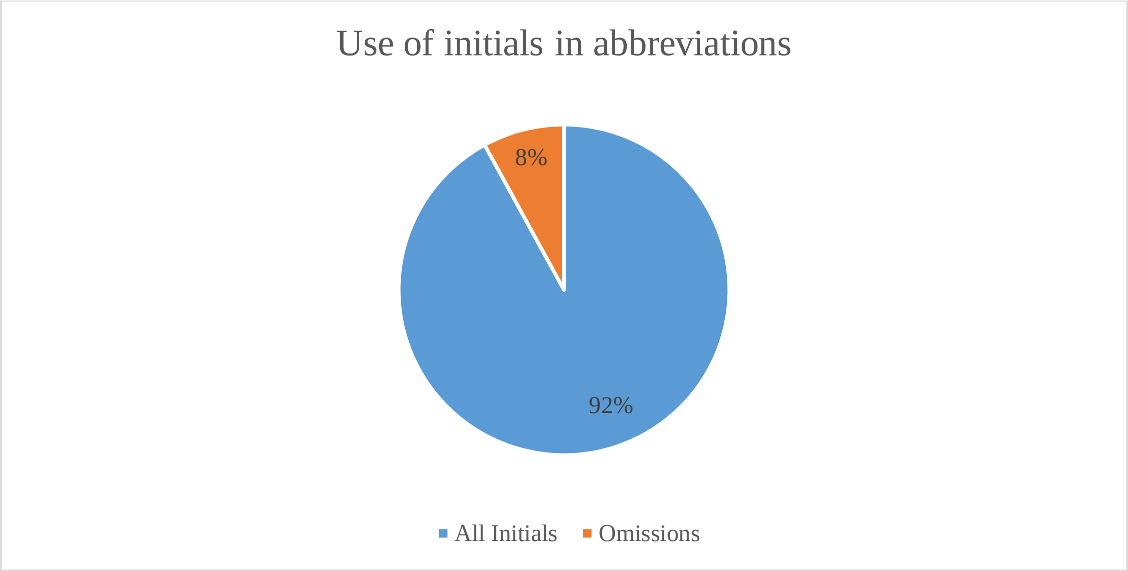


Figure 15. Use of initials in abbreviations.

(Example 7). FRAP = **F**KBP - **R**apamycin - **A**ssociated **P**rotein (FK506 Binding Protein - Rapamycin - associated protein).

(Example 8). NES = **N**ATO **E**lectronic Warfare Support Measures **S**ystem (North Atlantic Treaty Organization Electronic Warfare Support Measures System) [underlined letters were omitted from Abbreviation].

Type.

As mentioned before, an acronym might be distinguished from an initialism because of the pronunciation of all its constituting letters as a word and not as a sequence of letters. Nested abbreviations used in our research are distributed as follows: 262 cases (60.5%) corresponded to initialisms, and 171 cases to acronyms (39.5%). This information is shown in Figure 16. A similar tendency is presented in the study made by Cannon (1989:108), where the number of initialisms exceeded the number of acronyms (501 vs 130).

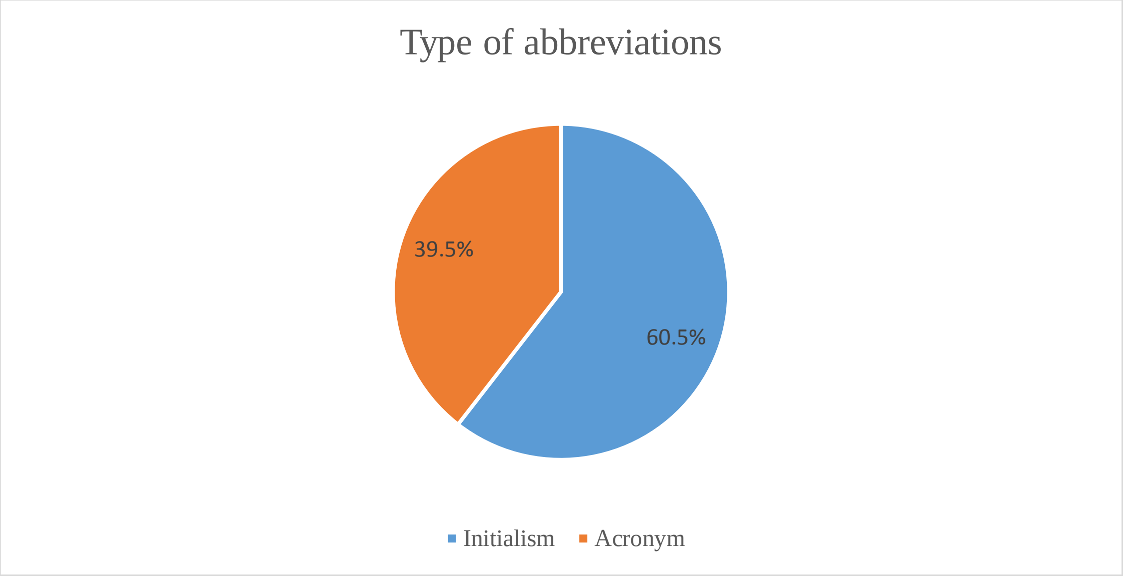


Figure 16. Type of abbreviations.

Length.

In relation to the number of initials of abbreviations from our corpus, it is relevant to highlight that it ranges from 2 to 15 letters, which corresponds to statements about regular abbreviations made by Grange & Bloom (2000:4); Alcaraz (2002:42) and Giraldo (2006:7). The higher frequencies are located between 3 and 5 initials with an accumulated frequency of 376, which represents 86.83% of the total of abbreviations. Conversely, abbreviations formed with 12 and 15 initials are the least frequent (2 occurrences accounting for 0.46%) as presented in Figure 17:

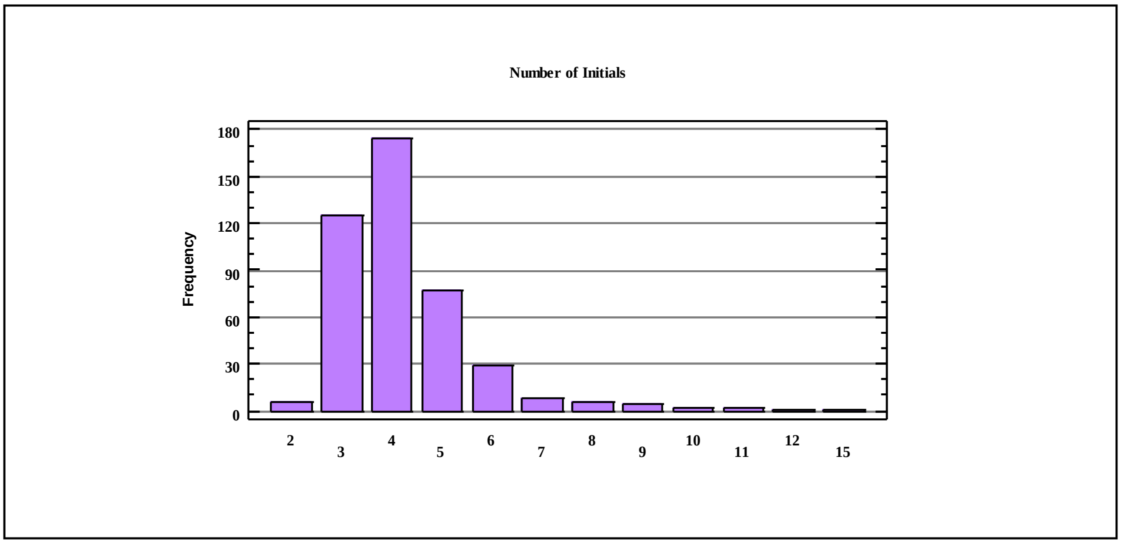


Figure 17. Number of initials.

The results are related to the number of tokens of each NP, considering that over 90% of the abbreviations used all in the initials in the formation process. This aspect can be compared with the results obtained by Quiroz (2006:378) who observed the length of 1724 NPs extracted from a corpus of 66534 words from 21 English texts.

In his study, the author shows that NPs with lengths that range from 3 to 5 words represented 96.6%, which is similar to the results of our study. This observation may be considered relevant taking into account that “this fact could lead to the stabilization and possible lexicalization of an NP, confirming the idea that a direct relationship exists between length, degree of specialization, and syntactic stabilization” (Quiroz, 2006:377). From these findings it could be interpreted that over 85% of terms extracted from our corpus are syntactically stable and, therefore lexicalized, that as explained by Cardero (2003) is “la cohesión de las partes del sintagma que hace que las partes que los constituyen sean indisociables, es decir, que el determinante y determinado no presenten variaciones” (Cardero, 2003:87).

Spelling.

Analyzing spelling patterns in abbreviations from our corpus, we observe that 406 cases, 93.7% of abbreviations, were written in uppercases only (see Example 9) and 27 abbreviated forms, 6.3% of 433 nested abbreviations, were written with upper and lowercases (see Example 10). Cannon (1989:110) exhibits a similar pattern when it comes to the spelling of the abbreviations of his corpus, 90% of abbreviations were spelled in uppercases, 6% in lowercases and 4% of the cases were written in both, upper and lowercases. Spelling information from our study is shown in Figure 18:

(Example 9). VDA = **V**irtual **D**NA **A**nalysis (Virtual Deoxyribonucleic Acid Analysis).

(Example 10). FLIP**s** = **F**ADD - **L**ike **I**CE Inhibitory **P**rotein**s** (Fas - Associated Protein via Death Domain - Like Interleukin-1β - Converting Enzyme Inhibitory Protein**s).**

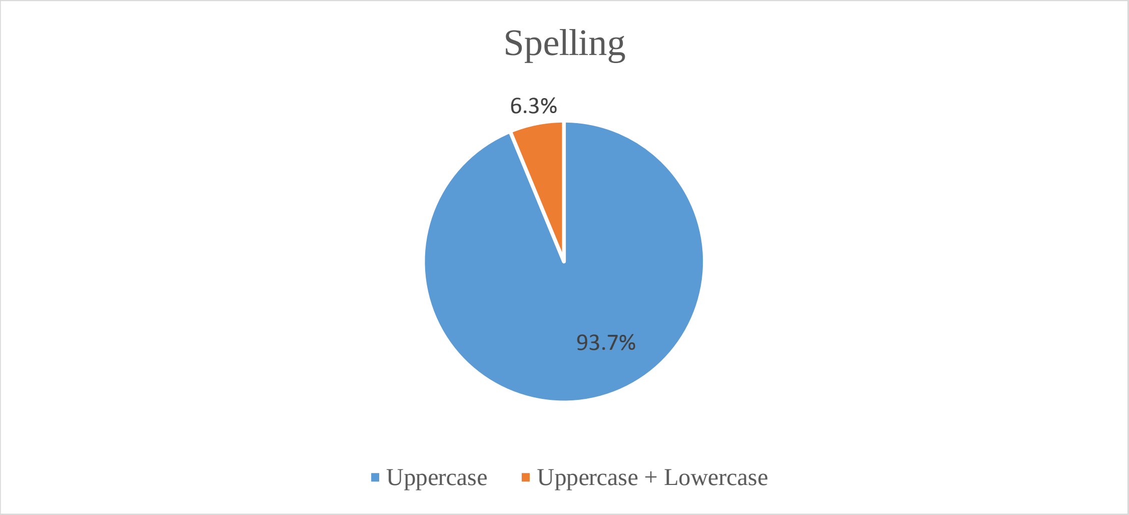


Figure 18. Spelling of abbreviations.

In this last group, where upper and lowercases were used, it is important to highlight that in 10 cases (33.3%) the lowercase was a “s” to show plural forms. In 7 cases (25.9%) the lowercase letter represented a determiner of the abbreviation as it is shown in Example 11:

(Example 11). CARE Act = **C**omprehensive **A**IDS **R**esources **E**mergency act (Comprehensive Acquired Immunodeficiency Syndrome Resources Emergency act).

We also consider relevant to point out the use of other typographic resources such as hyphens and parentheses in 4 cases each, 1 case had possessive (‘) and 1 case had the Greek character upsilon (Ɣ) (see Example 12). Neither abbreviation was written in lowercases only, nor were periods used in the formation of the abbreviations of our corpus.

(Example 12). **Ɣ**RRM = **Ɣ** **R**NA **R**ecognition **M**otif (Ɣ Ribonucleic Acid Recognition Motif).

The study of patterns of formation of nested abbreviations is important for several reasons, among others the application of these patterns by abbreviations recognition and extraction systems in order to improve their performance. Furthermore, when it comes to translation of nested abbreviations, knowledge of these patterns might help translators to apply them in the second language and to fulfill their task more accurately.

Number.

The abbreviations that constituted our corpus are distributed according to the number as shown in Figure 19. The first group is represented by 422 cases, 97.45% of total nested abbreviations, which were written in singular form, and a second group that contains 11 cases, accounting for 2.55%, which were written in plural forms. All plural forms were written using the letter "s", both in upper or lowercases as presented in subsection 5.1.1.4 of this chapter. Replication of initials was not used to show plural forms.

Nested abbreviations are present in singular form because they have a premodification function, and premodifications do not accept plural forms as stated by Sanz (2011): "El sustantivo premodificador, al actuar como un adjetivo en función atributiva adopta las características morfológicas del adjetivo, prescindiendo, por lo general, del morfema propio de plural" (Sanz, 2011:501).

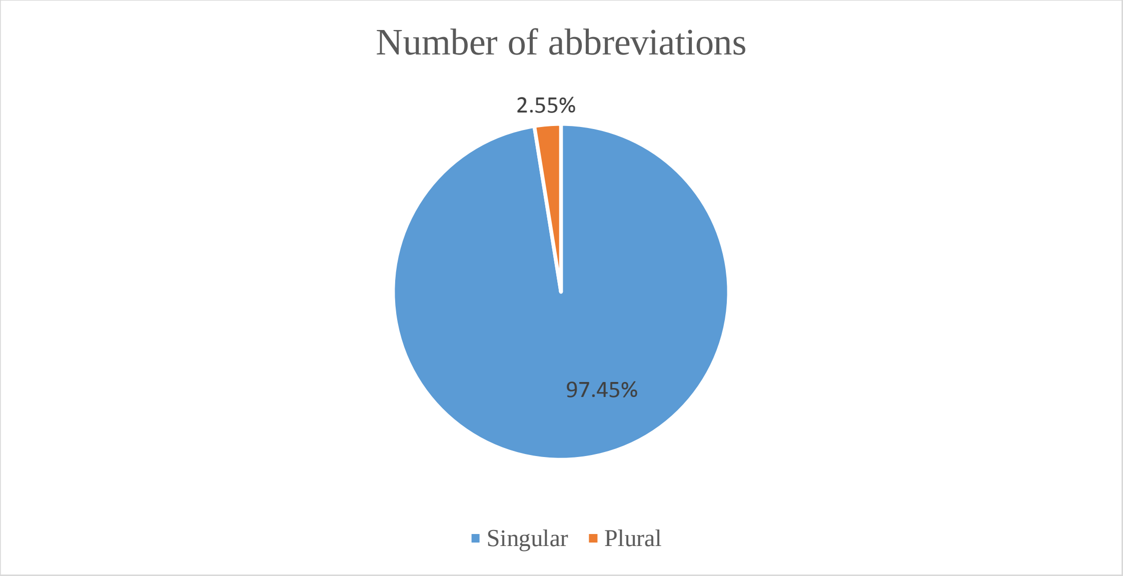


Figure 19. Number of abbreviations.

Use of articles.

Considering that the abbreviations from our corpus were extracted from specialized dictionaries, no contexts were provided to show the use of articles that precede our abbreviations. However, we searched online for some contexts that show this feature and it is possible to observe that articles can be either present or absent before the nested abbreviation as presented in the following examples (see Examples 13, 14 and 15):

(Example 13.) This exchange requires a molecule known as AAC (**A**DP/**A**TP **C**arrier). AAC is a membrane protein that acts like a revolving door – transporting ADP into mitochondria and ATP out of mitochondria and into the cytoplasm.[15] [No article preceded the abbreviation].

(Example 14.) I was the permanent representative in AAC (**A**CCS **A**dvisory **C**ommittee), and “Backsitter” to the NACM.[16] [No article preceded the abbreviation].

(Example 15.) This consensus sequence sometimes called the ACS (**A**RS **C**onsensus **S**equence) is the only homology between known ARS elements.[17] [Article preceded the abbreviation].

[15] Extracted from: https://books.google.com.co/books?id=dAAmwzbIyWcC&pg=PA148&lpg=PA148&dq=%22AAC+ADP+-+ATP+carrier%22&source=bl&ots=sfhxrCJYSj&sig=NQ3OOJHYT6Op_O7OdAt3mFWKQMs&hl=es&sa=X&ved=0ahUKEwjbyt31xarNAhUJ0h4KHVnTAAQ4ChDoAQhQMAk#v=onepage&q=%22AAC%20ADP%20-%20ATP%20carrier%22&f=false.

[16] Extracted from http://people.bayt.com/ioannis-giannopoulos/.

[17] Extracted from https://books.google.com.co/books?id=FzBs_QgihRIC&pg=PA301&lpg=PA301&dq=%22ACS+ARS+consensus+sequence%22&source=bl&ots=be56vNotpO&sig=bimf02KIi_FGXpLPSqeRpphbik4&hl=es&sa=X&ved=0ahUKEwjln5r4yqrNAhVGpR4KHROGAncQ6AEITjAI#v=onepage&q=%22ACS%20ARS%20consensus%20sequence%22&f=false.

Considering that the contexts presented above were chosen randomly just to provide some examples of this characteristic, it is important to remark that no further generalizations can be made and it would be convenient to study this feature in future researches involving different types of corpora.

Syntactic Aspects of Nested Abbreviations.

Lexical category.

Analyzing the meaning of nested abbreviations from our corpus, it is possible to determine that they all correspond to NPs. The number of tokens of each unit ranges from 2 to 24 and, as shown in Figure 20. The higher frequencies are located between 3 and 5 tokens with an accumulated frequency of 277, which represents 63.97% of the total abbreviations. On the contrary, 10 -or- more words NPs are the least frequent with 4 occurrences accounting for 0.92%.

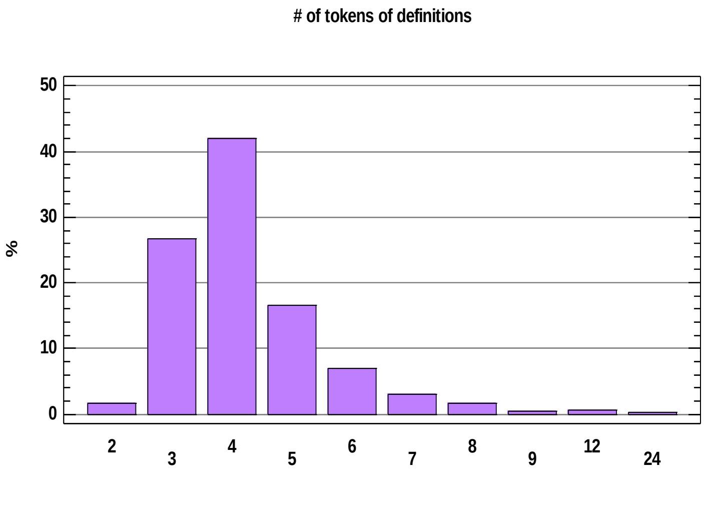


Figure 20. Number of tokens of definition.

Another aspect that we consider important to analyze refers to the most common lexical categories in the NPs from the corpus. In Table 6, we present the different lexical categories, the number of occurrences and the percentage of each category. Although, premodification is not the main topic of study; in our research we noted that 86.37% of our abbreviations have premodification (374 cases out of 433). These patterns can be compared to the ones presented in Quiroz (2006:379) who analyzed premodification in English, and explained how nouns are more frequent than adjectives in premodification because “specialized discourse uses nominalization as a discursive strategy to express impersonalization and objectivity” (Quiroz, 2006:379).

Table 6. Occurrences and percentages of POS.

POS	**Frequency**	**Percentages**
N (plus heads)	1473	76.72
Adj	212	11.04
PP	72	3.75
Adv	69	3.59
PPi	55	2.86
Prep	28	1.46
Conj	8	0.42
V	3	0.6
Total	1920	100

According to the data exposed in Table 6, nouns (N) are the most common lexical category in NPs extracted from our corpus with 1473 occurrences, accounting for 76.72% of all lexical forms. This finding is consistent with the statement presented above, considering that our data were extracted from specialized dictionaries and they represent the codification of a specialized terminology which is used in specialized discourse (Cabré, 1999:167). In addition, it is expected that nouns are prevalent in our corpus since abbreviations behave grammatically as nouns and each NP contains at least one abbreviation, as shown in Example 16 (Note abbreviations in bold in Examples 16 to 23):

(Example 16.) Noun: NIB = **NATO** **I**ntelligence **B**oard (**N** N N).

In second place we find adjectives (Adj), including their past participle (PP) and present participle (PPi) forms. Together, they account for 17.65% with 339 occurrences. In order to illustrate some adjectives, we present Examples 17, 18 and 19 along with the syntactic patterns provided by TreeTagger:

(Example 17.) Adjective: ACTG = **AIDS** **C**linical **T**rials **G**roup (**N** Adj N N).

(Example 18.) Past participle: AFP = **A**llied **FORACS** **P**ublication (PP **N** N).

(Example 19.) Present participle: DBD = **DNA** **B**inding **D**omain (**N** PPi N).

The remaining 5.63% is represented by other lexical categories such as adverbs (Adv), prepositions (Prep), conjunctions (Conj) and verbs (V). Samples of each category are exhibited in Examples 20, 21, 22 and 23:

(Example 20.) Adverb: RAPD = **R**andomly **A**mplified **P**olymorphic **DNA** (Adv PP Adj **N**).

(Example 21.) Preposition: AmFAR = **Am**erican **F**oundation for **AIDS** **R**esearch (Adj N Prep **N** N).

(Example 22.) Conjunction: HAMB = **HIV** and **AIDS** **M**alignancy **B**ranch (**N** Conj **N** N N).

(Example 23.) Verb: ACT UP = **AIDS C**oalition **t**o <u>Unleash</u> **P**ower (**N** N Prep <u>V</u> N).

Abbreviation position.

In order to identify the position of the abbreviation within the definition of the nested abbreviations, all extended forms with premodification were analyzed. As discussed in 5.1.2.1 they all represented Noun Phrases (NPs) and are constituted by 374 abbreviations, accounting for 86.37% of abbreviations.

If the abbreviation is located in the head noun position of the definition of the nested abbreviation, it is identified with the tag Initial (1) as seen in Example 24. If it is identified as the closest premodifier to the Head Noun (HN) it is tagged Medium (2) as presented in Example 25, and, if the abbreviation is located in a different position was tagged as Final (3) as shown in Example 26. NPs with pre and post modification, which represent 13.63% of total abbreviations, are not analyzed in this work and their study is proposed for further research.

(Example 24.) DEUS = **D**eveloping **Eu**ropean <u>SMEs</u> [Head Noun] → Initial (1).

(Example 25.) EBN = **E**uropean <u>**B**IC</u> [Abbreviation] **N**etwork [Head Noun] → Medium (2).

(Example 26.) BTSC = <u>**B**ICES</u> [Abbreviation] **T**eam **S**teering **C**ommittee [Head Noun] → Final (3).

