Description and representation in language resources of Spanish and English specialized collocations from Free Trade Agreements

Pedro Patiño García

Dissertation for the degree philosophiae doctor (PhD) at

NHH Norwegian School of Economics

Department of Professional and Intercultural Communication

Supervisors:

Prof. Dr. Gisle Andersen

Assoc. Prof. Dr. Marita Kristiansen

Prof. Dr. Koenraad De Smedt

Bergen / Medellín

September, 2016

Contents

List of Figures

List of Tables

Dedicatoria

Dedico este trabajo a mis padres, Juanita y Pedro, a quienes amo profundamente y este logro también es de ustedes pues han hecho posible que yo llegue hasta donde estoy ahora con todo el esfuerzo y sacrificio que han hecho desde siempre, trabajando desde su infancia en las montañas antioqueñas y luego en la ciudad "capital de la montaña" para que su hijo llegara muy lejos. A ustedes me debo y a ustedes me doy pues cada pequeño triunfo mío es de ustedes también.

¡Gracias!

"Derrotado por aquellas prácticas de consolación, José Arcadio Buendía decidió entonces construir la máquina de la memoria que una vez había deseado para acordarse de los maravillosos inventos de los gitanos. El artefacto se fundaba en la posibilidad de repasar todas las mañanas, y desde el principio hasta el fin, la totalidad de los conocimientos adquiridos en la vida. Lo imaginaba como un diccionario giratorio que un individuo situado en el eje pudiera operar mediante una manivela, de modo que en pocas horas pasaran frente a sus ojos las nociones más necesarias para vivir."

"Cien años de soledad", Gabriel García Márquez, 1967.

"Defeated by those practices of consolation, José Arcadio Buendía then decided to build the memory machine that he had desired once in order to remember the marvelous inventions of the gypsies. The artifact was based on the possibility of reviewing every morning, from beginning to end, the totality of knowledge acquired during one's life. He conceived of it as a spinning dictionary that a person placed on the axis could operate by means of a lever, so that in a very few hours there would pass before his eyes the notions most necessary for life."

"One Hundred Years of Solitude", Gabriel García Márquez, 1967.

Acknowledgements

The funding for this research was made possible thanks to the European Union's 7th Framework Program under the Marie Curie Initial Training Network No. 238405, project CLARA, Common Language Resources and their Applications, and the Department of Professional and Intercultural Communication, NHH Norwegian School of Economics, Bergen, Norway. Later, I received funding from the *Vicerrectoría de Docencia*, Universidad de Antioquia, Medellín, Colombia.

I would like to thank my advisors Gisle Andersen, Marita Kristiansen and Koenraad DeSmedt, who have been a great academic support and a helping hand throughout the years and since the very first day of this process. Your patience, thought-provoking and insightful comments and positive criticism have helped me a lot to clarify the scope of my project and to pursue a career as an *early stage researcher*. Thanks are due to Prof. Dr. Rita Temmerman from Erasmushogeschool Brussels for her valuable insights in the very beginning of this journey.

Also, thanks to all the colleagues at FSK, *Institutt for fagspråk og interkulturell kommunikasjon*, both the Faculty and the (past and present) fellow PhD students. Here, special thanks are due to Tove Skaar and Sally Tveit from the administrative staff for their diligence and willingness to assist me at the right moment. From the very beginning, back in April 29, 2010 when I arrived at the Department, every one at the institute made me feel

comfortable even though I was in a foreign country far away from home.

To my dear colleagues at the Escuela de Idiomas, Universidad de Antioquia, Medellín, Colombia, who provided insights, scripts, dictionaries, glossaries, criticism or valuable suggestions to improve my work: especially Gabriel Quiroz and Diego Burgos, John Jairo Giraldo, Ramiro Restrepo, Germán Mira, José Luis Rojas and Gustavo Zapata. Gabriel read early drafts of my work and offered great insights to improve it. Diego was willing to discuss some conceptual and technical aspects and kindly provided with me with useful scripts he had developed. They all gave me useful comments, suggestions, coffee and beer.

To my friends, who live in different parts of the world, among them Alejandro Uribe, Diana Giraldo, Bernardo Vélez, Silvia Flórez, Mauricio Cardona, and many other friends I shall not mention here.

A todos mis familiares y amigos en Medellín y otros sitios de Colombia, España y Noruega, que de diferentes maneras me apoyaron durante este proceso y mejor no enumero los nombres porque corro el riesgo de dejar muchos sin mencionar.

To other researchers who with their skills and insight greatly contributed to improve my work: Knut Hofland, Michael Oakes, Julián Cárdenas, Alejandro Uribe, Antonio Tamayo and Andrés Quintero.

Thanks to the researchers and enthusiastic developers who give away a lot of their time and knowledge and offer for free a plethora of open source tools I could use in this project.

Thanks to God.

Abstract

Description and representation in language resources of
Spanish and English specialized collocations
from Free Trade Agreements

Pedro Patiño García
pedro.patino@nhh.no - nel.patino@udea.edu.co

NHH Norwegian School of Economics
Department of Professional and Intercultural Communication
Bergen, Norway / Medellín, Colombia
September, 2016

Abstract

This thesis is concerned with specialized collocations, defined as a type of multiword expression composed of a term that serves as the node of the collocation. The collocates can be nouns, verbs, adjectives or adverbs in a direct syntactic relation with the node. These constituents make a lexical combination that can be unpredictable and semi-compositional and have an internal and statistical tendency of preference. The data was drawn from a

parallel corpus of English and Spanish texts taken from 16 official texts of Free Trade Agreements, hereinafter FTA. The present work offers a description and classification of English and Spanish specialized collocations from Free Trade Agreements that appear in the parallel corpus data. Besides, a proposal is presented for the computational representation of specialized collocations in schemes for linguistic annotation of terminological and computational lexicons. This proposal involves the use of annotations that can be used for encoding linguistic information for collocation information, such as the part of speech, the subject field to which these lexical units belong and morphosyntactic and semantic information. These schemes have been issued by standardization bodies such as the International Organization for Standardization. Specifically, the Terminological Markup Framework (TMF) ISO 16642:2003, TermBase eXchange (TBX) ISO 30042:2008, and Lexical Markup Framework (LMF) ISO 24613:2008.

Resumen

Esta tesis se trata de las colocaciones especializadas, definidas como un tipo de expresión poliléxica compuesta por un término que sirve como el nodo de la colocación. Los colocativos pueden ser nombres, verbos, adjetivos y adverbios en una relación sintáctica directa con el nodo. Estos constituyentes crean una combinación léxica que puede ser impredecible y semicomposicional y tienen una tendencia de preferencia estadística e interna. Los datos se obtuvieron de un corpus paralelo de textos en inglés y español extraídos de 16 textos oficiales de Tratados de Libre Comercio, en lo sucesivo llamados FTA (por sus siglas en inglés). Este trabajo ofrece una descripción y una clasificación de las colocaciones especializadas en inglés y en español de Tratados de Libre Comercio que aparecen en los datos del corpus paralelo.

Además, se presenta una propuesta para la representación computacional de las colocaciones especializadas en los esquemas existentes para la anotación lingüística de los lexicones terminológicos y computacionales. Esta propuesta involucra el uso de anotaciones que se pueden emplear para codificar los metadatos para la información colocacional, tales como la categoría gramatical, el área temática a la que estas unidades léxicas pertenecen

y la información morfosintáctica y semántica. Estos esquemas han sido emitidos por entidades normalizadoras tales como la Organización Internacional para la Estandarización. Específicamente, las normas *Terminological Markup Framework* (TMF) ISO 16642:2003, *TermBase eXchange* (TBX) ISO 30042:2008, y *Lexical Markup Framework* (LMF) ISO 24613:2008.

Keywords

specialized collocation, free trade agreement, phraseology, terminology, natural language processing, corpus linguistics, language for special purposes, harmonization of terminological resources.

Palabras clave

colocación especializada, tratado de libre comercio, fraseología, terminología, procesamiento del lenguaje natural, lingüística de corpus, lenguas para fines específicos, armonización de recursos terminológicos.

Foreword

0.1 Motivation for this study

Constantly, translators have to face the challenge of finding the right equivalent for the collocations that appear in their translation tasks. From my experience as a professional translator of scientific, technical and legal texts, working for more than a decade as a freelance translator and also with a group of colleagues, including sworn translators, I have developed an interest in the topic of how to handle the type of collocations that appear in specialized texts. The same challenge of dealing with the translation of collocations arises while teaching translation students how to identify and find the equivalent for collocations, both with general texts and with domain-specific texts. I also developed an interest in the topic of collocations that appear in specialized texts while working as a lecturer of the subjects Scientific and Technical Translation and the Translation Practicum. The latter subject is aimed at advanced students of the Bachelor of Arts in English-French-Spanish Translation taught in the School of Modern Languages, University of Antioquia, located in Medellín, Colombia.

The present work builds on previous work done during my master studies, which focused on the semi-automatic extraction of specialized idioms found in the Spanish subcorpus of economics developed at the University Institute

of Applied Linguistics (IULA), Pompeu Fabra University, Barcelona, Spain. In such a work, a set of morphosyntactic patterns that, according to the literature, form specialized Spanish idioms, was used as a starting point for the extraction of these lexical units, as discussed in Patiño (2010). My master's thesis also looked into the formalized representation of these idioms using the Lexical Markup Framework (LMF) for the constitution of interoperable language resources such as computational lexicons.

The present work is part of a PhD research project affiliated to the EU-funded project CLARA, Common Language Resources and their Applications, under the subproject Harmonization of Terminological Resources[1]. This European project is aimed at establishing a common set of language resources and their harmonization.

[1] `http://clara.uib.no/`

CHAPTER 1

Introduction

> One way of describing collocation is to say that the choice of one word conditions the choice of the next, and of the next again. (Sinclair et al., 1970, 19).

1.1 Introduction

The present work investigates the specialized lexical combinations that include a term and that appear in a specific kind of specialized texts from the field of international trade, namely, Free Trade Agreements (henceforth, FTAs).

The tendency of words to co-occur with a set of other words to form lexical combinations has been named collocation. It is a relevant and pervasive feature of all natural languages. In this thesis, collocations are understood as a subset of multiword expressions (henceforth MWEs), in harmony with Manning and Schütze (1999); Evert (2009); Baldwin and Kim (2010); Seretan (2011) and Seretan (2013). Baldwin and Kim (2010, 274) assert that "collocations form a proper subset of MWEs". The notion of MWE is defined in Section 2.8.1.

The phenomenon of collocation has been noted by many researchers for decades, who have studied that particularity of words both at the lexical and at the grammatical levels (Palmer and Hornby, 1933; Firth, 1957). Sinclair et al. (1970) named "collocability" the tendency of a lexical unit to be conditioned to combine with other words. In virtue of this tendency, both in spoken and written language, words are not combined randomly but are ruled by some patterns and preferences for their felicitous combination, as expressed in the quote at the beginning of the chapter from Sinclair et al. (1970).

To illustrate the phenomenon of collocation, I will take as an example the adjective "sharp". This adjective is defined in the first sense offered in the online version of the Free Merriam-Webster Dictionary as "adapted to cutting or piercing".[2] In the Corpus of Contemporary American English or COCA corpus (Davies, 2009), which as of July 2016 contains 520 million words from texts written from 1990 to 2015, the adjective "sharp" frequently collocates with the nouns *contrast, distinction*, and *knife*. In these examples, the adjective *sharp* adds something to the meaning of the noun which is not the same in the case of *knife* as compared to the other two nouns. The COCA corpus offers 541 instances of the collocation *sharp knife*. However, the adjective *trenchant*, which is synonymous with *sharp* does not appear at all, thus *trenchant knife* does not form a collocation. In contrast, the adjective *trenchant* collocates with the nouns *analysis, criticism* and *observation*. This suggests that, to gain specific lexical knowledge of a certain word, it is necessary to know which words accompany it and not only to know the word in isolation. Much research into this phenomenon of lexical units has been carried out for several decades, especially within general purpose texts. That is why Mel'čuk (1998, 24) says that "the literature on collocations is simply overwhelming". For example, the online bibliography database "Collocations and Idioms: An International Bibliography" presents a directory with more than 4,400 publications on the topic of collocations.[3]

One way of describing phraseology is found in the words of Gledhill (2000, 1). He defines it as "the preferred way of saying things in a particular dis-

[2] `http://www.merriam-webster.com/dictionary/sharp`
[3] `http://kollokationen.bbaw.de/bib/index_en.html`

course". From a broad conception of phraseology, one that includes both idioms and collocations, the latter are considered the most frequent subset among the lexical units that conform phraseology. This view is supported by Mel'čuk (1998, 24) who claims that "in any language i.e. in its lexicon, phrasemes outnumber words roughly ten to one. Collocations make up the lion's share of the phraseme inventory". Therefore, collocations are indeed relevant lexical units that merit being described and studied to gain specific knowledge on the vocabulary of any language.

1.2 A scientific map of specialized phraseology

In the present work, an interdisciplinary approach is assumed to address the study of specialized phraseology, specifically the lexical collocations that include a term and which appear in a specific domain. As a point of departure for the present research, several theoretical and practical notions, principles and procedures are taken from various subdisciplines pertaining to the field of linguistics, specifically from terminology, phraseology, corpus linguistics, lexicography and natural language processing (NLP). Besides linguistics, these subdisciplines base their founding principles on other sciences, including philosophy and information science in the case of terminology. In the case of NLP, its scientific background comes from the disciplines of computational linguistics, computer science and artificial intelligence. None of these disciplines can claim exclusive property over these notions but are rather used in several of these disciplines. The notions of *concept* and *term* are taken from the field of terminology. Terms provide valuable information about the salient concepts within a specific domain and are therefore crucial to determine a "domain-specificity". The particularity of a given set of words as being pertinent and salient in a determined subject field is what is meant in the present work by domain-specificity.

Since ancient times, the concept of *concept* has been an important issue to study within philosophy, as evidenced in Cratylus, a dialogue by Plato written approximately in 360 B.C.E. In this dialogue, the ancient Greek

philosopher discusses the nature of names and their relation to the things they designate. The notions of *collocation* and *idiom* are adopted from the linguistic subdisciplines of phraseology, corpus linguistics, terminography, lexicography and specialized lexicography. Collocations are important components for describing words besides terms, and occupy an important place in many dictionaries. They provide precise information about the words that co-occur in any given text.

Also from corpus linguistics, lexicography and terminography comes the method of working with *concordances* to analyze the linguistic behavior of words. There are different linguistic levels to perform this analysis in their context, especially syntax, morphology, semantics and pragmatics. Corpus linguistics advocates the use of examples derived from authentic data instead of merely relying on the linguist's intuition. In the case of lexicography, it is an applied subdiscipline of linguistics, related to lexicology and is concerned with making dictionaries for a variety of users and domains, besides general dictionaries.

NLP and other disciplines related to computational linguistics intend to develop methods and tools to allow and enhance the interaction between humans and between humans and computers, in an effort to overcome or at least reduce language barriers. These disciplines rely heavily on data and thus words and text are key components, like bricks and mortar necessary to build human language technologies.

This study stands in the arena of specialized phraseology, which some authors refer to as LSP phraseology, where LSP stands for Language for Special Purposes (Spang-Hanssen, 1983; Picht, 1987, 1990a; Budin, 1990; Thomas, 1993). In the case of corpus linguists and practitioners of natural language processing, terms are not the focus of their studies in the same sense as is done by terminologists. Thus, research that combines the disciplines mentioned above is not, using Gibbons et al. (1994) words, "located on the prevailing disciplinary map" of the terminological arena neither on that of corpus linguistics.

Figure 1.1 here, based on Kristiansen (2004, 35), illustrates the scientific map of specialized phraseology. In the figure, the arrows indicate linguistics subdisciplines and other disciplines as well, that provide specialized phrase-

ology with a theoretical and practical framework and which more directly contribute to the present work. The fields appearing on top outside the gray box provide a scientific basis for natural language processing, while philosophy is related to linguistics. Linguistics is the broad field in which specialized phraseology is grounded and its theoretical and practical frameworks stem from several linguistics subdisciplines, included inside the box, which contribute to delineate the scientific frontiers of specialized phraseology. Within specialized phraseology, the present thesis is focused on specialized collocations. The specialized features of this type of phraseology is discussed in section 2.12.

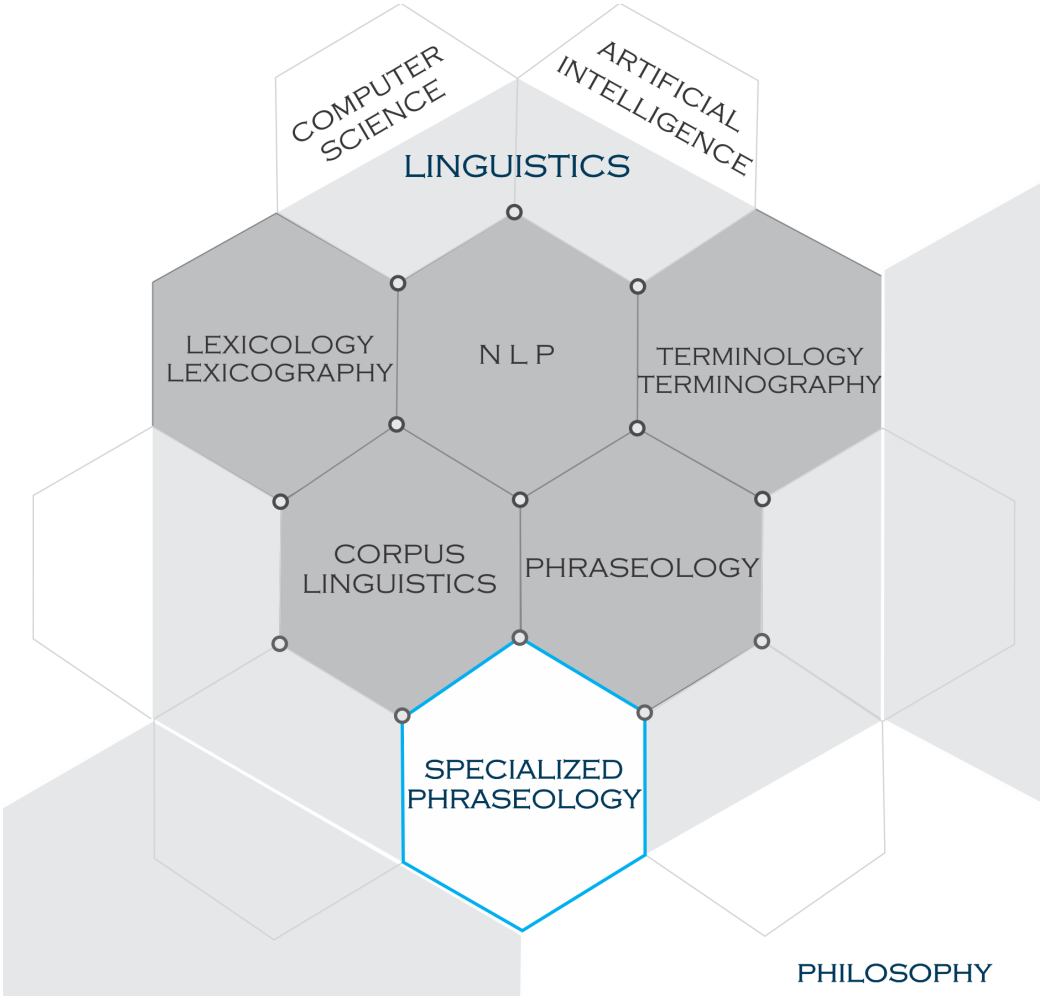

Figure 1.1: *A model of the scientific map of specialized phraseology*

In the following subsections I present the hypotheses and the objectives of this thesis.

1.3 Hypotheses

a) Specialized collocations contribute to delineating domain-specificity in a similar way as do the terms used in such a domain. Therefore, specialized collocations are part of specialized language.

b) Specialized collocations may be unpredictable and require idiomatic specialist knowledge.

c) The attribute of domain-specificity of specialized collocations is activated by some linguistic features of their constituents. The identification of these features can be useful to further describe the domain-specificity of phraseological units and also to represent specialized collocations for the creation of language resources.

1.4 Objectives

This thesis has a theoretical and an applied objective and some specific objectives:

- Theoretical objective: To determine how specialized collocations contribute to delineate the domain-specificity of English and Spanish texts that constitute Free Trade Agreements (FTAs).
- Applied objective: To assess the applicability of linguistic annotation schemes for the representation of specialized collocations in term bases and computational lexicons.

1.4.1 Specific objectives

To attain the theoretical and applied objectives, the following specific objectives are proposed:

- To perform a linguistic classification, description and comparison of FTA specialized collocations that appear in a corpus of English and Spanish from American and European FTA texts.
- To compare the characteristics of specialized collocations found in FTA texts with general and specialized English and Spanish corpora, dictionaries and term bases.

1.5 Thesis outline

In Chapter 2, I will introduce the notion of specialized collocation, the main features that distinguish collocations from other types of multiword expressions (MWEs) and a view on the concept of collocations from the perspective of several disciplines. Thus, Chapter 2 defines the object of study, namely, that of specialized collocation.

Following this, Chapter 3 describes the data that is included in the corpus, namely, supranational agreements. It also presents the countries and institutions involved in promoting free trade. Chapter 4 describes the compilation, preparation and processing of the data to constitute the FTA corpus. It also offers a description of the material and the method used to carry out the study.

Then, Chapter 5 presents the results and the analysis of the specialized collocations extracted from the corpus, which was carried out by using several tools and a combination of corpus-based and corpus-driven techniques. Next, Chapter 6 presents the proposal to represent lexical units such as specialized collocations in language resources such as computational lexicons. The aim of the proposal is to be able to process the data in such a way that it can be interchangeable, reusable and interoperable.

Finally, the conclusions of the study, its limitations and a perspective for future work are presented in Chapter 7.

CHAPTER 2

Theoretical foundations

> Collocations, even in specialized domains, are unpredictable combinations and should be described in dictionaries (L'Homme, 2006, 186).

2.1 Introduction

In this chapter, I introduce several theoretical notions which are central to this study, such as the concept of term, automatic term and collocation extraction, language resources and their standardization. Additionally, the differences among several types of MWEs are presented. Besides, I present several definitions of collocation according to representative authors from the field, with the aim of arriving at a definition of what constitutes a specialized collocation. To do this, it is important to adopt a definition of the notions of term and collocation and the features of both types of lexical units.

This chapter is aimed at approaching the study of the collocations that appear in specialized texts from the subject field of international trade, more specifically, in legal and economics texts written in English and Spanish and taken from official FTAs. The method used for the study of these specialized collocations is an interdisciplinary approach and it will be fully accounted

for in Chapter 4.

2.2 The concept of term

The field of terminology is concerned with concepts and these concepts are expressed linguistically by means of terms, which are carriers of specialized information in texts. In the literature there are many definitions of term, such as the following, to cite but a few.

In the International Standard ISO 704 Terminology work, Principles and methods, the International Organization for Standardization, ISO (2009, 34) offers this definition of term:

> A term is a designation consisting of one or more words representing a general concept in a special language in a specific subject field. A simple term contains only one root, while a term containing two or more roots is called a complex term.

This definition includes units which refer to concepts in a specific subject field and which are composed by one or more lexemes.

Gouadec (1990) offers another definition of term:

> Un terme est une unité linguistique désignant un concept, un objet ou un processus. Le terme est l'unité de désignation d'éléments de l'univers perçu ou conçu. Il ne se confond que rarement avec le mot orthographique.[4]

While Gouadec's definition emphasizes the cognitive attributes of terms, it is less linguistics-centered because it does not specify whether terms are composed by one or more lexemes. Also, in this definition an object or a process is subsumed by a concept.

For the purposes of this research, I adopt the definition of term presented by Lerat (1989):

[4] My translation: A term is a linguistic unit that designates a concept, an object or a process. The term is the unit to designate elements of the perceived or known universe. It is only rarely confused with the orthographical word.

> Une unité terminologique, ou terme, est un symbole conventionnel représentant une notion définie dans un certain domaine du savoir.[5]

This definition is pertinent for the current purposes because it associates a term, or terminological unit, to a specific domain. Besides, this definition includes the notion of terminological unit, which encompasses terms constituted by one or more lexemes.

At this point, a terminological clarification seems pertinent. Throughout the thesis, the terms *term* and *terminological unit* will be used interchangeably.

According to Cabré (1999), some of the features of terms used in specialized subject fields are:

- *Conciseness*: Terms are used as an attempt to avoid redundancy.
- *Preference for nominalization*: Nouns are preferred to express concepts over other lexical categories.
- *Impersonalization*: Terms are not emotive and the emphasis is set on the ideas and not on the source.

Similarly, Gotti (2003) describes the lexical features of specialized discourse, among them, the following:

- *Monoreferentiality*: Only one meaning is allowed.
- *Lack of emotion*: Terms have a purely denotative function.
- *Precision*: Every term points to its own concept.
- *Transparency*: The meaning of a term is accessed through its surface form.
- *Conciseness*: Concepts are expressed in the shortest possible form, including acronyms and abbreviations.
- *Conservatism*: Some concepts are expressed by means of classical languages and archaic formulae, which reinforces monoreferentiality.
- *Lexical productivity*: Some terms from a specialized setting are gradually adopted in everyday language.

[5] A terminological unit, or a term, is a conventional symbol that represents a concept defined within a particular field of knowledge. Translation from French by Cabré (1999).

Thus, specialized communication exhibits certain features that set it apart from general purpose communication. In specialized texts, terms contribute to the expression and transmission of these features, which enables domain specialists and terminologists to identify them in a specific domain associated with a set of terms, what in this thesis is named domain-specificity.

2.3 Term and collocation extraction

Manual acquisition of terms and their collocates from running text is not a trivial task. It is a slow process, it is time-consuming and prone to errors. Due to this, considerable research efforts have focused on the task of semi- or automatic candidate term extraction, which is called "automatic term extraction" (ATE) or "automatic term recognition" (ATR) (Foo, 2011).

Researchers in the field of NLP and related disciplines have explored different approaches and techniques to extract terms and collocations from corpora. They have implemented the use of statistical techniques along with the method of using linguistic knowledge in the form of morphosyntactic patterns. This has been done with the aim of performing this extraction in a more systematic and comprehensive manner, with varying results. ATE can be useful to disambiguate the sense of words, to identify the domain automatically and to improve systems of machine translation, among other applications.

In addition to ATE, recent research focuses on the fields of semi-automatic MWE (Ramisch, 2015) and collocation extraction (Seretan, 2011). As stressed by Seretan (2011, 2):

> As the compilation of such resources is increasingly corpus-based, automatic collocation extraction methods are being heavily used in many lexicographic projects for collecting the raw material to include in dictionaries, for validating the intuition of lexicographers, and for complementing collocation entries with additional corpus-based information such as frequency of use or usage samples.

The same can be said of the semi-automatic extraction of specialized collocations by means of electronic corpora as a means to gather the raw material

that later can be used for several language-related applications. Among one of the earliest approaches to identify collocations, the one employed by Sinclair et al. (1970) is based on studying each node word in a concordance and then manually scanning the text with a vertical view to identify significant collocates. By using a corpus, the researcher easily obtains a concordance of a given lexeme, what is also known as key word in context (KWIC). Subsequently, a careful vertical reading of the concordance reveals the words typically surrounding a particular lexeme and the collocational patterns involved in this occurrence in relation to other lexemes.

Oakes (1998, 149) remarks that collocations "can be extracted using purely syntactic criteria [...] by observing regular syntactic patterns which are known to be typical of idiomatic collocations or technical terms".

Other authors apply similar approaches to extract collocations (Seretan, 2011), complex specialized noun phrases (Quiroz, 2008) and simple or complex terms (Estopà, 1999; Burgos, 2014). Drouin (1997, 2004) describes two ways to identify terms: corpora comparison and comparison against non-terms as a way to detect features that can help to identify true terms. These approaches of using syntactic criteria besides concordance views to identify the collocates of a given term are also relevant for the acquisition of specialized collocations. The use of several corpora to compare with the FTA corpus by means of software as a means to identify relevant terms and specialized collocations occurring in the FTA corpus is described in Section 4.4.2.

2.3.1 Statistical measures used for collocation research

Researchers have also employed association measures (AMs) as a quantitative means to calculate collocation strength. An association measure is defined by Evert (2005) as a "formula that computes an association score from the frequency information in a pair type's contingency table. This score is intended as an indicator of how strong the association between the pair's components is, correcting for random effects." The logic behind the use of these AMs is the intention of answering a question: "to what extent do the occurrences of a word *w1* determine the occurrences of another word *w2*?" (Evert, 2009). Pecina and Schlesinger (2006) report that around 80 AMs

have been employed to extract collocations from text based on statistical information, with each AM having variable success or popularity over a period of time (Oakes, 1998; Evert, 2004). Each one of these AMs offers different results and the measures themselves are not comparable across them (Lyse and Andersen, 2012).

Krishnamurthy (2006) mentions two popular AMs that have been used for collocation extraction, namely mutual information (MI) score and t-score. Krishnamurthy compares these two AMs and concludes that "MI-score privileges lower-frequency, high-attraction collocates (e.g., dentist with hygienist, optician, and molar) while t-score favors higher-frequency collocates (e.g., dentist with chair), including significant grammatical words (e.g., dentist with a, and your)."

Besides using a concordance, other researchers have subsequently adopted a different approach, and implemented the use of NLP applications along with statistical AMs, combined with linguistic knowledge to extract collocations, in what is known as a "hybrid approach" (Church and Hanks, 1990; Daille, 1994; Orliac, 2004; Evert, 2004, 2005; Seretan, 2011).

These techniques used to extract terms are also useful to identify specialized collocations in a corpus. They offer the researcher the ability to perform a much faster retrieval and cover much bigger amounts of data, as compared to the manual identification of these specialized lexical units.

2.4 Language resources

Since the notion of language resources has been mentioned in the previous paragraphs, it is pertinent to define it at this point. In this work, language resources refer to sets of language data and descriptions in electronic form, used to build, improve or evaluate systems or algorithms for NLP (Godfrey and Zampolli, 1997).

Cunningham and Bontcheva (2006) call these resources "the raw material of language engineering" and differentiate between language resources and processing resources. Examples of language resources are dictionaries, term bases, corpora, treebanks and lexicons. Additionally, some examples of processing resources are part-of-speech (PoS) taggers, language generation

systems, automatic translators, parsers and speech recognition systems.

One of the most important aspects of NLP is that of lexical knowledge acquisition, since the performance of any system to process written or spoken text relies heavily on the degree of "knowledge" that the system incorporates on the linguistic data that is being processed (Grishman and Calzolari, 1997). Lexical knowledge acquisition is defined as "the production or augmentation of a lexicon for a natural language processing system" (McCarthy, 2006). Since the manual creation of these language resources is an extremely difficult task, modern lexicography and terminography rely on lexical acquisition. However, it is considered a bottleneck for the development of NLP tools, since the manual creation of a lexicon is expensive and requires a large team of qualified professionals, who are not always readily available. Furthermore, the manual creation of a lexicon is a tedious and time-consuming process, one that is prone to errors and inconsistencies, even though the same could be said of conventional printed dictionaries (Fontenelle, 1994; Matsumoto, 2003). Because of this, lexical acquisition has to be aided with automated tools to be feasible.

After processing the data, the resulting lexicon is a resource such as a dictionary or thesaurus in an electronic format but is presented in such a way that it is readable by a machine and not by a human only. This includes for example, the enrichment of a lexicon by the inclusion of the forms, meanings, synonyms, antonyms, hypernyms, and phraseological information (idioms and collocations) that a given word can take. Additional information includes the associated statistical information of their distribution, which may be of no interest for a human reader, but which proves vital for a computational system designed to perform complex operations such as word sense disambiguation, ATE, collocation extraction and similar tasks (Lyse, 2011).

Calzolari (1994) points out that it is almost a tautology to say that a good computational lexicon is an essential component of any linguistic application within the so-called "language industries", ranging from NLP systems to lexicographic projects. In other words, if an automated system for the processing of lexica is going to perform its tasks in an efficient and effective manner, it has to rely on the most complete repertoire of lexical information available (Pustejovsky, 1998).

Language resources are relevant for this project because with existing language processing tools, general and specialized lexicons and corpora, it is possible to find terms and the specialized collocations associated to these terms, which can in turn help create or improve other resources. The language resources used in this work are described under Section 4.4.2.

2.4.1 Dictionaries and Computational Lexicons

Currently, dictionaries are produced increasingly more in an electronic format, because of the clear advantages that it offers for a faster and more efficient retrieval of the desired information. Electronic dictionaries are simple to use and some of them allow the user to copy and paste the equivalents on a word processor or a translation memory software. In contrast, the traditional way of finding equivalents in a bulky printed dictionary can be cumbersome and demands more time from the user to find the precise information.

However, "traditional" dictionaries are not codified for computational processing, even though they might have been published in electronic format to be read online, because they are designed to be read by humans and not by machines. This means that initially, electronic dictionaries were a faithful transcription of its printed counterpart, yet with some added values such as the possibility of carrying out faster and more comprehensive searches, listening to the pronunciation of the entry through audio files, and gaining access to synonyms or additional information by means of hyperlinks.

Besides, electronic dictionaries are not bound to the space limitations of their paper versions and therefore, it is not necessary to save space by entering phraseological information as is normally done in paper dictionaries, for example by inserting a symbol such as ~ to replace the current entry. Nonetheless, if a processing task is intended, electronic dictionaries present disadvantages for their use as a repository from which to extract linguistic features from words, such as the lexical, semantic, phonological or morphosyntactic data (Hanks, 2003). One reason for this is the fact that in these dictionaries the data are not separated from the linguistic annotations, i.e., the linguistic information attached to each word. In other cases, there are no annotations at all because in certain types of dictionaries it could be re-

dundant, while a computer system needs the full explicitation of an entry to be able to process these annotations.

To overcome these problems, researchers and developers have proposed to standardize certain procedures for making electronic dictionaries in a more effective manner to be able to process the information adequately. This is described in the following section.

2.4.2 Standardization of language resources

The standardization of language resources is relevant for the present work. One of the objectives proposed in Chapter 1 is to assess the applicability of linguistic annotation schemes for the representation of specialized collocations in term bases and computational lexicons. This means that the protocols used to annotate the data should be in accordance with existing standards so that the data can be used, merged or imported into other resources that are based on the same standards.

Standardization emerged as a means to meet the need of producing reusable resources in electronic format. It is essential for creating a dictionary that can be processed computationally, and then it can be exchanged, updated or merged with other resources in a transparent way (Hanks, 2003; Calzolari et al., 2013).

If each project for the creation of language resources uses a particular annotation scheme to encode information, as has been the case over the years, at the moment of combining an existing resource with other resources or exporting or importing data, data reuse becomes difficult, to say the least, because the developers have to adapt their system to other data structures to be able to reuse the data.

Francopoulo et al. (2006b) suggest some benefits derived from the implementation of standards for linguistic resources. One of these is the possibility of having a stable foundation for their representation and being able to deploy a solid infrastructure for a network of language resources. Besides, it facilitates the reuse of software and data that is not tied to proprietary formats. This type of product is always subject to commercial issues and sometimes requires the use of a specific tool that could disappear from the market. This

would leave the data linked to that product, or would require the periodic renewal of an expensive license whenever a new version is launched.

According to Moreno (2000), two decades ago, researchers from the field of computational lexicography started to observe the importance of designing a set of standards for the creation of reusable and interoperable language resources. To this end, several projects have been undertaken to unify the coding of computational lexicons and terminologies through the creation of norms (Calzolari et al., 2013). Once the standard has been approved, one objective of the developers of these standards is to promote their implementation among organizations, research groups, companies and professionals of the field, for the sake of promoting the exchange of information without obstacles or loss in the transmission of data due to incompatibility by using dissimilar technologies or protocols.

Among these projects, several are worth mentioning:

- Preparatory Action for Linguistic Resources Organization for Language Engineering (PAROLE) (Zampolli, 1997);
- Generic model for reusable lexicons (GENELEX);[6]
- Multilingual Text Tools and Corpora (MULTEXT) (Ide and Véronis, 1994);
- Expert Advisory Group on Language Engineering Standards (EAGLES);[7]
- International Standards for Language Engineering (ISLE) (Calzolari et al., 2001) and
- Semantic Information for Multifunctional Plurilingual Lexica (SIMPLE).[8]

Regarding the information that is stored in computational lexicons, Maks et al. (2008), classify the information that is pertinent for three intended categories:

- *Humans*, such as definitions, lexicographic comments and descriptions;
- *Computational applications*, such as semantic information, examples and complementary patterns, and

[6] `http://llc.oxfordjournals.org/cgi/content/abstract/9/1/47`
[7] `http://www.ilc.cnr.it/EAGLES/browse.html`
[8] `http://www.ub.es/gilcub/SIMPLE/simple.html`

- *Relevant information for both*, where Maks et al. mention the lemma and word forms, part of speech, tagging of semantic and pragmatic information, phraseological units and translation equivalents.

Hanks (2003) argues that a dictionary in an electronic format that was originally meant for human reading, after an adequate preparation stage, can be an important data source. Similarly, Wilks et al. (2008) introduce the difference between dictionaries in an electronic format ("machine-readable dictionaries" or MRD) (Amsler, 1982), and processing-ready dictionaries ("machine-tractable dictionaries" or MTD), and present several strategies for the conversion from MRD to MTD. Likewise, Litkowski (2006) and McCarthy (2006) state that there are significant differences between the requirements of a lexicon meant for a computer system and the contents of a dictionary or thesaurus written for human readers.

For a dictionary to be prepared for computational processing, the metadata must be separated from the linguistic information. To solve this need, markup languages are used, such as the Standard Generalized Markup Language (SGML) and especially eXtensible Markup Language (XML). Initially, SGML was a popular choice, but over the last decade XML has become the most widely used option due to its versatility and capabilities for data manipulation (Litkowski, 2006).

Language resources designed specifically for NLP such as lexicons, dictionaries or thesauruses, should ideally include the lexical, syntactic, morphological, phonetic, semantic, pragmatic, phraseological and terminological information, besides examples, in a code processable by the machine. The most widely used machine-readable thesaurus to date is WordNet (Miller, 1995), according to McCarthy (2006).

2.5 Data representation

"Representation" refers in this context to the XML code that can be used to encode specialized collocational information in a computational lexicon. The aim of this representation is to prepare the data for machine-readable lexicons which can be interchanged across different language resources (Litkowski,

2006). This representation is carried out by means of linguistic annotations that are done automatically on the data after it has been prepared.

Wilcock (2009, 1) defines linguistic annotation in this way:

> Linguistic annotations are notes about linguistic features of the annotated text that give information about the words and sentences of the text.

This means that, ideally, these annotations are meant to be a formalized explicitation, one that is readable by a computer system, of the implicit knowledge that humans have of words at different linguistic levels: their phonetics, morphology, syntax, semantics and pragmatics. In addition to this, terminological and phraseological information should also be included.

To be able to represent information on specialized collocations in machine-readable dictionaries, there is some prior information that has to be taken into account.

Several questions arise regarding the issue of the computational representation of specialized collocations. To begin with, which constituent should include the collocation, the node or the collocate or both? In this regard, there is no standard procedure defined by current lexicographical practices. I agree with Thomas (1993), who argues that it is important to define consistent criteria to choose the headword or "entry point" for the storing of LSP collocations and terms made up of multiple lexical units for precision and time-saving.

L'Homme (2009, 239) asserts that "specialised dictionaries that take into account collocations differ with respect to the method chosen to list and represent them in entries". To illustrate, let us consider one example from two economics dictionaries, which employ different ways to list the related terms and their collocates. First, the *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English* (Alcaraz and Castro, 2007), under the entry for *tariff* offers a list of complex terms including the term *tariff*, which is frequent in FTA texts, plus another noun, such as *agreement, amendment, anomaly, barrier, benefit, classification* or *concession*. Also, the Routledge Spanish Dictionary of Business, Commerce and Finance (Routledge, 1998) provides several complex terms that also include

the same term, such as *agreement, barrier, concession, cut, expenditures, legislation* and *level*. The former dictionary includes all the related terms under the umbrella term *tariff* while the latter lists separate entries for each term. Unsurprisingly, a legal dictionary, the *Diccionario de Términos Jurídicos, Español-Inglés English-Spanish* (Ostojska-Asensio, 2002) offers the equivalent of *tariff* but does not provide any collocational information.

Which information should be included using tags to encode the linguistic data that is related to the collocational information? This information could include the morphosyntactic data, such as the part of speech, the subcategorization frame of the intervening lexical items, and the semantic information such as the domain(s) in which these lexical units are used. According to Matsumoto (2003), the subcategorization frame of a verb defines the set of syntactic constituents with which a certain verb can appear. These frames usually specify the syntactic constraints or preferences of a verb. Furthermore, information on the semantic constraints is not only desirable but mandatory.

How can specialized collocations be represented in schemes for linguistic annotation issued by the International Organization for Standardization (ISO), specifically standards for terminological and computational lexicons? Several of these schemes provide a model to represent phraseological information, such as the information contained in specialized collocations with varying degrees of detail. In contrast, other schemes were not designed for the transmission of phraseological information. These standards are discussed in Section 2.6.

2.6 Standards for computational lexicons

Several initiatives have been developed with the aim of establishing a standard for the interchange of lexical data, especially for machine translation purposes. The ISO website offers a catalogue of standards.[9]

Some of these initiatives are:

[9] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_ics_browse.htm?ICS1=01&ICS2=020&

- the Machine-Readable Terminology Interchange Format (MARTIF) ISO 12200:1999,
- the Open Lexicon Interchange Format (OLIF),[10]
- the Terminological Markup Framework (TMF) ISO 16642:2003,[11]
- the TermBase eXchange (TBX) ISO 30042:2008 and
- the Lexical Markup Framework (LMF) ISO 24613:2008.

Other newer standards, not directly relevant for this work, have been released from 2012 to 2016:

- the ISO 24615 Syntactic annotation framework (SynAF), composed of two parts,
- ISO 24612:2012, Language resource management - Linguistic annotation framework (LAF),[12]
- ISO 24611:2012, Language resource management - Morpho-syntactic annotation framework (MAF),[13] and
- the Semantic annotation framework (SemAF) ISO 24617, composed of eight parts (the third part is not yet available in the online ISO standards catalogue).

These standards are XML-compliant specifications for the implementation of a lexicon. Some of these standards, such as MARTIF, use an onomasiological or concept-oriented approach rather than a semasiological or lexically-oriented one, which, in my view, makes them unsuitable for representation in NLP or lexicographic applications.

The adoption of standards for the constitution of lexical and terminological resources raises several questions:

- How can language resources be encoded in an interoperable, scalable and interchangeable format? This would ensure that the data could be

[10] http://www.olif.net/

[11] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32347

[12] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?ics1=01&ics2=020&ics3=&csnumber=37326/

[13] http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?ics1=01&ics2=020&ics3=&csnumber=51934

merged with or exported to other language resources and that the data would not be lost due to technology incompatibilities, which is known as blind interchange.

- Are there commercial factors that affect the adoption and implementation of a given standard? This implies that the industry could prefer a certain technology while academia adopts a different protocol to store information but the two might be incompatible, which would hamper the development of language resources.

Some aspects of the LMF, TMF, OLIF and the TBX standards will be commented in subsection 6.1.1 and 6.1.2, with a focus on their suitability for the computational representation of MWEs, and specifically specialized collocations.

Corpora are another vital resource for NLP, and are described in the following section.

2.7 Corpus linguistics

The discipline of corpus linguistics provides a relevant methodology to study authentic texts in their context. According to Hunston (2006), a "corpus is an electronically stored collection of samples of naturally occurring language". McEnery (2003) asserts that a corpus is machine readable. He defines a corpus as "a body of machine-readable linguistic evidence, which is collected with reference to a sampling frame" (McEnery, 2003, 450). Corpus data are stored and indexed in such a way that they are searchable with computer software. Additionally, corpus data can be preprocessed and tagged with structural markers to identify documents, chapters, sections, paragraphs and sentences. Next, the data can be tokenized to identify each unit, then it can be annotated with part-of-speech tags, lemmatized and chunked. Other researchers prefer to store corpora without any of these annotations in an attempt to keep the data as close as possible to the original text. Besides, corpora can be monolingual, parallel or multilingual (McEnery, 2003; Aijmer, 2008).

Contrary to doing linguistic research by means of examples obtained by

the linguist through introspection, corpus linguistics relies heavily on finding real examples extracted from authentic material (McEnery and Wilson, 2001).

A corpus also allows researchers from other disciplines than linguistics, such as sociologists, lawyers, economists and anthropologists, to carry out studies based on authentic texts, such as the ones included in the corpus used for this research. However, users of corpora differ in their method and approach to the use of a corpus.

To carry out this study, a parallel and annotated corpus is a vital resource because it makes it possible to find the occurrences of FTA terms along with the collocates of these terms in their occurring context and not in isolation.

A corpus is an efficient tool to generate a concordance of the words under consideration, in order to perform a vertical and a horizontal examination of the words and their surrounding context, each one offering differing insights into these lexical units. Tognini-Bonelli (2001) explains that a horizontal reading enables to focus on larger units such as clauses, sentences and paragraphs. In contrast, a vertical reading is suitable to scan for patterns co-occurring with the node word. Thus, using a corpus-generated concordance to perform a vertical and horizontal reading of the words under consideration offers the researcher many advantages. According to (Wynne, 2009, 711)

> reading concordances allows the user to examine what occurs in the corpus, to see how meaning is created in texts, how words co-occur and are combined in meaningful patterns, without any fixed preconceptions about what those units are. It can be a method of approaching the corpus in a theory-neutral way. This is part of what Tognini-Bonelli (2001) calls corpus-driven linguistics.

Among corpus linguists there is not a single and unified method to do research using corpus linguistics. However, there are several approaches, which are supplementary methods for corpus exploitation, i.e. corpus-based, corpus-driven and corpus-assisted research.

2.7.1 Corpus-based vs. corpus-driven research

Tognini-Bonelli (2001, 2002) explains the difference between the two ap-

proaches to research done using corpus linguistics. These approaches have several common features while other features differ. Corpus-based refers to a type of research where the researcher uses a corpus as as test-bed. Instead of relying solely on his/her intuitions, the corpus provides examples to test or exemplify theories and descriptions that were formulated before the creation of large electronic corpora.

The second approach refers to a type of linguistic research in which the researcher lets the corpus "speak for itself" by using tools and techniques that exploit the frequency and other statistical information from the data with no pre-conceived idea on the theoretical constraints that might rule the types of possible queries. However, some authors express their criticism toward this approach because of its full reliance on data and claim that in the end all corpus methods are "corpus-based" (McEnery and Hardie, 2011).

In my view, no corpus research can claim a total adherence to any of the two approaches. Most modern approaches today use a combination of both approaches and thus are hybrid in nature. One approach uses linguistic knowledge expressed in the form of rules obtained from grammars while the other relies heavily on statistical data. Today, with the growing availability of computerized corpora and the production of corpus-aware grammars, linguists have more resources available to carry out research with the aid of corpora. Some linguists also use statistical methods applied to huge repositories of data, with excellent results. This way, a combination of both approaches gives the researcher more elements to process an amount of data that was not possible before.

In accordance with what is customary in corpus linguistics, lexicography and corpus-based terminology, I use a combination of both approaches for doing corpus linguistics. This work is corpus-based in the sense that morphosyntactic patterns that form collocations in English and Spanish are used to query a corpus that was previously lemmatized and annotated with part of speech tags. It is also corpus-based because a set of previously identified terms or candidate terms are used as "seeds" (Baroni and Bernardini, 2004). Other studies have used terms as seeds (Jacquemin et al., 1997; De Groc, 2011; Ljubešic et al., 2012; Burgos, 2014). In the case of this work, these seed terms serve as a starting point to identify semi-automatically the col-

locates found in the list of terms. However, this work is also corpus-driven because several applications and techniques that rely on statistics without *a priori* conceptions of what is in the corpus are used to calculate the collocability between a term and its collocates. These applications are explained in Chapter 4.

The remainder of this chapter is organized as follows. First, I present a theoretical background on collocations, followed by a review of the definitions proposed by representative authors in the field and the salient characteristics of collocations. Then, I present a view on collocations from different disciplinary perspectives. Before attempting to propose a definition of specialized collocation, I describe the criteria for collocability between two or more lexical units in Section 2.11. Then, in Section 2.12, I account for the features that give these units a specialized nature.

2.8 Definitions of collocation

This section presents the main features to identify collocations. Next, several definitions of the concept of collocation are provided, with the aim of arriving to a working definition of what constitutes a specialized collocation.

There is no general consensus on the definition of what a collocation is. The researchers that have done research on collocations have offered different definitions to characterize this phenomenon of lexical combinations. For example, Seretan (2011) presents a list with 21 definitions. In her view, even though collocations have attracted the attention of linguists for a long time, "they still lack a systematic characterization" (Seretan, 2011, 22).

Possibly, the first definition of collocation available is the one offered by Palmer and Hornby (1933). These authors had an interest in the teaching of collocations to students of English as a foreign language. They defined a collocation as "a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts." Stubbs (2009, 17) adds regarding Palmer and Hornby's definition of collocation that today we would "say that their semantics is non-compositional". In their work, Palmer and Hornby offered a report on English collocations with a pedagogical intention in mind.

Two decades later, Firth (1957, 11) published his famous sentence: "You shall know a word by the company it keeps", which has become the classical quote throughout the literature on collocations. However, besides this quote, Firth never actually defined the notion of what exactly constitutes a collocation.