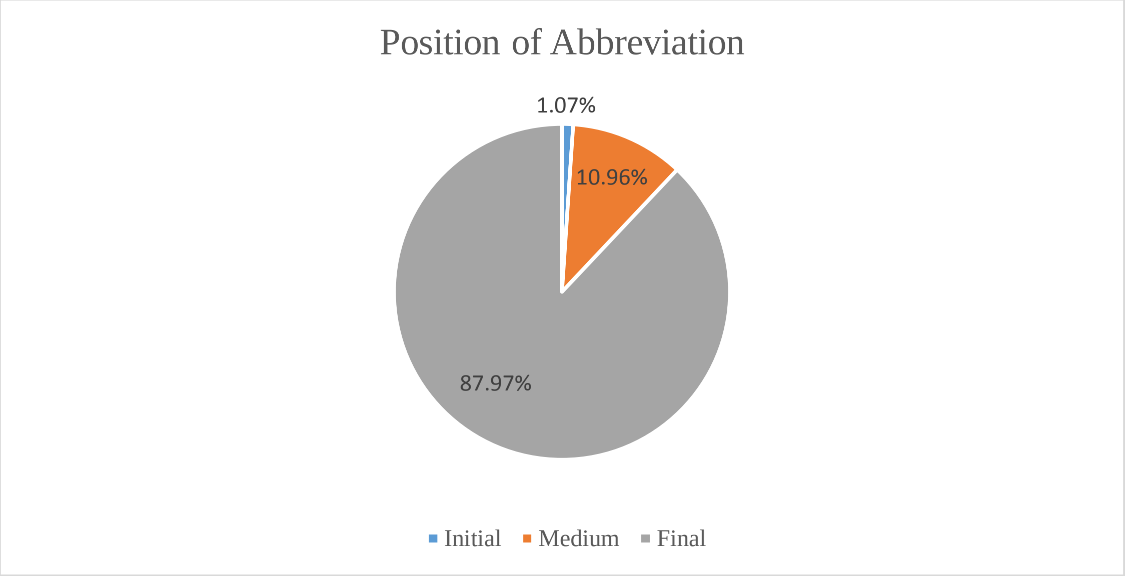


Figure 21. Position of abbreviation.

As shown in Figure 21, the most frequent position of the abbreviation in the definition of nested abbreviations is at the end of the NP, far from the head noun. This fact could be interpreted as that even though abbreviations behave grammatically like ordinary nouns; from a functional point of view, in the meaning of nested abbreviations they behave like adjectives. This phenomenon had been described by authors such as Arntz & Picht, (1995:147) and Rodríguez (1987:143) and denominated as *conversion*.

According to our data, we might say that this process occurred in 87.97% of the cases of nested abbreviations with premodification. Conversely, in only 1.07% of the cases the second abbreviation was located in the head noun position, and behaved functionally as a noun.

The phenomena presented above might be explained because abbreviations inside nested abbreviations behave as attributes of the head noun, specifying or restricting new features. This turns the NP in a more specialized term since it has more premodifiers.

Morphosyntactic patterns.

In order to analyze morphosyntactic patterns present in the NPs extracted from our corpus, data provided 149 surface patterns which represent our 433 NPs, but the first 24 patterns account for an important percentage of the NPs (292 NPs accounting for 67.44%) and are shown in Table 7. This subsample is the one considered in this part of the statistical analysis.

The remaining 125 patterns represent 141 NPs, accounting for 32.56% of total NPs from our corpus. This means that in this last group, the relation pattern: NP is 1.12:1 to be more precise. In other words, there is 1.12 pattern for each NP. We consider this fact a sign of the high syntactic variability and, as these 125 patterns do not allow us to make further generalizations, they will not be considered in this part of the analysis.

The length of abbreviations shown in this sample ranges from two to six words. However, 87.5% of the patterns contained between three and five words, 21 patterns out of 24. There were 8 patterns of three words, 8 patterns of four words and 5 patterns of five words. This information is shown in Tables 8, 9 and 10.

Table 7. The 24-most frequent morphosyntactic patterns of corpus.

Length	**Pattern**	**Example**	**Freq**	**%**
4	N N N N	ATAF airspace coordination centre	68	23.29
3	N N N	ACCS hardware committee	59	20.21
4	N Adj N N	LOCE mobile communications centre	34	11.64
3	N Adj N	ACCS logistic concept	12	4.11
3	N PPi N	CREB binding protein	11	3.77
4	Adj N N N	Interim JTIDS message standard	10	3.42
3	N - N N	ADP - ATP carrier	10	3.42
5	N Adj N N N	AIDS clinical trials information service	9	3.08
3	Adj N N	Deployable CIS module	9	3.08
5	N N Conj N N	NATO command and control system	8	2.74
5	N N N N N	NATO air force armaments group	8	2.74
3	PP N N	Conserved DNA elements	8	2.74
3	N - PPi N	AIDS - defining illness	6	2.05
4	N Prep N N	Center for EUV Astrophysics	5	1.71
4	Adj N Adj N	First GARP global experiment	4	1.37
5	Adj N N N N	Maritime CIS contingency assets pool	4	1.37
2	N N	NATO secret	4	1.37
4	N N Adj N	DNA damage responsive protein	4	1.37
3	N - PP N	AIDS - related encephalitis	4	1.37
5	Adj N Prep N N	Dynamic algorithm for NMR applications	3	1.03
4	N Adj Adj N	NATO multinational maritime force	3	1.03
6	N Adj N Conj N N	NATO joint communications and electronics committee	3	1.03
6	N N , N Conj N N	NATO consultation, command and control board	3	1.03
4	N PP N N	ACE centralized communications budget	3	1.03

Table 8. The most frequent 3-word patterns.

Length	**Pattern**	**Example**	**Freq**	**%**
3	N N N	ACCS hardware committee	59	20.21
3	N Adj N	ACCS logistic concept	12	4.11
3	N PPi N	CREB binding protein	11	3.77
3	N - N N	ADP - ATP carrier	10	3.42
3	Adj N N	Deployable CIS module	9	3.08
3	PP N N	Conserved DNA elements	8	2.74
3	N - PPi N	AIDS - defining illness	6	2.05
3	N - PP N	AIDS - related encephalitis	4	1.37

Table 9. The most frequent 4-word patterns.

Length	**Pattern**	**Example**	**Freq**	**%**
4	N N N N	ATAF airspace coordination centre	68	23.29
4	N Adj N N	LOCE mobile communications centre	34	11.64
4	Adj N N N	Interim JTIDS message standard	10	3.42
4	N Prep N N	Center for EUV Astrophysics	5	1.71
4	Adj N Adj N	First GARP global experiment	4	1.37
4	N N Adj N	DNA damage responsive protein	4	1.37
4	N Adj Adj N	NATO multinational maritime force	3	1.03
4	N PP N N	ACE centralized communications budget	3	1.03

Table 10. The most frequent 5-word patterns.

Length	**Pattern**	**Example**	**Freq**	**%**
5	N Adj N N N	AIDS clinical trials information service	9	3.08
5	N N Conj N N	NATO command and control system	8	2.74
5	N N N N N	NATO air force armaments group	8	2.74
5	Adj N N N N	Maritime CIS contingency assets pool	4	1.37
5	Adj N Prep N N	Dynamic algorithm for NMR applications	3	1.03

As presented in Table 7, the most common patterns are N N N N, N N N and N Adj N N with 161 occurrences accounting for 52.14%. These three patterns, which represent half of the sample, exhibit a high syntactic stability and higher probabilities of lexicalization (Sanz, 2011:480; Quiroz & Arroyave, 2014:143).

Two out of the three patterns mentioned above are consistent with some of the patterns found in Quiroz & Arroyave (2014:143). These authors aimed to study premodified terms in five specialized dictionaries of different academic fields: Medicine, Clinical Laboratory, Economy, Finances and Statistics. One of the findings of this research indicated that 3 to 5-word patterns were the most frequent and the following patterns: N N N, Adj N N, Adj Adj N, N Adj N and N N N N exhibited more syntactic stability and lexicalization. The patterns N N N and N N N N were mentioned in 1st and 5th place as the most common respectively.

The pattern N Adj N N was not mentioned in the five most frequent patterns by Quiroz & Arroyave (2014). However, it was found in 9th place in the Clinical Laboratory Dictionary IFCC, in 13th place in the Medical Dictionary Mosby and the Economy Dictionary IMF and in 14th place in the ISI Multilingual Glossary, a Statistics Dictionary.

Similar results were found in Quiroz (2008:133) in a study about long specialized NPs in the field of Genomics. The author aims to demonstrate the existence of these entities and to provide recommendations to treat them from a formal and semantic perspective. One of the findings of this research is the presence of patterns N N N and N N N N which were located in 1st and 8th place as the most common patterns of the corpus, and pattern N Adj N N was situated in 14th place.

Likewise, Sanz (2011:483) aims to make a contrastive analysis on the terminology of remote sensing and to study translation of syntagmatic compounds into Spanish. The results of her study showed that 2-word structures, N N and Adj N, were the most productive and relevant, accounting for 67.61% which differs from the findings of our study. Nonetheless, 3-word structures were found in 2nd place and, within this group, pattern N N N was located in 4th place. Pattern N Adj N N was situated in 14th place and, pattern N N N N was not mentioned at all.

Comparing the results from our study with the ones presented above, it is reasonable to find patterns like N N N N and N N N in 1st and 2nd place as the more frequent, since we are working with abbreviations, which are nouns. Abbreviations abound in NPs extracted from our corpus and saturate morphosyntactic patterns with nouns. In addition, it is also important to note that abbreviations inside the definition of nested abbreviations are not located in the head noun position but in a premodifier one generating this type of patterns.

The facts presented above might produce different results in our research from studies like those by Quiroz (2008:133) and Sanz (2011:483), which exhibited a high frequency of patterns containing adjectives such as Adj N N, Adj Adj N and N Adj N as the most frequent.

It is also relevant to highlight other structures in our corpus as patterns by its frequency N PPi N, PP N N, N – PPi N, N - PP N and N PP N N. Participle forms are studied with adjectives because “las formas en -ing y -ed desempeñan una función adjetiva cuando preceden al nombre” (Sanz: 2011:489). Patterns containing PP and PPi forms account for 10.96% with 32 occurrences as shown in Table 11:

Table 11. Patterns containing PP and PPi forms.

Length	Pattern	Example	Freq	%
3	N PPi N	CREB binding protein	11	3.77
3	PP N N	Conserved DNA elements	8	2.74
3	N - PPi N	AIDS - defining illness	6	2.05
3	N - PP N	AIDS - related encephalitis	4	1.37
4	N PP N N	ACE centralized communications budget	3	1.03

Syntactic relations.

As presented in subsection 3.5, a sample of 173 NPs out of 13 most common morphosyntactic patterns of our corpus was manually selected in order to perform an analysis on syntactic relations of morphosyntactic patterns in English. This sample is denominated *analysis corpus* and represents 86.30% of the most common patterns found in corpus.

The size of the sample was determined using a sample size calculator available online[18], setting a confidence level of 98% and a confidence interval of 5%. NPs were proportionally distributed according to the frequency of morphosyntactic pattern. The most prevalent patterns have more NPs assigned, and the least frequent the lowest number of NPs. For example, pattern N N N N is the most frequent and has 47 NPs assigned to analyze and pattern N – PPi N is the least frequent and has 4 NPs assigned.

In Table 12, the frequency of syntactic relations in morphosyntactic patterns is shown. Syntactic relation [C [B A]] is the most frequent with 35.84% of 13 syntactic relations present in the *analysis corpus* and included patterns like N N N, N Adj N, N PPi N, Adj N N and PP N N. It is important to remark this finding because this syntactic relation is the most prevalent in NPs formed with 3 tokens regardless the

[18] Calculator.net, Available at http://www.calculator.net/sample-size-calculator.html.

components of the NP and their location. Similar results were achieved by Quiroz (2008:142), where [C [B A]] accounted for 61.2% in the *analysis corpus*.

Table 12. Frequency of syntactic relations in *analysis corpus*.

Syntactic Relation	**Freq**	**%**	**Example**
[C [B A]]	62	35.84	YMCA cardiac therapy
[[D C] [B A]]	37	21.39	NAEW system improvement plan
[D [[C B] A]]	32	18.50	AIDS health services program
[[C - B] A]	7	4.05	AIDS - dementia complex
[[C B] A]	7	4.05	AWIPS program office
[D [C [B A]]]	7	4.05	NATO common interoperability standards
[[E [[D C]][B A]]]	5	2.89	AIDS clinical trials information service
[D [[C And B] A]]	5	2.89	NASA research and education network
[C - [B A]]	4	2.31	CAMP - binding protein
[[E [D C]] [B A]]	3	1.73	NATO air defense ground environment
[[E D] [C [B A]]]	2	1.16	NATO defense manpower audit authority
[[D C B] A]	1	0.58	MHC class II compartments
[E [D [[C B] A]]]	1	0.58	NATO initial data transfer system

In the syntactic relation [C [B A]], the HN and the first premodifier set an ensemble, which is called by Quiroz (2008:142) a *syntagmatic compound* and, it is modified by the second premodifier. This second modifier is in 82.26% of cases the abbreviation inside the meaning of each nested abbreviation (see Example 27). In 16.13% of cases the first premodifier is represented by an abbreviation (see Example 28) and in one case, 1.61% the abbreviation corresponded to HN.

(Example 27.) ACS = **A**RS **C**onsensus **S**equence → N N N → [C [B A]]

(Example 28.) DCM = **D**eployable **C**IS **M**odule → Adj N N → [C [B A]]

The syntactic relations [[D C] [B A]] and [D [[C B] A]] are located in 2nd and 3rd place with 21.39% and 18.50% respectively. This last fact can be explained because the high frequency of pattern N N N N which provided 47 NP to analyze. However, it is important to highlight that the syntactic relation [D [[C B] A]] not only included pattern N N N N, but also other patterns such as N Adj N N and Adj N N N.

Similar to what we presented above with [C [B A]], in 69 cases abbreviations inside the nested abbreviation were located in D, which means that they represent the third premodifier, this is exhibited in Example 29. There were 6 cases where the abbreviations were located in C, which is the second premodifier, and are related to pattern Adj N N N as shown in Example 30.

(Example 29.) NSIS = **N**ATO **S**ubject **I**ndicator **S**ystem → N Adj N N → [[D C] [B A]]

(Example 30.) MACS = **M**ulticenter **A**IDS **C**ohort **S**tudy → Adj N N N→ [D [[C B] A]]

Of the ten remaining syntactic relations, it is possible to infer that there is a relation 1:1 between the syntactic relation and the morphosyntactic pattern that originated it; this is why no further analysis can be made from them.

Now we present an individual analysis of the three most common morphosyntactic patterns of our corpus: N N N N, N N N and N Adj N N.

Pattern N N N N has three possible syntactic relations which are: [[D C] [B A]], [D [[C B] A]] and [[D C B] A]. Syntactic relation [[D C] [B A]], is the most frequent with 33 occurrences accounting for 70.21%. In second place, it is found [D [[C

B] A]] with 13 occurrences, accounting for 27.66%. Representing 2.13% with 1 occurrence is the relation [[D C B] A]. This information is summarized in Table 13 (note underlined abbreviation in Tables 13, 14 and 15):

Table 13. Syntactic relations of pattern N N N N.

Syntactic Relation	**Freq**	**%**	**Example**
[[D C][B A]]	33	70.21	NATO ammunition supply point
[D [[C B] A]]	13	27.66	TOGA heat exchange program
[[D C B] A]	1	2.13	MHC class II compartments

Pattern N N N has only two possible syntactic relations, which are [C [B A]] and [[C B] A]. The former is the most prevalent with 34 occurrences, accounting for 82.93% and the latter is located in 2nd place with 7 occurrences, representing 17.07% as presented in Table 14:

Table 14. Syntactic relations of pattern N N N.

Syntactic Relation	**Freq**	**%**	**Example**
[C [B A]]	34	82.93	DNA data bank
[[C B] A]	7	17.07	NADGE system stock

Pattern N Adj N N has three possible syntactic relations, which are: [D [[C B] A]], [D [C [B A]]] and [[D C] [B A]] accounting for 60.87%, 30.43% and 8.70% respectively as shown in Table 15:

Table 15. Syntactic relations of pattern N Adj N N.

Syntactic Relation	**Freq**	**%**	**Example**
[D [[C B] A]]	14	60.87	NATO maritime patrol aircraft
[D [C [B A]]]	7	30.43	NATO civil wartime agency
[[D C][B A]]	2	8.70	cAMP dependent protein kinase

Semantic Aspects of Nested Abbreviations.

Semantic type.

Regarding the semantic types of abbreviations, it is relevant to highlight that nominal abbreviations correspond to proper names of associations, political organizations, economic groups and diverse kinds of entities (see Example 31) and, conceptual abbreviations (see Example 32) are represented by abbreviated concepts of different fields (Figueroa & Silva, 2000:461). The expanded forms of abbreviations from our corpus provided the following distribution: 337 cases correspond to nominal abbreviations, this represents 77.83% of total nested abbreviations and 96 correspond to conceptual abbreviations, accounting for 22.17% as presented in Figure 22:

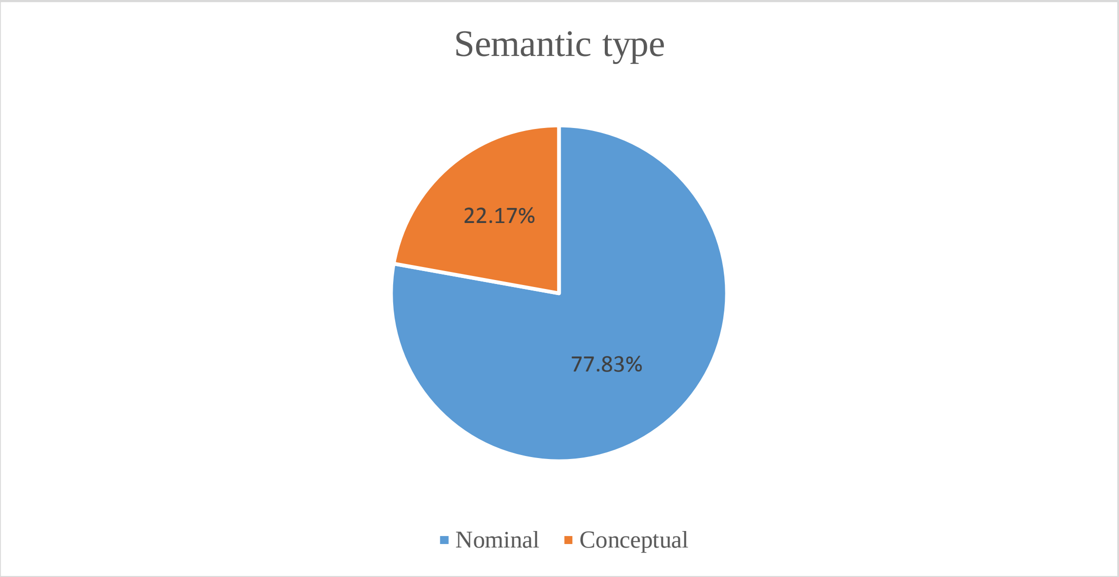


Figure 22. Semantic type of abbreviations.

(Example 31.) CPG = **C**NAD **P**artnership **G**roup [Field: Military Sciences].

(Example 32.) CIPs = **C**LOCK - **I**nteracting **P**roteins [Field: Molecular Biology].

In the first semantic group of abbreviations, which is denominated nominal abbreviations, is it important to remark that Military Sciences are the major contributing field of nested abbreviations in our corpus with 226 occurrences accounting for 67.06%. As shown in Example 33, in this subgroup we classified committees, groups, systems, forces, boards and agencies related to North Atlantic Treaty Organization (NATO). This finding is expected, considering that NATO is a very complex organization constituted by multiple groups essential for its internal operation.

(Example 33.) NADC = **N**ATO **A**ir **D**efense **C**ommittee [Field: Military Sciences].

AIDS, which is a very challenging disease with repercussions in several areas of health sciences, is considered in our study an academic field by itself. It is located in second place in the nominal abbreviations group with 49 occurrences. This means that 14.54% of nominal abbreviations are related to organizations for AIDS patients, studies and programs of the pathology as seen in Example 34:

(Example 34.) TASO = **T**he **A**IDS **S**upport **O**rganization [Field: AIDS].

Other fields that we consider relevant to point out as producers of nested abbreviations are Atmospheric Sciences with 3.86% and Communications accounting for 3.26% of nominal abbreviations. Other areas such as Astronomy, Cardiology, Economics, and Informatics were grouped together and they represent 11.28% in this semantic group (see Example 35).

(Example 35.) NAI = **N**ASA **A**strobiology **I**nstitute [Field: Astronomy].

All information presented above is exhibited in Figure 23:

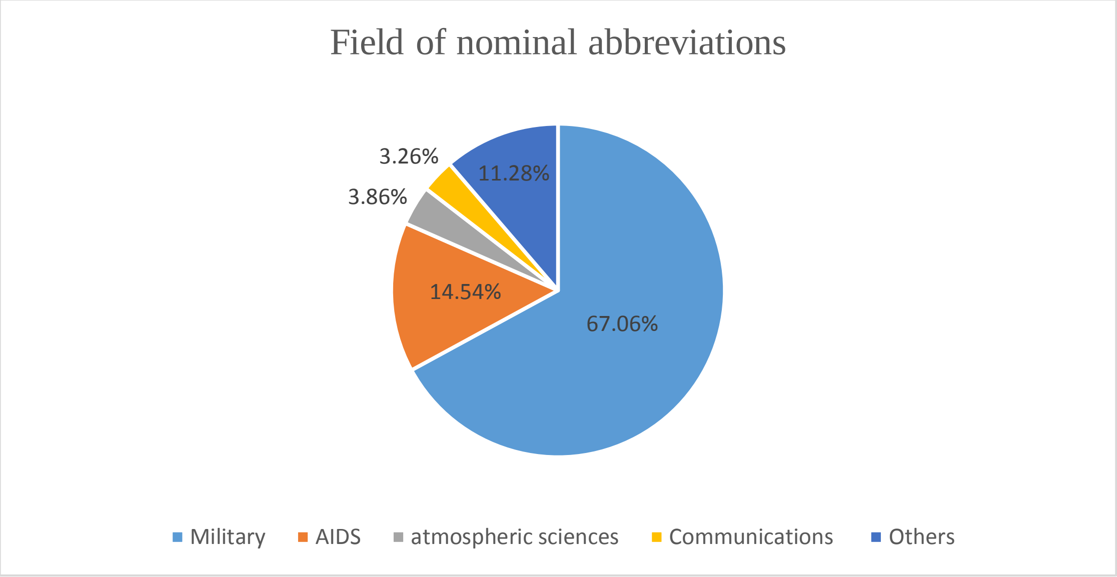


Figure 23. Field of nominal abbreviations.

In Figure 24, we present the group of conceptual abbreviations, which has Molecular Biology as its major contributing field with 80 occurrences out of 96 nested abbreviations classified in this collection, accounting for 83.33% (see Example 36). AIDS with its related pathologies also contributed with 8.33% of nested abbreviations with 8 occurrences (see Example 37). Other fields as Pharmacology, Cardiology, Neurology and Oncology were grouped together and sum up 8.33% to the conceptual abbreviations group (see Example 38).

(Example 36.) DRE = **D**NA **R**esponse **E**lements [Field: Molecular Biology].

(Example 37.) ARE = **A**IDS **- R**elated **E**ncephalitis [Field: AIDS].

(Example 38.) RSBD = **R**EM **S**leep **B**ehavior **D**isorder [Field: Neurology].

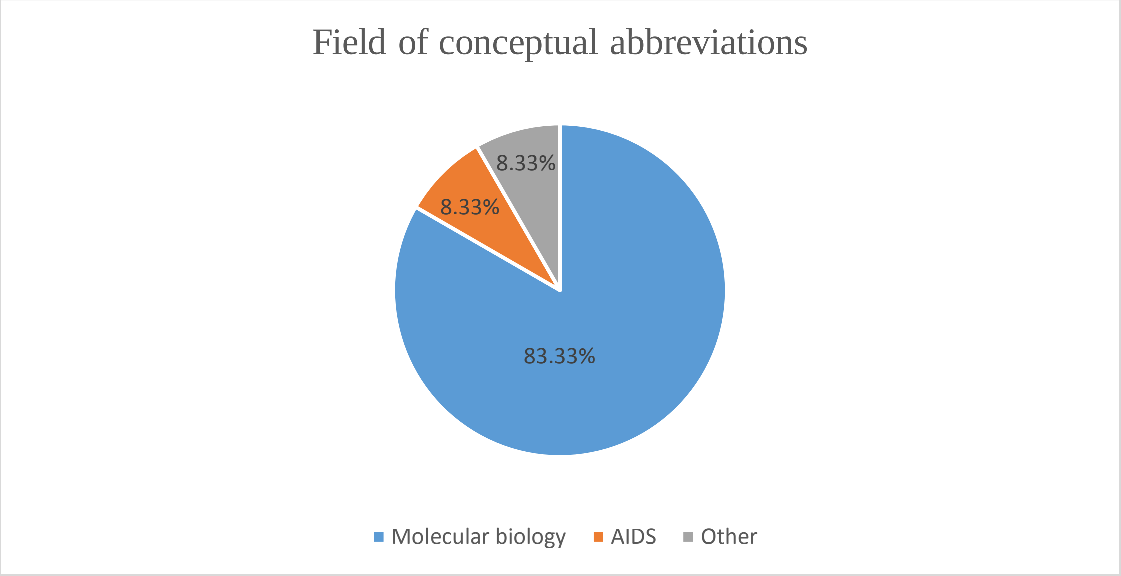


Figure 24. Field of conceptual abbreviations.