In more contemporary publications, for example in McKeown and Radev (2000, 507) a collocation is defined as "a group of words that occur together more often than by chance". These authors assert that collocations cover word pairs and phrases commonly used in language. However, they note that these words pose a challenge for their identification and classification because they are beyond the coverage of general syntactic or semantic rules.

From a theoretical linguistics perspective, collocations were, until relatively recently, not considered as an interesting subject of study, partly because, under the influence of Chomsky's generative grammar, "the lexicon was reduced to a mere list of fully interchangeable words" (Evert, 2004, 16) and these word combinations were explained merely as selectional restrictions.

Apart from the fact that there is not a unified definition of collocation in which authors agree, this linguistic phenomenon has also received different names, being "collocation" the most frequently used by authors (Firth, 1957; Halliday, 1961; Benson et al., 1986; Benson, 1990; Sinclair, 1991; Sinclair et al., 1970; Mel'čuk, 1998). However, other authors employ a different terminology such as "lexical combination" (L'Homme and Bertrand, 2000) and "frequent word combinations" (Cortes, 2004). In French there are other names that have been used in the literature, such as "groupements usuels" (Bally, 1932) and "formules langagières" (Clas, 1994). In Spanish "enlaces frecuentes" was used by lexicographer Moliner (1966).

Several features characterize the lexical phenomenon of collocation and will help us to differentiate it from other types of MWEs.

First, we can mention, from a statistical perspective, the probability that word x and word y co-occur, either adjacent or in a window of several intervening words. For example, Stubbs (2002, 30) asserts the following: "Collocation is a relation between words in a linear string: a node predicts that a preceding or following word also occurs." The statistical probability of two or

more words that tend to co-occur and their distributional characteristics has been studied by several researchers (Halliday, 1961; Evert, 2004; Pecina and Schlesinger, 2006; Seretan, 2011; Gries, 2013). Harris (1968) introduced the notion of distributional analysis which claims that the distributional characteristics of words give insights about the meaning of these words. Even though his theory has been questioned since then, it was a pioneering work in the field.

In the view of Halliday (2004), using a statistical and a linguistic standpoint, collocation is "a purely lexical relationship; [...] an association between one word and another, irrespective of what they mean. It can be defined quantitatively as the degree to which the probability of a word y occurring is increased by the presence of another word x". Halliday's definition puts emphasis on lexica and statistics but does not take syntax into account.

Second, collocations pose a challenge for their identification and classification because they are beyond the coverage of general syntactic or semantic rules (McKeown and Radev, 2000; Krishnamurthy, 2006). Because of this, I agree with Seretan (2011, 26), who claims that "providing a characterization of collocations in terms of syntactic behaviour seems very difficult".

Third, collocations are arbitrary and non-predictable groups of words that co-occur repeatedly in a language. Frequently, famous writers, journalists, politicians or other influential people popularize the use of one of these lexical combinations and speakers of a language adopt it in their everyday language (Benson, 1985; Bahns, 1993; Bossé-Andrieu and Mareschal, 1998a; Zuluaga, 2002; Krishnamurthy, 2006). As an example, Manning and Schütze (1999) offer "international best practice", an expression used by bureaucrats in Australia due to its repeated use and connotation. This collocation is also used in other varieties of English but with less frequency, as attested by the Corpus of Global Web-Based English (GloWbE),[14] with 1.9 billion words, which offers 280 occurrences of this collocation.

Researchers commonly set an arbitrary limit to the span of the units that are considered collocations. Some authors do not even include bigrams, i.e. two-word candidates, and prefer to focus on longer units. However, in my opinion, the exclusion of bigrams as candidate collocations would leave aside

[14] `http://corpus2.byu.edu/glowbe/`

a great amount of relevant collocations. Other researchers prefer to exclude longer collocations because they span two short units (Greaves and Warren, 2010), and yet others exclude non-lexical constituents from their descriptions of collocations (Bartsch, 2004).

Benson et al. (1986) and Benson et al. (2010, xix) classify collocations into two types: grammatical and lexical. For them, the term grammatical collocation refers to "a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or a clause" while "lexical collocations, in contrast to grammatical collocations, normally do not contain prepositions, infinitives or clauses. Typical lexical collocations consist of nouns, adjectives, verbs, and adverbs". This is the approach that I use in this work: only lexical words are considered as integral components of a specialized collocation. As a consequence, determiners, pronouns and other non-lexical constituents are left aside and are only included in certain patterns that are specified in Section 5.7.

For this work, following the definition presented by Bartsch (2004, 76) collocations will be understood as

> lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other.

This definition is adequate because it allows us to account for the different morpho-syntactic realizations of several lexical items and not only to two given words adjacent to each other. For example, in the FTA corpus, *adoptar un arancel aduanero* 'adopt a customs duty' can also appear as *adopción de un arancel aduanero* 'adoption of a customs duty' where the deverbal noun *adopción* 'adoption' also keeps a collocational relation with the term *arancel aduanero* 'customs duty'. Besides, this definition suggests that something else besides a syntactic relation holds between the constituents, such as lexical and pragmatic constraints.

2.8.1 Differences between several types of MWEs

Within a broad perspective of phraseology, there are three types of units: free combinations, collocations and idioms.

Multiword expression (MWE) is the hypernym which encompasses units such as multiword lexical unit, collocation, idiom, compound noun, lexical bundle, verb-particle construction, verbal expression and proverb (Seretan, 2011, 2013). In this thesis, collocations are understood as a subclass of MWE, in harmony with Baldwin and Kim (2010) and Seretan (2011, 2013).

Evert (2009, 1213-1214) explains a key difference between collocation and MWE:

> the former has a Neo-Firthian sense that alludes to lexical units of a semi-compositional and lexically determined nature whereas the latter has become the preferred form in the fields of computational linguistics and natural language processing.

MWEs are defined by Baldwin and Kim (2010, 269) based on Sag et al. (2002) as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity".

Sag et al. (2002, 197) themselves reserve the term collocation "to refer to any statistically significant co-occurrence, including all forms of MWE as described above and compositional phrases which are predictably frequent". Their definition is not entirely adequate for this work because I take into account the linguistic features of specialized collocations, not only their statistical significance.

All of these subclasses of MWEs exhibit different features and perform different functions. Figure 2.1 illustrates the subclasses of MWEs, the place that specialized collocations occupy in relation to other MWEs, their location regarding terminology and phraseology and how specialized collocations stand in the midst of both disciplines, indicated by the smaller inner hexagon in Figure 2.1.

Over the years, several names have been used to refer to this variety of multiword types. Within the field of NLP, researchers employ the term n-grams to refer to strings of two or more consecutive words calculated by means of statistical AMs.

Biber et al. (1999, 58) offer some clues to distinguish multi-word lexical units from collocations and from lexical bundles. According to these authors,

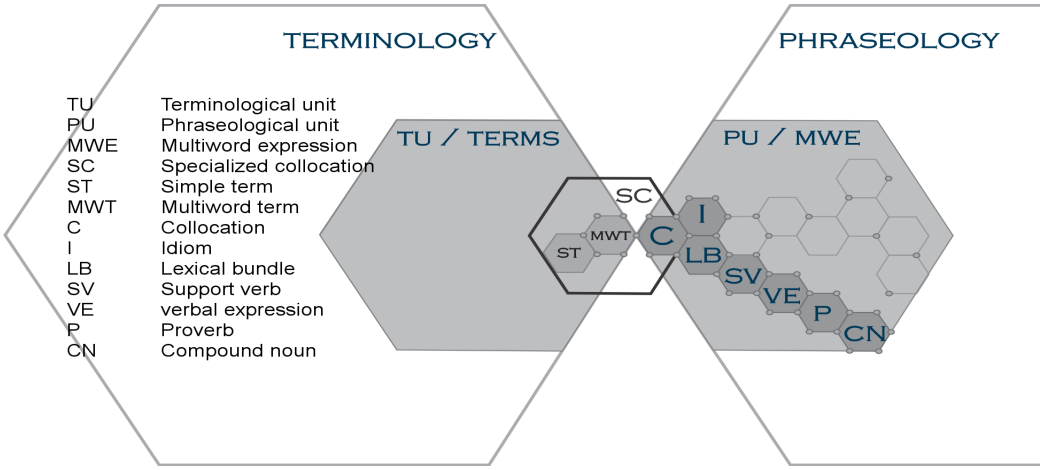

Figure 2.1: *A diagram representing the subclasses of MWEs and how specialized collocations are related to terminology and phraseology*

a multiword lexical unit is a lexicalized "sequence of word forms which functions as a single grammatical unit", e.g. *look into* which is used much the same way as *investigate*. Biber et al. (1999) group phrasal verbs (e.g. *point out*); prepositional verbs (e.g. *appear on*); complex prepositions (e.g. *except for*, *aside from*); correlative coordinators (e.g. *both ... and*, *either ... or*, *neither ... nor*) and complex subordinators (e.g. *as far as*; *given that*) as different types of multiword lexical units.

2.8.1.1 Lexical bundles

Lexical bundles are sequences of three or more words that tend to co-occur statistically in a register, irrespective of their idiomaticity and whether or not the sequence constitutes a grammatical unit (Biber et al., 1999; Cortes, 2004). In contrast, collocations consist of two or more lexical words with a tendency to co-occur. A lexical bundle is therefore a type of adjacent MWE considered as an extended collocation.

Cortes (2004) mentions two patterns that typically form lexical bundles in English, among others: Preposition + Determiner + Noun + Preposition and Determiner + Noun + Verb + Determiner. Thus, lexical bundles can provide valuable information about the lexis of a particular genre and its formulaic language but differ from collocations and idioms in several respects: lexical

bundles perform a grammatical and cohesive function, are adjacent MWEs and are syntactically fixed (Benson, 1985; Casares, 1992).

2.8.1.2 Differences between collocations and idioms

The criteria set to distinguish collocations from other types of MWEs are not clear-cut but are instead sometimes vague, confusing or contradictory among several researchers. Evert (2004) even holds that "the distinction between collocations and non-collocations is ultimately based on the intuition of a lexicographer, for instance, in contrast to the formal and unambiguous definitions that linguistic research aims for", which makes the scenario even more complicated.

Some authors (Thomas, 1993; Manning and Schütze, 1999) blur the line that separates idioms from collocations by using the two terms interchangeably. However, idioms differ from collocations and are either 'pure' phraseological units or relatively frozen expressions which exhibit distinct linguistic features. The most salient features that differentiate idioms from collocations are their degree of morphosyntactic fixedness, idiomaticity (also known as semantic opaqueness or fossilization) and non-compositionality. In contrast to idioms, collocations can be semantically transparent and semi-compositional. Manning and Schütze (1999) list non-compositionality, non-substitutability and non-modifiability as criteria for the linguistic treatment of collocations. However, accepting this view would contradict phraseologists, who assign the same features to idioms.

According to Saeed (2003), collocations can undergo a fossilization process until these lexical units become fixed expressions. Bahns (1993, 57) contrasts collocations with idioms and with free combinations. In his view, the "main characteristics of collocations are that their meanings reflect the meaning of their constituent parts (in contrast to idioms) and that they are used frequently, spring to mind readily, and are psychologically salient (in contrast to free combinations)". Figure 2.2 illustrates the degree of fixedness of free combinations or units, collocations and idioms, with total flexibility on the left and less possibility of flexibility on the right.

Collocations are not as syntactically fixed or semantically opaque as id-

Figure 2.2: *A diagram representing free combinations or units, collocations and idioms*

ioms but are non-predictable (Biber et al., 1999) and are found in a "transitional area approaching idiom" (Cruse, 1986, 41). Collocations, being more flexible, admit some transformations or operations while idioms, due to their fixedness and rigidity, only admit these morphosyntactic processes in exceptional cases. These are some examples taken from the FTA corpus (Patiño, 2013) to illustrate how the collocational relation is kept despite morphosyntactic changes: *aplicación de medidas no arancelarias*, 'application of non-tariff measures', *adoptar medidas arancelarias*, 'adopt tariff measures', *aplicar medidas de salvaguardia*, 'apply safeguard measures', *adoptar medidas provisionales oportunas*, 'take / adopt prompt interim measures', *adoptar medidas tributarias*, 'adopt taxation measures'. These examples are different morphosyntactic realizations of a collocation found through a Google search: *adopción de medidas tributarias*, 'adoption of taxation measures', *medidas tributarias adoptadas*, 'adopted taxation measures'. In these cases, the collo-

cational relation is still kept among the intervening constituents, even though some morphologically-related constituents occupy different grammatical categories, for example the deverbal noun *adopción* and the verb *adoptar*. To sum up, in addition to their semi-compositionality and frequency, collocations are found in a continuum, amidst free combinations and idioms.

2.8.1.3 Differences between collocations and free combinations

Koike (2001) presents several features as the most salient ones to distinguish collocations from free combinations. According to Koike, collocations exhibit the following features:

1. Frequent co-occurrence of lexical units.
2. Combinatory restrictions imposed by traditional use (*sharp distinction* and *trenchant analysis* form collocations whereas *trenchant knife* is an anti-collocation.)
3. Formal compositionality which allows for a certain formal flexibility. For example *adoption of taxation measures* and *taxation measures adopted* hold the same collocational relation.
4. Semantic precision of the combination. For example *safeguard measure* where the adjective adds semantic precision to the type of measure being adopted.

2.9 A look at collocations from different perspectives

Collocations are a relevant topic for several disciplines of the broader field of linguistics. The following subsections present a general overview of how collocations have been treated in several disciplines.

2.9.1 Collocations from the perspective of lexicography

In current lexicographic practice, there are no systematic criteria for the selection, presentation, inclusion or exclusion of collocations in general or specialized dictionaries, and because of this, some researchers argue that the treatment of collocations in general and specialized lexicography has been unsatisfactory (Benson, 1985; Cop, 1991; van Sterkenburg, 2003; Orliac, 2004; L'Homme, 2006; Aguado de Cea, 2007; Moon, 2008). For example, in a study of the treatment of collocations in several types of dictionaries, Moon (2008) compared the collocational information included in English and French dictionaries, both monolingual and bilingual, general and specialized. She reports that the language resources examined in her study offer partial phraseological information but most entries do not explicitly include the prepositional and adjectival combinations of the head words under scrutiny. The profile of a dictionary user affects the degree of information that is included in a dictionary. Learner's, translation and specialist dictionaries are aimed at different audiences and should therefore include phraseological information that is pertinent for the intended audience. McKeown and Radev (2000) argue that, given the fact that collocations are lexical in nature, they have been studied primarily by lexicographers, who are concerned with the identification of criteria to distinguish collocations from other lexical units, their characteristics and representation in dictionaries.

In the case of specialized collocations, this lack of coverage can have a negative impact for the work of language professionals that rely on dictionaries, such as translators, technical writers, lexicographers and terminologists. The same holds for language learners who want to learn how to successfully combine words in a foreign or second language and expect to rely on a dictionary to attain this end.

2.9.2 Collocations from the perspective of NLP

Collocations are crucial lexical units to improve the performance of NLP systems. In the words of (Gelbukh and Kolesnikova, 2013, iv), "Knowledge of collocation is important for natural language processing because collocation

comprises the restrictions on how words can be used together." This explains why a lot of efforts within this field have been devoted to the automatic or semi-automatic detection of collocations in NLP applications due to the relevance of collocations for NLP and their utility for statistical natural language paradigms (McKeown and Radev, 2000; Evert, 2004, 2009; McCarthy, 2006; Heid and Weller, 2008; Seretan, 2011, 2013).

These are some of the NLP tasks which would benefit greatly from a lexicon enriched with the collocational information of words:

- word sense disambiguation,
- optical character recognition,
- natural language generation,
- named entity recognition,
- morphological and syntactic analysis
- information retrieval,
- sentiment analysis,
- automatic topic identification,
- machine translation and
- text generation systems.

Collocations are not only relevant lexicographic information, they are particularly crucial for several NLP tasks. Collocations are useful for automatic topic identification since they provide useful information to disambiguate homographic and polysemous words, to distinguish quasi-synonyms and to remove syntactic ambiguities (Moon, 1998; McKeown and Radev, 2000; Seretan, 2011; Ramisch, 2015). Another factor that has to be taken into account is the fact that each domain has its idiosyncratic MWEs and therefore an NLP system should be enriched with this information. This is precisely what Sag et al. (2002, 2) hold:

> Specialized domain vocabulary, such as terminology, overwhelmingly consists of MWEs, and a system may have to handle arbitrarily many such domains. As each new domain adds more MWEs than simplex words, the proportion of MWEs will rise as the system adds vocabulary for new domains.

Besides, the same term can be used in several domains, with different senses, but the collocates of that term can help to determine the domain in which it is being used. It means that this type of lexical phenomenon is useful to discriminate among the several senses that a given word might take (McKeown and Radev, 2000; Stevenson and Wilks, 2001).

2.9.3 Collocations from the perspective of translation studies

An adequate handling of collocations is a key component for an optimal translation (Newmark, 1988; Heid and Freibott, 1991; Munday, 2016). For translators, dictionaries are a valuable support tool to find suitable equivalents for the words found in the texts they translate. However, dictionaries do not include the collocations of a given language in a systematic way. As a consequence, the translator faces many challenges when finding an equivalent for these lexical units, even more so when translating specialized texts (Benson, 1985; Heid and Freibott, 1991).

Similar to collocations found in general texts, the equivalents of specialized collocations also have an effect on the quality of a translation, as suggested by Oakes (1998, 159): "collocations tend to be specific to a domain sublanguage, and thus the collocations used in a sublanguage often have different translations to those in general usage". This implies that to attain accuracy a translator has to be aware of this type of lexical units depending on the subject field to which the translated text belongs. However, this is not easy because being a native speaker of a language does not necessarily entail that the translator has the competence to master the collocations that are typical of a particular domain. This view is supported by Baker (2011, 57), who argues that

> Being a native speaker of a language does not automatically mean that the translator can assess the acceptability or typicality of register-specific collocations. This is largely why courses in specialized and technical language form an important component of translation training syllabuses.

The challenge of correctly handling collocations when performing direct translation, i.e. towards the translator's mother tongue, is aggravated when dealing with inverse translation, that is, translating towards a non-mother tongue (Corpas and Seghiri, 2009). In this modality, if the translator does not have a ready-made equivalent, one that fully encompasses and expresses the same concept in the target language, especially when translating into a foreign language, he/she has to "guess" which is the right lexical unit to combine with another one. In this regard, Heid (2001, 788) asserts the following:

> Collocational word combinations are a problem for translation because, although many collocations of a foreign language are transparent so far as understanding is concerned and do not cause trouble in translation into one's mother tongue, it is impossible most of the time to "guess" the right word combinations when translating into a foreign language.

According to Heid, this happens in general as well as in specialized language, given the fact that collocations are not explicitly rule-governed but rather are to some extent a matter of convention.

Translators follow different strategies to translate collocations. Some of these strategies imply that the collocation is lost or "de-automatized" (Zuluaga, 1998), that is, the semantic link between the two intervening lexemes is not kept, or simply the collocation is not understood as such by the translator because she or he does not have the "phraseological competence" which for Corpas (2003) is still a pending subject for many translators.

Corpas (2003) labels these units as collocation translemes or translation units. She offers a classification of several cases that emerge in the translation of collocations:

1. Equivalent translation with idiosyncratic collocational feature: in this type of cases, only the base is translated independently from the collocate, while for collocates the translation equivalents can only be described according to the base that has determined the collocates (Heid and Freibott, 1991). Corpas offers the example *asignar recursos*, 'allocate resources'.

2. Undertranslation: this case emerges when in the target language there are no identical semantic features and therefore, when the collocation is translated, any of these aspects will be lost. For example *torrente sanguíneo*, 'bloodstream'.

3. Overtranslation: this case is the opposite to the previous one; it means that the target language collocation may present absent semantic features in the source language.

4. when there is a change in the register between the source language collocation and its target language equivalent. Corpas offers the example of *cálculos biliares*, 'gallstone disease' which has a specialized equivalent in the field of medicine, *litiasis biliar*.

The above might bear consequences for translators, who could easily ignore the collocational pattern of the target language and carry out a literal translation of the components of a collocation, by using a calque term instead of the customary equivalent in the target language. For example, in a movie, when an actor says "straight jacket" the Spanish translation of the subtitles read "chaqueta rígida", *rigid jacket*. This suggests that the translator was not aware of the phraseological relation between the two words. Baker (2011) estimates that the translator should re-read the first version of a translated text a few hours later with the aim of carrying out a reading closer to the collocational pattern of the target language. This way, the translator may overcome the obstacles which could otherwise emerge under the influence of the source language, such as proposed by the law of interference (Toury, 1995). For Baker, it is important to take into account the collocational meaning rather than doing a mere substitution of individual words with their dictionary equivalents. Baker (2011) argues that the task of identifying the collocational meaning is crucial at the first stage of translation, when the translator is interpreting the source text. She also holds the view that the different collocational patterns between the source language and the target language are a source of potential trouble when carrying out a translation task and thus this calls for special attention from the translator.

2.10 Syntactic patterns of collocations

Several syntactic combinations frequently form collocations. In the view of Manning and Schütze (1999), the two most frequent collocational patterns are those formed by Adjective + Noun and Noun + Noun. According to Maurer-Stroh (2004) many collocations are language-specific. Therefore, collocational patterns vary across language pairs. Benson et al. (1986) present a classification of the different types of collocations based on the constituents that fall into this linguistic phenomenon. Heid (1999, 2001) offers a syntactic classification of the most frequent collocational patterns for several Indo-European languages according to the two lexical items that make up the collocation. These patterns are constituted by:

1. Noun + Verb
2. Noun + Adjective
3. Noun + Noun
4. Verb + Adverb
5. Adjective + Adverb.

According to Heid, the first three types are much more common in specialized languages than the last two. He refers to this type of collocations as "multiword terms".

According to Koike (2001), Noun + Verb and Noun + Adjective are the most frequent collocations in Spanish. Koike offers a classification of simple and complex Spanish collocations. In his view, simple collocations are the ones formed by these patterns:

1. Noun + Verb
2. Verb + Noun (dependent clause)
3. Verb + Preposition + Noun
4. Noun + Adjective

5. Noun + Preposition (de) + Adjective

6. Verb + Adverb

7. Adverb + Adjective

8. Verb + Adjective.

For complex collocations, Koike proposes these patterns:

1. Verb + Noun phrase

2. Verb phrase + Noun

3. Noun + Adjectival phrase

4. Verb + Adverbial phrase

5. Adverbial phrase + Adjective.

Bossé-Andrieu and Mareschal (1998a) provide a similar classification of morphosyntactic patterns valid for the formation of collocations in French.

2.11 Criteria for collocability

Crystal (2008) defines collocability as "the potential of items to collocate" and provides *collocational range* as a synonym term. Some of the authors that have researched collocations offer criteria applied to determine the collocability between any two or more lexical units in general texts. These criteria are basically the same across research on collocations (Benson, 1985; Zuluaga, 2002; Evert, 2009). However, in my view some gray areas still persist and three main obstacles seem to obscure the notion of collocation, which poses a challenge for researchers.

First, after several decades of research, authors have not adopted a widely accepted definition that fully encompasses all the linguistic features of the collocational phenomenon.

Second, several types of MWE, despite their differences, are indistinctly called "collocation" or "idiom" or "phraseological unit" by some authors.

This terminological uncertainty can lead to confusion and does not help much to set a clearly delimited frontier to distinguish among different types of MWEs.

Third, researchers do not apply unanimous principles to establish collocability: the view on what constitutes a collocation held by one researcher is not necessarily shared by other researchers. As a consequence of the above, there are different and even contradictory criteria to distinguish collocations from other phraseological units.

Let us take as an example the multiword expression *preferential tariff treatment*, in Spanish *trato arancelario preferencial*. From a terminological point of view, this unit constitutes a multiword term. This can be validated internally by consulting the FTA data. For example, the FTA signed between Canada and the Republic of Peru, in Article 105, Definitions of General Application, reads:

> preferential tariff treatment means the application of the respective duty rate under this Agreement pursuant to the tariff elimination schedule to an originating good.

Besides, a specialized dictionary of international trade, the *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English* (Alcaraz and Castro, 2007, 475), includes this subentry:

> **preferential tariff arrangements or treatment** FISC régimen / tratamiento arancelario preferencial o preferente.

Since *preferential tariff treatment* is listed as an entry in the definitions section of an FTA and it is included in a specialist dictionary, it can be concluded that it indeed constitutes a term in the field of international trade. Besides, from a phraseological point of view, *preferential tariff treatment* also constitutes a specialized collocation. The criteria of collocability in specialized texts will be discussed in the following subsections using *preferential tariff treatment* and other collocations in FTAs as examples to test the claim that this term is also a specialized collocation.