Homonymy.

Although, the number of nested abbreviations that constituted our corpus is quite small, there is an evidence of 6 cases of homonymy, accounting for 1.38% of nested abbreviations and they are presented in Table 16. It is important to notice how abbreviations are the same but their meanings are completely different.

Table 16. Cases of homonymy in corpus.

AAC	ADP - ATP carrier
AAC	ACCS advisory committee
ARF	ACE reaction force
ARF	ASEAN regional forum
CAPS	Center for AIDS prevention studies
CAPS	CIAS1 - related autoinflammatory periodic syndromes
CBP	CREB binding protein
CBP	Camp - binding protein
CME	CNS midline element
CME	Combined METOC element
RISC	RNA - induced silencing complex
RISC	RNA interference specificity complex

Synonymy.

In Table 17, we present 2 cases that can be considered as synonyms. It is important to highlight that the Initialism ACTG stands for two different terms and the difference between them remains in the plural form of the word trial(s). A similar situation is exhibited in the second case; although the difference in the meaning of both initialisms is the number in the word fund(s), there is also the presence of a second “F” that stands for “fight” in GFFATM that is not present in GFATM producing two different initialisms.

Table 17. Cases of synonymy in corpus.

ACTG	AIDS Clinical Trials Group
ACTG	AIDS Clinical Trial Group
GFFATM	Global Funds to Fight AIDS, Tuberculosis and Malaria
GFATM	Global Fund to Fight AIDS Tuberculosis and Malaria

Validation of Terms in General and Specialized Corpus.

With the intention to explore the use of terms extracted from our corpus in real texts, we decided to search every intermediate and extended forms online, and verify their frequency in Google® and Ngram Viewer®. As stated before, intermediate forms are provided by dictionaries and extended forms are our creation, after developing the abbreviation within the meaning of nested abbreviations.

As presented in section 3.1, Google® represents the general corpus and Ngram Viewer® represents the specialized corpus. It is important to remark that the latter has a limit of five tokens to analyze terms. Therefore, all terms were verified in the general corpus, but only a percentage of terms, 80.37% of terms belonging to the intermediate forms group, were analyzed in specialized corpus.

General corpus.

In the intermediate forms group, 427 forms out of 433 were found in Google®, accounting for 98.61%. 6 NPs were not found, equivalent to 1.39% of cases. In the extended forms group, 224 forms were found, accounting for 51.73% and 209 NPs, accounting for 48.27% were not found. This information is summarized in Figure 25.

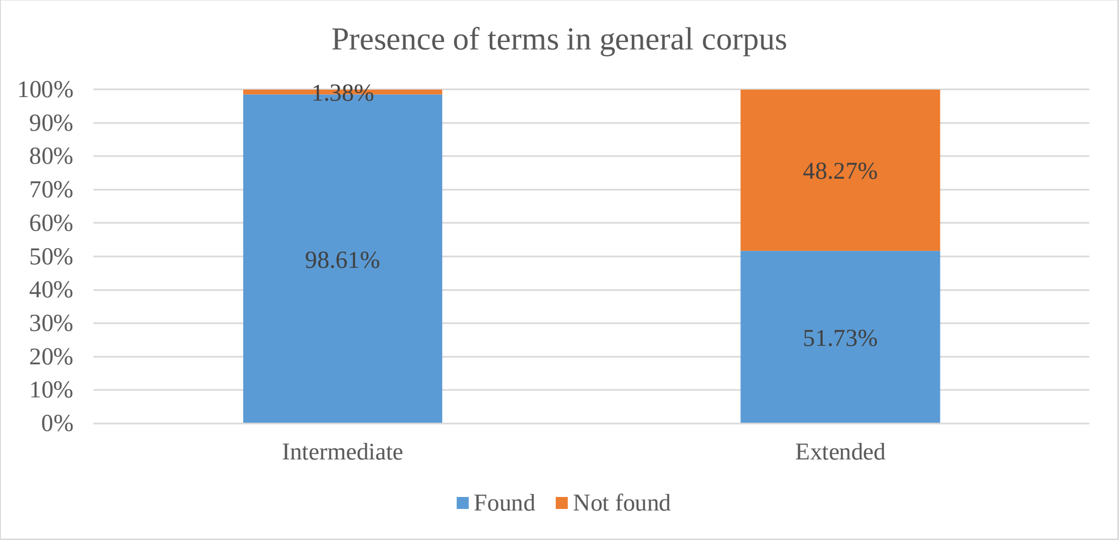


Figure 25. Presence of terms in general corpus.

From information presented in Figure 25, it can be inferred that intermediate forms are prevalent in texts. In almost half of texts, abbreviations inside the definition of nested abbreviations were not developed. Moreover, it can be interpreted from these results that in approximately 50% of cases involving nested abbreviations, one of the rules exposed in writing manuals about the use of abbreviations was not followed, which recommends to explain abbreviations the first time they are introduced in texts to improve readers comprehension (Sabin, 2004:147; Alred, Brusaw, & Oliu, 2006:2; American Medical Association, 2010:1275; Oxford, 2014:2).

Frequencies of first group range from a minimum value of 1 and a maximum of 5120000, with a mean of 72298 and a median of 827. Frequencies of second group exhibit a lower maximum value, when compared to the first group: 199000. In this group, a reduction in mean and median values is also present, which are 3339.9 and 139.5 respectively. In Tables 18 and 19, the first 20 intermediate and extended forms with higher frequencies in general corpus are presented.

Table 18. 20 intermediate forms with higher frequency in general corpus.

Nested	**Intermediate Form**	**Frequency**	**Semantic**	**Field**
ƔRRM	Ɣ RNA recognition motif	5120000	Conceptual	Molecular Biology
PLWHA	People living with HIV / AIDS	3710000	Conceptual	AIDS-related pathology
DBD	DNA binding domain	3180000	Conceptual	Molecular Biology
DDR	DNA damage response	2640000	Conceptual	Molecular Biology
2DDB	DNA data bank	1770000	Conceptual	Molecular Biology
DDBJ	DNA data bank of Japan	1770000	Nominal	Molecular Biology
ANNA	Army, Navy, NASA, Air Force	1320000	Nominal	Military
ABC	ATP - binding cassette	530000	Conceptual	Molecular Biology
IGS	IBM global services	488000	Nominal	Informatics
cAK	cAMP dependent protein kinase	441000	Conceptual	Molecular Biology
NUC	NATO - Ukraine Commission	410000	Nominal	Military
EXODUS	Experiments on the development of UMTS	398000	Nominal	Communications
GFATM	Global fund to fight AIDS Tuberculosis and Malaria	389000	Nominal	AIDS organization
CRE	cAMP response element	364000	Conceptual	Molecular Biology
ARF	ASEAN regional forum	344000	Nominal	Politics
DSB	DNA double - strand break	342000	Conceptual	Molecular Biology
RdRP	RNA - dependent RNA polymerase	333000	Conceptual	Molecular Biology
RBD	RNA binding domain	325000	Conceptual	Molecular Biology
NSN	NATO stock number	321000	Nominal	Military
IAS	International AIDS society	312000	Nominal	AIDS organization

In Table 18, it is important to highlight that 11 abbreviations are conceptual and 10 of them are related to Molecular Biology, while 9 abbreviations are nominal and are related to several fields. Among them it is possible to identify: 3 in Military

Sciences, 2 organizations for patients with AIDS and 1 in Communications, Politics, Molecular Biology and Informatics.

Table 19. 20 extended forms with higher frequency in general corpus.

Nested	Extended Form	Frequency	Semantic	Field
DARPA	Department of defense advanced research projects agency	199000	Nominal	Military
LBP	Lipopolysaccharide - binding protein	84200	Conceptual	Molecular Biology
ISGs	Interferon - stimulated genes	73000	Conceptual	Molecular Biology
RBD	Ribonucleic acid binding domain	48200	Conceptual	Molecular Biology
ABBQ	Acquired immunodeficiency syndrome beliefs and behavior questionnaire	47500	Nominal	AIDS study
ABC	Adenosine Triphosphate - binding cassette	24700	Conceptual	Molecular Biology
ABCA1	Adenosine Triphosphate - binding cassette A1	24300	Conceptual	Molecular Biology
DBD	Deoxyribonucleic acid binding domain	21000	Conceptual	Molecular Biology
TLF	Tata - binding protein - like factor	14400	Conceptual	Molecular Biology
ACTG	Acquired immunodeficiency syndrome clinical trial group	13400	Nominal	AIDS organization
ACTIS	Acquired immunodeficiency syndrome clinical trials information service	11900	Nominal	AIDS organization
IDS	Inhibitor of deoxyribonucleic acid synthesis	11900	Conceptual	Molecular Biology
CAK	Cyclin - dependent kinase activating kinase	10600	Conceptual	Molecular Biology
ADC	Acquired immunodeficiency syndrome - dementia complex	10600	Conceptual	AIDS-related pathology
RBM	Ribonucleic acid binding motif	8960	Conceptual	Molecular Biology
ADE	Advanced large - scale integrated computational environment differencing engine	8550	Nominal	Informatics
ADI	Acquired immunodeficiency syndrome - defining illness	6970	Conceptual	AIDS-related pathology
AMF	Allied command Europe mobile force	6240	Nominal	Military
GASP	G-Protein - coupled receptor - associated sorting protein	5920	Conceptual	Molecular Biology
ARRC	Allied command Europe rapid reaction corps	5600	Nominal	Military

Similar to what is presented in Table 18, Table 19 shows that 13 abbreviations are conceptual and eleven of them are related to Molecular Biology and 2 are AIDS-related pathologies, while seven abbreviations are nominal and are related to three fields in particular: 3 abbreviations are related to Military Sciences, 2 organizations for patients with AIDS and 1 related to an AIDS study. Finally, 1 term is related to Informatics.

These results are consisting with data presented in 5.1.3.1, where it is shown that Molecular Biology is the most productive field of conceptual nested abbreviations. It is possible to infer that terms from this field are the most prevalent in the texts of the general corpus.

Specialized corpus.

Considering limitations in specialized corpus, 80.37% of terms belonging to the intermediate forms group were validated in Ngram Viewer®, since 85 NPs have more than 5 tokens. Of the remaining 348 terms that were searched in specialized corpus, 269 cases, accounting for 72.30%, were not found. 68 NPs were found (19.54% of cases), and 11 terms, 8.16% of cases, required some kind of modification in order to be found. This modification consisted on addition of spaces between hyphens, backslashes and words.

In the group of extended forms, 404 cases, accounting for 93.30%, had more than 5 tokens and were not searched in Ngram viewer®. Only 29 extended forms qualified to verification in specialized corpus. However, none of them were found. This information is presented in Figure 26.

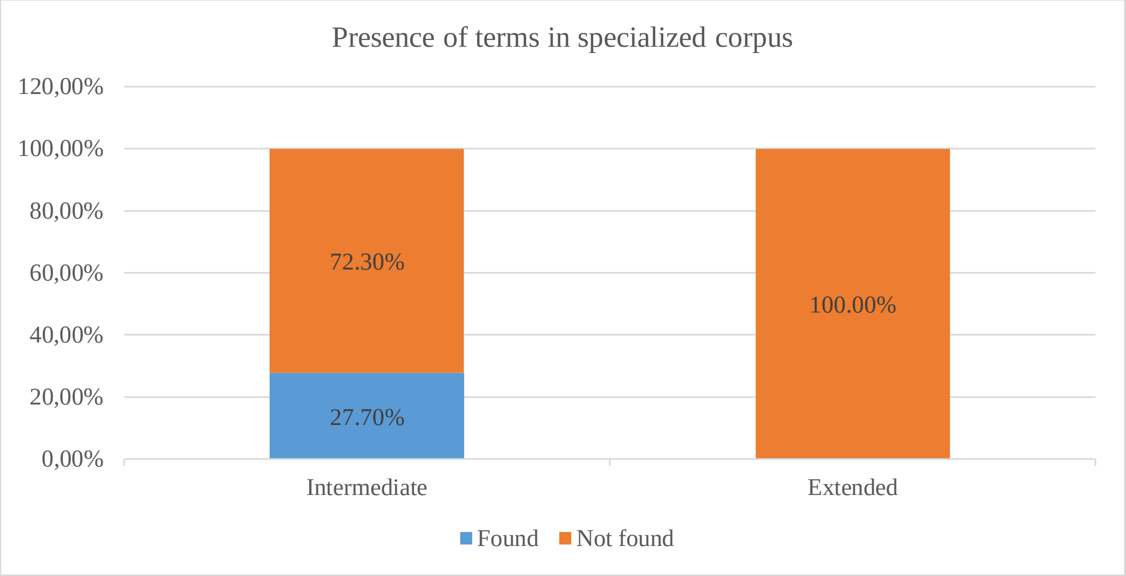


Figure 26. Presence of terms in specialized corpus.

As shown in Figure 26, there is a high percentage of terms that were not found in specialized corpus, this finding is quite surprising since nested abbreviations were extracted from specialized dictionaries. One might expect higher frequencies of these terms in specialized corpora.

From this information, it can be inferred that the use of nested abbreviations and the terms they stand for is merely beginning to be noticed in academic books. Therefore, it is important to expand this type of analysis in further research with different kinds of corpus.

Frequencies of intermediate forms in specialized corpus have a minimum value of 879 and a maximum value of 3180000, a mean of 203222 and a median of 45100. In Table 20, the first 20 intermediate forms nested with higher frequencies in specialized corpus are presented.

Table 20. 20 intermediate forms with higher frequency in specialized corpus.

Nested	**Intermediate Form**	**Frequency**	**Semantic**	**Field**
DBD	DNA binding domain	3180000	Conceptual	Molecular Biology
DDR	DNA damage response	2640000	Conceptual	Molecular Biology
2DDB	DNA data bank	1770000	Conceptual	Molecular Biology
DDBJ	DNA data bank of Japan	1770000	Nominal	Molecular Biology
ABC	ATP - binding cassette	530000	Conceptual	Molecular Biology
IGS	IBM global services	488000	Nominal	Informatics
cAK	cAMP dependent protein kinase	441000	Conceptual	Molecular Biology
CRE	cAMP response element	364000	Conceptual	Molecular Biology
ARF	ASEAN regional forum	344000	Nominal	Politics
RBD	RNA binding domain	325000	Conceptual	Molecular Biology
IAS	International AIDS society	312000	Nominal	AIDS organization
IDS	Inhibitor of DNA synthesis	286000	Conceptual	Molecular Biology
CBP	CREB binding protein	223000	Conceptual	Molecular Biology
ARC	AIDS - related complex	216000	Conceptual	AIDS-related pathology
PWA	Person with AIDS	180000	Conceptual	AIDS-related pathology
RBM	RNA binding motif	167000	Conceptual	Molecular Biology
cGK	cGMP dependent protein kinase	164000	Conceptual	Molecular Biology
aaRS's	Aminoacyl - tRNA synthetases	154000	Conceptual	Molecular Biology
ACTG	AIDS clinical trials group	132000	Nominal	AIDS organization
NRF	NATO response force	131000	Nominal	Military

As presented in Tables 18 and 19, Table 20 also exhibits a high frequency of conceptual abbreviations, 14 out of 20 abbreviations. From this group, 12 are related to Molecular Biology and 2 are AIDS-related pathologies. The 6 remaining cases of nominal abbreviations are distributed as follows: 2 related to organizations for patients with AIDS and one for each organization related to Military, Molecular Biology, Informatics and Politics. From the information presented above, it is possible to infer that terms related to Molecular Biology are also the most prevalent in texts of specialized corpus.

Nesting.

After having characterized nested abbreviations from a morphological, syntactic and semantic perspective, it is important to define and classify them according to our data. Thus, a nested abbreviation may be defined as *an abbreviated form, either an initialism or acronym, which has within its meaning another abbreviation*. Furthermore, nesting is defined as *the minor-word formation process in which an abbreviated form, either an initialism, acronym or other, is within the meaning of another abbreviation in order to form a new one*.

Nesting is presented in Figure 27 and located into the minor-word formation processes in the complex shortening group. Nested abbreviations are shown in purple because we consider relevant to clarify that nested abbreviations are a different type of abbreviation, which involve more than one abbreviation.

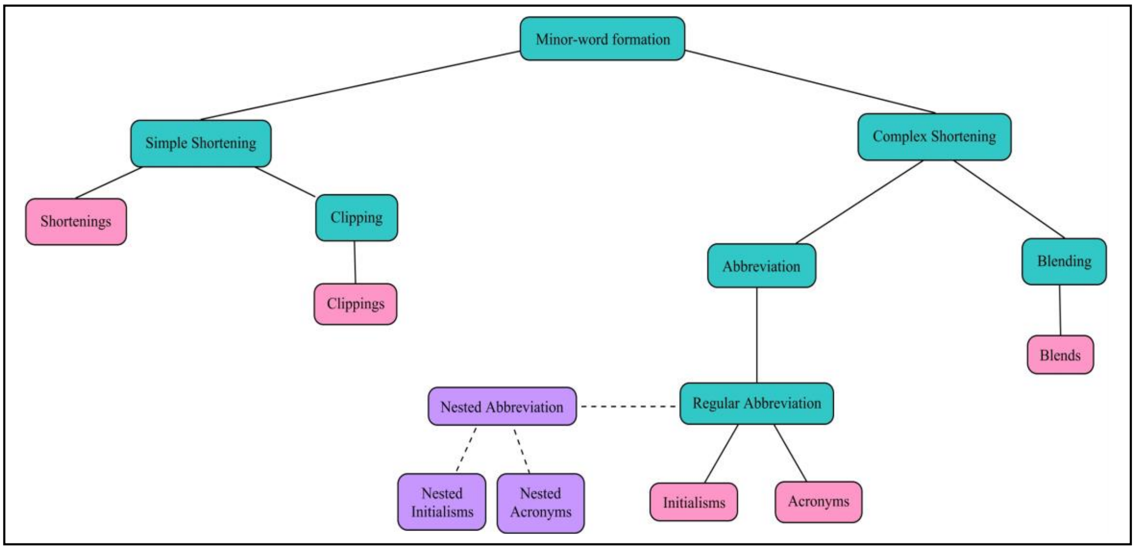


Figure 27. Minor-word Formation with Nested Abbreviation. (Adapted from López, 2004).

From the data we have analyzed, nesting can be divided into five categories according to different patterns on which abbreviations are introduced within the definition of other abbreviations. These categories are: simple, complex, double complex, atypical and hybrid nesting, they are exhibited in Figure 28. Each category is explained ahead:

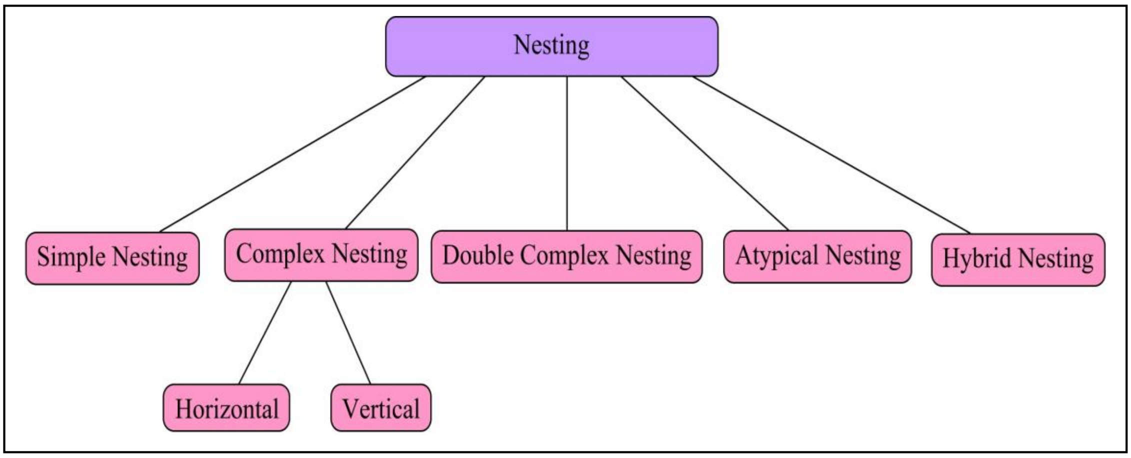


Figure 28. Types of Nesting.

Types of nesting.

Simple nesting. In this type of minor-word formation process only one abbreviation is involved in the developed form. This means that one abbreviation is inside the original abbreviated form.

As it is seen in Example 39, left column contains the nested abbreviation and right column, which is denominated second abbreviation, contains the abbreviation within the meaning of nested abbreviation.

(Example 39.)

Nested	**Second abbreviation**
AAC	ACCS Advisory Committee
AACC	ATAF Airspace Coordination Center
AADGE	ACE Air Defense Ground Environment
ABC	ATP - Binding Cassette

Complex nesting. It is the minor-word formation process that exhibits more than one shortened form inside the nested abbreviation's definition, usually two. They can be introduced at the same level as in the *complex horizontal nesting*, or in a deeper level as in the *complex vertical nesting*.

Complex horizontal nesting. This process shows that the two abbreviations are present in the same level in the meaning of the nested one. Each abbreviation is represented by an initial letter in the original shortened form. Some Examples are shown in (40). It contains the two abbreviations within the meaning of nested abbreviation:

(Example 40.)

Nested	**Second and Third Abbreviations**
A**A**C	ADP – **ATP** Carrier
A**A**IA	ACE **ACCIS** Implementation Architecture
CH**A**MP	Children with HIV and **AIDS** Model Program
PLWH**A**	People Living with HIV/**AIDS**

Complex vertical nesting. In this case, the second and third abbreviations appear in different levels in the meaning, which means that the third abbreviation is inside the second one. Second and third abbreviations are represented by one initial letter in the nested abbreviation. Examples of this type of nesting are presented in (41):

(Example 41.)	
a. Nested	NTS
	↓
Second Abbreviation	**NAMSA** Transportation System
	↓
Third Abbreviation	**NATO Maintenance and Supply Agency** Transportation System

b. Nested	NIP
	↓
Second Abbreviation	**NADGE** Improvement Plan
	↓
Third Abbreviation	**NATO Air Defense Ground Environment** Improvement Plan

Double complex nesting. In this type word formation process, both vertical and horizontal complex nesting are present in the developed form as shown in Example 42:

(Example 42.)			
a. Nested	NNCC		
	↓		(Vertical Nesting)
Second Abbreviation	*NAC*OSA Network Control Center		
	↓	→	(Horizontal Nesting)
Third and Fourth Abbreviations	***NATO* and *CIS* Operating and Support Agency** Network Control Center		

b. Nested	**Second and Third Abbreviations**
N**I**G → Nesting)	*N*CN – **IVSN** Gateway (Horizontal
	↓ (Vertical Nesting)
Fourth Abbreviation	*NATO* Circuit Number – **Initial Voice Switched**

In Example 42, it is important to highlight that a fourth abbreviation is exhibited in the process and it is located at the bottom line of each example. This phenomenon happens because horizontal nesting involves two abbreviations in the definition and vertical nesting involves one more.

Atypical nesting. The defining aspect in this kind of word formation process is the presence of characters that do not exist originally in the meaning of nested abbreviations, such as numbers. They are introduced in nested abbreviations in order to facilitate the pronunciation and the abbreviation itself as presented in Example 43:

(Example 43.)

Nested + Number	**Second Abbreviation**
AD**C2**S	ACE Deployable **C**ommand and **C**ontrol System
N**C3**A	NATO **C**onsultation, **C**ommand and **C**ontrol Agency
N**C3**B	NATO **C**onsultation, **C**ommand and **C**ontrol Board
N**C**3O	NATO **C**onsultation, **C**ommand and **C**ontrol Organization

Hybrid nesting. This process exhibits other forms of abbreviation, such as clipping and blending, which are present in the meaning. Examples of this last type of nesting are shown in (44):

(Example 44.)

Nested	**Second Abbreviation**	
CMFU	Combined METOC Forecast Unit	(Blending)
EMAS	ECO Management and Audit Scheme	(Clipping)
Met**R**S	Methionyl – **tRNA** Synthetase	(Clipping + Initialism)
Glu**R**S	Glutamyl – **tRNA** Synthetase	(Clipping + Initialism)

The distribution of each type of nesting in the abbreviations extracted from our corpus is exhibited in Figure 29.

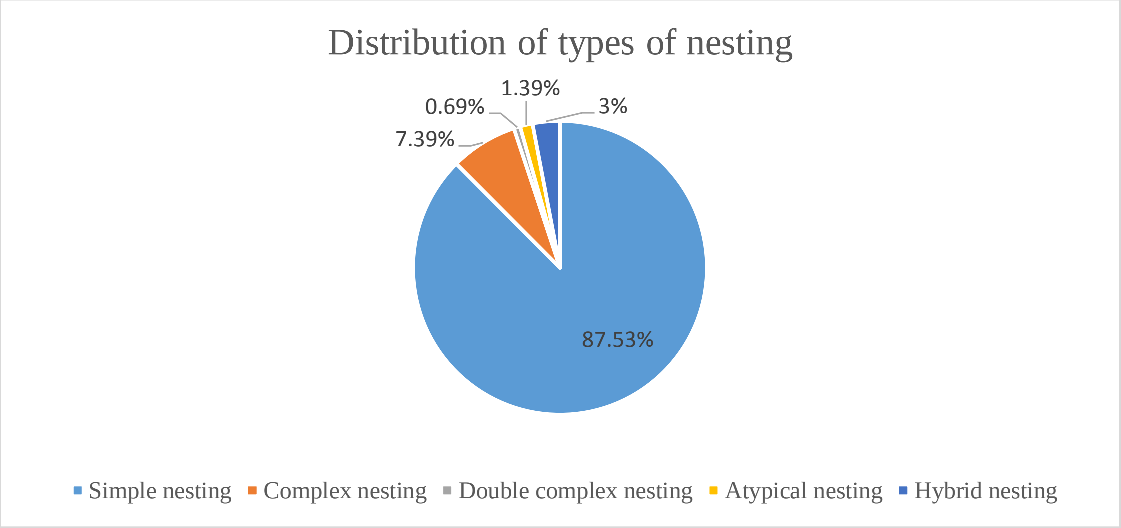


Figure 29. Distribution of types of nesting.

As seen in Figure 29, simple nesting is the most frequent type with 379 occurrences accounting for 87.53% of total nesting processes in abbreviations extracted from our corpus. This means that, in a great majority of cases, only one abbreviation is within the definition of the nested abbreviation. It is also important to remark that in 8.08% of the cases there was more than one abbreviation introduced in the definition as they were complex and double complex nesting processes. Other types of minor-word formation were present in 3% of the cases with 13 occurrences.

Translation Analysis

As mentioned in section 3.1, in order to analyze nested abbreviations extracted from our corpus for translation purposes, 433 English abbreviations were translated using terminology online databases such as IATE, UNTerm, HonSelect, Snomed, Termium Plus and NATO bilingual glossaries (NATO, 2000, 2005, 2010, 2013). Three Colombian professional translators validated Spanish translations of nested abbreviations and their meanings.

According to our data, only one nested abbreviation was translated, accounting for 0.23% of cases. The remaining 432, accounting for 99.77%, rested untranslated. This is consistent with the second solution to translate abbreviations shown in section 3.4, where the definition is translated into a second language and the abbreviation remains in the source language.

In our corpus, nested abbreviation PLWHA (People Living With HIV/AIDS) was translated into PVVS (Personas que Viven con VIH/SIDA) using the first translation solution of these entities stated by Fijo (2003:115), which translates the definition of the abbreviation and a new abbreviation is created based on the initials of the resulting term.

However, it is important to remark that abbreviations inside the definition were translated in a meaningful way. Table 21 shows all English abbreviations from the corpus that were translated into Spanish along with their corresponding definitions and Spanish translations.

Table 21. Translated abbreviations in corpus.

EN	**Definition**	**Freq**	**ES**	**Definición**
NATO	North Atlantic Treaty Organization	155	OTAN	Organización del Tratado del Atlántico Norte
AIDS	Acquired immunodeficiency syndrome	50	SIDA	Síndrome de inmunodeficiencia adquirida
DNA	Deoxyribonucleic acid	23	ADN	Ácido desoxirribonucleico
RNA	Ribonucleic acid	21	ARN	Ácido ribonucleico
HIV	Human immunodeficiency virus	15	VIH	Virus de la inmunodeficiencia humana
cAMP	Cyclic adenosine monophosphate	5	AMPc	Adenosina monofosfato cíclico
ACE	Angiotensin converting enzyme	3	ECA	Enzima convertidora de Angiotensina
ATM	Asynchronous transmission mode	3	MTA	Modo de transferencia asíncrono
UMTS	Universal mobile telecommunications system	2	SUTM	Sistema universal de telefonía móvil
ATAF	Allied tactical air force	1	FATA	Fuerza aérea táctica aliada
ATC	Air traffic control	1	CTA	Control del tráfico aéreo
CNS	Central nervous system	1	SNC	Sistema nervioso central
COPD	Chronic obstructive pulmonary disease	1	EPOC	Enfermedad pulmonar obstructiva crónica
EDP	Electronic data processing	1	PED	Procesamiento electrónico de datos
ESA	European space agency	1	AEE	Agencia espacial Europea
EUV	Extreme Ultraviolet	1	UVE	Ultravioleta Extrema
MHC	Major histocompatibility complex	1	CMH	Complejo mayor de histocompatibilidad
NMR	Nuclear magnetic resonance	1	RNM	Resonancia nuclear magnética
OCD	Obsessive - compulsive disorder	1	TOC	Trastorno obsesivo compulsivo
PSYOS	Psychological operations	1	OPSIS	Operaciones sicológicas
SMEs	Small and medium enterprises	1	PYME	Pequeñas y medianas empresas

As presented in Table 21, abbreviations like NATO, AIDS, DNA, ARN and HIV were translated extensively in our corpus with 264 occurrences. They were translated using the first solution of translation described above. It is possible to infer that these abbreviations, and the terms they stand for, are as important in Spanish as they are in English as stated by Newmark (1988:148).

Moreover, the fact that 11 out of 21 abbreviations in this group are related to Molecular Biology and Medicine is also relevant.

Another aspect that is important to point out is that abbreviations related to NATO, which were within definitions of nested abbreviations such as NACOSA[19], NADGE[20], NAEW[21] and NAMSA[22] remain untranslated as observed in NATO glossaries (NATO, 2000, 2005, 2010, 2013).

This demonstrates that the formation of nesting abbreviations is a very complex process (and not a sign of mental laziness as stated in Fijo (2003:91), and so it is the translation of these entities. Therefore, there are no regularities when it comes to this matter. While some abbreviations are translated others remain in their original language.

As mentioned earlier, only one nested abbreviation was translated into Spanish and translation activities were focused on definitions. Therefore, it is not worth a morphological analysis of translated nested abbreviations, since they are exactly the same as their English equivalents. The same situation takes place with a semantic analysis, considering that no substantial changes were found in semantic aspects analyzed in this study.

This is why our analysis on translations is focused on syntactic aspects such as lexical category, morphosyntactic patterns and syntactic relations of definitions of nested abbreviations, as explained below.

[19] NACOSA: NATO communications and information systems operating and support agency.

[20] NADGE: NATO air defense ground environment.

[21] NAEW: NATO airborne early warning.

[22] NAMSA: NATO maintenance and supply agency.

Syntactic Aspects of Translated Definitions.

Lexical category.

As stated in 5.1.2.1, the meaning of nested abbreviations corresponded to NPs. Thus, all translated definitions are considered NPs as well. The number of tokens of each unit ranges from 2 to 13. As shown in Figure 30, higher frequencies are located between 3 and 5 tokens with an accumulated frequency of 376, which represents 86.83% of total abbreviations.

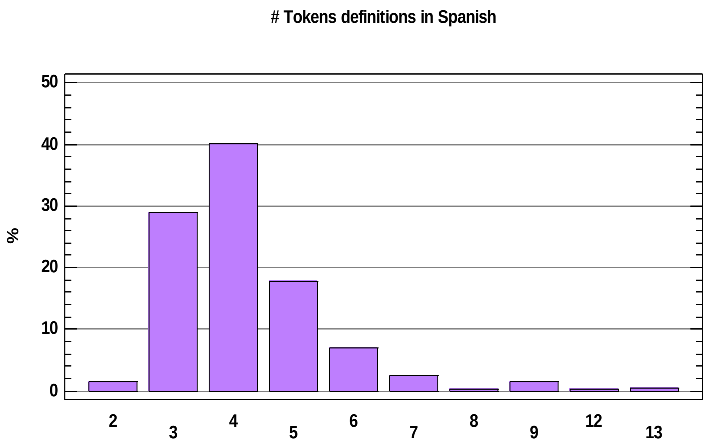


Figure 30. Number of tokens of definition in Spanish.

Another aspect to analyze refers to the most common lexical categories in the NPs in Spanish. In Table 22, we present the different lexical categories, the number of occurrences and the percentage of each category.

Table 22. Frequency and percentages of POS in Spanish.

POS	**Frequency**	**Percentages**
N	1459	55.35
Prep	770	29.21
Adj	291	11.04
Conj	60	2.28
PP	51	1.93
V	3	0.11
Adv	2	0.08
PPi	0	0.00
Total	2636	100

According to data exposed in Table 22, noun (N) is the most common lexical category in translated NPs with 1459 occurrences, accounting for 55.35% of all lexical forms. This finding is consistent with the results of English NPs, since translations were originated from NPs extracted from specialized dictionaries and, as stated by Quiroz (2006:379), nominalization is a strategy used in specialized discourses to express objectivity. Besides, as mentioned before, abbreviations behave grammatically as nouns and each NP contains at least one abbreviation.

Conversely to what happens in English, in 2nd place we find prepositions (Prep) with 770 occurrences, accounting for 29.21%. Leaving in 3rd place adjectives (Adj) with 291 occurrences, accounting for 11.04%. Similar results were obtained in Quiroz (2008), which shows how prepositions appeared in second place after nouns in most of the frequent lexical categories, moving adjectives to 3rd place.

Past participle (PP) forms moved from 3rd place in English NPs to 5th place with 1.93% of lexical forms in translations. PPi forms of English disappeared in

Spanish translations. In Table 23, we present a comparison between lexical categories of English NPs extracted from our corpus and their translations into Spanish.

Table 23. Comparison between English and Spanish lexical categories.

	English		**Spanish**	
POS	**Frequency**	**Percentage**	**Frequency**	**Percentage**
N	1473	76.72	1459	55.35
Adj	212	11.04	291	11.04
PP	72	3.75	51	1.93
Adv	69	3.59	2	0.08
PPi	55	2.86	0	0
Prep	28	1.46	770	29.21
Conj	8	0.42	60	2.28
V	3	0.16	3	0.11
Total	1920	100	2636	100

Morphosyntactic patterns.

In order to analyze morphosyntactic patterns of translations into Spanish and compare them with patterns obtained in NPs extracted from our corpus, 168 surface patterns were obtained. This can be interpreted as a higher syntactic variability and less lexicalization of NPs in Spanish when compared with the ones in English.

Conversely to what is observed in NPs in English, the first 24 patterns of translations account for a lower percentage of the NPs (265 NPs accounting for 61.20%) and those are shown in Table 24.

Table 24. The 24-most frequent morphosyntactic patterns of translations into Spanish.

Length	**Pattern**	**Example**	**Freq**	**%**
3	N Prep N Prep N	Proteína de unión a EPO	46	17.36
4	N Prep N Prep N Prep N	Polígono de tiro de misiles de la OTAN	31	11.70
3	N Adj Prep N	Servicios globales de IBM	22	8.30
4	N Prep N Adj Prep N	PCR con molde específico de ARN	17	6.42
4	N Prep N Prep N N	Motivo de conocimiento de ARN Ɣ	16	6.04
4	N Adj Prep N Prep N	Programa TOGA sobre intercambio de calor	14	5.28
3	N Prep N N	Secuencia de consenso ARS	13	4.91
3	N Adj N	Red europea BIC	13	4.91
4	N Prep N Adj N	Grupo de ensayos clínicos del SIDA	11	4.15
5	N Prep N Prep N Adj Prep N	Servicio de información sobre los ensayos clínicos de SIDA	9	3.40
4	N Prep N N N	Plan de mejoramiento del sistema NAEW	7	2.64
3	N PP Prep N	Proteína asociada a SLAM	7	2.64
3	N - N N	Metionil - ARNt sintetasa	6	2.26
4	N Adj Adj Prep N	Proteína cinasa dependiente de AMPc	6	2.26
4	N Adj PP Prep N	Síndromes periódicos asociados a criopirinas	6	2.26
4	N Adj Prep N N	Células repobladoras de ratones SCID	6	2.26
5	N Adj Prep N Prep N Prep N	Ley integral de emergencia de recursos para el SIDA	5	1.89
3	N N Prep N	Virus ARN de la Leishmania	5	1.89
2	N Prep N	Persona con SIDA	5	1.89
4	N Prep N Conj N Prep N	Sistema de mando y control de la OTAN	5	1.89
5	N Prep N Prep N Prep N Prep N	Comisión de evaluación de inversión en seguridad de la OTAN	4	1.51
3	N PP N	Demencia asociada al VIH	4	1.51
3	N N – N	Complejo demencia – SIDA	4	1.51
4	N N Adj Prep N	Satélite reconocimiento Oceánico por ELINT	3	1.13

The length of abbreviations shown in this sample ranges from two to five words. 83.33% of patterns contained three to five words, 20 patterns out of 24. This is similar to results obtained in Quiroz (2008:340). The author states that 3-word patterns are the most frequent in postmodification of long specialized NPs in Spanish in the field of genomics, they represented over 80% of the sample. In our study, 4-word patterns are the most common with 11 occurrences, accounting for 45.83% of the sample, and 3-word patterns represent 37.5%.

Data from our translations show 9 patterns of three words, 11 patterns of four words, 3 patterns of five words. This information is shown in Tables 25, 26 and 27:

Table 25. The most frequent 3-word patterns in Spanish.

Length	**Pattern**	**Example**	**Freq**	**%**
3	N Prep N Prep N	Proteína de unión a EPO	46	17.36
3	N Adj Prep N	Servicios globales de IBM	22	8.3
3	N Prep N N	Secuencia de consenso ARS	13	4.91
3	N Adj N	Red Europea BIC	13	4.91
3	N PP Prep N	Proteína asociada a SLAM	7	2.64
3	N - N N	Metionil - ARNt Sintetasa	6	2.26
3	N N Prep N	Virus ARN de la Leishmania	5	1.89
3	N PP N	Demencia asociada al VIH	4	1.51
3	N N – N	Complejo demencia – SIDA	4	1.51

Table 26. The most frequent 4-word patterns in Spanish.

Length	**Pattern**	**Example**	**Freq**	**%**
4	N Prep N Prep N Prep N	Polígono de tiro de misiles de la OTAN	31	11.7
4	N Prep N Adj Prep N	PCR con molde específico de ARN	17	6.42
4	N Prep N Prep N N	Motivo de conocimiento de ARN Ɣ	16	6.04
4	N Adj Prep N Prep N	Programa TOGA sobre intercambio de calor	14	5.28
4	N Prep N Adj N	Grupo de ensayos clínicos del SIDA	11	4.15
4	N Prep N N N	Plan de mejoramiento del sistema NAEW	7	2.64
4	N Adj Adj Prep N	Proteína cinasa dependiente de AMPc	6	2.26
4	N Adj PP Prep N	Síndromes periódicos asociados a Criopirinas	6	2.26
4	N Adj Prep N N	Células repobladoras de ratones SCID	6	2.26
4	N Prep N Conj N Prep N	Sistema de mando y control de la OTAN	5	1.89
4	N N Adj Prep N	Satélite reconocimiento oceánico por ELINT	3	1.13

Table 27. The most frequent 5-word patterns in Spanish.

Length	**Pattern**	**Example**	**Freq**	**%**
5	N Prep N Prep N Adj Prep N	Servicio de información sobre los ensayos clínicos de SIDA	9	3.4
5	N Adj Prep N Prep N Prep N	Ley integral de emergencia de recursos para el SIDA	5	1.89
5	N Prep N Prep N Prep N Prep N	Comisión de evaluación de inversión en seguridad de la OTAN	4	1.51

As presented in Table 24, the most common patterns are N Prep N Prep N, N Prep N Prep N Prep N and N Adj Prep N with 99 occurrences accounting for 37.36% of the sample. All three patterns were described in Quiroz (2008:189). Pattern N Prep N Prep N was found in 3rd place in genomics corpus in Spanish, pattern Prep N Prep N Prep N was located in 12th place and pattern N Adj Prep N in 2nd place.

These patterns were also identified in Cardero (2003:95) in a study about terminology of satellite control in Mexico with a corpus constituted in Spanish.

However, the author does not show frequencies and percentages of patterns present in corpus, she shows a list of 29 different patterns found in corpus. As a consequence, further comparisons with our data are not possible.

The importance of analyzing morphosyntactic patterns and their relation to translation does not remain in the pattern itself, it is significant to associate these patterns with their syntactic relations. Because they provide information on how each component impacts others within the NP, helping translators to perform a more accurate task, especially during translation of specialized texts.

Syntactic relations.

As presented in subsection 4.5, a sample of 173 NPs out of 13 most common morphosyntactic patterns of our corpus was manually selected in order to perform an analysis on syntactic relations of morphosyntactic patterns in English. This sample is denominated *analysis corpus* and represents 86.30% of the most common patterns found in the corpus. This sample is compared with the morphosyntactic patterns and the syntactic relations of their translation into Spanish, in order to look for regularities in translation solutions.

In Table 28, frequency of syntactic relations of the morphosyntactic patterns group in Spanish is shown:

Table 28. Frequency of syntactic relations in Spanish.

Syntactic Relation	**Freq**	**%**	**Example**
[[A B] C]	62	35.84	Elementos de respuesta a ADN
[[A B][C D]]	43	24.86	Plan de mejoramiento del sistema NAEW
[[A [B C]] D]	30	17.34	Archivo de ataques nucleares del ACO
[[[A B] C] D]	9	5.20	Estudio multicéntrico de cohortes de SIDA
[[A B][C D] E]]	6	3.47	Servicio de información sobre los ensayos clínicos de SIDA
[A [B C]]	5	2.89	Oficina del programa AWIPS
[[A [B C][D E]]	4	2.31	Cuestionario sobre creencias y conductas relacionadas con el SIDA
[A [B - C]]	3	1.73	Comisión OTAN – Ucrania
[[A - B] C]	3	1.73	Aminoacil - ARNt Sintetasa
[[A B][[C [D E] F]	2	1.16	Centro de información sobre las municiones de riesgo atenuado de la OTAN
[[A B][C [D E]]]	2	1,16	Sistema inicial de transferencia de datos de la OTAN
[A [B C D]]	1	0.58	Compartimiento del CMH clase II
[[A B][[C D] E]]	1	0.58	Centro de coordinación del espacio aéreo de la FATA
[[A [B C D] E]]	1	0.58	Grupo sobre el armamento de las fuerzas terrestres de la OTAN
[[[A [B C]] D] E]	1	0.58	Conferencia de los altos responsables de la logística de la OTAN
[A [B / C]]	1	0.58	Transportador ADP/ATP

Syntactic relation [[A B] C] is the most frequent with 35.84% of 16 syntactic relations present in the analysis corpus. It is related to several morphosyntactic patterns such as N Prep N Prep N, N Adj Prep N, N Prep N N, N Adj N, N PP Prep N, N N Adj, N N Prep N and N PP N. Moreover, this relation is a translation solution of NPs with syntactic relations [C [B A]] in 55 cases, accounting for 88.71%, [C - [B A]] in four cases, equivalent to 6.45% and [[C B] A] in 3 cases (4.84%). This information is summarized in Table 29.

Table 29. Syntactic relations in English providing relation [[A B] C] as solution in Spanish.

Syntactic relation EN	Freq	%
[C [B A]]	55	88.71
[C - [B A]]	4	6.45
[[C B] A]	3	4.84

In the syntactic relation [[A B] C], the first premodifier acts on the HN and, this compound is modified by the second premodifier. This second modifier is in 96.77% of cases the abbreviation inside the meaning of nested abbreviations (see Example 45). In 1.61% of cases, the first premodifier is represented by an abbreviation (see Example 46). In one case, 1.61%, the abbreviation corresponded to HN (see Example 47).

(Example 45.) DCM = Módulo Desplegable del CIS→ N Adj N→ [[A B] C].

(Example 46.) NRA = Agencia de la OTAN para los Refugiados → N Prep N Prep N → [[A B] C].

(Example 47.) hTR = ARN Telomerasa Humana → N N Adj → [[A B] C].

Comparing our results with the ones presented in Quiroz (2008:197), the relation [[A B] C] is located in 2nd place in a corpus of genomics with 45.5% of cases. In the study mentioned above, the relation [A [B C]] is the most prevalent in the corpus with 50.5%. This finding differs from our study since this relation is found in 6th position with 5 occurrences, accounting for 2.89%.

Syntactic relations [[A B] [C D]] and [A [[B C] D]] are located in 2nd and 3rd place with 24.86% and 17.34% respectively. The former was found in Quiroz

(2008:197) in 3rd place in corpus with 3% of occurrences. The latter was not mentioned at all.

On the one hand, syntactic relation [[A B] [C D]] is linked to an important amount of 4-word morphosyntactic patterns, which are shown in Table 30, in order to facilitate their identification of pattern and frequencies in analysis corpus. As presented in Table 30, patterns N Prep N Prep N Prep N, N Prep N Prep N N and N Prep N N N account for 73.81% of cases.

Table 30. Morphosyntactic patterns related to syntactic relation [[A B] [C D]] in Spanish.

Patterns	**Freq**	**%**
N Prep N Prep N Prep N	15	35.71
N Prep N Prep N N	10	23.81
N Prep N N N	6	14.29
Adj Prep N N Prep N	1	2.38
N Adj Adj Prep N	1	2.38
N Adj Prep Adj Prep N	1	2.38
N Adj Prep N N	1	2.38
N Adj Prep N Prep N	1	2.38
N N N Adj	1	2.38
N N N N	1	2.38
N N Prep Adj N	1	2.38
N N Prep N Prep N	1	2.38
N Prep N N Prep N	1	2.38
N Prep N Prep N Conj N	1	2.38

In addition, this syntactic relation is the result of translation of NPs associated syntactic relations such as [[D C] [B A]] in 35 cases, accounting for 81.40%, [D [[C B] A]] in 5 cases, equivalent to 11.63%, [C [B A]] in 2 cases, representing 4.65% and [D [C [B A]]] in one occasion, accounting for 2.33% as presented in Table 31.

Table 31. Syntactic relations in English providing relation [[A B] [C D]] as solution in Spanish.

Syntactic relation EN	**Freq**	**%**
[[D C] [B A]]	35	81.40
[D [[C B] A]]	5	11.63
[C [B A]]	2	4.65
[D [C [B A]]]	1	2.33

On the other hand, syntactic relation [A [[B C] D]] is linked to 9 morphosyntactic patterns of three words. They are presented in Table 32.

Table 32. Morphosyntactic patterns related to syntactic relation [A [[B C] D]] in Spanish.

Patterns	**Freq**	**%**
N Prep N Adj Prep N	12	40.00
N Prep N Adj N	7	23.33
N Prep N Prep N Prep N	3	10.00
N Prep N Prep N N	2	6.67
N Prep N Conj N Prep N	2	6.67
N Prep N PP N	1	3.33
N Prep N N Adj	1	3.33
N Prep Adj N N	1	3.33
N N Adj Prep N	1	3.33

Furthermore, this syntactic relation is the result of NPs translated into Spanish associated with syntactic relations [D [[C B] A]] in 23 cases, accounting for 76.67%, [C [B A]] in 4 cases, equivalent to 13.33%, [[C B] A] in 2 cases, accounting for 6.67% and [D [C [B A]]] in one occasion, representing 3.33%. This information is shown in Table 33.

Table 33. Syntactic relations in English providing relation [A [[B C] D]] as solution in Spanish.