2.11.1 Frequency of co-occurrence

An empirical quantitative study using a corpus-linguistic method allows us to establish how often two or more lexical units from the open lexical categories co-occur in running text, which is an indicative factor of recurrent word association among these units (Benson et al., 1986, 2010). For example, in the FTA corpus, the Adjective + Noun + Noun collocation *preferential tariff treatment* has a high frequency in the domain of FTAs because it appears 70 times in 1.37 million words, or with a relative frequency of 51 times per million words. The term *tariff treatment* also enters into a collocation with other adjectives besides *preferential*: *current tariff treatment, preferred tariff treatment, differential tariff treatment, favorable tariff treatment* and *free tariff treatment*. In contrast, the Corpus of Contemporary American English (COCA) (Davies, 2009) does not offer any occurrence of *preferential tariff treatment* even though this corpus contains 520 million words. It only has 5 occurrences of *preferential tariff*, all of them extracted from economic newspapers discussing free trade topics. The differing proportions in the frequency of occurrence between the two corpora suggest that this particular collocation is only used in a restricted subject field.

However, even though frequency might be an important factor for the semi-automatic identification and extraction of collocations, it is not necessarily a determining factor in the case of terms, because even huge corpora might include a term from a specific domain only once or less than five times in a corpus of millions of words, as evidenced with the term *preferential tariff*. This implies that setting a minimum frequency threshold might work for the automatic extraction of collocations, which is common practice in corpus linguistics and NLP, but not necessarily for extracting terms.

2.11.2 Combinatory restrictions

According to several researchers (Firth, 1957; Sinclair et al., 1970; Kilgarriff, 2005) words have a tendency to co-occur with other words with some restrictions set upon them. For example, in the FTA corpus, the term *arancel aduanero*, 'customs duty' collocates with *tasa*, 'rate' and forms the collocation *tasa de arancel aduanero*, 'customs duty rate'. There is a preference

for *tasa*, 'rate' and not for other synonyms such as *proporción*, 'proportion', *medida*, 'measure' or *nivel*, 'level' and this preference is distinct for every subject field, arbitrary and sometimes imposed by traditional use or by convention. The same holds for other combinations besides Noun + Adjective in the case of Spanish or Adjective + Noun in English.

2.11.3 Degree of composionality

The principle of compositionality, usually attributed to Frege, is defined as follows: "The meaning of an expression is a function of the meanings of its immediate syntactic components plus their syntactic mode of composition", as is implicit in Frege's work on the philosophy of language (Van Eijck and Unger, 2010, 150). Idioms are non-compositional, i.e. the meaning of the whole unit is not simply a sum of the meaning of component words, whereas collocations are semi-compositional but "nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" (Cruse, 1986, 40). Therefore, collocations, as semi-compositional word pairs, are conformed by two parts, the *node* and the *collocate*. The node is a free element that retains its independent meaning. The collocate is lexically determined by the node and adds to the combination of the two elements a meaning that it cannot have on its own (Sinclair, 1991; Stubbs, 2002; Evert, 2004). For example, in the specialized collocation *preferential tariff treatment*, the adjective in isolation does not have a meaning related to any of the two nouns.

In addition to the varied terminology used to refer to collocations, other authors have given different names to the constituents that make up a collocation: "node" and "collocate" (Sinclair, 1991; Stubbs, 2002) and "base" and "collocator" (Mel'čuk, 1998). Throughout this thesis, *node* and *collocate* will be used to refer to the constituents of a collocation.

The above does not mean that a collocation only has two elements, as sometimes they also span more than two constituents. For example, the multiword term *service supplier* collocates with the adjective *financial* and forms a collocation of the type Adjective + Noun + Noun: *financial service supplier*. Other multiword terms might be formed by two or even three

collocations that are subsumed and form another one. It is the case of *preferential tariff treatment* where *preferential tariff*, *preferential treatment* and *tariff treatment* are also collocations.

2.11.4 Degree of transparency

In the case of idioms, all of the constituents can be semantically opaque, that is, metaphorical, e.g. *kick the bucket* where both *to kick* and *bucket* are opaque because neither the verb nor the noun have any literal meaning related to death. In other idioms, only one of the constituents is opaque, such as *cocodrilo*, 'crocodile' in *lágrimas de cocodrilo*, 'crocodile tears', while *lágrimas*, 'tears' is transparent. In contrast, collocations can be semi-opaque or fully transparent. It means that one constituent can be opaque but the meaning can still be inferred from the other constituent. For example, in the Noun + Noun collocation (*mass transit*) or the Verb + Noun collocation (*breach an obligation*) the first constituents, that is, the collocates, are idiomatic whereas the second component, the nodes, are fully transparent. In other cases, both constituents are semantically transparent, such as in the verbal collocation *constituir una expropiación indirecta*, 'constitute an indirect expropriation' or the Noun + Noun collocation *tariff preference*.

2.11.5 Adjacency vs. span of words between node and collocate

In the view of some authors, the constituents of a collocation are adjacent to each other (Choueka, 1988) while for others one of the constituents, either the node or the collocate, can co-occur some words after or before the other one (Sinclair et al., 1970). Sinclair (1991, 170) argues that "collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening." Thus, for Sinclair, adjacency is not a defining feature of collocations. In contrast, Choueka (1988) defines a collocation as "a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived

directly from the meaning or connotation of its components". In this way, Choueka's definition rules out as collocations the combinations formed by two co-occurring words but which are not found consecutively in a text.

2.12 Specialized features

In specialized texts, the same phenomenon of collocation explained in Section 2.8 is present. Specialist dictionaries and term bases include terms, normally nouns or noun phrases. However, these resources do not customarily include the collocational relations of these terms with other lexical units from the open lexical categories, namely nouns, verbs, adjectives or adverbs which tend to co-occur with these nouns or noun phrases (Wanner et al., 2007).

Unsurprisingly, collocations pose a challenge for translators, interpreters and other language professionals, besides language learners. One of the outstanding reasons is because collocations are idiomatic and unpredictable, i.e. they cannot be predicted solely on syntactic grounds, as pointed out by the quote from L'Homme at the beginning of the chapter, a view that is supported by other authors (Pavel, 1993a; Matsumoto, 2003; Nugues, 2006). For example, IATE,[15] InterActive Terminology for Europe, the online term repository of the European Union, offers 65 entries that include the term *arancel aduanero* 'tariff'.[16] However, these entries do not offer much phraseological nor collocational information that include this term, except for some entries such as *establecimiento de un arancel aduanero común* 'establishment of a common customs tariff' or *arancel aduanero preferente* 'preferential customs tariff', but it is not explicitly identified as a collocation that includes this term.

Several lexicographical projects have tried to fill this gap and have produced monolingual dictionaries of collocations in several languages with varying degrees of success and coverage. Cowie (1986) and Cop (1990) present an overview of these efforts.

In the view of Pavel (1993b, 29) regarding LSP phraseology, "the interphrasal combinations of terms and words in actual LSP discourse, it is at best

[15] `http://iate.europa.eu`

[16] According to a query performed on August 17, 2016.

given cursory consideration. More often that not, it is completely ignored on the mistaken assumption that LSP collocations are not unlike common language ones."

I agree with Bartsch (2004, 20), who claims that in a specialized context, terminology alone is not enough, since it is also necessary to master the collocations that are used with those terms: "in specialist communication, it does not suffice to acquire command of the relevant terminology, command of the domain-specific collocations is the key to mastery of specialist communication". Additionally, Cortes (2004) argues that the use of MWEs, in the forms of collocations and fixed expressions associated with particular registers and genres, are markers of proficient language use in that particular register or genre. Along a similar line of thought, L'Homme (2009, 238) asserts that "non-experts may have difficulties producing the correct verb, noun or adjective that is typically found in combination with a specific term". Hence, it is relevant to collect and study the collocations that are common in specialized texts, such as the ones found in FTA texts. This in turn can help characterize the collocations in a broader domain such as economics.

Certain multiword terms from a specialized subject field along with the phraseological units that include these terms can gradually be adopted in the general language (Zuluaga, 2002; Tecedor, 1998). Other times, a phraseological unit including a term, is transferred from one field to others. For Zuluaga, these terminological units keep their specialized status while at the same time exhibit the features of collocations. Zuluaga presents several Spanish examples such as *comercio internacional, interés compuesto* and *impuesto sobre la renta*. After their adoption, according to Tecedor, some terms amplify their meaning because they are used in general language. In Tecedor's words, "El trasvase de términos de la lengua común a las lenguas especiales produce una especialización o particularización del significado de los términos trasvasados".[17] (Tecedor, 1998, 129) She illustrates her study with Spanish idioms that originated in bull fighting, but which are now broadly used in other fields, namely politics, sport and show business.

In the view of Heid (2001), a linguist may be more interested in describ-

[17] The transfer of terms from common language to specialized language produces a specialization or particularization of the meaning of the transferred terms (My translation).

ing the collocational behavior of a set of lexical items, stating which noun or adjectives will select other lexical items, while the terminologist is more concerned with how a term and its collocate can be the denomination of a new concept.

It seems pertinent to pose the following question: What differentiates specialized collocations from multiword terms? Heid (2001) asserts that "[w]e are not aware of any broadly agreed standard for distinguishing noun-noun and adjective-noun collocations from multiword terms" Heid (2001, 788-789). In my opinion, both types of MWEs may sometimes exhibit the same features and the same units can be considered either as multiword terms or specialized collocations, depending on the theoretical stance of the researcher, as pointed out earlier with the example *preferential tariff treatment*.

Several authors have explored the relation between terms co-occurring with other lexical units that make up phraseological units in this kind of texts. (Picht, 1987, 1990a,b; Kjær, 1990; Budin, 1990; Galinski, 1990; Thomas, 1993; Pavel, 1993b; Clas, 1994; Bossé-Andrieu and Mareschal, 1998b; Cabré, 1999; Heid, 2001; Lorente, 2002a,b; Tognini-Bonelli, 2002; Orliac, 2004; Bevilacqua, 2004; Kjær, 2007; Fernández, 2008; Méndez, 2008). Some of them have called the phenomenon "LSP phraseology". LSP phraseology is at the crossroads between two disciplines, terminology and phraseology. The object of study of terminology is constituted by concepts and terms while phraseology is concerned with phraseological units. Therefore, much research into the phraseological units that include terms is still needed in these fields, a view that is supported by Tognini-Bonelli (2002) and Kjær (2007).

Studies so far have typically focused on the existence of collocations that surround terms and offer examples from dictionaries. Some of these authors (L'Homme, 1998; L'Homme and Bertrand, 2000; L'Homme, 2001; Heid, 2001; Orliac, 2004; L'Homme, 2009; L'Homme and Leroyer, 2009) have carried out studies and have proposed a theoretical and practical framework for the automatic extraction of these units from texts. The interest of these authors has been mostly of an applied nature, to acquire these units automatically or semi-automatically with the intention of improving the lexicons of specific domains, especially in technical texts and texts related to the stock market. For example, Cohen (1986) and Meynard (2000) compiled English-French termi-

nologies that include the specialized collocates of terms in the stock market and the Internet respectively. Similarly, Coxhead (2007) reports recent work toward the compilation of discipline-specific or technical vocabulary that includes collocation lists that can be used as pedagogical resources for several subject fields. Coxhead points out that there is still a need for more of these collocation lists across a wider range of academic disciplines.

Several terms have been employed by researchers to refer to the lexical units relevant for LSP phraseology. L'Homme (1998); L'Homme and Bertrand (2000) use the term "specialized lexical combination" to refer to the collocations that co-occur with terms and have studied the phenomenon in texts related to computers. Orliac (2008) uses the term "specialized collocation" and "specialized lexical combination" interchangeably. Heid and Weller (2008) also use the term "specialized collocation". Gozdz-Roszkowski (2011) calls these units "terminological bundles" while Kjær (1990) uses "terminological phrases".

The specialized phraseology of a specific field is a relevant linguistic subfield that should be accounted for by terminology and LSP studies because it offers insights on "the preferred way of saying things" (Gledhill, 2000, 1), in this case in the field of international trade. Other authors also agree that every specialized field has its particular and peculiar phraseology (Martin, 1992; Aguilar-Amat Castillo, 1994; Gaussier and Langé, 1994; Manning and Schütze, 1999; Oakes, 2009; Gozdz-Roszkowski, 2011). Stubbs (2002, 29) points out that "collocations may differ quite sharply in different text types. Many text-types are specialized in their uses of language, and no corpus can fairly represent every one of them."

Therefore, since phraseology is domain-specific, the knowledge of a language, whether it is the mother tongue or a foreign language, is not enough. It is also necessary to acquire a command of the particular and peculiar phraseology unique to a specific domain, which is normally acquired and commanded only by experts in such domain (Bartsch, 2004). Consequently, research in the fields of terminology and specialized phraseology can benefit from using a corpus to study terms and phraseological units, such as idioms and collocations, in these resources.

For Picht (1990a), LSP phraseology should be included in dictionary en-

tries, but in his opinion it is "unexplored territory", especially in the case of term banks. Pavel (1993a) also holds that there is an under-representation of these units in dictionaries, and later (Pavel, 1993b, 29) even claims that the phraseology that appears in specialized texts is "terminology in the making".

In a study on legal language, Gozdz-Roszkowski (2011, 125) comments that "terminological bundles represent highly technical and specialized vocabulary the occurrence of which is confined to the legal discipline".

Manning and Schütze (1999, 185-186) stress the relevance of multiword terms which they classify as a subclass of collocations:

> Terminological expressions or phrases refer to concepts and objects in technical domains. Although they are often fairly compositional (hydraulic oil filter), it is still important to identify them to make sure that they are treated consistently throughout a technical text.

Furthermore, Heid (1999, 242) provides a list of possible information that is conveyed by means of these lexical combinations and which is highly relevant for terminologists, terminographers, and specialized phraseologists and lexicographers:

> [O]ther collocations provide information that is relevant for definitions, hyponyms or subtypes, components or actions concerning the object or concept denoted by the "single word term" which is the base of the collocation.

L'Homme (1998) argues that in terminologically relevant collocations of the type Noun + Verb, the nominal component is usually a term and acts as the node or base of the collocation, while in Noun + Noun collocations, the node is a term and the collocate is the nominalization of a verb or an adjective. L'Homme also asserts that verbs and adjectives provide valuable information regarding the meaning of terms and that is why they should be taken into account by terminographers. Additionally, if a dictionary is supposed to cover in an exhaustive way the vocabulary of a domain, then the most representative among these combinations should also be listed in the dictionary (L'Homme, 2002; Bossé-Andrieu and Mareschal, 1998b). Regarding the inventory of collocations included in specialist dictionaries and term bases, Heid (1999, 241) argues that, in

> most terminological data collections, it is normal to have entries consisting of more than one word form: although few term bank models seem to have explicit devices to deal with collocations, some collocational material is present in most terminological data collections.[18]

This reinforces the view that even though some specialized collocations are covered, the criteria of inclusion or exclusion are rather arbitrary.

Heid emphasizes the relevance of collocations for terminology work and lists these patterns as the most interesting types of collocations for terminology work: Noun + Verb, Noun + Adjective and Noun + Noun which are divided into Noun + Preposition + Noun and Noun + Noun (in genitive) collocations as "multiword terms". (Heid, 1999).

Later, Heid (2001, 794) offers a relevant insight which is a central notion for this study. Heid holds that even though partial compositionality is quite often listed as a relevant criterion to define general language collocations, it plays a much less important role in specialized language. He also adds that "from the point of view of concept-based terminological description, one of the two components of the collocation must be a term for which a conceptual description is (or at least may be) available" (Heid, 2001, 788-789).

An example to illustrate this may be the term *customs duty*, which occurs 8 times in the COCA corpus, constituted by 520 million words, with a relative frequency of 0.017 times per million words and 311 times in the FTA corpus, with a relative frequency of 226 times per million words. Thus, it is one of the most frequent terms in the FTA corpus. In the *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English* (Alcaraz and Castro, 2007), in the entry for *Customs*, there is a subentry for *customs duty*. The information offered in the subentry includes a tag to identify the subject field, the Spanish equivalent and an English example of the term, with no collocational information given. Also, in the Dictionary of Banking and Finance (Russell, 2005) there is an entry for *customs duty*. This entry only offers the reader the phonetic transcription and the definition but no collocational information is provided. Thus, dictionary

[18] An example of a term portal which includes terms along with collocations is Termportalen from the CLARINO project http://www.terminologi.no

users, be it translators, LSP learners or technical writers, are left without collocational information about these terms and have to look somewhere else for clues regarding the verbs, nouns, adjectives and adverbs that idiomatically combine with the term in question. In the FTA corpus, the term *customs duty* co-occurs with 28 different verbs at position T -1, T -2 and T -3, where T is the term and the collocate is found one word (-1), two words (-2) or three words (-3) to the left of the verb.

Table 2.1: *Candidate specialized collocations of English term customs duty at position T -1 extracted with IMS CWB*

4	apply custom duty
4	impose custom duty
2	calculate custom duty
1	assess custom duty
1	collect custom duty
1	concern custom duty
1	eliminate custom duty
1	include custom duty
1	increase custom duty
1	refund custom duty

These verbs can be seen in Table 2.1 for the 10 verbs co-occurring at position T -1. Similarly, Table 2.2 presents the 17 verbs co-occurring at position T -2 and Table 2.3 presents the 12 verbs co-occurring at position T -3. In Table 2.2, the most frequent verb is *increase* with 10 occurrences, followed by *apply* with 9 occurrences, *eliminate* with 6, and *favour*, *pay* and *raise* with 3 occurrences each. This example suggests that there is a greater verbal diversity at position T -2. However, most of the verbs found at position T -1 do not occur at position T -2 or T -3. In the verbs occurring at position T -3, the term *customs duty* is part of a multiword term with another lexical item such as the adjectives *existing, new* and *applicable*.

2.13 Relevant specialized collocations for this work

The type of collocations relevant for this research can be either:

Table 2.2: *Candidate specialized collocations of English term customs duty at position T -2*

10	increase a custom duty
5	apply a custom duty
4	apply to custom duty
4	eliminate its custom duty
3	favour nation custom duty
3	pay any custom duty
3	raise a custom duty
2	assess the custom duty
2	dismantle its custom duty
2	eliminate all custom duty
1	adopt any custom duty
1	apply the custom duty
1	decide that custom duty
1	determine applicable custom duty
1	exceed the custom duty
1	favor nation custom duty
1	impose the custom duty
1	increase any custom duty
1	maintain any custom duty
1	mean the custom duty
1	reduce a custom duty
1	reduce its custom duty
1	reduce such custom duty

Table 2.3: *Candidate specialized collocations of English term customs duty at position T -3*

3	accelerate elimination of custom duty
3	adopt any new custom duty
3	eliminate its respective custom duty
3	increase any existing custom duty
1	be subject to custom duty
1	decide to apply custom duty
1	evidence payment of custom duty
1	grant waiver of custom duty
1	involve exemption from custom duty
1	pay the corresponding custom duty
1	stage rate of custom duty
1	waive otherwise applicable custom duty

1. two or more consecutive words as the n-grams relevant in NLP applications, or

2. two or more non-consecutive words, i.e. the collocate can be located several words to the right or to the left of the node.

For example, in this clause from the FTA corpus there is a verbal collocation: *as if the safeguard measure had never been applied*. In this clause, the term *safeguard measure* enters into a collocation with the verb *apply* which occurs four words to the right of the term, or as expressed in corpus linguistics terms, in position n + 4.

There are two of the factors of collocability mentioned above that stand out as criteria for the lexical units relevant to my research. The first one is frequency of co-occurrence that holds among two or more lexical items within a specific subject field in comparison to another field. The second one is the combinatory restrictions that two or more lexical items exhibit.

In the present work adjacency is not a definitive feature to attest collocability. Consequently, I have set an extension to the window to look for specialized collocations: the collocate can occur in a window of three words to the left or to the right of the node word. Sinclair et al. (1970) found that most collocates are found within a span of five tokens on either side of the node word. A high percentage of terms is made up by two or three words. With an extension of three tokens on either side, an important amount of multiword terms and their collocates span five tokens.

2.14 Definition of specialized collocation

In Section 2.1, it was stated that collocations can be unpredictable word combinations and Section 2.11 discussed how the node lexically determines the other lexemes that can co-occur with it. The latter section also discussed the definition of a multiword expression and its subsets. For the purposes of my PhD research, a definition of a specialized collocation is hereby proposed:

> A specialized collocation is a type of multiword expression composed of at least a term that serves as the node of the collocation.

> The collocates of the term can be nouns, verbs, adjectives or adverbs in a direct syntactic relation with the node and they can be found either adjacent to the term or within one or more intervening words. The combination of the term and the collocates constitutes a lexical combination that can be unpredictable and semi-compositional and have an internal and statistical tendency of preference.

Figure 2.3 represents a specialized collocation, with the lexical words that form collocates and the term that can form the node. Figure 2.4 illustrates the notion of specialized collocation when the term has the object role in relation to a verb. It is meant to indicate that in a corpus of specialized texts, any term, whether it is composed by one or more lexemes, may enter into a specialized collocation with a restricted set of other nouns or with other adjectives, adverbs or verbs that are in direct syntactic relation to the noun acting as the term. This lexical relationship held among the constituents of the specialized collocation adds linguistic features to the term that serves as the node of this type of collocation. In specialized texts, the same term can enter into a specialized collocation with several lexical units.

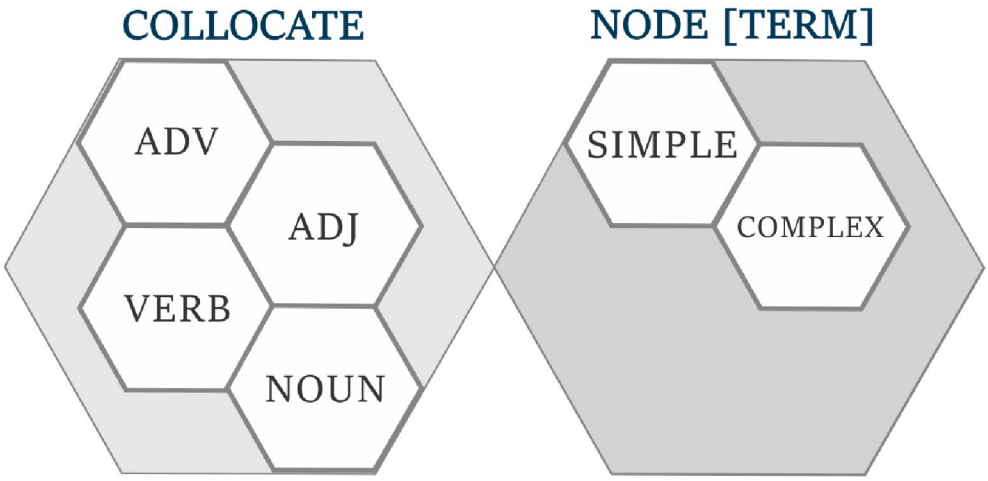

Figure 2.3: *A diagram representing a specialized collocation, with the lexical words that form collocates and the type of terms that can form the nodes*

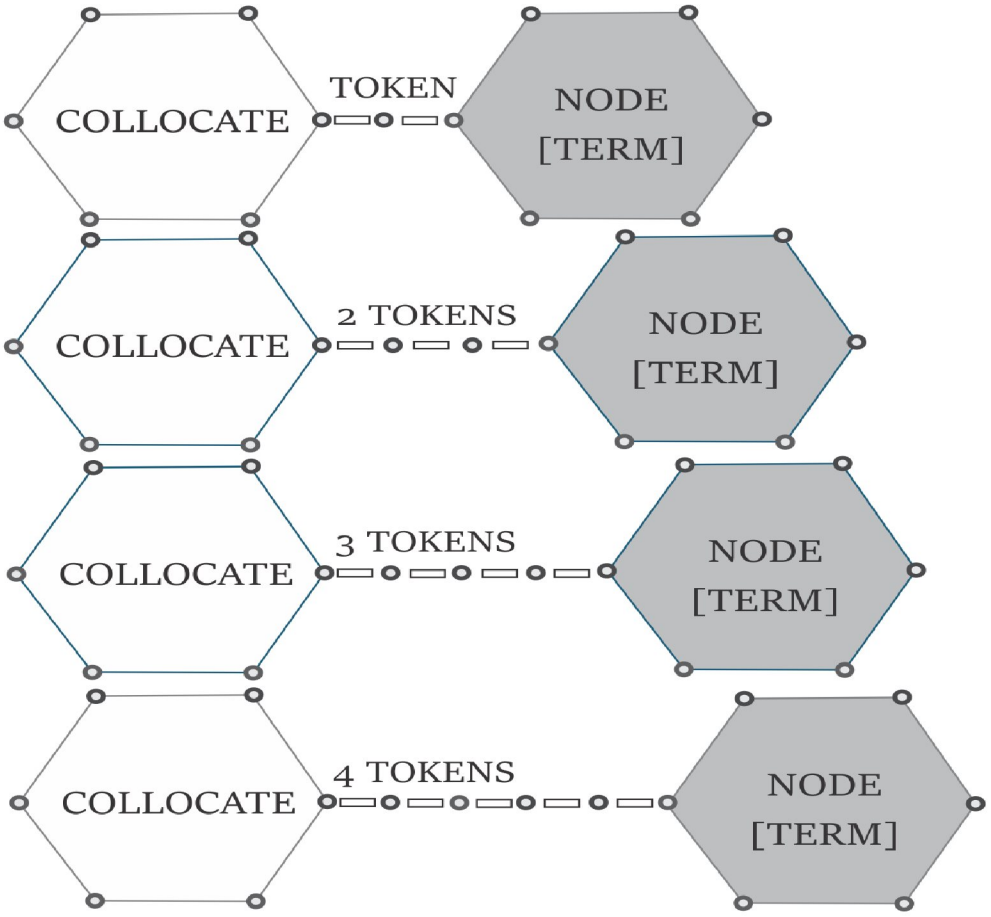

Figure 2.4: *A diagram representing specialized collocations when the term has the object role in relation to a verb*

2.15 Criteria for the selection of a specialized collocation

The criteria that have been established to consider a lexical unit as a specialized collocation are listed as follows:

- the node of a specialized collocation is a simple or a complex term, i.e. a one-word term or a term composed of two or more words whose termhood is evidenced from their use in the context of a specialized subject field;
- the collocate of a specialized collocation can be any word from the

lexical categories, namely verb, noun, adjective or adverb ending with the suffix *-ly* for English and *-mente* for Spanish. This morphological restriction is aimed at detecting only lexical adverbs.