Syntactic relation EN	Freq	%
[D [[C B] A]]	23	76.67
[C [B A]]	4	13.33
[[C B] A]	2	6.67
[D [C [B A]]]	1	3.33

In syntactic relation [A [[B C] D]], 96.66% of times the abbreviation inside the meaning of nested abbreviations is located in D, which means the third modifier (see Example 48). In one case, accounting for 3.34%, abbreviation was located as a second premodifier (see Example 49). In syntactic relation [[A B] [C D]], 88.37% of cases the abbreviation inside the meaning is located in D (see Example 50). In 6.98% of cases, abbreviation is located as the second premodifier (see Example 51) and in 4.65% of cases is the first premodifier (see Example 52).

(Example 48.) CME = Elemento de la Línea Media del SNC → N Prep N Adj N → [[A [B C]] D].

(Example 49.) IJMS = Patrón de Mensaje JTIDS Provisional → N Prep N N Adj → [[A [B C]] D].

(Example 50.) NSIP = Plan de Mejoramiento del Sistema NAEW → N Prep N N N → [[A B] [C D]].

(Example 51.) Ɣ RRM = Motivo de Conocimiento de ARN Ɣ → N Prep N Prep N N → [[A B] [C D]].

(Example 52.) SEMM = Memoria MOS de Único Electrón → N N Prep Adj N → [[A B] [C D]].

Now we present an individual analysis of three most common morphosyntactic patterns of NPs translations into Spanish extracted from our corpus: N Prep N Prep N, N Prep N Prep N Prep N and N Adj Prep N.

Pattern N Prep N Prep N, has two possible syntactic relations which are: [A [B C]] and [[A B] C]. Syntactic relation [[A B] C], is the most frequent with 28 occurrences accounting for 93.33%. In second place, it is found [A [B C]] with 2 occurrences, accounting for 6.67%. This information is summarized in Table 34 (note underlined abbreviation in Tables 34, 35 and 36):

Table 34. Syntactic relations of pattern N Prep N Prep N.

Syntactic relation	**Freq**	**%**	**Example**
[[A B] C]	28	93.33	Comisión de recursos de la OTAN
[[A [B C]	2	6.67	Helicóptero para la fragata de la OTAN

This pattern in Spanish is the result of translation of NPs in English containing morphosyntactic patterns such as N N N, N N N N, N PPi N and N – PPi N which are associated with syntactic relations [C [B A]], [[C B] A], [[D C] [B A]], and [C - [B A]].

Pattern N Prep N Prep N Prep N has three possible syntactic relations, which are [[A B] [C D]], [[A [B C]] D] and [[[A B] C] D]. The former is the most prevalent with 15 occurrences, accounting for 78.95%. The second relation has 3 occurrences, accounting for 15.79%. The latter, with one occurrence, accounts for 5.26% as it is presented in Table 35:

Table 35. Syntactic relations of pattern N Prep N Prep N Prep N.

Syntactic relation	**Freq**	**%**	**Example**
[[A B][C D]]	15	78.95	Unidad de evaluación de la vacuna contra el SIDA
[[A [B C]] D]	3	15.79	Seguridad de los campos de tiro de la OTAN
[[[A B] C] D]	1	5.26	Polígono de tiro de misiles de la OTAN

This pattern in Spanish is the result of translation of NPs in English containing morphosyntactic patterns such as N N N N, N N N, N Adj N N and N Adj N which are associated with syntactic relations [[D C] [B A]], [D [[C B] A]], [C [B A]] and [[C B] A].

Pattern N Adj Prep N has only one syntactic relation, which is: [[A B] C] and is the result of translation of NPs containing morphosyntactic patterns such as N N N, N Adj N, N PPi N, Adj N N and PP N N which are related with pattern [C [B A]]. Information is shown in Table 36:

Table 36. Syntactic relations of pattern N Adj Prep N.

Syntactic relation	**Freq**	**%**	**Example**
[[A B] C]	14	100	Necesidad operacional de la OTAN

From the information presented above, it is possible to infer that when it comes to translation of meaning of nested abbreviations there are no regular solutions, since the amount of morphosyntactic patterns and syntactic relations is as wide as the number of terms existing in specialized languages.

However, some tendencies in translation were found and described. For instance, the fact that NPs related to several morphosyntactic patterns in English were translated into Spanish and provided only one morphosyntactic pattern in this language, and that there is an inverse relation between syntactic relations in English and their translations into Spanish. Therefore, this information might be useful for machine translation software, terminology extraction systems and specialized translators as well.

This is why it is important to continue the study of translation of these entities from different perspectives, exploring other fields and languages, and using different types of corpora such as parallel corpus.

Conclusions

The purpose of this research was to describe nested abbreviations from a linguistic perspective, with the intention of identifying linguistic aspects to be analyzed during translation of these entities. In order to fulfill our goal, a literature review on specialized languages, minor-word formation processes and translation of abbreviations was performed. As a result, it was found that nested abbreviation is a phenomenon that has not been described in English grammars, terminology and translation manuals. However, nested abbreviations are used by scholars in several fields to condense information and save space in academic texts.

Moreover, our literature review also showed that there are no regularities about translation of abbreviations, neither regular nor nested. Translating solutions are left to specialists or translators' criteria, since there are no solid recommendations on how to proceed in these cases. Therefore, all kind of solutions are found in translated texts, creating real difficulties to identify or disambiguate abbreviations in systems working in languages different from English.

From a methodological perspective, two abbreviation dictionaries were selected to identify nested abbreviations. Once they were identified, extraction, tokenization and syntactic tagging processes were involved. Translation of abbreviations was performed and validated by three professional translators.

In order to answer the first research question: Which are the linguistic features of abbreviations when nesting phenomena are involved? A linguistic description of the phenomenon was conducted and the most relevant conclusions of this first analysis are:

1. Although, the percentage of nested abbreviations obtained from dictionaries is quite low, less than 1% of total abbreviations, it is highly relevant to study this

phenomenon. Since this low percentage could mean that nested abbreviation is a developing process which can continue to grow, as the amount of abbreviations in specialized languages does. This can be inferred from the results mentioned before, which showed that over 85% of terms extracted from our corpus are syntactically stable and, therefore lexicalized.

In addition, the study of linguistic aspects of nested abbreviations is important to help improve not only the performance of abbreviations recognition, extraction and disambiguation systems but also the work of technical translators.

2. Nested abbreviations share with regular abbreviations a considerable percentage of morphological features. For instance, a great majority of nested abbreviations are formed only with the initial letters of a NP. Likewise, length of nested abbreviations is basically 3-5 letters, exhibiting a high level of lexicalization. Besides, they are related to the number of tokens of the NP that gives origin to the abbreviation. Additionally, spelling of nested abbreviations is written mainly in uppercases and singular forms are predominant.

3. The most important difference between regular and nested abbreviations is related to syntactic aspects. The most frequent position of the abbreviation in the definition of nested abbreviations is at the end of the NP, acting as an attribute of the HN. From this fact it could be inferred that abbreviations behave grammatically like ordinary nouns and are tagged as nouns by tools such as TreeTagger. However, from a functional point of view, they behave like adjectives in the meaning of nested abbreviations.

4. Furthermore, the most common morphosyntactic patterns in English were: N N N N, N N N, N Adj N N, N Adj N and N PPi N and the most prevalent lexical

category is noun, indicating that this category is highly important to communicate specialized knowledge and nominalization is used to express impersonalization and objectivity. In Spanish, patterns with higher frequencies were: N Prep N Prep N, N Prep N Prep N Prep N, N Adj Prep N, N Prep N Adj Prep N and N Prep N Prep N N, which are closely related to patterns presented in English. Prepositional phrases (Prep N) are particularly relevant in translated NPs.

5. In addition, syntactic relation [C [B A]] was the most frequent in the *analysis corpus* and was associated with patterns like N N N, N Adj N, N PPi N, Adj N N and PP N N in English. This is worth noticing since possibilities to face 3-word patterns during translation tasks are high and this syntactic relation is the most prevalent regardless of the lexical components of the NP and their location.

Syntactic relation [[D C] [B A]] is the most prevalent in pattern N N N N and relation [D [[C B] A]] not only included pattern N N N N, it involved other patterns such as N Adj N N and Adj N N N as well, which is especially meaningful when it comes to translation of these types of NPs.

6. From a semantic perspective, proper names of associations, political organizations, economic groups and diverse kinds of entities are an important source of nested abbreviations and Military Sciences is an especially productive field. However, AIDS as an entity inside Health Sciences is a relevant source of nested abbreviations itself.

7. Regarding the use of nested abbreviations in texts, it was found that intermediate forms are prevalent. In almost half of texts, abbreviations inside the definition of nested abbreviations were not developed. Moreover, it is important to notice that in approximately 50% of cases involving nested abbreviations, one of the

rules exposed in writing manuals about the use of abbreviations was not followed, that which recommends to explain abbreviations once they are introduced in texts to improve readers' comprehension.

8. Likewise, the abbreviations identified in our corpus showed that conceptual abbreviations were prevalent in texts of general and specialized corpora, and a significant amount of these abbreviations were related to Molecular Biology. It is also worth remarking the high percentage of terms that were not found in the specialized corpus. Nonetheless, the limitation in number of tokens in the specialized corpus left a large amount of terms unexplored. Therefore, it is not possible to affirm if nested abbreviations are not used in academic books or if their use is merely beginning to be noticed.

In order to fulfill our second specific objective, development of abbreviations inside definitions of nested abbreviations allowed the identification of different types of nesting including: simple, complex, double complex, atypical and hybrid nesting.

Simple nesting is the most prevalent form, since it involves only one abbreviation in the definition. However, other types of nesting show how authors are becoming even more creative and find different ways to compress larger amounts of information, producing highly specialized terms. It can be inferred that these terms are restricted to an expert audience, since their formation processes and comprehension of their full meaning are highly complex.

In order to answer the second research question: What is the behavior of nested abbreviations when translated from English into Spanish? 433 nested abbreviations in English were translated into Spanish and three Colombian professional translators validated these translations.

According to our data, translation of nested abbreviations is focused on translation of the definition, including the abbreviation within in most of cases. This finding is consistent with the second solution to translate abbreviations proposed by Fijo (2003:115), where the definition is translated into a second language and the abbreviation remains in the source language.

The translation of all abbreviations inside definitions involved the translation of the definitions into a second language and the formation of new abbreviations using the initial letters of the resulting term. Molecular Biology provided half of translated abbreviations into Spanish.

From a translational perspective, a comparison between definition of nested abbreviations in English and their translations into Spanish showed that the number of tokens of each NP remained similar in both languages. Moreover, nouns are also the most frequent lexical category in Spanish, consistent with nominalization strategies of specialized discourse.

As stated before, the analysis of morphosyntactic patterns and their relation to translation is more significant when their syntactic relations are considered too. Because they provide information on how each component within the NP impacts others. This information might be useful in translation tasks as it provides translators a better understanding of terms, helping them to perform a more accurate task, especially during the translation of specialized texts.

Based on our data, it is possible to infer that when it comes to translation of abbreviations, nested or regular, there are no standard translation solutions since the number of morphosyntactic patterns and syntactic relations is as diverse as the number of terms existing in specialized languages. However, some regularities were found, and these regularities improve understanding and provide evidence of some translation solutions. For instance, it was found that several morphosyntactic patterns in English were related to one pattern in Spanish:

1. Morphosyntactic patterns in English such as N N N, N PPi N, N – PPi N and N N N N (in one case) which are associated with syntactic relations [C [B A]], [[C B] A], [C - [B A]] and [[D C] [B A]], were translated into Spanish to pattern N Prep N Prep N, which has two possible syntactic relations which are: [A [B C]] and [[A B] C].

2. Patterns N N N N, N N N, N Adj N N and N Adj N which are associated with syntactic relations [[D C] [B A]], [D [[C B] A]], [C [B A]] and [[C B] A] were translated into Spanish to pattern N Prep N Prep N Prep N , which has three possible syntactic relations: [[A B] [C D]], [[A [B C]] D] and [[[A B] C] D].

3. 3-word patterns such as N N N, N Adj N, N PPi N, Adj N N and PP N N were associated with syntactic relation [C [B A]] were translated into Spanish to pattern N Adj Prep N and were associated with syntactic relation [[A B] C].

Although, the number of patterns analyzed is little and no further generalizations are possible, it is relevant to show that certain regularities were found, even in complex processes as nested abbreviation. Therefore, these trends might help translators to find solutions based on evidence during the performance of translation of specialized texts.

Recommendations and Lines for Future Work

We hope that our description of nested abbreviations and types of nesting might be used in terminology and translation manuals, since this type of minor-word formation is used by specialists in academic texts and it is not described yet. Therefore, translators and terminologists may not know how to proceed when they have to face these entities.

As the percentage of nested abbreviations found in abbreviation dictionaries is still small, we recommend continuing the identification of these abbreviations in dictionaries, since there are higher possibilities to find them in these kinds of corpus. However, bilingual abbreviation dictionaries are not easy to find online and this might cause difficulties in the collection of a parallel corpus in order to perform translation analysis.

An alternative to solve this inconvenient might be the use of bilingual glossaries like the ones used by NATO, which are an excellent source of nested abbreviations. It is important to consider that these dictionaries manage English and French and, this may open an opportunity to expand the study of nested abbreviations into other languages. Nonetheless, these glossaries will limit analysis to Military Sciences and we consider important to diversify the analysis into other fields in order to search for more regularities that allow us make generalizations.

Another difficulty in our research was the presence of a 5-token limit in the specialized corpus leading to a significant number of extended forms unexplored. These limits were discovered during the use of the Ngram Viewer® website, since the creators did not inform about this feature of the tool. Therefore, we recommend the use of another specialized corpus such as Google Scholar® in order to maintain similar patterns of analysis considering that it provides the number of occurrences of terms.

As presented before, in English some of the abbreviations require the article when functioning as head in the NP structure; however, proper name abbreviations that stand as full NPs are used without the definite article (Huddleston & Pullum, 2002).

The contexts presented in chapter 4 were searched and chosen randomly in order to provide some examples of this feature. Therefore, it would be convenient to use textual corpora in order to have access to contexts surrounding nested abbreviations.

Literature review and translation analysis of nested abbreviations showed that there were no regularities on translation of this type of abbreviations. It is important that the decision of translating or not abbreviations within texts is no longer left to the preferences of specialist in certain fields, since they seem not to know, most of the time, the reasons for the choices they make.

In my personal experience as a Medical Doctor with 8-year practice and 7-year training, I might say that a great percentage of specialists act as their colleagues do in other areas and other countries, that is without a real conscience about linguistic aspects of medical abbreviations and terms.

As researchers, it is important to continue this type of work, since it may provide information that can be used by technical committees and terminology commissions in order to create more solid policies about translation of these entities. As a consequence, terminologists and translators might have more foundations to assist specialists during the production of academic texts in different languages.

For further researches, it would be convenient to deepen the analysis of the meaning of nested abbreviations from the syntactic perspective. For instance, studying NPs with pre and postmodification, and the location of abbreviations in this type of NPs.

From a translational perspective, the use of parallel corpus will provide important information on translation strategies for the implementation of contrastive analysis on the translation of nested abbreviations.

Another area of interest for future studies would be the implementation of morphologic features, morphosyntactic patterns and syntactic relations found in this work to existing extraction and disambiguation of abbreviations systems or the creation of a new one with this information. This might be done in order to demonstrate if systems' performance can be improved by applying what has been found in our research.

REFERENCES

Alcaraz, M. Á. (2002). Las siglas en el discurso biomédico escrito en Inglés: análisis y aplicaciones didácticas. *The ESPecialist, Pesquisa Em Línguas Para Fins Específicos. Descrição, Ensino E Aprendizagem*, *23*(1), 37–51. Retrieved from revistas.pucsp.br/index.php/esp/article/download/9388/6960

Alonso, A. (2002). La abreviación en los libros de texto y en los medios de comunicación. In *Actas Del V Simposio Regional De Actualización Científica Y Didáctica de Simposio Regional De Actualización Científica Y Didáctica De Lengua Española Y Literatura*. Sevilla: Asociación Andaluza de profesores de Español "Elio Antonio de Nebrija." Retrieved from http://www.todostuslibros.com/libros/actas-del-v-simposio-regional-de-actualizacion-cientifica-y-didactica_978-84-88842-11-4

Alred, G. J., Brusaw, C. T., Oliu, W. E., Goossens, M., Mittelbach, F., Braams, J., & Lamport, L. (2006). Handbook of technical writing. St. Martin's.

American Medical Association. (2010). *Chicago Manual of Style*. Chicago: The University of Chicago Press.

Arntz, R., & Picht, H. (1995). *Introducción a la Terminología*. (A. Irazazábal, M. Jiménez, E. Schwarz, S. Yunquera, Trads) Madrid, España: Ediciones Pirámide, S.A. Retrieved from http://ocw.um.es/cc.-sociales/terminologia

Betancourt, B., Treto, L., & Fernández, A. (2013). Traducción de acrónimos y siglas en textos médicos de cardiología. *CorSalud*, *5*(1), 93–100.

Biber, D., Stig, J., Leech, G., Conrad, S., & Edward, F. (1999). *Longman Grammar of Spoken and Written English*. Londres: Longman.

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. (C. Tebé, Trad.) Barcelona: Antártida.

Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*, ed Juan C. Sager, Amsterdam, Netherlands: John Benjamins.

Cannon, G. (1989). Abbreviations and acronyms in English word-formation. *American Speech*, 64(2), 99-127.

Cardero, A. (2003). *Terminología y Procesamiento*. México, D.F: Universidad Nacional Autónoma de México.

Cardero, A. (2004a). *El terminólogo, la lingüística y la terminología. Una experiencia*. México, D.F: Universidad Nacional Autónoma de México.

Cardero, A. (2004b). *Lingüística y Terminología*. México, D.F: Universidad Nacional Autónoma de México.

Ciapuscio, G. (2003). Textos especializados y terminología. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cuadrado, L. A. (1998). Sobre la formación de palabras en español. *Lengua y cultura en la enseñanza del español a extranjeros*, Cuenca: Ediciones de la Universidad de Castilla-La Mancha, 257-263.

Dannélls, D. (2005). Recognizing Swedish acronyms and their definitions in biomedical literature. Retrieved from http://www.biomedcentral.com/1471-2015-6-103

Dribniuk, V. T. (2009). English Medical Abbreviations. *Наукові записки [Національного університету* , (11), 206-209.

Eatock, R. A., Fay, R. R., & Popper, A. N. (2006). *Vertebrate hair cells,* 348-442. New York: Springer.

Felber, H. (1984). *Terminology Manual*. Paris: Unesco.

Figueroa, B., & Silva, T. (2000). Un diccionario de especialidad: las siglas. Casal ML, editora. *La lingüística francesa en España camino del siglo XII. En: Casal ML, Conde G, Lago J, Pino L, Rodríguez N, editores. La lingüística francesa en España camino del siglo XXI.* Arrecife: España, 455-467.

Fijo, M. I. (2003). *Las siglas en el lenguaje de la enfermería: análisis contrastivo inglés-español por medio de fichas terminológicas* (Doctoral dissertation). Universidad Pablo de Olavide, Sevilla, España.

Frantzi, K. T., & Ananiadou, S. (1996). Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics* (1), 41-46. Association for Computational Linguistics.

Giraldo, J. J. (2006). Sistemas de detección y extracción semiautomática de siglas Estado de la cuestión. Retrieved from: http://www.recercat.cat/bitstream/handle/2072/9083/2006%20BE%2000357%20(mem%C3%B2ria).pdf?sequence=1.

Giraldo, J. J. (2008). *Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente*. (Doctoral dissertation) Universitat Pompeu Fabra, Barcelona, España.

Gómez, J. (1992). Las siglas en el lenguaje de la economía. *Revista de filología románica*, (9), 267-274.

Goodman, N. W., & Edwards, M. B. (1997). *Medical Writing: A Prescription for Clarity*. Cambridge University Press.

Gotti, M. (2003). *Specialized discourse: Linguistic features and changing conventions*. Berna: Peter Lang.

Grange, B., & Bloom, D. A. (2000). Acronyms, abbreviations and initialisms. *BJU international*, 86(1), 1-6.

Gutiérrez, B. M. (1998). *La ciencia empieza en la palabra. Análisis e historia del lenguaje científico*. Barcelona: Península.

Halliday, M. A., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar*. New York: Routledge.

Hoffmann, L. (1998). *Llenguatges d'especialitat*. Barcelona, España: IULA.

Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge: Cambridge University Press.

Huth, E. J. (1990). *How to write and publish papers in the medical sciences*. Williams & Wilkins.

Hutzler, W. P., Geriner, P. T., & Gulledge, T. R. (Eds.). (1994). *Software Engineering Economics and Declining Budgets: With 63 Figures*. Springer.

ISO, I. (2000). 704: Terminology work–Principles and methods. *International Organization for Standardization.*

ISO 1087-1 (2000), Terminology work–Vocabulary–Part 1: Theory and application. *International Organization for Standardization.*

ISO 12620 (1999), Computer applications in terminology. *International Organization for Standardization.*

ISO 17241 (2000), Computer applications in terminology — Generic model — (GENETER)for SGML - based representation of terminological data. *International Organization for Standardization.*

Jablonski, S. (2009). *Jablonski's Dictionary of Medical Acronyms & Abbreviations (Ed. 6)*. Elsevier Health Sciences.

Lang, U., & Schreiner, R. (2002). *Developing secure distributed systems with CORBA*. Artech house.

Lerat, P. (1997). *Las Lenguas Especializadas* (A. Ribas, Trad). Barcelona: Ariel Lingüística.

Lombard, J., & Kotzé, H. (2013). English Style Guide A handbook for authors and translators in the European Commission. *Up.Ac.Za*, 94. Retrieved from http://web.up.ac.za/sitefiles/file/1 Tanya/web office/UP English Style Guide_2011.pdf

López, P. (2004). Acronyms & Co.: A typology of typologies. *Estudios Ingleses de la Universidad Complutense, 12*, 109-129.

Lušicky, V., & Wissik, T. (2015). *Procedural Manual on Terminology. Translation-Oriented Terminology Work.* Retrieved January 10th 2016, from http://www.sep.gov.mk/data/file/Preveduvanje/Procedural_Manual_on_Terminology_final_version.pdf

Mattia, F. B. (2003). *Elsevier's dictionary of acronyms, initialisms, abbreviations and symbols*. Elsevier.

Microsoft Corporation. (2004). *Microsoft manual of style for technical publications* (3rd ed). Washington: Microsoft Press.

Miller, D. F. (1995). Guidelines for Creating and Using Abbreviations and Acronyms. *Jet Propulsion Laboratory*, *95*(9).

Muoniz Castro, E. G. (1997). *Routledge Spanish Dictionary of Business, Commerce and Finance Diccionario Inglés de Negocios, Comercio y Finanzas: Spanish-English/English-Spanish*. Londres: Routledge

NATO. (2000). NATO Glossary of abbreviations used in NATO documents and publications. Retrieved from http://seb.brc.free.fr/ressources/aap-15 complet.pdf

NATO. (2005). NATO Glossary of abbreviations used in NATO documents and publications. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a574310.pdf

NATO. (2010). Nato Glossary of Terms and Definitions. Retrieved from http://static.lexicool.com/dictionary/GR2PJ13289.pdf

NATO. (2013). NATO Glossary of abbreviations used in NATO documents and publications, 330. Retrieved from http://www.dtic.mil/doctrine/doctrine/other/aap15.pdf

Navarro, F. (2000). *Diccionario Crítico de Dudas Inglés-Español de Medicina* España: McGraw-Hill/Interamericana. Retrieved from http://www.casadellibro.com/libro-diccionario-critico-de-dudas-ingles-espanol-de-medicina-2-ed/9788448198084/1043688

Newmark, P. (1979). A layman's view of medical translation. *British medical journal*, 2(6202), 1405.

Newmark, P. (1988). *A textbook of translation* (1), 988. New York: Prentice Hall.

Nossum, V. (2012). *SAT-based preimage attacks on SHA-1*. (Doctoral dissertation) University of Oslo, Oslo, Norway.

Okazaki, N., Ananiadou, S., & Tsujii, J. (2010). Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, *26*(9), 1246–1253.