- the collocate must be in a syntactic relation with the node.
- the collocate has to be found in a window of maximum five tokens to the left or to the right from the node term, in harmony with Sinclair et al. (1970);
- this combination of a term with a collocate has to occur at least once in the FTA corpus.

2.16 Research questions

This theoretical background gives rise to some research questions that are now presented:

1. What lexical and terminological information do specialized collocations provide about specialized texts such as the ones found on the FTA corpus?

2. Which morphosyntactic and semantic features are inherent to the constituents that serve as collocates of terms in specialized texts such as the ones found on the FTA corpus?

3. What information do the linguistic features of the constituents, i.e. the node and its collocates, that make up specialized collocations offer in FTA texts?

4. Which grammatical categories are or can be used to tag specialized collocations in the encoding of language resources?

5. How can the notion of specialized collocation be incorporated into the theory of terminology while using a corpus linguistics methodology?

6. How can specialized collocations be represented in term bases and computational lexicons in such a way that the data can be reusable, scalable and interoperable?

2.17 Basic assumptions

For this research, I take the following claims:

- Despite their widespread presence and use, specialized collocations are not covered systematically in specialist dictionaries, neither in terminological resources nor in human- or machine-readable dictionaries (Moon, 2008).
- Specialized collocations that appear in FTA texts have the same linguistic characteristics as general language collocations but their terminological value is evidenced by their use in context, where such value is activated (Cabré, 1999).

Hence, the inclusion in language resources of the collocates of a term, such as the ones found with the term *customs duty* would allow to create more comprehensive resources, whether they are meant to be read by human users, such as translators or LSP language learners, or by a machine in a MRD.

To sum up, this chapter has shown that the linguistic phenomenon of collocation is present both in general as well as in specialized texts. Yet, even though specialist dictionaries and term bases include terms, these resources do not habitually include the collocational relations of these terms. For terminological theory, nouns are considered as the prototypical lexical unit for the study of concepts, as can be seen on any terminographical repertoire, where nouns are much more frequent than verbs and adjectives (Cabré, 1993). However, researchers from the field of terminology have also explored other lexical units that appear in a syntactic relation with these nouns as a means to expand the specialist knowledge that can be conveyed through these units.

The next chapter offers a description of Free Trade Agreements, how they affect international trade and the institutions that are involved in the enactment of such agreements.

CHAPTER 3

Free Trade Agreements

> The Republic of Colombia (hereinafter referred to as "Colombia") on one part, and the Republic of Iceland, the Principality of Liechtenstein, the Kingdom of Norway and the Swiss Confederation (hereinafter referred to as "the EFTA States") : [...]
>
> AIMING to create new employment opportunities, improve health and living standards and to ensure a large and steadily growing volume of real income in their respective territories through the expansion of trade and investment flows, thereby promoting broad-based economic development in order to reduce poverty; (Free Trade Agreement between The Republic of Colombia and the EFTA States, 2008).

3.1 Introduction

Free Trade Agreements (FTAs) regulate the trade of goods and services among nations throughout the world. FTAs are contractual documents, thus belonging to the legal domain. The negotiators of these agreements are concerned with several key subjects that have to be negotiated and then written in the agreements. These subjects include the technical barriers to trade, government procurement, intellectual property rights, national treatment,

Most Favoured Nation status, dispute settlement, antidumping and customs valuation (WTO, 2015).

For example, the multilateral and supranational WTO agreement includes several sections on the following topics: agriculture, the application of sanitary and phytosanitary measures, textiles and clothing, technical barriers to trade, trade-related investment measures, preshipment inspection, rules of origin, import licensing procedures, subsidies and countervailing measures and safeguards.

Regarding their level of specialization, FTAs are specialized texts aimed at expert to expert communication (Spang-Hanssen, 1983; Pearson, 1998). FTAs are specialized official documents that set the norms for the trade of goods among two or more parties and are thus a rich repository for the terminology and phraseology that is used in different fields of business activity throughout the world. As regards scientific domain, FTA texts belong to the field of international trade, which is a branch of macroeconomics which belongs to the broader field of economics. From another viewpoint, FTA texts are part of international law, which stems from business law. In turn, this field is a subfield of the broader field of law.

FTA texts provide a relevant context for the study of specialized collocations because they include terms from a variety of disciplines besides law itself. Thereby, they offer challenges for translators and other language professionals who work with texts related to the above-mentioned disciplines.

This chapter describes the trade agreements included in the corpus and associations or blocs of countries which are signataries of these agreements.

3.2 The parallel corpus of Free Trade Agreements

Most of the agreements included in the FTA corpus (the only exception being the FTAA) have been officially signed and ratified by several national and supranational organizations, countries and multilateral bodies in the last two decades. Specifically, these entities are the World Trade Organization (WTO), the European Free Trade Association (EFTA), the European Union

(EU), the Free Trade Area of the Americas (FTAA), Canada, the United States of America, Mexico, the Caribbean Community, the Dominican Republic, Colombia, Peru and Chile. Therefore, the corpus includes texts from different language variants, as it comprises texts written in English from the United States and Canada, European Union texts, as well as Spanish from many Latin American countries and blocs of countries. The FTAA agreement has not yet been ratified, thus, it remains as a draft version since November 2003 and the parties have completely stopped discussing the negotation due to political differences (Schott, 2005).

These FTA texts were drafted in English and Spanish by the negotiating teams of the parties involved. Since the FTAs in principle have equal legal status, they are not considered as translations of each other. However, they were produced in different forms: some of the agreements were written in English and then translated into Spanish or vice versa. In other cases, as the negotiation advanced, normally using English as a lingua franca, the teams of free trade experts wrote a bilingual draft (Šarcevic, 2000), with each team writing in its mother tongue. Alexander (1999, 1470) claims that "English is optional or necessary for international business transactions, among non-native speakers."

Whether it is Norwegian salmon, Colombian coffee, Peruvian avocado, Chilean wine or any other product, the trade of these goods among nations is regulated by a trade agreement. Trade is a very important human activity that has emerged since the beginning of civilization. It has promoted not only economic exchange, but also cultural and political integration among nations. Nowadays, in an allegedly globalized world, trade continues to be an activity of great relevance for economic development and several supranational organizations have developed a framework to regulate international trade.

The growth of world merchandise exports has been exponential over the last decades, as illustrated in Figure 3.1. In 1948, as the world was recovering from the Second World War, merchandise exports amounted to USD 59 billion. Twenty five years later, it had increased to USD 579 billion. Then, in 2010, according to WTO data, world merchandise exports amounted to USD 14,851 billion, and the European Continent alone had a share of 37.9%.

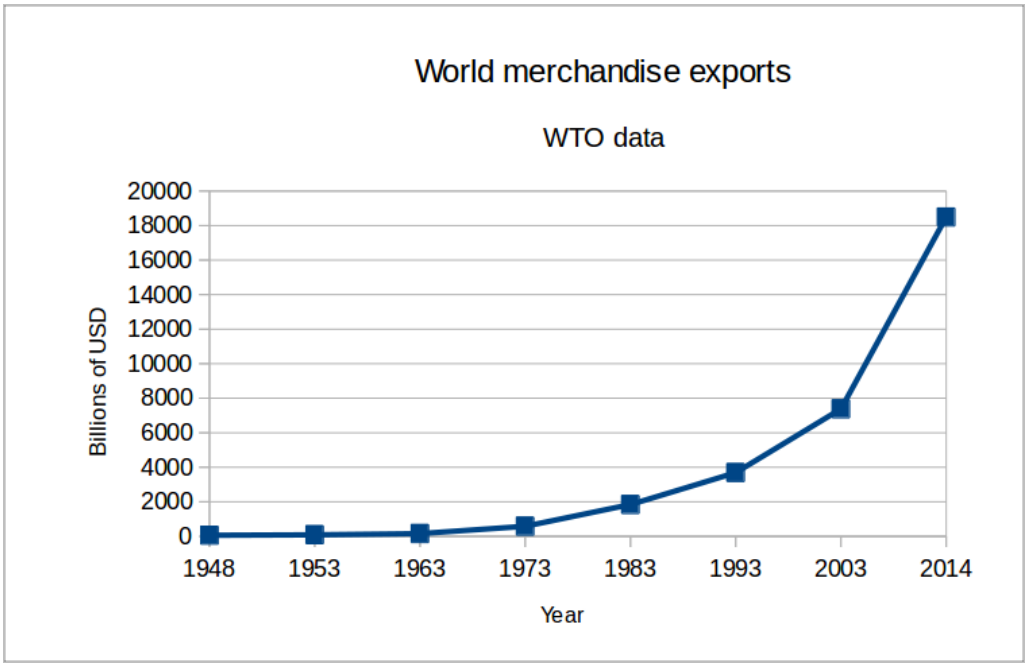

Figure 3.1: *World merchandise exports in billions of USD from 1948 to 2014 according to WTO data*

In 2014, exports peaked USD 18,494 billion worldwide, with Europe as the stronger party, and within Europe, Germany, Netherlands, France and Italy were the main players involved in export activities. Thus, as of 2014, the main exporting region is Europe with 36.8% of the total share, followed closely by Asia with 32% whereas North America, which comprises the USA, Mexico and Canada, comes third with 13.5% of the share of exports.[19]

Worldwide trade policies are regulated mainly by three supranational economic organizations, namely, the World Bank, the International Monetary Fund (IMF) and the WTO. By far, the WTO is the youngest of the three (Narlikar, 2005). These supranational entities are described more specifically in Section 3.3.

FTAs were preceded by other trade regulation systems. Shortly after the Second World War, there was an initiative to promote free trade globally as a strategy to foster economic development. This initiative was the Interna-

[19] For detailed data, see `https://www.wto.org/english/res_e/statis_e/its2015_e/its15_world_trade_dev_e.htm`

tional Trade Organization (ITO). However, the ITO never entered into force. Then, in 1947, the General Agreement on Tariffs and Trade (GATT) was established. This agreement set the norms and regulations for the growth of trade among nations in the postwar period. The GATT lasted almost 48 years until it was absorbed into the WTO.

Some definitions related to trade seem relevant at this point. Free trade refers to the "absence of government policies designed to regulate international trade, especially import limitations such as tariff or quotas" (Moon, 2000a, 574). A free trade area is defined as "a regional bloc made up of two or more countries which agree to liberalize their bilateral trade, while maintaining their restrictions on trade within third countries" (Nicolaides, 2000, 575). Last, a trade agreement is defined by Moon (2000b, 1570) as

> any initiative involving the cooperation of two or more governments to facilitate or regulate trade between their economies. It can take several forms: bilateral, minilateral (or regional) and multilateral (or global)".

The corpus data for this research includes all these types of agreements.

The nature of FTA texts is appropriate for the presence of interdisciplinary terminology from the fields of law (e.g. *customs legislation, procuring entity*), economics (e.g. *unilateral tariff reduction*, *preferential tariff treatment*), subdomains involved in the goods subject to trade (e.g. *freight brokerage services*, *on-line data processing and information*) or specific products (e.g. *fine animal hair, textured polyester filaments*).

According to Gamero (2001), prestigious international entities such as UNESCO, offer validated criteria to classify specialized texts such as the different scientific and technical fields. Therefore, from an onomasiological perspective, the texts containing FTAs can be delimited conceptually by using the UNESCO nomenclature.[20] Under heading 53, this nomenclature lists Economic Sciences: 5307: Economic theory, 5307.12 International Trade theory, 5310 International economics: 5310.04 International business, 5310.07 International investment and, perhaps the most relevant for FTAs, 5310.09 International trade relations.

[20] `http://unesdoc.unesco.org/images/0008/000829/082946eb.pdf`

3.3 Supranational entities involved in world trade

The following subsections list several supranational entities that have engaged in free trade. All of these institutions are not included in the FTA corpus which was used for this work. They are listed here because they represent Latin American countries most of which are represented in the FTA corpus or because they represent alternative projects involved in free trade. They offer a glimpse of the most representative worldwide associations that engage in free trade with other entities.

3.3.1 World Trade Organization

The WTO was established in 1995 by 128 member countries (Narlikar, 2005). Its headquarters are located in Geneva and as of July 2016 there are 164 member countries.[21] The policies of this body have served as a model for the other FTAs. In other cases, some developed countries have resorted to signing individual FTAs with developing countries when the WTO regulations are not convenient to their interests.

3.3.2 European Union

The institutions that decades later led to the European Union (EU) were established shortly after the Second World War. Today, there are 28 member countries and other countries are pending to enter into this treaty. The EU establishes economic, social and political norms for its member countries. It was preceded by other pioneering entities such as the European Economic Community. The Maastricht Treaty or the Treaty on European Union was signed on February 7, 1992 and it officially created the EU.[22]

[21] http://wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm

[22] Retrieved from Encyclopædia Britannica http://www.britannica.com/EBchecked/topic/196399/European-Union-EU/224464/The-Maastricht-Treaty

3.3.3 European Free Trade Association

The European Free Trade Association (EFTA) is a regional association that includes Iceland, Liechtenstein, Norway, and Switzerland. It is operative since 1960 but several of the founding countries left the EFTA and joined the European Economic Community (EEC) instead. Since 1994, this bloc of countries also implemented a free trade zone with the EU.[23]

3.3.4 Free Trade Area of the Americas

The Free Trade Area of the Americas (FTAA) is a proposed trade agreement covering all countries in North, Central and South America and the Caribbean, with the exception of Cuba, modeled after the NAFTA and WTO agreements. The FTAA has not entered into force and the 2005 deadline originally proposed was not met.[24] A major obstacle has been the negotiation between Latin American and North American countries regarding the agricultural subsidies paid to farmers in the United States and Canada. These subsidies set an obstacle for less developed countries to compete against lower prices for agricultural products coming from developed countries.

3.3.5 Andean Community (CAN)

In 1969, several Andean countries established an agreement to promote trade and also to foster industrial, agricultural and social cooperation, among other aspects. The original member countries were Bolivia, Colombia, Ecuador, Peru, Venezuela and Chile. Later, Venezuela and Chile withdrew from the group and Peru suspended its membership for some years but later rejoined. The headquarters of the CAN are located in Lima, Peru. Beginning in July 2004, the CAN countries implemented a free trade area.[25]

[23]Retrieved from Encyclopædia Britannica `http://www.britannica.com/EBchecked/topic/196231/European-Free-Trade-Association-EFTA`

[24]Retrieved from Encyclopædia Britannica `http://www.britannica.com/EBchecked/topic/1015476/Free-Trade-Area-of-the-Americas-FTAA`

[25] Retrieved from Encyclopædia Britannica `http://www.britannica.com/EBchecked/topic/744592/Andean-Community`

3.3.6 Caribbean Community (CARICOM)

Following the Caribbean Free Trade Association (CARIFTA) that was established in 1968, the CARICOM agreement was established in 1973. The entity is concerned with economic integration and development planning of the involved parties. Its headquarters are based in Georgetown, Guyana. Its member countries include Antigua and Barbuda, The Bahamas, Barbados, Belize, Dominica, Grenada, Guyana, Haiti, Jamaica, Montserrat, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, and Trinidad and Tobago. Other countries have an associate member status. These countries are Anguilla, Bermuda, the British Virgin Islands, the Cayman Islands, and the Turks and Caicos Islands. Other countries with coasts on the Caribbean only maintain an observer status: Aruba, Colombia, the Dominican Republic, Mexico, Puerto Rico and Venezuela.[26]

3.3.7 Mercosur

This entity is known in Spanish as *Mercado Común del Sur* 'Common Market of the South'. The Mercosur is a South American initiative for economic integration created in 1991 by the Treaty of Asunción. It is composed by four countries, namely Argentina, Brazil, Uruguay and Venezuela, after Paraguay was suspended in 2012.[27] Mercosur was preceded by the Latin American Free Trade Association (1960) and the Latin American Integration Association (1980).[28] Other countries currently have an associate member status: Bolivia, Chile, Colombia, Ecuador, Guyana, Peru and Suriname.

3.3.8 Unasur

Unasur stands for *Unión de Naciones Suramericanas* 'Union of South American Nations'. It is aimed at attaining South American integration as a continuation of the CAN and Mercosur initiatives. It was constituted in 2008

[26] Retrieved from Encyclopædia Britannica `http://www.britannica.com/EBchecked/topic/95814/Caribbean-Community-CARICOM`

[27] Retrieved from Encyclopædia Britannica `http://global.britannica.com/EBchecked/topic/375563/Mercosur`

[28] Retrieved from Encyclopædia Britannica `http://www.britannica.com/EBchecked/topic/375563/Mercosur`

and its headquarters are located in Quito, Ecuador. Its member countries are also part of the CAN, as in the case of Bolivia, Colombia, Ecuador and Peru or are members of Mercosur, in the case of Argentina, Brazil, Paraguay, Uruguay and Venezuela. Other members of Unasur are Chile, Guyana and Suriname, while Mexico and Panama have an observer status.[29]

3.3.9 ALBA

The *Alianza Bolivariana para los Pueblos de Nuestra América* (ALBA) 'The Bolivarian Alliance for the Peoples of Our Americas' is based on the political ideology of Simón Bolívar, the 19th century Andean independence leader who dreamed about the idea of a great Latin American homeland. The ALBA initiative arises from the leftist governments of Venezuela and Cuba and was founded in late 2004 as an alternative to other FTAs, allegedly conceived by and directed from the United States. Its nine member countries are Antigua and Barbuda, Bolivia, Cuba, Dominica, Ecuador, Nicaragua, Saint Vincent and the Grenadines, Venezuela and Saint Lucia, while Suriname has a guest country status and Haiti has an observer status. One of the official objectives of the ALBA countries is to create a common currency, the Sucre.[30]

3.3.10 Alianza del Pacífico

Over the last two decades, several of the Latin American countries have been intensely participating in the establishment of FTAs with other nations in an effort to expand their economies and gain broader market access. Chile has signed 22 FTAs, Mexico has signed 19, Peru has signed 18 and Colombia has signed 14.[31] These four countries signed the Pacific Alliance in 2014, in an attempt to strenghten their capacity with the aim of acting as one bloc to trade with other blocs or stronger countries such as the US, the EU and

[29] Retrieved from Encyclopædia Britannica `http://global.britannica.com/EBchecked/topic/1496583/UNASUR`

[30] Retrieved from Encyclopædia Britannica, `http://global.britannica.com/EBchecked/topic/1271045/Bolivarian-Alliance-for-the-Peoples-of-Our-America-ALBA`

[31] `http://www.semana.com/economia/articulo/acuerdos-comerciales-se-cierra-un-ciclo/359157-3`

China. The Pacific Alliance entered into force on May 1, 2016. Combined, the economies of the four countries that integrate the Pacific Alliance, with a population of 210 million inhabitants, represent 36% of Latin America's GDP and as a bloc would be the world's ninth largest economy. Other Latin American countries are interested in joining this pact, such as Costa Rica and Panama.

3.4 Advantages and disadvantages of free trade agreements

FTAs offer several advantages to the countries that enter into this kind of agreements. In 1776, the Scottish economist Adam Smith published his famous book *The Wealth of Nations*, where he advocated the advantages of economic liberty where free trade was an important component (Irwin, 2009). According to the Australian Department of Foreign Affairs and Trade, trade liberalization and tariff reductions allow countries abiding by these agreements to engage in business with the other parties with less restrictions (APEC Study Centre, 2001).[32] This way, countries can expand their access to other markets and thereby produce and export more goods instead of relying solely on the national market, with its inherent limitations. This in turn can generate additional jobs or at least absorb the jobs lost in other economic sectors that are affected by the implementation of FTAs. Besides, countries abiding by FTAs can in theory optimize their economy by specializing their production. As a consequence, they are able to focus their export efforts on the economic sectors in which they have a relative strength compared to other countries.

In contrast, critics of FTAs point to the fact that jobs are lost in some economic sectors where a country does not have a relatively stronger position compared to another country. When a country has to abide by a FTA, its economy becomes more dependent on trade partners. Thus, it can be affected by the economic fluctuations of another economy that receives its exports. In the last two decades, many developing countries have been entering into

[32] `http://www.dfat.gov.au/publications/aus_us_fta_mon/`

FTAs with developed countries (Crump, 2007). This way, developing countries, which typically export commodities and raw materials, gain access to trade with markets. However, they are also affected by scale economies that produce many goods at a cheaper price, while these raw materials are processed in developed countries and then the finished products are imported to less developed countries with much higher prices. This added new competition can destroy jobs in the less developed countries, which are condemned to import, with the result that their local industry is progressively diminished.

Developing countries tend to export commodities and raw materials. This is what for a long time happened with Colombian coffee and still happens to a lesser degree. Coffee beans were exported to North America and Europe without any added value and then reimported as a finished product. This way, local farmers would only get a tiny fraction of the total income of the coffee industry, whereas a handful of multinational companies would get hold of the vast majority of wealth. These issues are discussed in detail in Graham (2004) and Irwin (2009).

This chapter has offered a description of the institutions that regulate world trade policies, the various FTAs that are included in the FTA corpus and has offered a glimpse of the complex geopolitical and economic interests that affect these policies and their implementation.

The following chapter describes how this texts have been integrated into the corpus and the method and materials used to carry out the research using the FTA data.

CHAPTER 4

Material and Methods

4.1 Material

Existing corpora are not always pertinent to address the particular questions that a researcher aims to investigate. As a consequence, when a linguist needs data that fit a very specific and restricted purpose in terms of text genre, time period or content of the data, with restrictions related to a narrow domain or subdomain, the researcher often finds that no corpus is available and necessarily has to build his/her own corpus to fit his/her particular needs. Thus, because of the particular needs of this project, and since there was no corpus available dealing exclusively with FTA data, it was necessary to first gather texts from several FTAs and then to process the data to compile a parallel corpus with FTA texts.

To carry out the research, test the hypotheses set in Section 1.3 and attain the objectives set in Section 1.4, a specialized parallel corpus with FTA texts in two languages has been compiled. The aim of building the corpus is to study the terms of legal and economic domains in this genre and in particular the specialized collocations that include these terms.

The examples were extracted semi-automatically from a parallel corpus

of English and Spanish official texts from FTAs comprising approximately 1.5 million words in each language (Patiño, 2013). These examples reflect the usage of specialized collocations by experts in the subject field of international trade. The content of the corpus will be described further in Section 4.2.2.

Additionally, for contrast purposes, a collection of reference lexical resources was also compiled. These lexical resources that were used as reference are composed by four bilingual and two monolingual dictionaries from the specialist domain of economics and a subdomain of economics, namely, international trade. These resources will be described beginning in Section 4.2.4 until Section 4.3.

Two approaches were employed to carry out the research. At the macro level, a corpus-driven approach was used to investigate the frequency and representativeness of the specialized collocations found in these texts, by using tools which employ a combination of both methods, i.e. statistical techniques and linguistic rules. Later, at the micro level, a comparison of specialized collocations from English and Spanish texts found on the FTA corpus was made. Then, an analysis was carried out in order to identify the linguistic clues that these lexical units provide which is useful to establish a domain-specificity. This information can be used to model the metadata and linguistic annotation for processing these lexical units.

4.2 Methodology and workflow

Figure 4.1 illustrates the main steps that make up the methodological workflow that was used to carry out the study of specialized collocations:

1. *FTA corpus construction.* In this step, the data was preprocessed and prepared to be aligned at the sentence level for the two languages to create a parallel corpus. This stage is described in Subsections 4.2.4 and 4.2.5.

2. *Candidate term and collocation extraction.* Several software tools were used to extract the candidate specialized collocations (CSC) of the identified terms. The tools are described in Section 4.4.2.

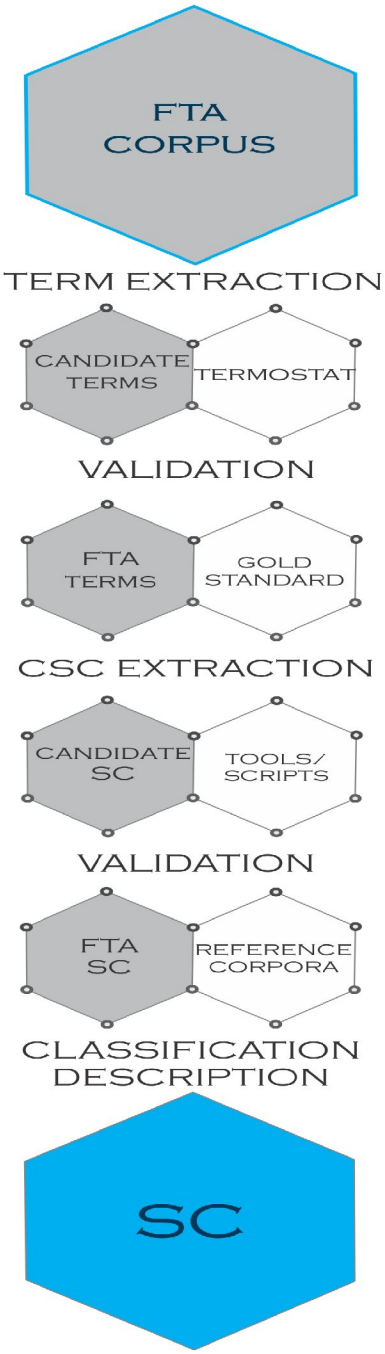

Figure 4.1: *Methodology workflow for extraction of CSCs*

3. *Validation of candidate terms and CSCs.* Reference corpora, dictionaries and glossaries were used to perform the validation. The aim of this

step was to find out which terms and CSCs appear both in the FTA corpus and the other lexical resources, and which ones occur only in the latter resource.

4. *Proposal for the representation of specialized collocations in computational lexical resources.* The Lexical Markup Framework (LMF), described in Section 6.1.1, was used to represent specialized collocations in such a way that the data could be reused in machine-readable and human-readable dictionaries. In this context, representation means a formal way to annotate specialized collocations using XML code. It can be used to display the data in a form readable for humans as well as for machines. Furthermore, the formal definition of such representation includes the possibility that the data will be reusable, interoperable and mergeable with existing lexical and terminological resources. More details on this subject are offered in Section 2.5.

4.2.1 Construction of the aligned FTA corpus

Specialized corpora are useful tools to investigate in context how language operates in a restricted domain (Flowerdew, 2004). A parallel corpus is understood here in the same sense as the one employed by McEnery and Xiao (2007) and Aijmer (2008), i.e. as a collection of source texts and their translations, aligned at the sentence level. Johansson (2007) prefers to use the term *translation corpora* to avoid confusion between *comparable* and *translation corpora* with the less precise term *parallel corpora*. Since the FTA corpus is made up of English and Spanish aligned texts that are not necessarily translations of each other, but not in the sense of comparable corpora, and both texts have in principle equal legal status, I avoid the term translation corpora.

The corpus is specialized because it exclusively contains texts from a specific domain, in this case FTAs. This facilitates the study of the terms and the collocations that include these terms found in this type of texts (Koester, 2010). An advantage of using a corpus is that the terms and their collocates can be found in their context and not as a mere list of disconnected items.

As mentioned above, FTAs are specialized texts for specialist communication (Spang-Hanssen, 1983; Pearson, 1998). The nature of these texts results in the presence of terminology from several domains, especially the domains of law as illustrated in these English examples: (*arbitral tribunal*), economics (*issuance of bills of lading*) and the fields concerning the goods subject to trade (*multimodal transport operator*) or as evidenced in the Spanish texts where there are some terms used in law and international trade: *derechos antidumping* 'countervailing duties', *procedimientos judiciales civiles* 'civil judicial proceedings', *derecho internacional consuetudinario* 'customary international law' and *arancel aduanero* 'customs duty'.

4.2.2 Description of the FTA corpus

Groom (2007) lists two features of a specialized corpus which are necessary for it to be adequate. A specialized corpus should

1. be constructed in a way that it can provide empirical support for the researcher's claims about the particular language variety that the corpus aims to represent and

2. be amenable to the particular method of investigation that the researcher wishes to apply.

I will argue that the FTA corpus meets both of these requirements. First, the FTA corpus aims to represent FTA texts in English and Spanish from countries in Europe and the Americas where these languages are the official language(s) used by government bodies. Second, the data are annotated with parts of speech and lemmatized. This linguistic annotation allows the user to perform queries based on morphosyntactic patterns. Besides, the selected software tools were chosen or in some cases prepared to process the data and compute statistics on the lexical units and their distribution and frequencies.

The English-Spanish FTA corpus consists of 233 XML source files in each language. The corpus contains approximately 1,370,000 million words in the English section and 1,483,000 million words in its Spanish counterpart. Compared to the "mega-corpora" being built nowadays (e.g. the COW – Corpora from the Web with 16.8 billion tokens in the English Section (Schäfer, 2015)

or the Global Web-Based English (GloWbE) with 1.9 billion words from 20 English-speaking countries[33]) or other corpora comprising up to 500 million words, a corpus with 1.5 million words is relatively small. There is a general agreement that a small corpus contains from 20,000 to 250,000 words (Flowerdew, 2004). However, due to the increased size of corpora in the last few years, for Koester (2010), a written corpus is considered small when it holds less than five million words. Also, certain types of corpora such as spoken or multimodal corpora are much smaller than written corpora.

Flowerdew (2004) argues that specialized corpora are quite useful to perform research on specific types of academic and professional language. One important reason is because specialized corpora include terms and phraseology that are used in specific domains. Flowerdew offers a list of parameters to define a corpus as specialized:

1. Specific purpose for compilation,
2. contextualization,
3. size,
4. genre,
5. type of text / discourse,
6. subject matter / topic and
7. variety of English.

According to these criteria, the FTA corpus is specialized because of its size, genre, subject matter, topic and language variety. For such a specialized domain as FTAs, 1.5 million words is considered to provide sufficient data for the study of terms and collocations typical of this text genre. The texts included in the FTA corpus comprise the Spanish and English versions of the agreements signed by several countries or blocs of countries, as shown in Table 4.1.

[33] `http://corpus2.byu.edu/glowbe/`

The website of the Foreign Trade Information System of the Organization of American States (OAS)[34] lists 51 FTAs signed and ratified in English and Spanish. The full text of twelve of these agreements has been aligned and included in the FTA corpus, besides the EU texts and the pending FTAA draft. By number of words, four of the FTAs account for half of the data: First, the EU texts alone, with 196,494 words, account for 14.3% of the data. Then comes the NAFTA agreement comprising 182,990 words, which represent 13.3% of the data. Third, we find the draft FTAA agreement with 179,747 words, which means 13.1% of the data. In the fourth position, with 160,091 words, we find the agreement signed by Colombia and the USA, representing 11.6% of the data. In terms of the date when the data included in the FTA corpus was officially approved and published, the data spans almost two decades, from 1992 to 2011, with 2003 as the average year for its publication.

The oldest texts are the EU and NAFTA agreements, both published in 1992 and the most recent FTA text was published in 2011, namely the FTA signed by Colombia, Peru and the EU and operable since August 1, 2013. Obviously, there is FTA data published before these years but it was not taken into account when building the corpus.

4.2.3 Copyright issues

When building a corpus, researchers often have to take into account the copyright and legal terms for the use of the data. The data for the FTA corpus was downloaded from two sources. The first was the webpage of the Foreign Trade Information System of the Organization of American States (OAS).[35] The second source was the European Union Law (EUR-LEX).[36] EUR-LEX is an official website that provides free access to European Union law and other public documents. The documents are available in 24 official languages of the EU. As of May 2016, EUR-LEX stores more than 3 million

[34] `http://www.sice.oas.org/agreements_e.asp`
[35] `http://www.sice.oas.org`
[36] `http://eur-lex.europa.eu/en/treaties/`

Countries	English words	Percentage	Year
Canada – Peru	69,930	5.10	2008
CARICOM – Dominican Rep.	9,458	0.69	1998
CARIFORUM – EU	51,483	3.76	2008
Chile – EU	34,381	2.51	2002
Chile – Australia	64,841	4.73	2008
Chile – EFTA	16,671	1.22	2003
Chile – USA	86,112	6.28	2003
Colombia – Peru - EU	121,003	8.83	2011
Colombia – USA	160,091	11.68	2006
Colombia – EFTA	69,569	5.08	2008
EFTA - Peru	24,201	1.77	2010
EU	196,494	14.34	1992 / 2007
FTAA (draft)	179,747	13.12	2003
Mexico – EFTA	14,862	1.08	2000
NAFTA	182,990	13.35	1992
World Trade Organization	88,548	6.46	1994
Total	1,370,381	–	–

Table 4.1: *Components of the English-Spanish section of the FTA corpus*

documents with texts dating back to 1951. This database is updated on a daily basis and every year around 12,000 documents are included.

With regard to copyright issues for building the FTA corpus, two segments taken from two of the FTA guarantee the availability of the data for academic research. First, the OAS website includes this statement:

> The General Secretariat of the Organization of American States (GS / OAS) holds copyright on the information available on this website, unless otherwise stated. Copyright in any third-party materials found on this website must also be respected. Anyone may use or reproduce any information presented on this website for educational and other non-commercial purposes, provided that the use of such information is accompanied by an acknowledgement of the GS/OAS as the source.

In the case of EUR-LEX, this website states the following regarding data use and reproduction:

> Except where otherwise stated, downloading and reproduction, for personal use or for further non-commercial or commercial dissemination, of legal texts and other documents publicly available on the EUR-Lex website are authorised provided appropriate acknowledgement is given as follows: '© European Union, http://eur-lex.europa.eu/'

Therefore, since the objective of the corpus is academic and non-commercial, to the best knowledge of the author, the corpus does not infringe the copyright laws.

4.2.4 Corpus pre-processing

Most of the original files were downloaded as PDF, HTM or RTF files. Some of the PDF files were scanned copies of the hard copy. Therefore, they were first processed with Abbyy Fine Reader 9.0,[37] a software for Optical Character Recognition to convert the files to MS Word DOC format. Also, for ease of processing, the longer FTA files were segmented semi-automatically into several smaller files to facilitate alignment. Some in-house MS Word macros and Sed commands in a Linux shell were used to convert the files to an XML format that is readable by the Translation Corpus Aligner 2 (TCA2) software (Hofland and Johansson, 1998). TCA2 is a Java application designed for the alignment of parallel data and its exportation as XML files compliant with the Text Encoding Initiative (TEI).[38] The TEI is a consortium that develops and maintains a set of guidelines that serve as a standard for archiving machine-readable data useful for research, interchange and data preservation.

4.2.5 Sentence alignment

Subsequently, the data was aligned using the 2010 version of TCA2. This program presents the user with three procedural alignment options: "One at a time", "Skip 1-1" and "Automatic". The second option was chosen to proceed with the alignment in an efficient manner. The program uses an anchor file to improve the alignment process. The anchor file is a bilingual lexicon used to compute equivalent words between sentences appearing in a pair of bilingual files being aligned. Each line of this file includes a source word and its equivalent target word, separated by a slash, using this format:

```
Free Trade Agreement / Tratado de Libre Comercio
```

[37] `http://www.abbyy.com`
[38] `http://www.tei-c.org`

To improve the alignment of FTA texts, several of the most frequent English terms and their Spanish equivalents were introduced in the anchor file. Figure 4.2 illustrates the alignment process using TCA2.

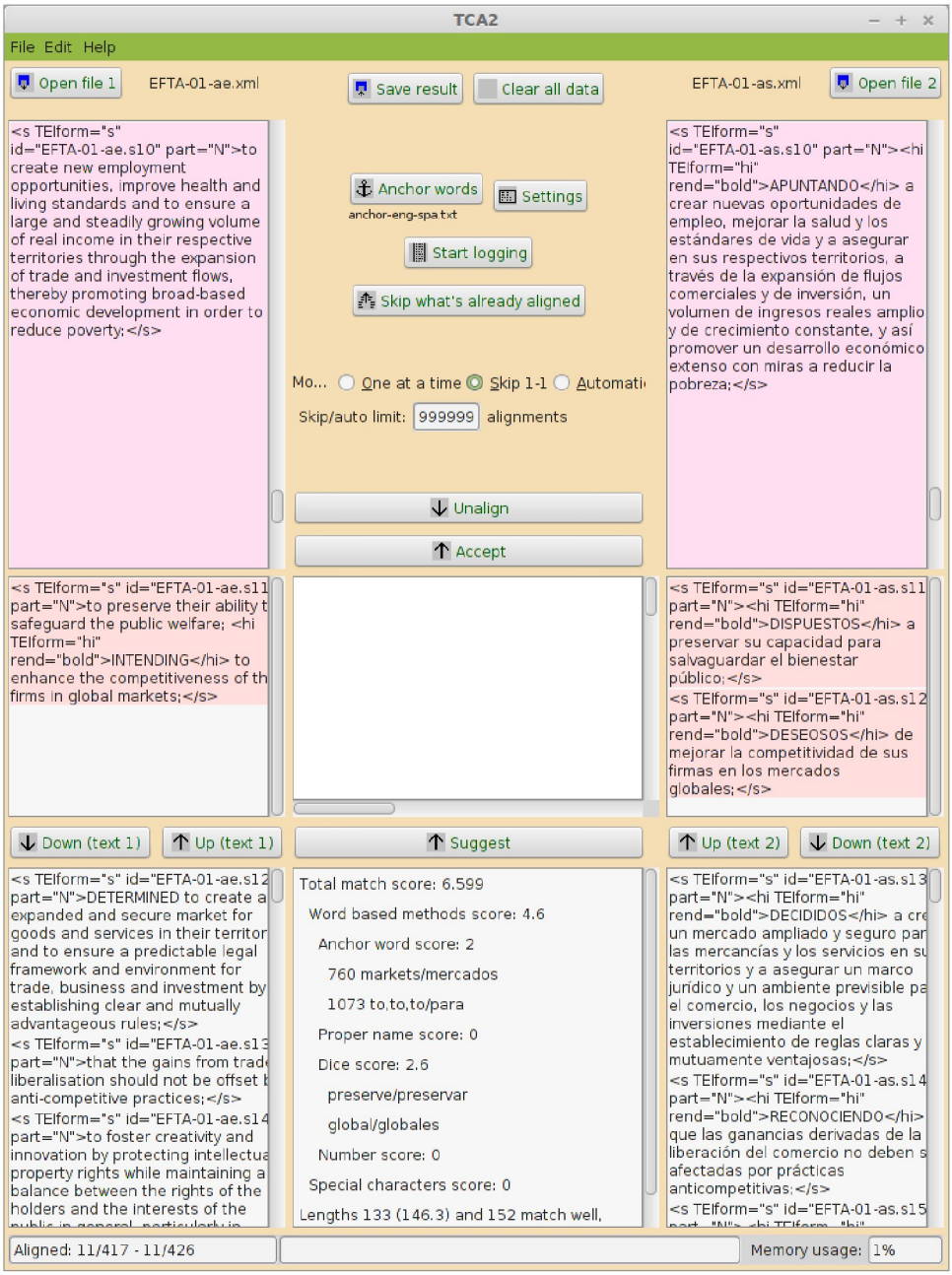

Figure 4.2: *Parallel corpus alignment using TCA2*

All the FTAs included in the corpus contain a section with definitions organized in alphabetical order according to the English text. Therefore, to facilitate the alignment of the files, the Spanish files that include a section with definitions had to be manually edited and rearranged in such a way that each entry in that section would match alphabetically its English counterpart with definitions. This was a requisite prior to the alignment of these segments.

4.2.6 PoS tagging of data

The data was first tokenized and then processed with the TreeTagger (Schmid, 1994), a software that lemmatizes each word form and annotates it with Part-of-Speech tags. This software supports both English and Spanish, among other languages. This way, it is possible to perform queries by using morphosyntactic patterns.

The IMS Corpus Query Processor (CQP) from the Corpus Workbench (CWB) toolkit (Christ, 1994) was used to extract the list of all the lexical units which are relevant for candidates of specialized collocations: the lexemes that were annotated as nouns, adjectives, verbs and adverbs. It was also used to calculate their frequencies. The CWB system is a collection of open-source tools that allows to encode and query large text corpora with linguistic annotations.

The output of the TreeTagger is printed as three tab-separated columns where the first column corresponds to the word form, the second to the part of speech and the third to the lemma, as shown in the following lines:

```
maintain VV maintain
or CC or
increase VV increase
a DT a
customs NNS custom
duty NN duty
as RB as
authorized VVD authorize
by IN by
the DT the
Dispute NP
```

```
Settlement NP Settlement
Body NP Body
of IN of
the DT the
WTO NP WTO
. SENT .
```

The TreeTagger developers claim that this system attains 95% of accuracy (Schmid, 1994). The TreeTagger assigns the tag *unknown* to the lemmas that are not recognized. In order to obtain a better output from the queries, the scripts were prepared to match the word form instead of the lemma to reduce noise from words tagged as *unknown*. An approach that relies on corpus-driven collocation extraction is used with this program. This is done with the CQP program by using this command to find, for instance, the adjectival collocates of the term *agreement*:

```
cwb-scan-corpus -C FTA-EN pos+0=/JJ/ word+0 word+1=/agreement/
```

These commands, which allow to extract morphosyntactic patterns and count their frequencies, were used to retrieve the collocates of CSCs:

```
FTA-EN>
[pos="V.*"] [] "agreement";
sort by word;
count by lemma;
```

These are some examples of CSCs that were retrieved from the English subcorpus by following these commands, in this case with the candidate term *tariff*:

```
16      determine the tariff  [#2-#17]
5       satisfy the tariff  [#33-#37]
3       file a tariff  [#20-#22]
%3       mean the tariff  [#26-#28]
3       regard the tariff  [#30-#32]
%3       take the tariff  [#39-#41]
2       eliminate the tariff  [#18-#19]
```

```
2       follow the tariff  [#23-#24]
1       apply a tariff  [#0]
1       apply the tariff  [#1]
1       include the tariff  [#25]
1       raise a tariff  [#29]
1       suspend the tariff  [#38]
```

4.2.7 Query interface

The FTA corpus can be queried using the IMS CWB interface, as shown in Figure 4.3. The web interface is currently set for queries spanning three adjacent words or part of speech tags. This method is useful to identify the collocates of the terms that appear in this specialized corpus. It is also possible to exclude stop words in the query. Besides, the interface includes an option to select whether words should be case-insensitive or not, to perform specific queries to match proper nouns or acronyms. More information on the FTA corpus is offered in Patiño (2013).

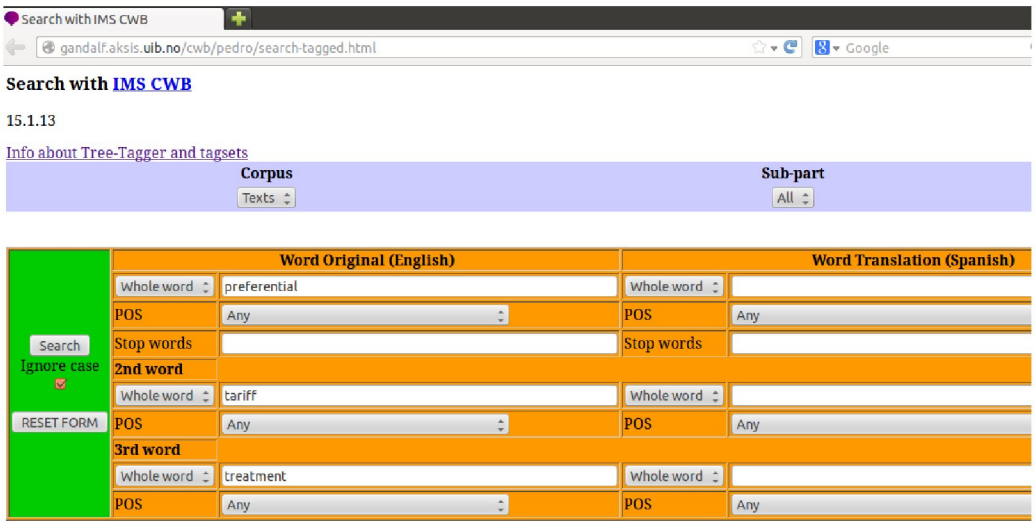

Figure 4.3: *IMS CWB online interface to query the corpus*

Figure 4.4 shows the results of a query of the complex term *preferential tariff treatment*. It is optional to include the PoS or to use stop words. This way, using the example mentioned before, it is possible to find the frequent

collocates of the term *preferential tariff treatment*. The results are presented as a table, where column A corresponds to the English sentence and column B to its Spanish equivalent sentence. The queried expression in the source language is displayed in bold letters. Using the corpus interface, it is possible to compare, with the aid of the parallel corpus, the terms and their context in both languages.

preferential tariff treatment means the application of the respective duty rate under this Agreement pursuant to the tariff elimination schedule to an originating good ; (CAN_PER-1e.s60)	trato arancelario preferencial significa el arancel aplicable bajo este Tratado de conformidad con el cronograma de eliminación de aranceles a una mercancía originaria. (CAN_PER-1s.s60)
For purposes of determining whether a good is an originating good, the production of the good in the territory of one or both of the Parties by one or more producers shall, at the choice of the exporter or producer of the good for which **preferential tariff treatment** is claimed, be considered to have been performed in the territory of either of the Parties by that exporter or producer, provided that: (CAN_PER-1e.s392)	Para efectos de determinar si una mercancía es originaria, la producción de esa mercancía en el territorio de una o ambas Partes por uno o más productores, a elección del exportador o productor de la mercancía para la cual se solicita trato arancelario preferencial, será considerada como realizada en el territorio de cualquiera de las Partes por ese exportador o productor, siempre que: (CAN_PER-1s.s400)
require an exporter in its territory to complete and sign a Certificate of Origin for any exportation of a good for which an importer may claim **preferential tariff treatment** upon importation of the good into the territory of the other Party ; and (CAN_PER-1e.s518)	exigirá a un exportador en su territorio que complete y firme un Certificado de Origen para cualquier exportación de una mercancía para la cual un importador pudiera solicitar trato arancelario preferencial en la importación de la mercancía en el territorio de la otra Parte ; y (CAN_PER-1s.s527)
Except as otherwise provided in this Chapter, each Party shall require an importer in its territory that claims **preferential tariff treatment** for a good imported into its territory from the territory of the other Party to: (CAN_PER-1e.s528)	Salvo que se disponga algo distinto en este Capítulo, cada Parte exigirá a un importador en su territorio que solicite trato arancelario preferencial para una mercancía importada a su territorio proveniente del territorio de la otra Parte, que: (CAN_PER-1s.s537)
Where an importer claims **preferential tariff treatment** for a good imported from the territory of the other Party: (CAN_PER-1e.s534)	Cuando un importador solicite trato arancelario preferencial para una mercancía importada del territorio de la otra Parte: (CAN_PER-1s.s543)

Figure 4.4: *Results of the query* preferential tariff treatment

4.3 Reference lexical and terminological resources

To contrast the candidate specialized collocations against other sources, some reference material was needed. This material comprises several general and specialized dictionaries and glossaries as well as corpora that are described below. A total of 69,643 terms that appear in nine specialist dictionaries were included in this study to compare against the candidate terms extracted from the FTA corpus. These terms come from the domains of international trade, economics, business and finance. These resources are listed in Table 4.2. As can be seen, they include several glossaries and terminology compilations from FTA-related institutions, plus specialist dictionaries from the broader field of economics.

Resource	**Available languages**	**Number of Terms**
WTO Glossary	EN, FR - ES	10,854
IMF Terminology	EN, FR, DE, RU	4,098
SICE-OAS online Dictionary of Trade Terms	EN, ES, FR, PT	416
Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English, (Alcaraz 2007)	EN, ES	6,404
Routledge Spanish Dictionary of Business, Commerce and Finance, (Routledge 1998)	EN, ES	29,893
Pocket Business Spanish Dictionary, English-Spanish/Spanish-English, (Collin 2003a)	EN, ES	5,895
Dictionary of Banking and Finance, (Collin 2003b)	EN	3,206
Routledge Dictionary of Economics, (Rutherford 2002)	EN	4,200
Dictionary of International Business Terms (Capela 2000)	EN	4,677

Table 4.2: *Specialist reference dictionaries*

Table 4.3 presents the reference corpora used in this work. The general corpora were used as reference material to compare the relative frequency of words found in the FTA corpus in comparison with general texts, using the Wordsmith tools (Scott, 2007), as described in Section 4.4.2. These corpora resources are relevant for research purposes because of their size and

Corpus	Author	Millions of words
Corpus of Contemporary American English http://www.americancorpus.org	Davies, 2009	520
Corpus del Español http://www.corpusdelespanol.org	Davies, 2002	100
OpenSubtitles 2011 parallel corpus of English and Spanish movie subtitles http://www.opensubtitles.org	Tiedemann, 2009	267

Table 4.3: *English and Spanish reference corpora*

representativity.