Ordóñez, A. (1992). *Lenguaje Médico. Estudio Sincrónico de una Jerga*. Madrid: Editorial Universidad Autónoma de Madrid.

Oxford, U. of. (2014). University of Oxford Style Guide. Retrieved from https://www.ox.ac.uk/public-affairs/style-guide?wssl=1

Pavel, S., & Nolet, D. (2001). *Handbook of Terminology*. (C. Leonhardt, Trad.) Quebec: Translation Bureau.

Plag, I. (2003). *Word-Formation in English (review)*. *Cambridge Textbooks in Linguistics*. Cambridge University Press. http://doi.org/10.1353/lan.2006.0013

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1990). *A Comprenhensive Grammar of the English Language*. London: Longman.

Quiroz, G. (2006). Using an English-Spanish Parallel Corpus to Solve Complex Premodification in Noun Phrases. *Insights into Specialized Translation. Bern: Peter Lang,* 367-390.

Quiroz, G. (2008). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma*. (Doctoral dissertation) Universitat Pompeu Fabra, Barcelona, España. Retrieved from http://tdx.cat/handle/10803/7509.

Quiroz, G., & Arroyave, A. (2014). On Premedified Terms in Five Specialized Dictionaries. In G. Quiroz & P. Patiño (Eds.), *LSP in Colombia : Advances and Challenges* (p. 355). Peter Lang.

Rodríguez, F. (1987). Naturaleza sintáctica de las formas siglares. El cambio funcional. *Estudios de lingüística* 4, 139-148.

Rodríguez, F. (1993). Las siglas como procedimiento lexicogenésico. *Estudios de lingüística* 9, 9-24.

Rojas, J. L. (2014). Etiquetaje y descripción de unidades fraseológicas especializadas en un diccionario bilingüe de comercio internacional. (Masters Thesis) University of Antioquia, Medellín, Colombia. Retrieved from http://tesis.udea.edu.co/dspace/bitstream/10495/1905/1/JoseLuisRojas_trabajodegrado_maestria.pdf.

Sabin, W. (2010). *The Gregg Reference Manual: A Manual of Style, Grammar, Usage, and Formatting.* New York: McGraw-Hill.

Sánchez-Gijón, P. (2004). L'ús de corpus en la traducció especialitzada. Barcelona: Institut Universitari de Lingüística Aplicada.

Sanz, M. L. (2011). *Análisis contrastivo de la terminologıa de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al espanol* (Doctoral dissertation) Universidad de Salamanca; Salamanca, España.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. http://doi.org/10.1.1.28.1139.

Schwager, E. (1991). *Medical English Usage and Abusage.* Phoenix: Oryx Press.

The Economist. (2005). *Style Guide*. London: Profile Books.

Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. In *Workshop Description*, 54–57. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.5316&rep=rep1&type=pdf.

Williams, B., Gage, K., Arvai, A., Baski, M., & Cooke, W. (2012). Risk Assessment and Mission Planning During Heightened Meteoroid Activity: The Evolution and Current State of Art Adapted for the Chandra Mission. *SpaceOps 2012 Conference*, 1–12. http://doi.org/10.2514/6.2012-1285451.

Witten, I. H., & Bainbridge, D. (2002). *How to Build a Digital Library*. Morgan Kaufmann. Retrieved from https://books.google.com/books?id=mrlUvcs9koAC&pgis=1.

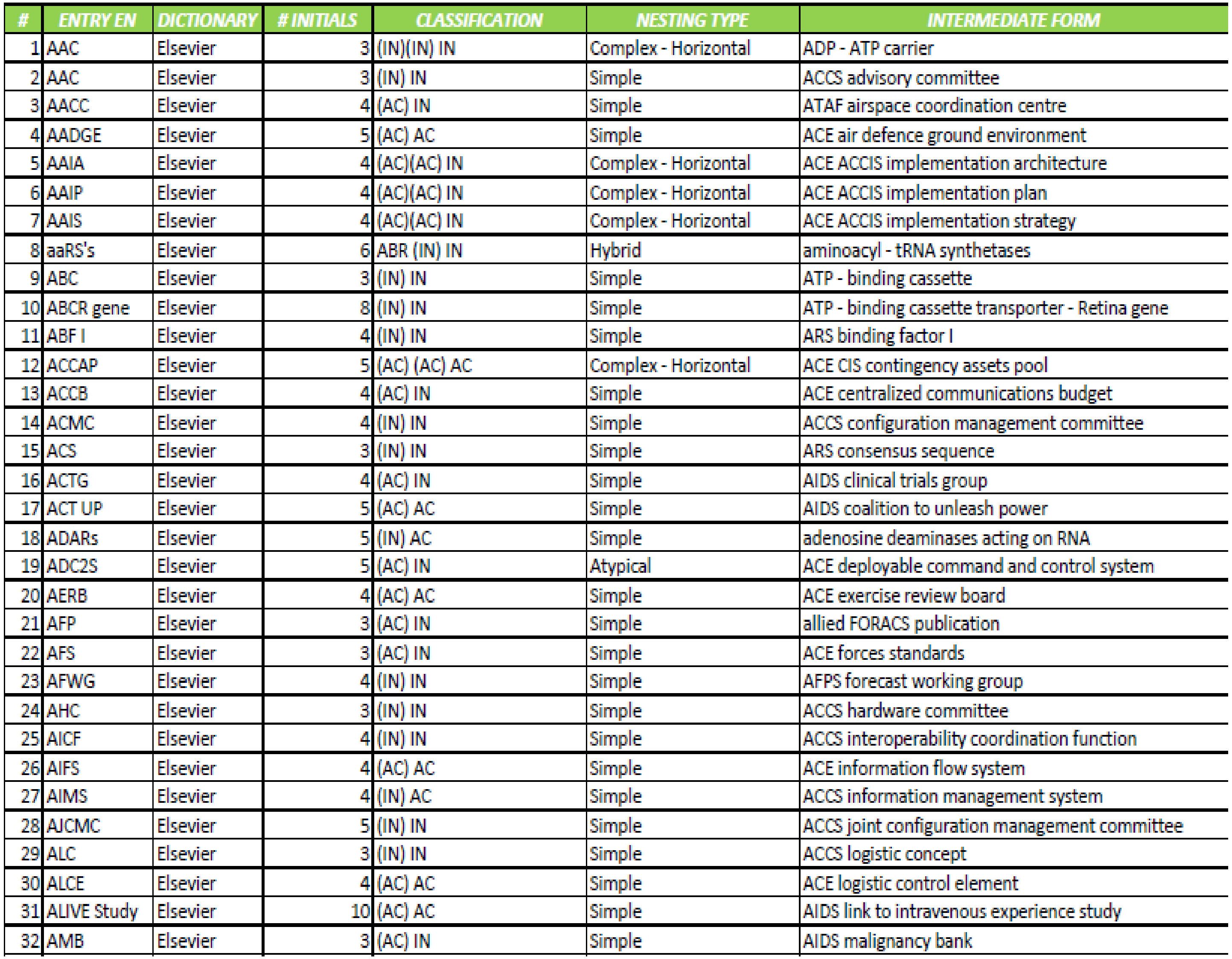
Wren, J. D. (2003). *The IRIDESCENT System: An Automated Data-Mining Method to Identify, Evaluate, and Analyze Sets of Relationships Within Textual Databases* (Doctoral dissertation). University of Texas Southwestern Medical Center, Dallas, United States.

Yetano, J., & Alberola, V. (2004). *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias*. (M. de S. y Consumo, Ed.). Madrid: Solana e hijos. Retrieved from papers://0d26f653-cfd4-4933-b3b6-e1aeec47660b/Paper/p1071.

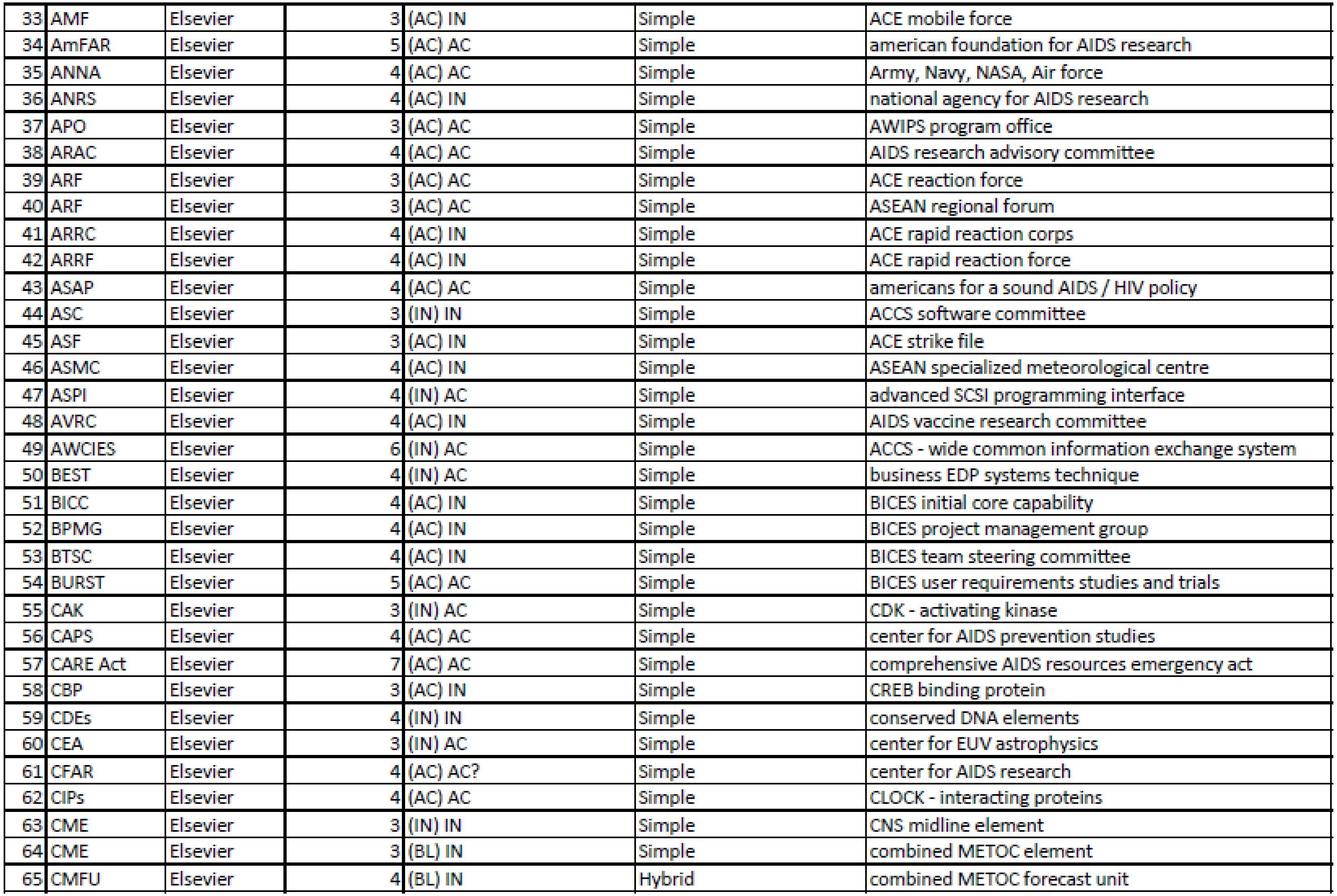
Zolondek, D. (1991). La siglaison. *Terminogramme* 62, 1-5.

APPENDIX

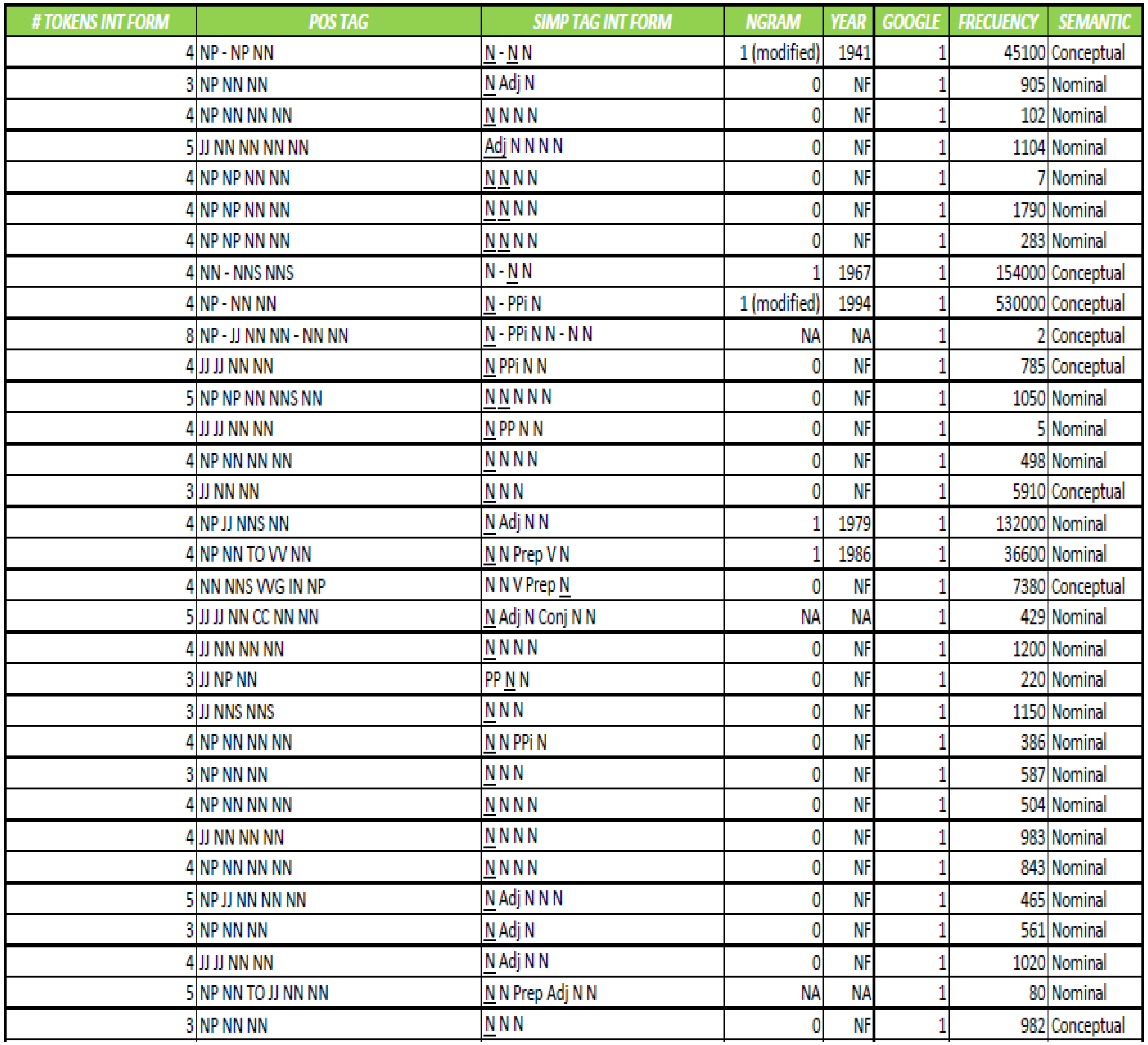
A.1 Extract of database in English



#	ENTRY EN	DICTIONARY	# INITIALS	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM
1	AAC	Elsevier	3	(IN)(IN) IN	Complex - Horizontal	ADP - ATP carrier
2	AAC	Elsevier	3	(IN) IN	Simple	ACCS advisory committee
3	AACC	Elsevier	4	(AC) IN	Simple	ATAF airspace coordination centre
4	AADGE	Elsevier	5	(AC) AC	Simple	ACE air defence ground environment
5	AAIA	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation architecture
6	AAIP	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation plan
7	AAIS	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	ACE ACCIS implementation strategy
8	aaRS's	Elsevier	6	ABR (IN) IN	Hybrid	aminoacyl - tRNA synthetases
9	ABC	Elsevier	3	(IN) IN	Simple	ATP - binding cassette
10	ABCR gene	Elsevier	8	(IN) IN	Simple	ATP - binding cassette transporter - Retina gene
11	ABF I	Elsevier	4	(IN) IN	Simple	ARS binding factor I
12	ACCAP	Elsevier	5	(AC) (AC) AC	Complex - Horizontal	ACE CIS contingency assets pool
13	ACCB	Elsevier	4	(AC) IN	Simple	ACE centralized communications budget
14	ACMC	Elsevier	4	(IN) IN	Simple	ACCS configuration management committee
15	ACS	Elsevier	3	(IN) IN	Simple	ARS consensus sequence
16	ACTG	Elsevier	4	(AC) IN	Simple	AIDS clinical trials group
17	ACT UP	Elsevier	5	(AC) AC	Simple	AIDS coalition to unleash power
18	ADARs	Elsevier	5	(IN) AC	Simple	adenosine deaminases acting on RNA
19	ADC2S	Elsevier	5	(AC) IN	Atypical	ACE deployable command and control system
20	AERB	Elsevier	4	(AC) AC	Simple	ACE exercise review board
21	AFP	Elsevier	3	(AC) IN	Simple	allied FORACS publication
22	AFS	Elsevier	3	(AC) IN	Simple	ACE forces standards
23	AFWG	Elsevier	4	(IN) IN	Simple	AFPS forecast working group
24	AHC	Elsevier	3	(IN) IN	Simple	ACCS hardware committee
25	AICF	Elsevier	4	(IN) IN	Simple	ACCS interoperability coordination function
26	AIFS	Elsevier	4	(AC) AC	Simple	ACE information flow system
27	AIMS	Elsevier	4	(IN) AC	Simple	ACCS information management system
28	AJCMC	Elsevier	5	(IN) IN	Simple	ACCS joint configuration management committee
29	ALC	Elsevier	3	(IN) IN	Simple	ACCS logistic concept
30	ALCE	Elsevier	4	(AC) AC	Simple	ACE logistic control element
31	ALIVE Study	Elsevier	10	(AC) AC	Simple	AIDS link to intravenous experience study
32	AMB	Elsevier	3	(AC) IN	Simple	AIDS malignancy bank

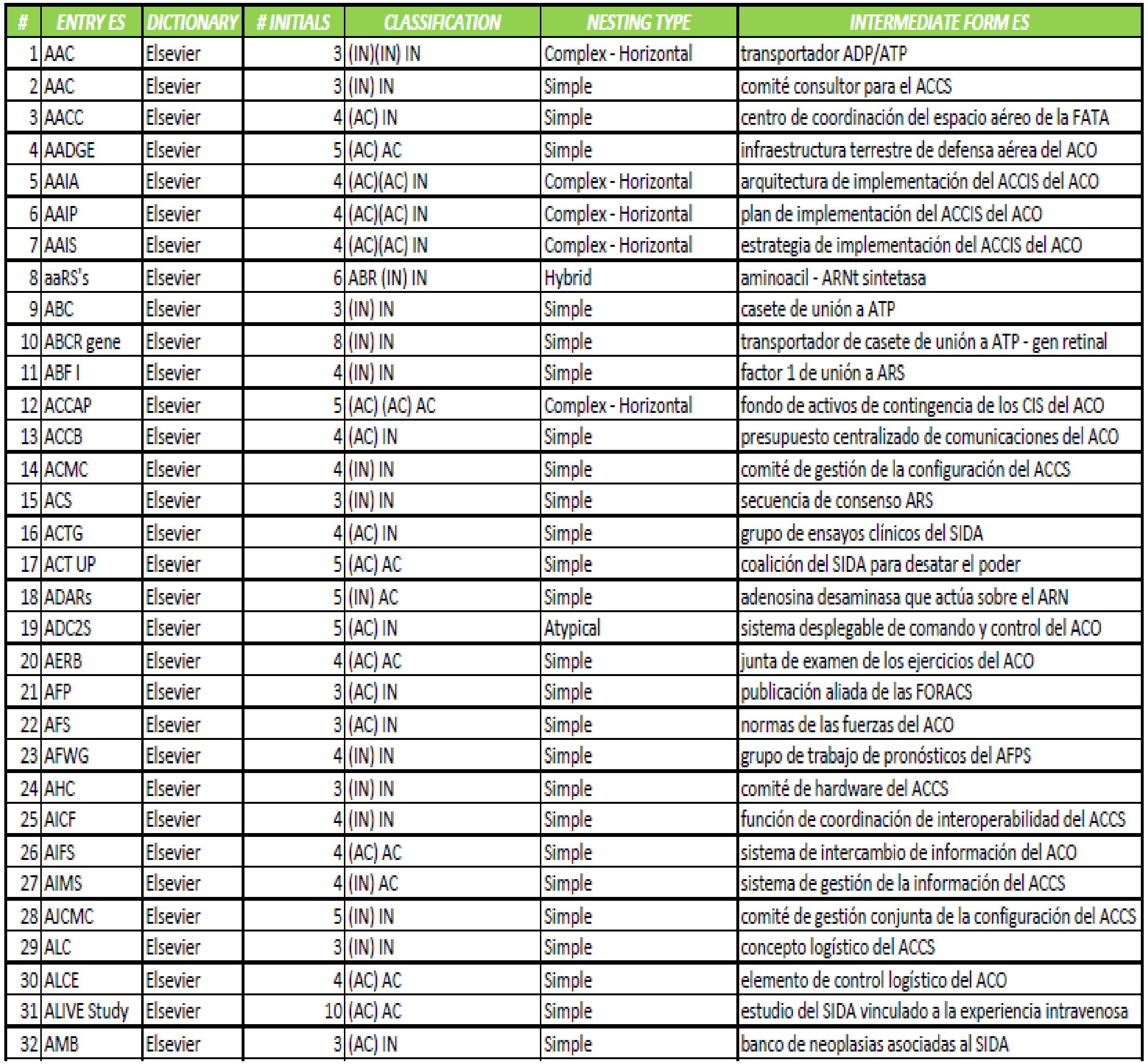


33	AMF	Elsevier	3	(AC) IN	Simple	ACE mobile force
34	AmFAR	Elsevier	5	(AC) AC	Simple	american foundation for AIDS research
35	ANNA	Elsevier	4	(AC) AC	Simple	Army, Navy, NASA, Air force
36	ANRS	Elsevier	4	(AC) IN	Simple	national agency for AIDS research
37	APO	Elsevier	3	(AC) AC	Simple	AWIPS program office
38	ARAC	Elsevier	4	(AC) AC	Simple	AIDS research advisory committee
39	ARF	Elsevier	3	(AC) AC	Simple	ACE reaction force
40	ARF	Elsevier	3	(AC) AC	Simple	ASEAN regional forum
41	ARRC	Elsevier	4	(AC) IN	Simple	ACE rapid reaction corps
42	ARRF	Elsevier	4	(AC) IN	Simple	ACE rapid reaction force
43	ASAP	Elsevier	4	(AC) AC	Simple	americans for a sound AIDS / HIV policy
44	ASC	Elsevier	3	(IN) IN	Simple	ACCS software committee
45	ASF	Elsevier	3	(AC) IN	Simple	ACE strike file
46	ASMC	Elsevier	4	(AC) IN	Simple	ASEAN specialized meteorological centre
47	ASPI	Elsevier	4	(IN) AC	Simple	advanced SCSI programming interface
48	AVRC	Elsevier	4	(AC) IN	Simple	AIDS vaccine research committee
49	AWCIES	Elsevier	6	(IN) AC	Simple	ACCS - wide common information exchange system
50	BEST	Elsevier	4	(IN) AC	Simple	business EDP systems technique
51	BICC	Elsevier	4	(AC) IN	Simple	BICES initial core capability
52	BPMG	Elsevier	4	(AC) IN	Simple	BICES project management group
53	BTSC	Elsevier	4	(AC) IN	Simple	BICES team steering committee
54	BURST	Elsevier	5	(AC) AC	Simple	BICES user requirements studies and trials
55	CAK	Elsevier	3	(IN) AC	Simple	CDK - activating kinase
56	CAPS	Elsevier	4	(AC) AC	Simple	center for AIDS prevention studies
57	CARE Act	Elsevier	7	(AC) AC	Simple	comprehensive AIDS resources emergency act
58	CBP	Elsevier	3	(AC) IN	Simple	CREB binding protein
59	CDEs	Elsevier	4	(IN) IN	Simple	conserved DNA elements
60	CEA	Elsevier	3	(IN) AC	Simple	center for EUV astrophysics
61	CFAR	Elsevier	4	(AC) AC?	Simple	center for AIDS research
62	CIPs	Elsevier	4	(AC) AC	Simple	CLOCK - interacting proteins
63	CME	Elsevier	3	(IN) IN	Simple	CNS midline element
64	CME	Elsevier	3	(BL) IN	Simple	combined METOC element
65	CMFU	Elsevier	4	(BL) IN	Hybrid	combined METOC forecast unit