4.4 Method

A quantitative and qualitative study of the most frequent morphosyntactic patterns that occur in both the Spanish and the English data was carried out to determine how specialized collocations behave across languages, with the aim of making a selection of criteria for the extraction of CSCs. As it was defined in Section 2.14, a specialized collocation is a type of MWE composed of at least a term that serves as the node of the collocation. Its collocates can be nouns, verbs, adjectives or adverbs in a direct syntactic relation with the node. This way, the terms, along with its co-occurring constituents, make up a lexical combination that can be unpredictable and semi-compositional and have an internal and statistical tendency of preference.

In the extraction, non-relevant words from the closed lexical categories, such as determiners, prepositions and pronouns were excluded by using a list of stop words. This procedure made it easier to see the collocational relations of terms with other lexical items co-occurring with the terms. Moreover, in harmony with the criteria set forth in Section 2.15, to operationalize the extraction of CSCs, alphabetical symbols such as punctuation marks were excluded from the search window; therefore, whenever a period, a comma or a question mark co-occurs between the node and the collocate, the sample was discarded.

4.4.1 Gold standard of Free Trade terms and collocations

Since one of the objectives is to find CSCs in the FTA corpus, I consider that a gold standard of terms is useful to attain this end. The termhood of the items that make up the gold standard of terms is externally determined by the document authors. To create a gold standard of terms it was taken into account that all FTAs include a 'Definitions' section. Such a section sets a common vocabulary for trade experts to negotiate each agreement and to agree upon a common set of concepts. By appearing in that section as the entry for each definition, these terms are thus *a priori* validated as terms in the data. For example, in the FTA between the EFTA states and the Republic of Colombia, Article 1.9 (b) states:

> "juridical person" means any legal entity duly constituted or otherwise organised under applicable law, whether for profit or otherwise, and whether privately-owned or governmentally-owned, including any corporation, trust, partnership, joint venture, sole proprietorship or association.

Juridical person is thus a valid term in the FTA texts. A list of 441 terms found in the definition section of each FTA was extracted semi-automatically by using regular expressions. The regular expressions that were used match a pattern such as the following: *X* means *Y*, where *X* is a term. Section 5.2 provides further information on the gold standard of terms from a linguistic and terminological point of view.

The preliminary detection was first done through a corpus-based approach, carried out by means of scripts that run the program *cwb-scan-corpus* from the IMS CWB toolkit,[39] in order to identify the lexical units that co-occur with a term in a predefined window, from 1 to 5 tokens to the right of a term acting as the subject and from 1 to 5 tokens to the left of a term acting as the object. For example, this command was used in a batch mode to extract the candidate verbal collocates found at 1 token to the right of

[39] For a complete tutorial on the use of the IMS CWB toolkit, see the program documentation at `http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf`

the term "date of entry" in the English subcorpus, using the program *cwb-scan-corpus*:

```
cwb-scan-corpus -C FTA-EN word+0=/date/word+1=/of/
word+2=/entry/ lemma+3 pos+3=/V.*/ > outputFile
```

Software for language processing was used to extract CSCs or to automatically perform a semantic tagging of a data sample. These software tools are described in more detail in Section 4.4.2.

4.4.2 Extraction of CSCs

Several software packages were used to follow the method set forth in the present chapter in harmony with the research questions regarding the nature of specialized collocations and their linguistic features.

First, Termostat (Drouin, 2003)[40] was used with the raw data in each language to extract the candidate terms found in the FTA corpus, extract some preliminary specialized collocations that include the most relevant terms and a list of keywords. Currently, Termostat's online version is prepared for the extraction of candidate terms in French, English, Spanish, Italian and Portuguese. Termostat is suitable for corpus-driven research supported by "seed" terms because it is a hybrid system for term extraction that incorporates statistical measures and linguistic rules for several languages. For each language, the system compares the data against a reference corpus to generate a list of keywords relevant for the FTA data.

Once the terms were extracted, the list was manually cleaned to discard non-terms or false positives. The criteria that were used to discard as non-terms some of the candidates extracted with Termostat are:

- one word from the candidate term was mistakenly split into two words in the source file or was split by Termostat, e.g. *euro peo, estableci miento* while the correct form should be *europeo* and *establecimiento*.
- Abbreviations or acronyms were tagged as nouns by the TreeTagger and were then included in the candidate term list. Examples of discarded

[40] `http://olst.ling.umontreal.ca/~drouinp/termostat_web`

candidate terms that were actually abbreviations are *"html", "http", "ex", "kg"* and terms including these abbreviations or acronyms as a constituent.

- FTA texts include texts in several languages, especially when institutions or products from several countries are mentioned. For example, some Spanish words were extracted by Termostat along with adjacent English or French words as candidate terms, e.g. *"eaux"* and were therefore discarded manually.
- Termostat might have some defective morphosyntactic rules to extract terms, and thus, a truncated chunk of text was extracted as a candidate term while one or more adjacent components were excluded, as seen in the following incomplete Spanish noun phrases, *"cariforum en cuestión", "cariforum en virtud"* which were erroneously tagged as candidate terms. To prevent this problem, Termostat should include some rules to expand morphosyntactically the candidate term extraction until reaching the leftmost or rightmost constituent or a noun-phrase delimiter such as a determiner or a punctuation sign as explained by Jacquemin et al. (1997) and Burgos (2014).

By using Termostat, a preliminary list of 10,743 candidate terms in Spanish was automatically retrieved. For the Spanish data, after the list of candidate terms extracted with Termostat was manually cleaned, 307 non-terms were discarded, corresponding to 2.85% of the candidate terms. The remaining candidate terms were 10,436. In turn, for the English data, the preliminary list of 6,464 candidate terms extracted with Termostat was also manually cleaned. After this was done, 179 non-terms (2.76% of the candidate terms) were discarded from the list of candidate terms. This left 6,285 remaining English candidate terms. The above numerical difference indicates that for the Spanish subcorpus there are more term variants extracted, possibly because translators might have offered a new translation for previously translated terms. Even though the cross-language analysis of these terms and their variants is a relevant and interesting topic, it is beyond the scope of this thesis.

Termostat extracts the terms based on morphosyntactic patterns and statistical association measures. For the English data, the extraction is carried

out based on 54 patterns. For the extraction of the Spanish data, 12 patterns are used. These patterns are listed in the Appendix in Tables 1 and 2, respectively.

Section 5.3 presents the most frequent patterns that form English terms. A sample of four patterns that form 5,028 terms, equivalent to 80% of the terms, was selected to query the corpus for specialized collocations. Statistically, the sample of 5,028 terms obtained with these four frequent patterns represents a confidence level of 95% and a confidence interval of 0.62.[41]

To extract the CSC, the list of "seed" terms was taken from the candidate terms extracted with Termostat after the list was manually cleaned. To carry out the extraction of the CSCs, several in-house *Sed* scripts were prepared. The scripts invoke a component of the IMS CWB toolkit, the *cwb-scan-corpus* program, which was used to retrieve candidate specialized collocations. For each term from the list, each script extracts all lexical verbs that co-occur in a position of 1, 2 and 3 tokens to the left and to the right of the term, as explained in the criteria set forth in Section 2.13. A percentage of 51.5% of the English candidate terms includes two or three word terms, with structures such as Adjective + Noun, Noun + Noun and Noun + Prep + Noun. Therefore, two or three tokens to the right or the left of the syntactic head of the term are covered by this span.

By using morphosyntactic patterns and code such as the following, it is possible to extract all the verbs that appear, in this example, three tokens before the Spanish term *parte*:

```
cwb-scan-corpus -C FTA-ES lemma+0 pos+0=/V.*/ lemma+3=/parte/ >
candSpCo-FTA-ES-verb3-termN-termostat
```

To create a concordance to scan the terms and their collocates, the CQP tool of IMS CWB toolkit was used. For example, by using the query pattern below in a window of four tokens with only one token intervening between the verb and the term "custom duty":

```
[pos="V.*"] [ ] "custom" "duty";
```

[41] The calculations were obtained from the Sample Size Calculator available at `http://www.surveysystem.com/sscalc.htm` Thanks are due to Assist. Prof. Dr. Julián Cárdenas from Universidad de Antioquia for his timely advisory.

Researchers on collocations have noted that these lexical combinations cannot be entirely explained by assuming exclusively a syntactic approach (McKeown and Radev, 2000; Krishnamurthy, 2006). Thus, it seems adequate to review the semantic features of specialized collocations in an attempt to characterize them and predict them with the intention of carrying out semi-automatic extraction of candidate specialized collocations. Several tools were used to perform this process semi-automatically. First, the Natural Language Toolkit (NLTK) (Bird et al., 2009) was used, which is an open source Python-based platform to run and build natural language applications. The program incorporates functions to process linguistic data that are useful for the purposes of this research. Then, the NLTK was used in combination with other tools to calculate the semantic features of a list of specialized collocations, specifically, with the suite of tools called Freeling (Padró and Stanilovsky, 2012), which is an "open source language analysis tool suite".[42] Freeling was used to perform the semantic annotation of nouns, verbs, adjectives and adverbs occurring in a set of 1,589 specialized collocations with the terms from the gold standard, in combination with a Python script along with NLTK and Princeton's Wordnet,[43] (Miller, 1995) a lexical reference system, which was used to annotate the specialized collocations with semantic information.

One of the linguistic tools included with Freeling is executed using this command, where the file *analyzer.cfg* incorporates the parameters chosen for a particular task:

```
analyze analyzer.cfg input > output
```

Wordnet uses a tagset of 45 lexicographer files to annotate the lexical units.[44] These were used to categorize the nouns, verbs, adjectives and adverbs occurring in the extracted specialized collocations.

Table 4.4 shows the tags that have been used to extract CSCs from the English subcorpus, as well as their verb form and examples for each tag.[45] In contrast, Table 4.5 shows the tags used to annotate the verbs *to be* and

[42] `http://nlp.lsi.upc.edu/freeling/`

[43] `http://wordnetweb.princeton.edu/perl/webwn`

[44] `http://wordnet.princeton.edu/man/lexnames.5WN.html`

[45] Adapted from `http://www.sketchengine.co.uk/documentation/wiki/tagsets/penn`

Table 4.4: *TreeTagger tags used for collocation extraction from the English data*

Tag	**Verb form**	**Example**
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
JJ	adjective	green
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
RB	adverb	usually, naturally

to have. Since only lexical verbs in English and Spanish are relevant for the extraction of specialized collocations, the tags included in Tables 4.5 and 4.6 for English and Spanish respectively, were discarded from the queries.

Table 4.5: *TreeTagger tags excluded from collocation extraction from the English data*

Tag	**Verb form**	**Example**
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has

In addition to the above mentioned tools, the Wordsmith Tools (Scott, 2007), a well-known suite of programs for lexical analysis, was used to generate concordances and to extract word lists and keywords, for using the

Table 4.6: *TreeTagger tags excluded from collocation extraction from the Spanish data*

Tag	**Verb form**	**Example**
VEadj	Verb estar. Past participle	estado
VEfin	Verb estar. Finite	esté
VEger	Verb estar. Gerund	estando
VEinf	Verb estar. Infinitive	estar
VHadj	Verb haber. Past participle	habida
VHfin	Verb haber. Finite	haya
VHger	Verb haber. Gerund	habiendo
VHinf	Verb haber. Infinitive	haber
VMadj	Modal verb. Past participle	debido
VMfin	Modal verb. Finite	podrá
VMger	Modal verb. Gerund	pudiendo
VMinf	Modal verb. Infinitive	poder
VSadj	Verb ser. Past participle	sido
VSfin	Verb ser. Finite	sea
VSger	Verb ser. Gerund	siendo
VSinf	Verb ser. Infinitive	ser

reference corpora to contrast against FTA candidate terms and specialized collocations.

Another tool, Xaira, which stands for XML Aware Indexing and Retrieval Architecture,[46] an open source software package was also used to extract candidate collocations. It supports indexing and analysis of corpus data. The system is designed to use Z-score and MI to extract collocations. One advantage of this system is its ability to calculate the collocates of a given term. However, its MS Windows version is only capable of performing this extraction on an individual term-by-term basis, which considerably slows down the process. Xaira's Linux version can allegedly perform collocation extraction from a list of candidate terms. However, its installation presented several dependency issues with obsolete packages, which posed problems to install it successfully and this tool was therefore discarded for use in batch mode.

[46] `http://projects.oucs.ox.ac.uk/xaira/index.xml?ID=body.1_div.1`

4.5 Morphosyntactic patterns for the extraction of specialized collocations

Using the criteria set forth in Section 2.15, several morphosyntactic patterns, which, according to the literature, are frequent in the formation of terms in English were used to extract candidate specialized collocations from the corpus. This was made in harmony with the research on term extraction carried out by authors such as Daille (1994); Gaussier and Langé (1994); Estopà (1999); Heid (1999, 2001); Daille (2001); Drouin (2003); Orliac (2008); De Groc (2011); Ljubešic et al. (2012) and Burgos (2014).

Both for English and Spanish, the code presented below looks for terms found to the right of the collocate. To look for terms found to the left of a verbal collocate, i.e. when the term is the subject, the term is expressed as

```
[word=<term>]
```

where *term* corresponds to an entry from a) the list of 441 terms that make up the gold standard or b) the terms extracted semi-automatically with Termostat. The tags in parentheses are assigned by the TreeTagger to identify these parts of speech, using the tags defined in the Penn Treebank Tag Set for the English language:[47]

```
Adjective ([pos=‘‘JJ’’]) + Term  [word=<term>]
Noun ([pos=‘‘NN.*’’]) + Term  [word=<term>]
Adverb ([pos=‘‘RB’’]) + Term  [word=<term>]
Verb ([pos=‘‘VV.*’’]) + Term  [word=<term>]
```

These are the patterns used for the extraction of the candidate specialized collocations from the Spanish subcorpus. The TreeTagger uses a different tagset for Spanish,[48] as follows:

```
Noun ([pos=‘‘NC’’]) + [word=<term>]
Adjective  ([pos=‘‘ADJ’’])  + [word=<term>]
```

[47] http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

[48] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt

```
Adverb ([pos=‘‘ADV’’])  + [word=<term>]
Verb ([pos=‘‘VL.*’’])  + [word=<term>]
```

The Spanish tagset includes the following tags to annotate lexical verbs, which were used for the extraction:

```
VLadj (Lexical verb. Past participle)
VLfin (Lexical verb. Finite)
VLger (Lexical verb. Gerund)
VLinf (Lexical verb. Infinitive)
```

To summarize, in order to identify the specialized collocations that include a verb and one of the terms from the above mentioned lists, the collocates were extracted in a window that includes the term and spans three tokens to the right from the rightmost constituent of the term, i.e. when the term is the syntactic subject. Then, the collocates of the terms were extracted in a window that includes the term and spans three tokens to the left from the leftmost constituent of the term, i.e. when the term is the syntactic object.

Chapter 5 offers an analysis and description of the CSCs obtained after applying the queries mentioned above to all the terms in the gold standard and the terms extracted with Termostat. These analyses are carried out from the morphosyntactic, semantic and terminological perspectives.

4.6 Representation of specialized collocations in language resources

Finally, in Chapter 6 I will present a proposal to represent these units in computational lexicons. That chapter describes how several initiatives have been conceived to represent data in MRDs. However, most of these initiatives are not fully prepared to represent phraseological data. Therefore, the proposal is presented using the LMF standard. LMF code is XML-compliant and therefore it is fully interchangeable and mergeable with existing or future language resources. This representation aims at being useful both for humans and for computers.

CHAPTER 5

Results and analysis

5.1 Introduction

This chapter presents the candidate specialized collocations that have been extracted following the method described in Chapter 4. I begin by presenting the gold standard of terms and the candidate terms extracted semi-automatically with Termostat (Drouin, 2003). To do so, first, the most frequent morphosyntactic patterns found in the two sets of terms are presented and exemplified for both languages. Second, I continue with a linguistic, quantitative and qualitative classification and description of the candidate specialized collocations that include these terms. Special attention is given to the verbs that collocate with these terms in the specialized context of FTAs.

The two sets consist of a) a gold standard of 441 terms retrieved by means of regular expressions from the English subcorpus of FTA texts and b) the candidate terms extracted semi-automatically with Termostat. By using a combination of corpus-based and corpus-driven approaches, as described in Section 2.7.1, these two sets of terms are used as "seeds" (Baroni and Bernardini, 2004) in each query to the corpus with the aim of finding the

collocates that usually co-occur with these terms, using the criteria defined in Section 2.15.

5.2 Description of the gold standard of terms

A gold standard composed of 441 terms was retrieved semi-automatically using the criteria set forth in Section 4.4.1.

From a morphosyntactic point of view, the three most frequent patterns that compose the gold standard of terms together account for 44% of the terms found in the gold standard. In the case of the English data, some examples of such terms are, in the first place the ones following the pattern Adjective + Noun, with 77 occurrences and which account for 17.46% of the terms, for example: *commercial presence, competent authority, administrative refusal, electronic auction, financial service* and *procedural provision.* The second type of terms are simple, that is, terms formed by one noun, with 66 occurrences, equivalent to 14.97% of the data. Some examples are *commission, enterprise, importer, investment* and *measure.* The third pattern is occupied by terms formed by two consecutive nouns, with 53 occurrences, that is, 12.02% of the total number of terms found in the gold standard. Some examples of the terms that correspond to this pattern are *applicant authority, competition law, customs legislation, government monopoly* and *investment agreement.*

The frequency of the patterns Noun + Adjective or Adjective + Noun, for Spanish or English, respectively, agrees with the findings of Cartagena (1998), who found a high frequency of these patterns in specialized corpora. According to Cartagena (1998, 287),[49]

> Desde luego que existe una relación directa entre la longitud, el grado de especialización y la estabilidad sintáctica del término; a mayor longitud, mayor especialización e inestabilidad.

In contrast, for the gold standard of terms in English, some of the less frequent patterns that compose the gold standard of terms, along with their

[49] Of course there is a direct relationship among length, degree of specialization and syntactic stability of a term; the longer a term, the greater specialization and instability. (My translation).

occurrences and percentages are terms composed by the pattern Adjective + Noun + Noun, with 20 occurrences, accounting for 4.54% of the data. Examples are *agricultural export subsidy, agricultural safeguard measure, collective investment scheme, economic integration agreement* and *financial service supplier*. Next come the terms formed by the pattern Noun + Preposition + Noun, with 10 occurrences, equivalent to 2.27% of the terms. Some examples of terms formed with this pattern are *agreement on subsidies, claim of origin, conditions for participation, country of importation* and *items of correspondence*.

Finally, terms formed by the pattern Noun + Preposition + Adjective + Noun, with 2 occurrences, accounting for 0.45% of the total term count in the gold standard of terms. These are the two cases with this pattern: *notice of intended procurement, threat of serious injury*.

Table 5.1: *The top 10 most frequent terms and their verbal collocates*

Term	**Freq.**	**Verbal collocate**
information	99	provide, include, protect, disclose, submit, contain
good (noun)	64	advertise, supply, produce, originate, provide, include, transport
national	57	comply, mean, impair, accord, ensure, appoint, forward
measure	42	maintain, apply, adopt, impose, enforce, execute
decision	40	adopt, issue, take, reach, follow, implement
service	33	supply, provide, permit, govern, accord, withdraw, include
supplier	32	preserve, allow, exclude, enable, inform, provide, recognize, require
value	30	estimate, calculate, include, exclude, denote, convert, recognize, declare
procurement	29	apply, cover, regard, describe, conduct, govern, develop, cancel, use, relate, divide
production	27	use, include, initiate, undergo, require, distort, determine, apply
entry	26	grant, follow, seek, prevent, authorize, request

The most frequent terms of the gold standard form a series of specialized collocations. Table 5.1 presents the first 10 most frequent terms and their verbal collocates. In this particular case, these lexical items co-occur when the verb is at position -2 from the term. All the 10 frequent terms exemplified here are simple lexemes.

This list of the 10 most frequent terms from the gold standard and the collocates they take to form a specialized collocation in FTA texts suggest their relevance for this type of agreements. The terms themselves, related to economics, contract law and legal issues, emphasize the importance of information, services, measures, suppliers, procurement, national boundaries and production in the field of FTAs.

The relevance of the term *information* is highly evident, with five verbs frequently co-occurring with that term.

The collocation formed by *maintain measure* is highly frequent compared to the other cases, with 88 occurrences in the English subcorpus of FTA texts.

Table 5.2: *Top verbal specialized collocations from the terms found in the gold standard where the verb is at position -2 in relation to the term*

Verbal collocate	**Term**	**Freq.**
maintain	measure	88
provide	information	54
adopt	decision	38
submit	claim	35
apply	procurement	22
cover	procurement	21
supply	service	21
apply	measure	15
include	information	14
adopt	measure	14
relate	qualification	13
grant	entry	12
relate	investment	12
make	claim	11
request	establishment	11
protect	information	11
disclose	information	11
indicate	sector	11
issue	decision	10
provide	service	10
submit	information	10

Table 5.2 presents the top specialized collocations including terms from the gold standard when the verb is found at position -2 in relation to the term. The table presents the lemma for both the verb and the term that co-occurs with the verb. Any other lexical item occurring between the verbal

collocate and the term is omitted from this list.

5.3 Description of the candidate terms extracted with Termostat

Complex terms, specifically two-word terms are the most prevalent in the English data. In detail, one-word terms account for 19.6% in the gold standard of terms and 16.8% in the list of candidate terms, while two-word terms correspond to 44% of the first subset of the data and 51.5% in the list of candidate terms; three-word terms represent 15.3% and 19.61%, respectively, while four-word terms account for 10.6% in the gold standard and 9.4% in the case of the candidate terms. In other words, terms are more often composed by multiword strings than by simple lexemes. The token count distribution of the English gold standard and the candidate terms is presented in Table 5.3.

Terms made up by 1 to 4 tokens were included in the extraction, while terms composed by 5 to 7 tokens were not taken into account because of their low frequency.

Table 5.3: *Word count distribution of the English gold standard and the candidate terms*

Words	**Gold st. of terms**	**%**	**Cand. terms**	**%**
1	87	19.6	1060	16.8
2	195	44.0	3238	51.5
3	68	15.3	1232	19.6
4	47	10.6	595	9.4
5	27	6.0	120	1.9
6	11	2.4	31	0.4
7	5	0.6	7	0.0

In previous works done in the field of ATE, other authors have excluded units longer than 4 words, due to their low frequency (Daille, 1994), while other researchers have presented lists of morphosyntactic patterns to extract English and Spanish candidate terms that span up to 9 words (Quiroz, 2008; Burgos, 2014).

Table 5.4: *Distribution of patterns for the English candidate terms*

Pattern	**Examples**	**Percentage**	**Freq.**
Adj N	financial service intellectual property competent authority financial institution	33.2	2105
N N	service supplier custom duty property right woven fabric	18.3	1165
N	party service agreement measure	16.9	1073
N Prep N	date of entry rule of origin period of time certificate of origin	10.8	685
Adj N N	regional value content financial service supplier economic need test intellectual property right	3.6	234
N Prep Adj N	supplier of public telecommunication notice of intended procurement enforcement of intellectual property form of numerical quota	3.3	211
Adj Adj N	national central bank ordinary / special legislative procedure equal annual stage relevant international standard	2.6	170
Adj Conj Adj N	sanitary or phytosanitary measure sanitary and phytosanitary measure arbitrary or unjustifiable discrimination natural or legal person	2.5	164

Figure 5.1 illustrates the word-count distribution for both the English gold standard and the candidate terms extracted with Termostat. As is evident from the figure, in both datasets, two-word terms are the most frequent type. Of these, terms with the pattern Adjective + Noun are the most frequent ones.