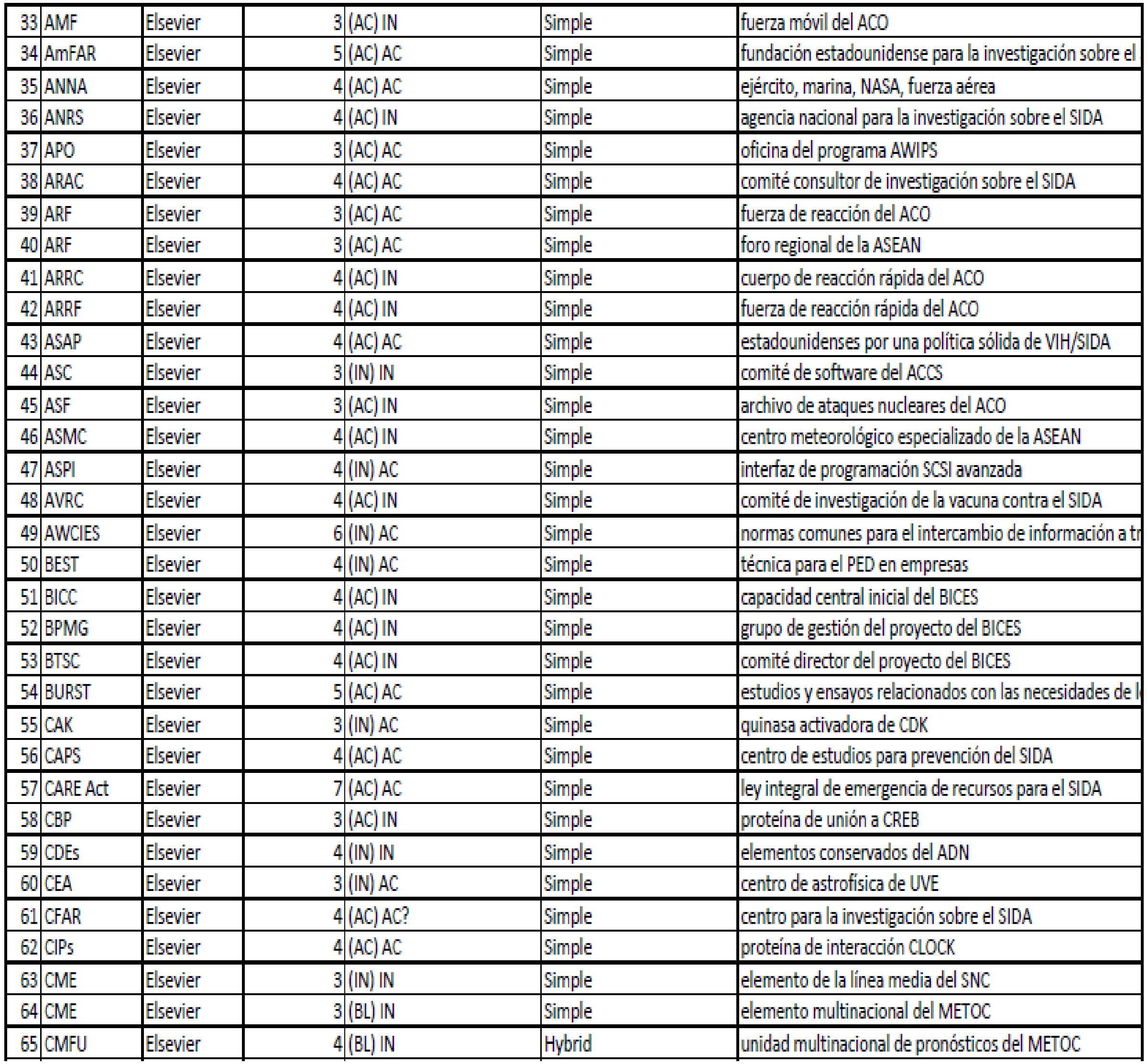


# TOKENS INT FORM	POS TAG	SIMP TAG INT FORM	NGRAM	YEAR	GOOGLE	FRECUENCY	SEMANTIC
4	NP - NP NN	N - N N	1 (modified)	1941	1	45100	Conceptual
3	NP NN NN	N Adj N	0	NF	1	905	Nominal
4	NP NN NN NN	N N N N	0	NF	1	102	Nominal
5	JJ NN NN NN NN	Adj N N N N	0	NF	1	1104	Nominal
4	NP NP NN NN	N N N N	0	NF	1	7	Nominal
4	NP NP NN NN	N N N N	0	NF	1	1790	Nominal
4	NP NP NN NN	N N N N	0	NF	1	283	Nominal
4	NN - NNS NNS	N - N N	1	1967	1	154000	Conceptual
4	NP - NN NN	N - PPi N	1 (modified)	1994	1	530000	Conceptual
8	NP - JJ NN NN - NN NN	N - PPi N N - N N	NA	NA	1	2	Conceptual
4	JJ JJ NN NN	N PPi N N	0	NF	1	785	Conceptual
5	NP NP NN NNS NN	N N N N N	0	NF	1	1050	Nominal
4	JJ JJ NN NN	N PP N N	0	NF	1	5	Nominal
4	NP NN NN NN	N N N N	0	NF	1	498	Nominal
3	JJ NN NN	N N N	0	NF	1	5910	Conceptual
4	NP JJ NNS NN	N Adj N N	1	1979	1	132000	Nominal
4	NP NN TO VV NN	N N Prep V N	1	1986	1	36600	Nominal
4	NN NNS VVG IN NP	N N V Prep N	0	NF	1	7380	Conceptual
5	JJ JJ NN CC NN NN	N Adj N Conj N N	NA	NA	1	429	Nominal
4	JJ NN NN NN	N N N N	0	NF	1	1200	Nominal
3	JJ NP NN	PP N N	0	NF	1	220	Nominal
3	JJ NNS NNS	N N N	0	NF	1	1150	Nominal
4	NP NN NN NN	N N PPi N	0	NF	1	386	Nominal
3	NP NN NN	N N N	0	NF	1	587	Nominal
4	NP NN NN NN	N N N N	0	NF	1	504	Nominal
4	JJ NN NN NN	N N N N	0	NF	1	983	Nominal
4	NP NN NN NN	N N N N	0	NF	1	843	Nominal
5	NP JJ NN NN NN	N Adj N N N	0	NF	1	465	Nominal
3	NP NN NN	N Adj N	0	NF	1	561	Nominal
4	JJ JJ NN NN	N Adj N N	0	NF	1	1020	Nominal
5	NP NN TO JJ NN NN	N N Prep Adj N N	NA	NA	1	80	Nominal
3	NP NN NN	N N N	0	NF	1	982	Conceptual

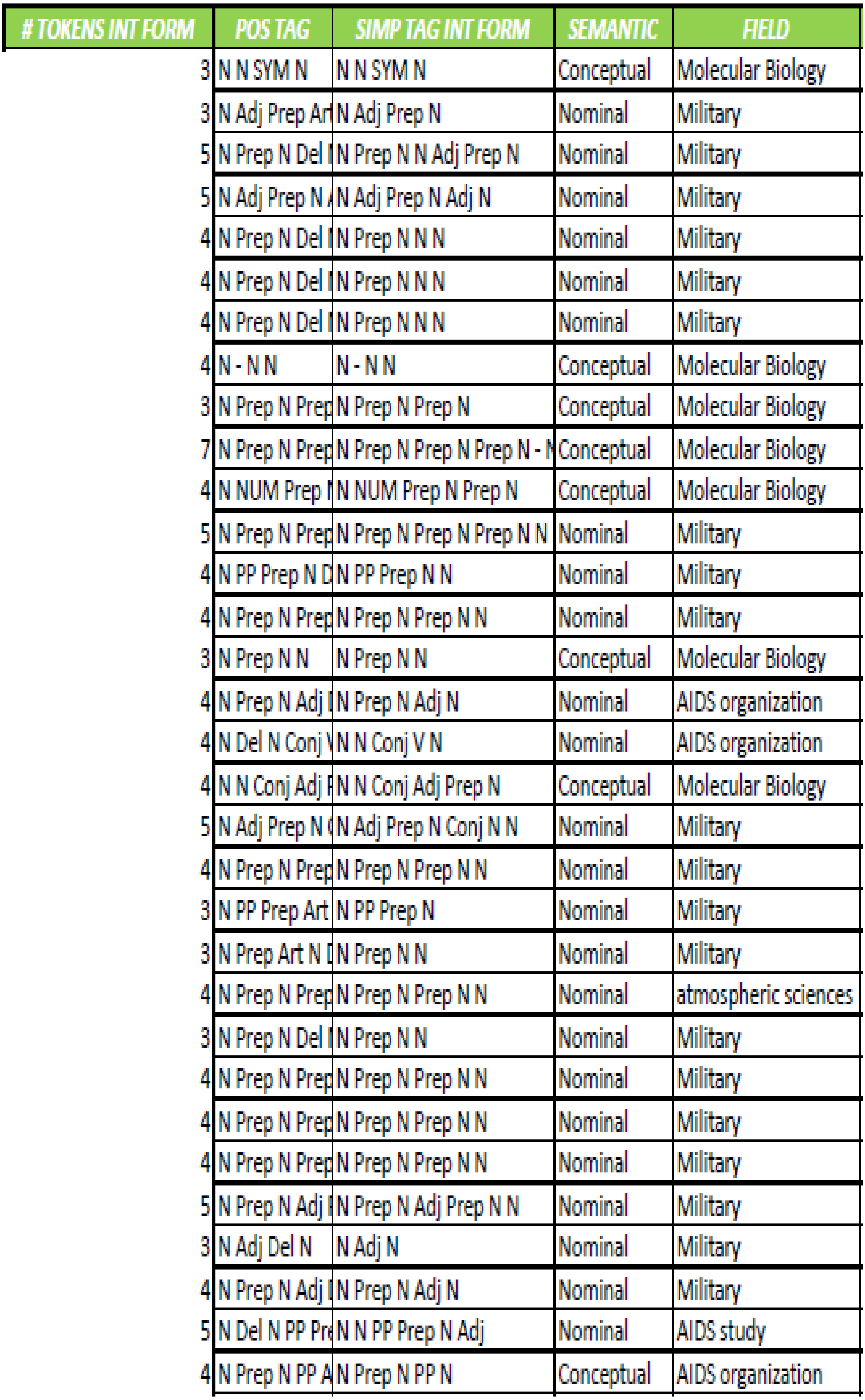
A.2 Extract of database in Spanish



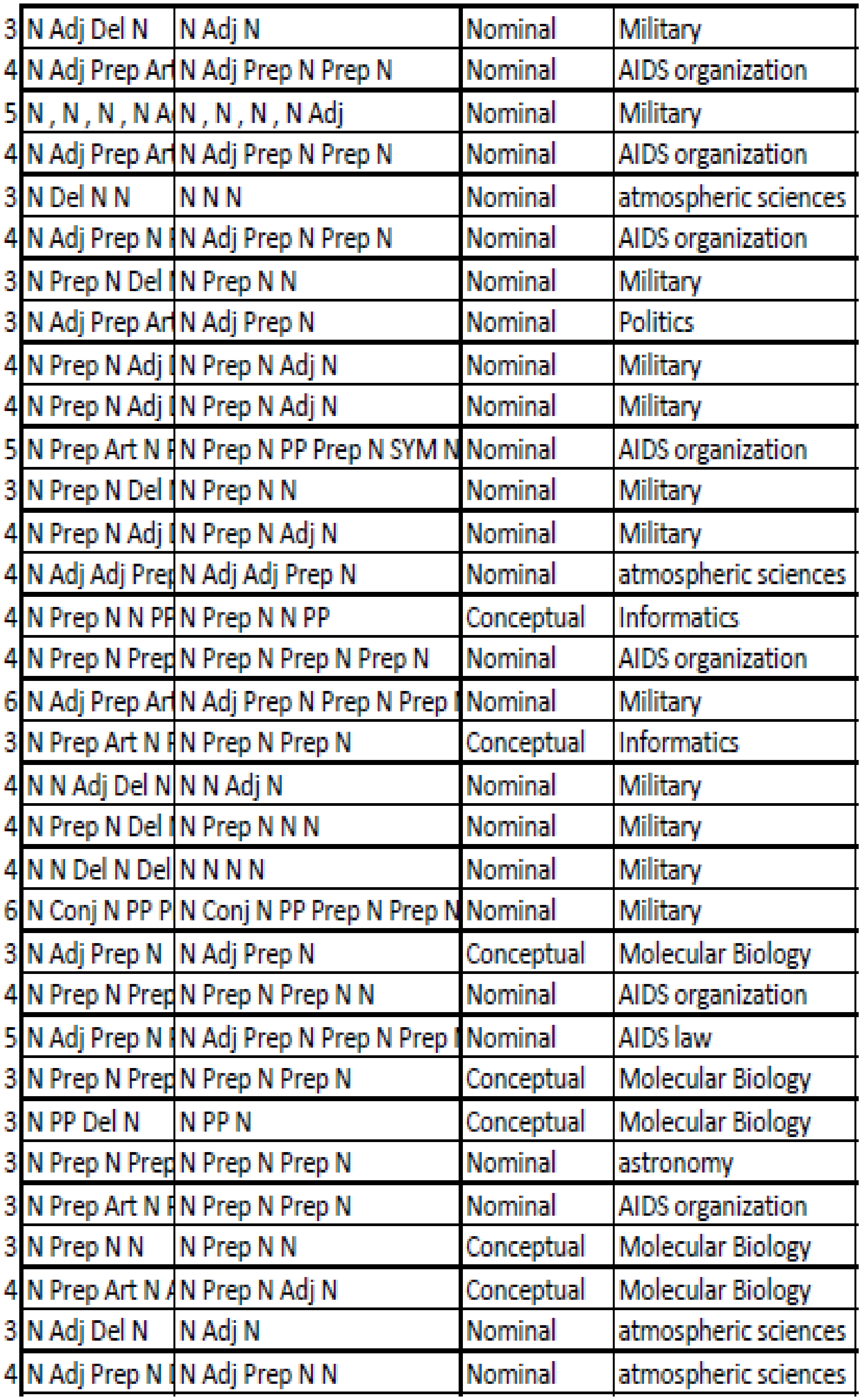
#	ENTRY ES	DICTIONARY	# INITIALS	CLASSIFICATION	NESTING TYPE	INTERMEDIATE FORM ES
1	AAC	Elsevier	3	(IN)(IN) IN	Complex - Horizontal	transportador ADP/ATP
2	AAC	Elsevier	3	(IN) IN	Simple	comité consultor para el ACCS
3	AACC	Elsevier	4	(AC) IN	Simple	centro de coordinación del espacio aéreo de la FATA
4	AADGE	Elsevier	5	(AC) AC	Simple	infraestructura terrestre de defensa aérea del ACO
5	AAIA	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	arquitectura de implementación del ACCIS del ACO
6	AAIP	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	plan de implementación del ACCIS del ACO
7	AAIS	Elsevier	4	(AC)(AC) IN	Complex - Horizontal	estrategia de implementación del ACCIS del ACO
8	aaRS's	Elsevier	6	ABR (IN) IN	Hybrid	aminoacil - ARNt sintetasa
9	ABC	Elsevier	3	(IN) IN	Simple	casete de unión a ATP
10	ABCR gene	Elsevier	8	(IN) IN	Simple	transportador de casete de unión a ATP - gen retinal
11	ABF I	Elsevier	4	(IN) IN	Simple	factor 1 de unión a ARS
12	ACCAP	Elsevier	5	(AC) (AC) AC	Complex - Horizontal	fondo de activos de contingencia de los CIS del ACO
13	ACCB	Elsevier	4	(AC) IN	Simple	presupuesto centralizado de comunicaciones del ACO
14	ACMC	Elsevier	4	(IN) IN	Simple	comité de gestión de la configuración del ACCS
15	ACS	Elsevier	3	(IN) IN	Simple	secuencia de consenso ARS
16	ACTG	Elsevier	4	(AC) IN	Simple	grupo de ensayos clínicos del SIDA
17	ACT UP	Elsevier	5	(AC) AC	Simple	coalición del SIDA para desatar el poder
18	ADARs	Elsevier	5	(IN) AC	Simple	adenosina desaminasa que actúa sobre el ARN
19	ADC2S	Elsevier	5	(AC) IN	Atypical	sistema desplegable de comando y control del ACO
20	AERB	Elsevier	4	(AC) AC	Simple	junta de examen de los ejercicios del ACO
21	AFP	Elsevier	3	(AC) IN	Simple	publicación aliada de las FORACS
22	AFS	Elsevier	3	(AC) IN	Simple	normas de las fuerzas del ACO
23	AFWG	Elsevier	4	(IN) IN	Simple	grupo de trabajo de pronósticos del AFPS
24	AHC	Elsevier	3	(IN) IN	Simple	comité de hardware del ACCS
25	AICF	Elsevier	4	(IN) IN	Simple	función de coordinación de interoperabilidad del ACCS
26	AIFS	Elsevier	4	(AC) AC	Simple	sistema de intercambio de información del ACO
27	AIMS	Elsevier	4	(IN) AC	Simple	sistema de gestión de la información del ACCS
28	AJCMC	Elsevier	5	(IN) IN	Simple	comité de gestión conjunta de la configuración del ACCS
29	ALC	Elsevier	3	(IN) IN	Simple	concepto logístico del ACCS
30	ALCE	Elsevier	4	(AC) AC	Simple	elemento de control logístico del ACO
31	ALIVE Study	Elsevier	10	(AC) AC	Simple	estudio del SIDA vinculado a la experiencia intravenosa
32	AMB	Elsevier	3	(AC) IN	Simple	banco de neoplasias asociadas al SIDA



33	AMF	Elsevier	3	(AC) IN	Simple	fuerza móvil del ACO
34	AmFAR	Elsevier	5	(AC) AC	Simple	fundación estadounidense para la investigación sobre el
35	ANNA	Elsevier	4	(AC) AC	Simple	ejército, marina, NASA, fuerza aérea
36	ANRS	Elsevier	4	(AC) IN	Simple	agencia nacional para la investigación sobre el SIDA
37	APO	Elsevier	3	(AC) AC	Simple	oficina del programa AWIPS
38	ARAC	Elsevier	4	(AC) AC	Simple	comité consultor de investigación sobre el SIDA
39	ARF	Elsevier	3	(AC) AC	Simple	fuerza de reacción del ACO
40	ARF	Elsevier	3	(AC) AC	Simple	foro regional de la ASEAN
41	ARRC	Elsevier	4	(AC) IN	Simple	cuerpo de reacción rápida del ACO
42	ARRF	Elsevier	4	(AC) IN	Simple	fuerza de reacción rápida del ACO
43	ASAP	Elsevier	4	(AC) AC	Simple	estadounidenses por una política sólida de VIH/SIDA
44	ASC	Elsevier	3	(IN) IN	Simple	comité de software del ACCS
45	ASF	Elsevier	3	(AC) IN	Simple	archivo de ataques nucleares del ACO
46	ASMC	Elsevier	4	(AC) IN	Simple	centro meteorológico especializado de la ASEAN
47	ASPI	Elsevier	4	(IN) AC	Simple	interfaz de programación SCSI avanzada
48	AVRC	Elsevier	4	(AC) IN	Simple	comité de investigación de la vacuna contra el SIDA
49	AWCIES	Elsevier	6	(IN) AC	Simple	normas comunes para el intercambio de información a tr
50	BEST	Elsevier	4	(IN) AC	Simple	técnica para el PED en empresas
51	BICC	Elsevier	4	(AC) IN	Simple	capacidad central inicial del BICES
52	BPMG	Elsevier	4	(AC) IN	Simple	grupo de gestión del proyecto del BICES
53	BTSC	Elsevier	4	(AC) IN	Simple	comité director del proyecto del BICES
54	BURST	Elsevier	5	(AC) AC	Simple	estudios y ensayos relacionados con las necesidades de l
55	CAK	Elsevier	3	(IN) AC	Simple	quinasa activadora de CDK
56	CAPS	Elsevier	4	(AC) AC	Simple	centro de estudios para prevención del SIDA
57	CARE Act	Elsevier	7	(AC) AC	Simple	ley integral de emergencia de recursos para el SIDA
58	CBP	Elsevier	3	(AC) IN	Simple	proteína de unión a CREB
59	CDEs	Elsevier	4	(IN) IN	Simple	elementos conservados del ADN
60	CEA	Elsevier	3	(IN) AC	Simple	centro de astrofísica de UVE
61	CFAR	Elsevier	4	(AC) AC?	Simple	centro para la investigación sobre el SIDA
62	CIPs	Elsevier	4	(AC) AC	Simple	proteína de interacción CLOCK
63	CME	Elsevier	3	(IN) IN	Simple	elemento de la línea media del SNC
64	CME	Elsevier	3	(BL) IN	Simple	elemento multinacional del METOC
65	CMFU	Elsevier	4	(BL) IN	Hybrid	unidad multinacional de pronósticos del METOC

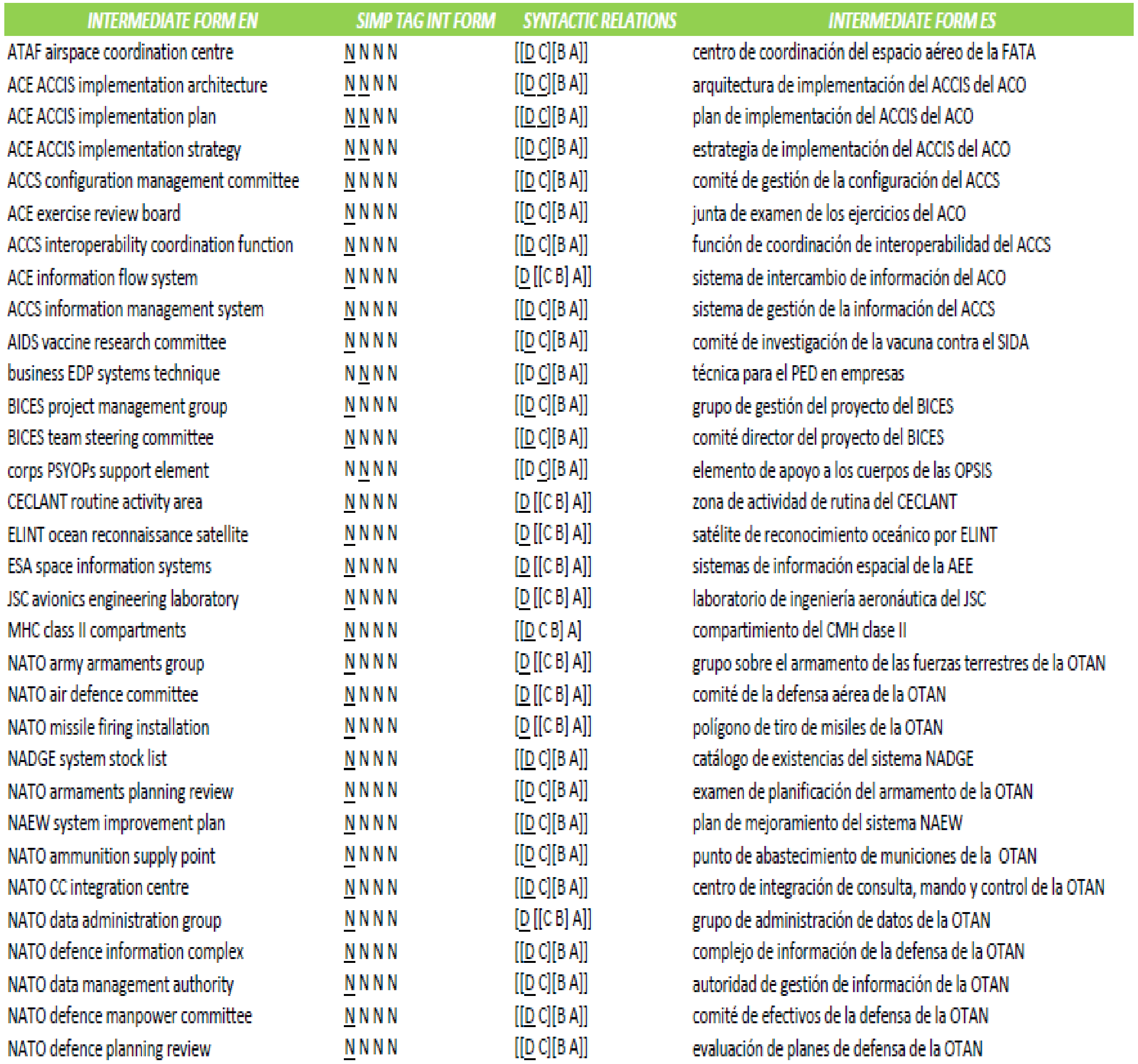


# TOKENS INT FORM	POS TAG	SIMP TAG INT FORM	SEMANTIC	FIELD
3	N N SYM N	N N SYM N	Conceptual	Molecular Biology
3	N Adj Prep Ar	N Adj Prep N	Nominal	Military
5	N Prep N Del I	N Prep N N Adj Prep N	Nominal	Military
5	N Adj Prep N /	N Adj Prep N Adj N	Nominal	Military
4	N Prep N Del I	N Prep N N N	Nominal	Military
4	N Prep N Del I	N Prep N N N	Nominal	Military
4	N Prep N Del I	N Prep N N N	Nominal	Military
4	N - N N	N - N N	Conceptual	Molecular Biology
3	N Prep N Prep	N Prep N Prep N	Conceptual	Molecular Biology
7	N Prep N Prep	N Prep N Prep N Prep N - I	Conceptual	Molecular Biology
4	N NUM Prep I	N NUM Prep N Prep N	Conceptual	Molecular Biology
5	N Prep N Prep	N Prep N Prep N Prep N N	Nominal	Military
4	N PP Prep N D	N PP Prep N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	Military
3	N Prep N N	N Prep N N	Conceptual	Molecular Biology
4	N Prep N Adj I	N Prep N Adj N	Nominal	AIDS organization
4	N Del N Conj \	N N Conj V N	Nominal	AIDS organization
4	N N Conj Adj F	N N Conj Adj Prep N	Conceptual	Molecular Biology
5	N Adj Prep N (	N Adj Prep N Conj N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	Military
3	N PP Prep Art	N PP Prep N	Nominal	Military
3	N Prep Art N I	N Prep N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	atmospheric sciences
3	N Prep N Del I	N Prep N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	Military
4	N Prep N Prep	N Prep N Prep N N	Nominal	Military
5	N Prep N Adj I	N Prep N Adj Prep N N	Nominal	Military
3	N Adj Del N	N Adj N	Nominal	Military
4	N Prep N Adj I	N Prep N Adj N	Nominal	Military
5	N Del N PP Pre	N N PP Prep N Adj	Nominal	AIDS study
4	N Prep N PP A	N Prep N PP N	Conceptual	AIDS organization

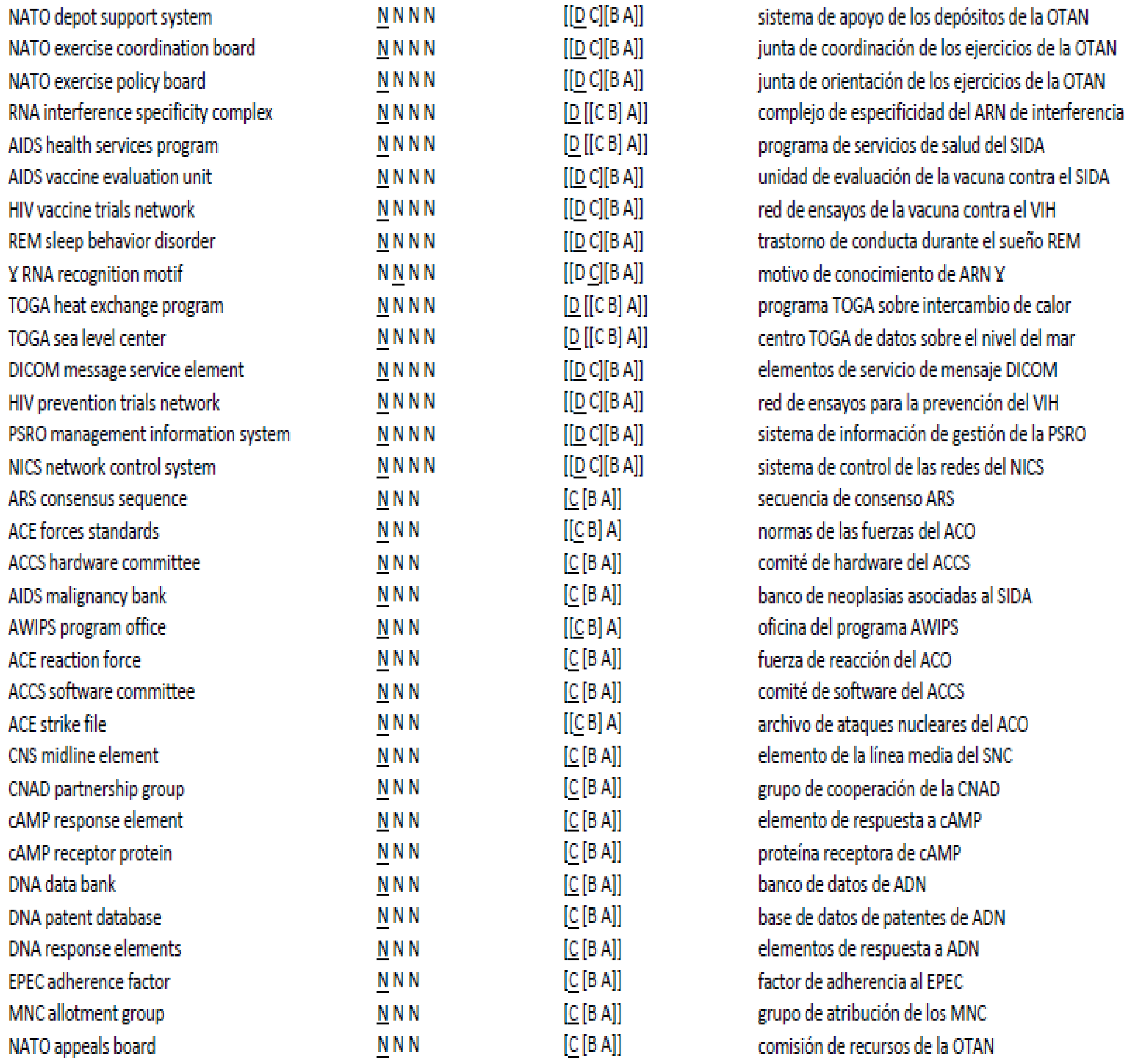


3	N Adj Del N	N Adj N	Nominal	Military
4	N Adj Prep Ar	N Adj Prep N Prep N	Nominal	AIDS organization
5	N , N , N , N A	N , N , N , N Adj	Nominal	Military
4	N Adj Prep Ar	N Adj Prep N Prep N	Nominal	AIDS organization
3	N Del N N	N N N	Nominal	atmospheric sciences
4	N Adj Prep N	N Adj Prep N Prep N	Nominal	AIDS organization
3	N Prep N Del	N Prep N N	Nominal	Military
3	N Adj Prep Ar	N Adj Prep N	Nominal	Politics
4	N Prep N Adj	N Prep N Adj N	Nominal	Military
4	N Prep N Adj	N Prep N Adj N	Nominal	Military
5	N Prep Art N	N Prep N PP Prep N SYM N	Nominal	AIDS organization
3	N Prep N Del	N Prep N N	Nominal	Military
4	N Prep N Adj	N Prep N Adj N	Nominal	Military
4	N Adj Adj Pre	N Adj Adj Prep N	Nominal	atmospheric sciences
4	N Prep N N PF	N Prep N N PP	Conceptual	Informatics
4	N Prep N Prep	N Prep N Prep N Prep N	Nominal	AIDS organization
6	N Adj Prep Ar	N Adj Prep N Prep N Prep	Nominal	Military
3	N Prep Art N	N Prep N Prep N	Conceptual	Informatics
4	N N Adj Del N	N N Adj N	Nominal	Military
4	N Prep N Del	N Prep N N N	Nominal	Military
4	N N Del N Del	N N N N	Nominal	Military
6	N Conj N PP P	N Conj N PP Prep N Prep N	Nominal	Military
3	N Adj Prep N	N Adj Prep N	Conceptual	Molecular Biology
4	N Prep N Prep	N Prep N Prep N N	Nominal	AIDS organization
5	N Adj Prep N	N Adj Prep N Prep N Prep	Nominal	AIDS law
3	N Prep N Prep	N Prep N Prep N	Conceptual	Molecular Biology
3	N PP Del N	N PP N	Conceptual	Molecular Biology
3	N Prep N Prep	N Prep N Prep N	Nominal	astronomy
3	N Prep Art N	N Prep N Prep N	Nominal	AIDS organization
3	N Prep N N	N Prep N N	Conceptual	Molecular Biology
4	N Prep Art N	N Prep N Adj N	Conceptual	Molecular Biology
3	N Adj Del N	N Adj N	Nominal	atmospheric sciences
4	N Adj Prep N	N Adj Prep N N	Nominal	atmospheric sciences

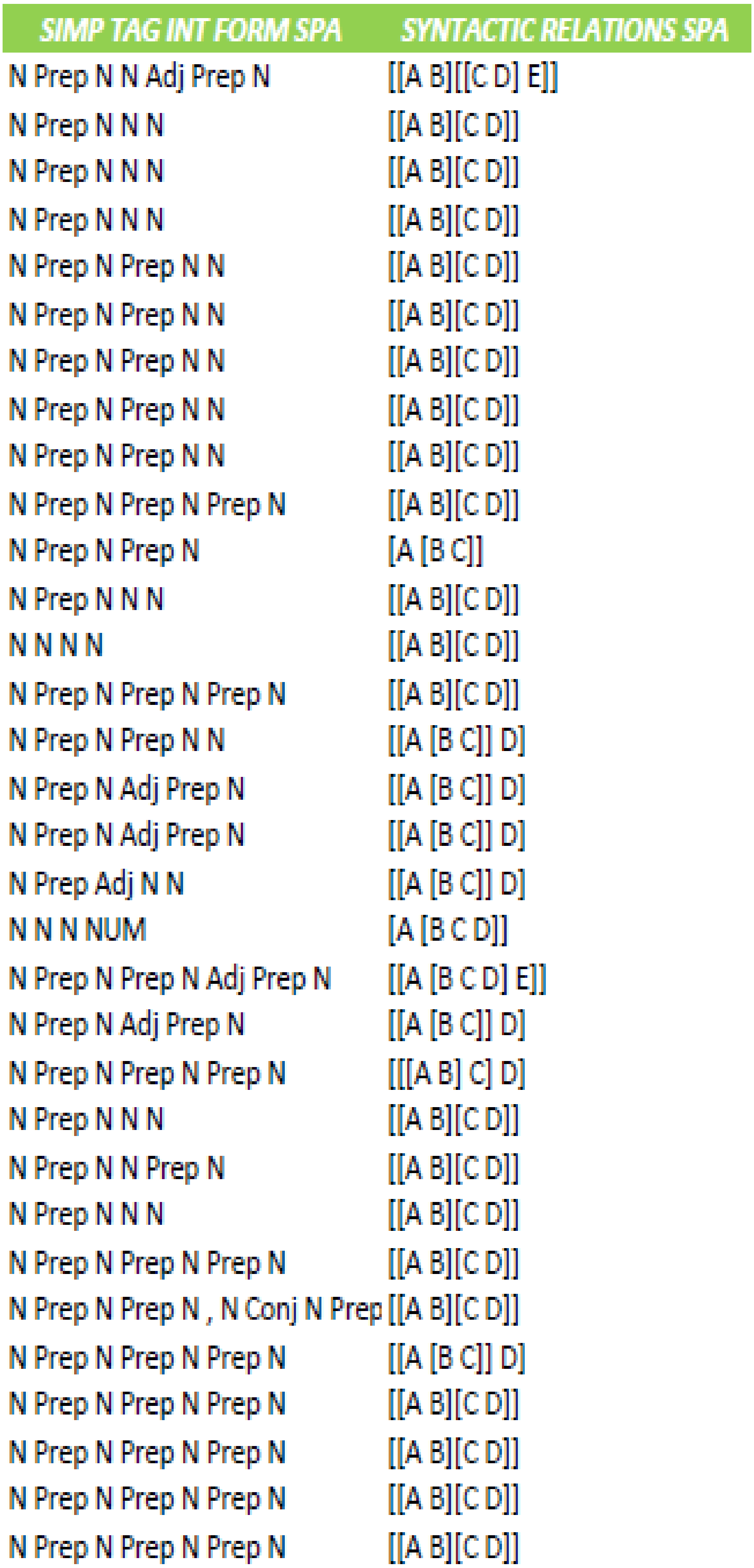
A.3 Extract of the *analysis corpus*



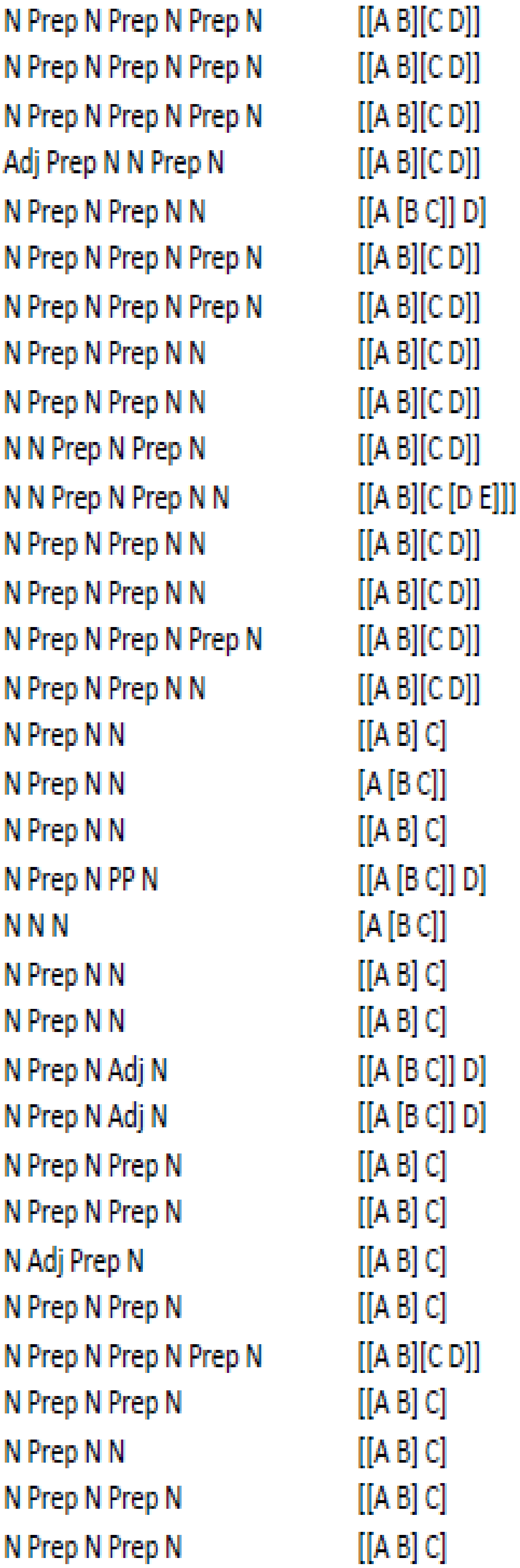
INTERMEDIATE FORM EN	SIMP TAG INT FORM	SYNTACTIC RELATIONS	INTERMEDIATE FORM ES
ATAF airspace coordination centre	N N N N	[[D C][B A]]	centro de coordinación del espacio aéreo de la FATA
ACE ACCIS implementation architecture	N N N N	[[D C][B A]]	arquitectura de implementación del ACCIS del ACO
ACE ACCIS implementation plan	N N N N	[[D C][B A]]	plan de implementación del ACCIS del ACO
ACE ACCIS implementation strategy	N N N N	[[D C][B A]]	estrategia de implementación del ACCIS del ACO
ACCS configuration management committee	N N N N	[[D C][B A]]	comité de gestión de la configuración del ACCS
ACE exercise review board	N N N N	[[D C][B A]]	junta de examen de los ejercicios del ACO
ACCS interoperability coordination function	N N N N	[[D C][B A]]	función de coordinación de interoperabilidad del ACCS
ACE information flow system	N N N N	[D [[C B] A]]	sistema de intercambio de información del ACO
ACCS information management system	N N N N	[[D C][B A]]	sistema de gestión de la información del ACCS
AIDS vaccine research committee	N N N N	[[D C][B A]]	comité de investigación de la vacuna contra el SIDA
business EDP systems technique	N N N N	[[D C][B A]]	técnica para el PED en empresas
BICES project management group	N N N N	[[D C][B A]]	grupo de gestión del proyecto del BICES
BICES team steering committee	N N N N	[[D C][B A]]	comité director del proyecto del BICES
corps PSYOPs support element	N N N N	[[D C][B A]]	elemento de apoyo a los cuerpos de las OPSIS
CECLANT routine activity area	N N N N	[D [[C B] A]]	zona de actividad de rutina del CECLANT
ELINT ocean reconnaissance satellite	N N N N	[D [[C B] A]]	satélite de reconocimiento oceánico por ELINT
ESA space information systems	N N N N	[D [[C B] A]]	sistemas de información espacial de la AEE
JSC avionics engineering laboratory	N N N N	[D [[C B] A]]	laboratorio de ingeniería aeronáutica del JSC
MHC class II compartments	N N N N	[[D C B] A]	compartimiento del CMH clase II
NATO army armaments group	N N N N	[D [[C B] A]]	grupo sobre el armamento de las fuerzas terrestres de la OTAN
NATO air defence committee	N N N N	[D [[C B] A]]	comité de la defensa aérea de la OTAN
NATO missile firing installation	N N N N	[D [[C B] A]]	polígono de tiro de misiles de la OTAN
NADGE system stock list	N N N N	[[D C][B A]]	catálogo de existencias del sistema NADGE
NATO armaments planning review	N N N N	[[D C][B A]]	examen de planificación del armamento de la OTAN
NAEW system improvement plan	N N N N	[[D C][B A]]	plan de mejoramiento del sistema NAEW
NATO ammunition supply point	N N N N	[[D C][B A]]	punto de abastecimiento de municiones de la OTAN
NATO CC integration centre	N N N N	[[D C][B A]]	centro de integración de consulta, mando y control de la OTAN
NATO data administration group	N N N N	[D [[C B] A]]	grupo de administración de datos de la OTAN
NATO defence information complex	N N N N	[[D C][B A]]	complejo de información de la defensa de la OTAN
NATO data management authority	N N N N	[[D C][B A]]	autoridad de gestión de información de la OTAN
NATO defence manpower committee	N N N N	[[D C][B A]]	comité de efectivos de la defensa de la OTAN
NATO defence planning review	N N N N	[[D C][B A]]	evaluación de planes de defensa de la OTAN



NATO depot support system	N N N N	[[D C][B A]]	sistema de apoyo de los depósitos de la OTAN
NATO exercise coordination board	N N N N	[[D C][B A]]	junta de coordinación de los ejercicios de la OTAN
NATO exercise policy board	N N N N	[[D C][B A]]	junta de orientación de los ejercicios de la OTAN
RNA interference specificity complex	N N N N	[D [[C B] A]]	complejo de especificidad del ARN de interferencia
AIDS health services program	N N N N	[D [[C B] A]]	programa de servicios de salud del SIDA
AIDS vaccine evaluation unit	N N N N	[[D C][B A]]	unidad de evaluación de la vacuna contra el SIDA
HIV vaccine trials network	N N N N	[[D C][B A]]	red de ensayos de la vacuna contra el VIH
REM sleep behavior disorder	N N N N	[[D C][B A]]	trastorno de conducta durante el sueño REM
Ɣ RNA recognition motif	N N N N	[[D C][B A]]	motivo de conocimiento de ARN Ɣ
TOGA heat exchange program	N N N N	[D [[C B] A]]	programa TOGA sobre intercambio de calor
TOGA sea level center	N N N N	[D [[C B] A]]	centro TOGA de datos sobre el nivel del mar
DICOM message service element	N N N N	[[D C][B A]]	elementos de servicio de mensaje DICOM
HIV prevention trials network	N N N N	[[D C][B A]]	red de ensayos para la prevención del VIH
PSRO management information system	N N N N	[[D C][B A]]	sistema de información de gestión de la PSRO
NICS network control system	N N N N	[[D C][B A]]	sistema de control de las redes del NICS
ARS consensus sequence	N N N	[C [B A]]	secuencia de consenso ARS
ACE forces standards	N N N	[[C B] A]	normas de las fuerzas del ACO
ACCS hardware committee	N N N	[C [B A]]	comité de hardware del ACCS
AIDS malignancy bank	N N N	[C [B A]]	banco de neoplasias asociadas al SIDA
AWIPS program office	N N N	[[C B] A]	oficina del programa AWIPS
ACE reaction force	N N N	[C [B A]]	fuerza de reacción del ACO
ACCS software committee	N N N	[C [B A]]	comité de software del ACCS
ACE strike file	N N N	[[C B] A]	archivo de ataques nucleares del ACO
CNS midline element	N N N	[C [B A]]	elemento de la línea media del SNC
CNAD partnership group	N N N	[C [B A]]	grupo de cooperación de la CNAD
cAMP response element	N N N	[C [B A]]	elemento de respuesta a cAMP
cAMP receptor protein	N N N	[C [B A]]	proteína receptora de cAMP
DNA data bank	N N N	[C [B A]]	banco de datos de ADN
DNA patent database	N N N	[C [B A]]	base de datos de patentes de ADN
DNA response elements	N N N	[C [B A]]	elementos de respuesta a ADN
EPEC adherence factor	N N N	[C [B A]]	factor de adherencia al EPEC
MNC allotment group	N N N	[C [B A]]	grupo de atribución de los MNC
NATO appeals board	N N N	[C [B A]]	comisión de recursos de la OTAN



SIMP TAG INT FORM SPA	SYNTACTIC RELATIONS SPA
N Prep N N Adj Prep N	[[A B][[C D] E]]
N Prep N N N	[[A B][C D]]
N Prep N N N	[[A B][C D]]
N Prep N N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N	[A [B C]]
N Prep N N N	[[A B][C D]]
N N N N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N N	[[A [B C]] D]
N Prep N Adj Prep N	[[A [B C]] D]
N Prep N Adj Prep N	[[A [B C]] D]
N Prep Adj N N	[[A [B C]] D]
N N N NUM	[A [B C D]]
N Prep N Prep N Adj Prep N	[[A [B C D] E]]
N Prep N Adj Prep N	[[A [B C]] D]
N Prep N Prep N Prep N	[[[A B] C] D]
N Prep N N N	[[A B][C D]]
N Prep N N Prep N	[[A B][C D]]
N Prep N N N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N , N Conj N Prep	[[A B][C D]]
N Prep N Prep N Prep N	[[A [B C]] D]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]



N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
Adj Prep N N Prep N	[[A B][C D]]
N Prep N Prep N N	[[A [B C]] D]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N N Prep N Prep N	[[A B][C D]]
N N Prep N Prep N N	[[A B][C [D E]]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N N	[[A B][C D]]
N Prep N N	[[A B] C]
N Prep N N	[A [B C]]
N Prep N N	[[A B] C]
N Prep N PP N	[[A [B C]] D]
N N N	[A [B C]]
N Prep N N	[[A B] C]
N Prep N N	[[A B] C]
N Prep N Adj N	[[A [B C]] D]
N Prep N Adj N	[[A [B C]] D]
N Prep N Prep N	[[A B] C]
N Prep N Prep N	[[A B] C]
N Adj Prep N	[[A B] C]
N Prep N Prep N	[[A B] C]
N Prep N Prep N Prep N	[[A B][C D]]
N Prep N Prep N	[[A B] C]
N Prep N N	[[A B] C]
N Prep N Prep N	[[A B] C]
N Prep N Prep N	[[A B] C]