Table 5.4 presents the distribution of the eight most salient morphosyntactic patterns for the candidate terms. It also offers some examples for the candidate terms in English extracted semi-automatically with Termostat, af-

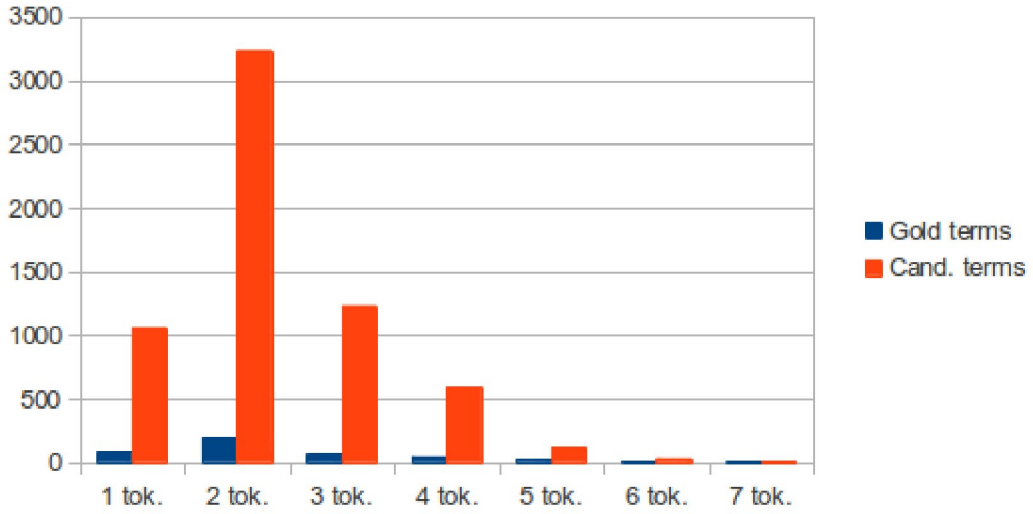

Figure 5.1: *Word count distribution of English gold standard and candidate terms*

ter the list was manually cleaned to discard non-candidate terms. These eight patterns account for 91.6% of the whole list of candidate terms.

Out of this list, the first two patterns in frequency are Adjective + Noun with 33.2% and 2,105 occurrences out of 6,285 terms, and Noun + Noun with 18.3% and 1,165 occurrences. In the third place come terms composed by a noun with 16.9% and 1,073 cases in the English data. The fourth most frequent pattern is Noun + Preposition + Noun with 685 occurrences which represents 10.8% of the candidate terms. Therefore, these four patterns which account for 80% of the whole list of candidate terms were selected as the primary target to query the corpus to search for candidate specialized collocations.

These phraseological units are used in different disciplines. Some of the terms are mostly used in macroeconomics and finance, such as *collective investment, debt instrument* and *service supplier*. Other terms are more commonly associated to international trade, a subdomain of macroeconomics that comes from economics, such as *cross-border supply, customs duty* and *preferential tariff*, while other terms are related to law such as *intellectual property, domestic law, domestic legislation, legal entity, legal person* and *legislative act*. Other terms refer to the goods that are included in the agree-

ments, such as *animal hair, man-made fibre, milk powder, woven fabric* and *agricultural product*.

These findings document the most productive patterns in term formation for this domain. This suggests that extraction efforts should prioritize these highly productive patterns. This finding is also useful for the teaching of LSP, specialized translation and specialized phraseology, where future practitioners should be taught to focus on these patterns as the most frequent carriers of specialized information in highly specialized texts from the domain of economics, including international trade.

For the Spanish data, the morphosyntactic distribution of the list of 10,436 candidate terms extracted with Termostat is illustrated in Table 5.5. The four more frequent patterns account for 87.4% of the list of candidate terms and were therefore selected to query the corpus to find the verbal collocates that these terms take in the FTA corpus. These patterns are relevant for term extraction besides their interest in the teaching of LSP, terminology, specialized translation and phraseology. Combined, the patterns Noun + Preposition + Noun and Noun + Adjective, the two most frequent patterns for the Spanish candidate terms, account for 60.81% of the units. Next come two other frequent patterns. In the first place appear simple terms composed by a noun and then come complex terms consisting of four words: Noun + Preposition + Noun + Adjective, with roughly 14% and 12% respectively.

5.4 Frequent Spanish and English verbs

As a preliminary step to focus the extraction efforts in finding the most relevant verbs that form specialized collocations in the FTA corpus, the most frequent verbs appearing in the corpus were identified and ranked according to their frequency. First, 1,205 lexical verbs were extracted. The most frequent are 214 verbs, which occur from 2,900 to 100 times in the Spanish data. Their frequency suggests that these verbs are thus the most representative ones that form specialized collocations in Spanish FTA texts.

Table 5.6 presents the top-20 Spanish and English lexical verbs in the data along with their frequencies. They are not translations of each other. Rather, they are the most frequent verbs in decreasing order of frequency.

Table 5.5: *Distribution of patterns for the Spanish candidate terms*

Pattern	**Examples**	**Percentage**	**Freq.**
N Prep N	proveedor de servicio fecha de entrada medida de salvaguardia solución de controversia derecho de propiedad	31.17	3253
N Adj	parte contendiente contratación pública entidad contratante servicio financiero arancel aduanero	29.64	3093
N	parte mercancía proveedor servicio entidad	14.02	1463
N Prep N Adj	derecho de propiedad intelectual valor de contenido regional proveedor de servicio financiero prueba de necesidad económica rama de producción nacional	12.57	1,312
N Adj Adj	procedimiento legislativo ordinario transporte marítimo internacional trato arancelario preferencial información comercial confidencial tratamiento arancelario preferencial	5.66	591
N N	nota número artista intérprete año calendario mercancía objeto derecho antidumping	5.03	525
N Adj Coord Conj Adj	medida sanitaria y fitosanitaria asunto exterior y político disposición legal y reglamentaria derecho antidumping y compensatorio fibra artificial y sintética	1.72	179

For the English data, 1,555 unique lexical verbs were extracted and are also the most frequent verbs that form specialized collocations in English FTA texts. The most frequent of these lexical verbs are 258 and occur from 5,435 to 100 times in the English subcorpus.

Table 5.6: *Top 20 verbs for the Spanish and English data*

Freq	**Spanish Verbs**	**Freq**	**English Verbs**
2,904	establecer	5,436	provide
2,367	incluir	4,052	include
2,203	aplicar	3,208	apply
1,812	adoptar	2,960	make
1,481	disponer	2,366	establish
1,134	significar	2,286	take
1,113	considerar	2,261	refer
1,004	relacionar	2,054	mean
1,001	realizar	1,953	relate
984	presentar	1,813	require
941	prever	1,716	adopt
938	referir	1,624	set
914	mantener	1,542	use
852	otorgar	1,461	ensure
849	tratar	1,437	agree
845	cumplir	1,422	follow
821	utilizar	1,334	consider
801	solicitar	1,286	maintain
784	determinar	1,228	concern
732	indicar	1,171	cover

5.4.1 Candidate terms found in the FTA corpus

A list of the 100 most frequent candidate terms that were extracted automatically was processed into a "cloud" of words by Termostat. The size of the font indicates the frequency of the term in the subcorpus. Figure 5.2 shows the 100 most frequent candidate terms in the English component of the FTA corpus, which highlights salient terms such as *agreement, measure, service, procedure* and *supplier*. Later, Figure 5.3 presents the 100 most frequent candidate terms found in the Spanish component of the FTA corpus, which presents relevant terms such as *mercancía, proveedor, servicio, subpartida* and *parte contendiente*. Regarding their morphosyntactic composition, 86 out of the 100 most frequent candidate terms found in the cloud of words by Termostat are simple terms. Thus, only 14 are complex terms, where one corresponds to the pattern Noun + Preposition + Noun, 8 correspond to the pattern Adjective + Noun and 5 to the pattern Noun + Noun.

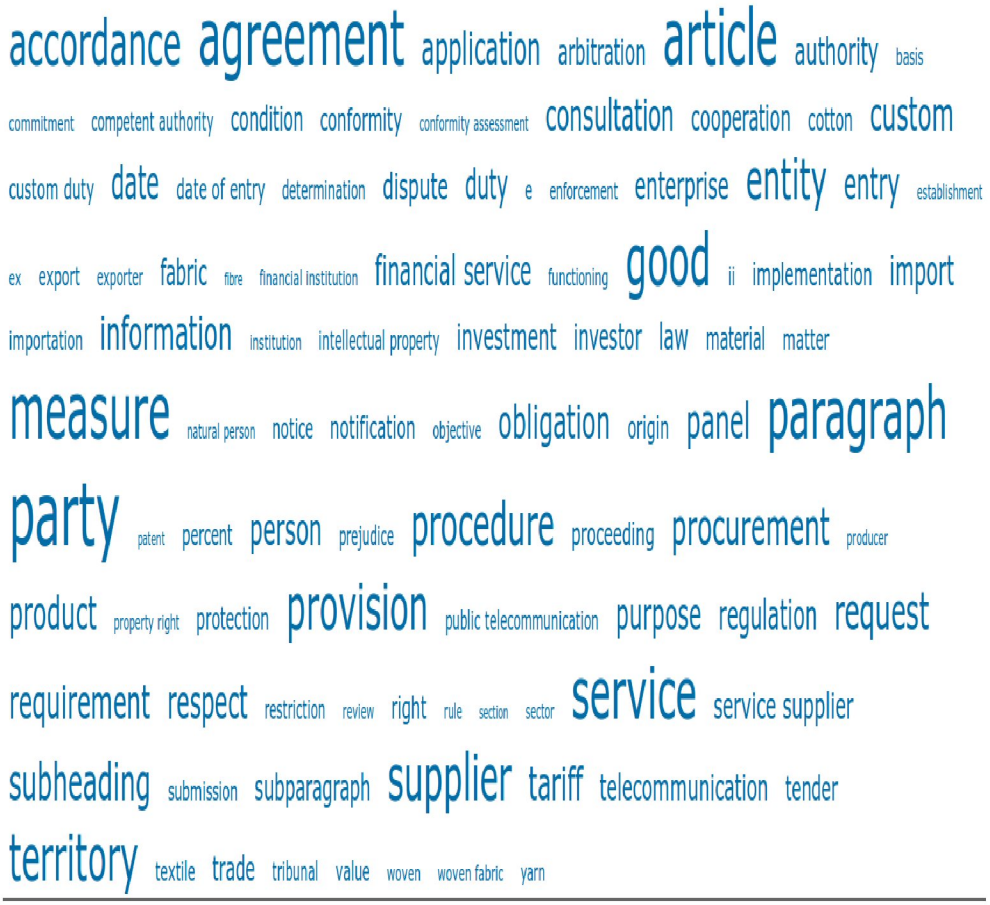

Figure 5.2: *Top 100 terms in the FTA English subcorpus*

5.5 Candidate specialized collocations in the FTA corpus

A list of candidate specialized collocations (CSC) was extracted semi-automatically for the different subsets of the corpus, by using scripts with IMS CWB, which are described in Section 4.4.2. The corpus was queried to look for verbal CSCs in English and Spanish, both when the term is the subject as well as when it is the object of the verb forming the collocational relation with the term.

Tables 5.7 and 5.8 respectively, present the Spanish and English patterns that were used to query the corpus using the *CWB-Scan-Corpus* program from IMS CWB. The patterns include a term that is realized in several

Figure 5.3: *Top 100 terms in the FTA Spanish subcorpus*

morphosyntactic patterns and the verbal collocates that co-occur with that term, both when the term is the subject or the object of the co-occurring verb. These patterns can be used to develop a semi-automatic system to detect CSCs in a tagged corpus in English or Spanish.

Table 5.9 presents the distribution of the most frequent patterns of CSCs extracted from the English data, when the verb is found at 1, 2 or 3 tokens from the term, while Table 5.10 presents the same distribution of terms and their verbal collocates in the case of the Spanish data. These tables suggest that for English data, the patterns listed as Term + Verb 2 (that is, the verb is found two tokens to the right of the term) and Term + Verb 3 and especially Verb + Term 2 and Verb + Term 3 are the ones where most specialized collocations are formed. This indicates that terms in an object

Table 5.7: *Patterns used to extract CSCs in Spanish*

Slot 1	Slot 2	Example
Term (N + Adj + Adj)	+ Verb	tasa arancelaria aplicable proveer
Term (N + Adj)	+ Verb	servicio financiero excluir
Term (N + Prep + N + Adj)	+ Verb	derecho de propiedad intelectual cubrir
Term (N + Prep + N)	+ Verb	otorgamiento de licencia certificar
Term (N)	+ Verb	derecho adoptar
Verb +	Term (N + Adj + Adj)	negar trato arancelario preferencial
Verb +	Term (N + Adj)	autorizar entrada temporal
Verb +	Term (N + Prep + N + Adj)	ofrecer proveedor de servicio financiero
Verb +	Term (N + Prep + N)	determinar valor en aduana
Verb +	Term (N)	mantener medida

Table 5.8: *Patterns used to extract CSCs in English*

Slot 1	Slot 2	Example
Verb +	Term (Adj + N)	provide judicial authority
Verb +	Term (N + N)	apply taxation measure
Verb +	Term (N + Prep + N)	accrue date of expropriation
Verb +	Term (N)	maintain measure
Term (Adj + N)	+ Verb	applicable tariff provide
Term (N + N)	+ Verb	tariff classification require
Term (N + Prep + N)	+ Verb	restitution of property provide
Term (N)	+ Verb	tariff provide

Table 5.9: *CSC patterns extracted from the English data*

English data	N	A+N	N+N	N+P+N	Total
Term+Verb 1	3,221	646	362	107	4,336
Term+Verb 2	5,998	1,093	504	150	7,745
Term+Verb 3	5,533	905	376	122	6,936
Verb+Term 1	3,534	862	373	85	4,854
Verb+Term 2	8,012	1,478	614	360	10,464
Verb+Term 3	9,230	1,574	708	357	11,869

role in a direct syntactic relation with a verb more frequently form specialized collocations. The same observation holds for the Spanish data.

In the case of ATE, it is important to notice that adjectives that appear

Table 5.10: *CSCs extracted from the Spanish data*

Spanish data	**N**	**N+A**	**N+A+A**	**N+P+N**	**N+P+N+A**	**Total**
Term+Verb 1	2,571	977	76	388	96	4,108
Term+Verb 2	4,735	1511	107	579	67	6,999
Term+Verb 3	4,408	1,006	56	570	38	6,078
Verb+Term 1	1,688	336	17	164	22	2,227
Verb+Term 2	6,377	1,224	87	734	152	8,574
Verb+Term 3	7,354	1,355	114	764	137	9,724

closer to the term, usually a noun, might have more terminological relevance. In the case of the semi-automatic extraction of specialized collocations, verbs found at position -2 and -3 from the term are more likely to enter into a specialized collocation. That is why adjacency is not a definitive factor to identify a specialized collocation, as is the case for term extraction.

Following this, Table 5.11 presents the distribution of CSCs for the English and Spanish data after the tags that signal non-lexical verbs were excluded, such as the tags used for modal verbs. This table also provides evidence that most specialized collocations in the FTA corpus are formed when a verb appears three tokens before a term.

Table 5.11: *Cleaned list of CSC in English and Spanish*

Pattern	**Spanish**	**English**
Verb1+term	1,806	4,468
Verb2+term	7,786	9,812
Verb3+term	8,571	10,832
Term+verb1	3,189	3,577
Term+verb2	5,574	6,537
Term+verb3	4,602	5,980

In the case of the English data, the patterns and the top frequencies of verbal CSCs formed by a term in the subject role are described as follows. In the first case, the verb is found at one token to the right of the rightmost constituent of the term, expressed as Term + Verb 1 in Table 5.12.

When the verb is found two tokens to the rightmost constituent of the term, the distribution of CSCs is exemplified in Table 5.13, while Table 5.14

Table 5.12: *CSCs (Term + Verb1) extracted from English data*

Term type	Term+Verb 1	Frequency
N		3,221
	procedure refer, measure adopt, good provide, procedure lay, information provide	
A+N		646
	non-conforming measure refer, similar good use, national value add	
N+N		362
	calendar year specify, tariff rate provide, tariff classification set, investment scheme locate	
N+P+N		107
	notice of arbitration give, appointment of personnel recruit, term of office begin	
Total		4,336

illustrates the frequencies when the verb is three tokens from the term.

For the English data, the patterns and the top frequencies of verbal CSCs are described as follows. Table 5.15 presents examples and frequencies of CSCs when the verb is found one token to the left of the term. Table 5.16 applies to verbs found two tokens to the left of the term, while Table 5.17 exemplifies CSCs and their frequencies when the verb is found three tokens to the left of the term.

For the Spanish subcorpus, these are morphosyntactic patterns that form terms that have been queried to find a term to the left from the verb:

- Noun
- Noun + Adjective
- Noun + Preposition + Noun
- Noun + Preposition + Noun + Adjective.

Table 5.18 presents some examples and the frequencies with the most frequent CSCs including these patterns when the term is found one token to the left of the verb.

Further, Table 5.19 presents some examples with the most frequent CSCs including these patterns when the term is found two tokens to the left of the verb.

Table 5.13: *CSC (Term + Verb2) extracted from the English data*

Term type	Term+Verb 2	Frequency
N		5,176
	right to adopt, classification require, tariff provide	
A+N		1,093
	visible lining contain, applicable tariff provide, exclusive right authorize	
N+N		504
	tariff classification required, apparel article satisfy, animal hair knit, conformity assessment locate	
N+P+N		150
	term of office engage, restitution of property provide, period of time require, term of office save	
Total		6,923

Table 5.14: *CSC (Term + Verb3) extracted from the English data*

Term type	Term+Verb 3	Frequency
N		5,533
	date enter, tariff require, apparel satisfy, measure maintain	
A+N		905
	aggregate quantity enter, legislative procedure adopt, qualified majority define, legislative procedure establish	
N+N		376
	market access list, foreign person undertake, state enterprise maintain, market value expropriate	
N+P+N		122
	level of government set, level of government schedule, term of protection grant	
Total		6,936

Table 5.15: *CSC (Verb + Term1) extracted from the English data*

Term type	Verb+Term 1	Frequency
N		3,534
	cover investment, cover procurement, propose measure, import good	
A+N		862
	associate traditional knowledge, submit responsive tender, afford adequate opportunity, remedy serious injury, identify individual sector, pass specific examination	
N+N		373
	impede law enforcement, countervail duty law, countervail duty investigation, develop country party, maintain price stability	
N+P+N		85
	apply rate of duty, obtain recognition of qualification, restrict sale of good, follow rate of duty, develop exchange of information, submit statement of case	
Total		4,854

Table 5.16: *CSC (Verb + Term2) extracted from the English data*

Term type	Verb+Term 2	Frequency
N		7,512
	adopt measure, indicate note, maintain measure, reserve right, satisfy requirement, supply service	
A+N		1,345
	take necessary measure, enter aggregate quantity, govern public law, appoint common accord, calculate regional value	
N+N		582
	indicate note number, apply taxation measure, determine tariff classification, relate qualification requirement, apply custom duty, require business person, deny tariff treatment	
N+P+N		84
	apply rate of duty, obtain recognition of qualification, restrict sale of good, follow rate of duty, develop exchange of information	
Total		9,523

Table 5.17: *CSC (Verb + Term3) extracted from the English data*

Term type	Verb+Term 3	Frequency
N		8,435
	apply measure, export territory, set paragraph, use production	
A+N		1,392
	act qualified majority, provide judicial authority, define relevant law, prepare responsive tender	
N+N		651
	calculate value content, apply safeguard measure, introduce export subsidy, provide tariff item, confirm government share	
N+P+N		334
	arise list of commitment, describe list of commitment, dump amount of subsidy, favour allocation of resource, apply rate of duty, accrue date of expropriation	
Total		10,812

Table 5.18: *CSC (Term + Verb 1) extracted from the Spanish data*

Term type	Term+Verb 1	Frequency
N		2,070
	procedimiento prever, medida adoptar, capital suscribir, autoridad requerir, procedimiento establecer	
N+A		760
	parte contendiente acordar, producto originario comprender, propiedad intelectual relacionar, persona natural domiciliar, parte contendiente entregar	
N+P+N		302
	servicio de apoyo relacionar, requisito de capital contemplar, suministro del servicio integrar, suministro de servicio relacionar, regla de origen establecer	
N+P+N+A		66
	accionista de entidad financiera constituir, reserva del compromiso horizontal contraer, valor de contenido regional expresar, principio de trato nacional establecer, monto de arancel aduanero pagar	

Table 5.19: *CSC (Term + Verb 2) extracted from the Spanish data*

Term type	Term+Verb 2	Frequency
N		4,058
	derecho adoptar, fin garantizar, prenda satisfacer, medida otorgar, calendario especificar	
N+A		1,234
	trato especial diferenciar, tasa arancelaria proveer, servicio financiero excluir, derecho exclusivo autorizar, parte contendiente presentar	
N+P+N		453
	principio de contabilidad aceptar, derecho de propiedad relacionar, miembro del consejo representar, otorgamiento de licencia certificar, lista de compromiso figurar	
N+P+N+A		54
	operador de transporte multimodal efectuar, derecho de propiedad intelectual cubrir - pagar - condicionar - proporcionar - derivar, valor de contenido regional determinar, tasa de arancel aduanero corresponder	

Finally, Table 5.20 presents some examples with the most frequent CSCs including these patterns when the term is found three tokens to the left of the verb.

In the case of the Spanish subcorpus, the distribution of patterns and top frequencies of verbal CSC are described below.

When the verb is found at one token to the left of the leftmost constituent of the term, expressed as Verb 1 + Term in Table 5.11, the distribution and examples are presented in Table 5.21. Hyphens are used to separate the most relevant verbs that alternate with the same term to form a specialized collocation.

In the second place, when the verb is found at two tokens to the left of the leftmost constituent of the term, expressed as Verb 2 + Term in Table 5.11, the distribution and examples are presented in Table 5.22.

In third place, when the verb co-occurs at three tokens to the left of the leftmost constituent of the term, expressed as Verb 3 + Term in Table 5.11, the distribution and examples are presented in Table 5.23.

Table 5.20: *CSC (Term + Verb 3) extracted from the Spanish data*

Term type	Term+Verb 3	Frequency
N		3,784
	perjuicio conformidad, arreglo disponer, arreglo prever, material utilizar, derecho autorizar, tasa proveer	
N+A		813
	arquitecto extranjero requerir, lista permanente calificar, ventaja relativa ofertar, procedimiento legislativo adoptar, responsabilidad civil derivar	
N+P+N		439
	nave de bandera prestar, persona de negocio afectar, servicio de transporte definir, ejercicio de facultad contemplar, instrumento del mercado incluir	
N+P+N+A		30
	servicio de transporte parcial integrar, operador de transporte multimodal entender, agente de carga internacional actuar, valor de contenido regional especificar, tipo de servicio universal definir	

Table 5.21: *CSC (Verb 1 + term) extracted from the Spanish data*

Term type	Verb 1 + term	Frequency
N		1,570
	prestar - incluir servicio, adoptar - incluir - mantener medida, otorgar trato, suspender beneficio, realizar consulta, otorgar derecho	
N+A		303
	formar parte integrante, incluir medida relativa, otorgar trato nacional, solicitar entrada temporal, constituir parte integrante, solicitar trato arancelario, establecer acuerdo comercial	
N+P+N		130
	aplicar procedimiento de licencia, codificar portadora de programa, realizar despacho de aduana, aplicar prueba de necesidad, asumir compromiso de conformidad, prestar - suministrar servicio de transporte	
N+P+N+A		19
	conferir igualdad de oportunidad competitiva, utilizar nave de bandera colombiana, incluir medida en materia ambiental, formar parte del costo total, ejercer derecho de propiedad intelectual	
N+A+A		17
	otorgar - solicitar - negar - conseguir - obtener tratamiento arancelario preferencial	

Table 5.22: *CSC (Verb 2 + term) extracted from the Spanish data*

Term type	Verb2 + term	Frequency
N		6,033
	reservar derecho, mantener - aplicar - adoptar medida, certificar origen, establecer conformidad	
N+A		1,118
	tomar medida razonable, adoptar medida necesaria, autorizar entrada temporal, tratar personal extranjero	
N+P+N		675
	afectar comercio de servicio, suministrar ejercicio de facultad, calcular valor de contenido, incluir servicio de transporte	
N+P+N+A		145
	suministrar ejercicio de facultad gubernamental, calcular valor de contenido regional, infringir derecho de propiedad intelectual	
N+A+A		80
	derivar acuerdo comercial internacional, negar - suspender - solicitar trato arancelario preferencial, suspender procedimiento legislativo ordinario	

Table 5.23: *CSC (Verb 3 + term) extracted from the Spanish data*

Term type	Verb3 + term	Frequency
N		6,618
	indicar nota, cumplir valor, aplicar medida, establecer anexo, disponer - contemplar - prever apartado, cumplir requisito, acondicionar venta	
N+A		1,355
	consolidar establecimiento directo, eliminar etapa anual, aplicar medida tributaria	
N+P+N		637
	cumplir valor de contenido, garantizar - ofrecer - otorgar proveedor de servicio	
N+P+N+A		103
	cumplir valor de contenido regional, permitir - ofrecer - otorgar - exigir - autorizar proveedor de servicio financiero	
N+A+A		91
	eliminar - reducir etapa anual igual, asignar banco central nacional	

5.5.1 List of terms that appear in the top-100 list of specialized collocations

A comparison of the terms that appear in the top distribution of specialized collocations from 1 to 3 tokens from a verb to the left or to the right, reveals that 54 terms are common to these lists. All of the terms identified in the comparison are simple lexemes. These terms are:

> action, agreement, application, arbitration, authority, body, capital, case, contract, cooperation, country, decision, dispute, duty, enterprise, entity, good, information, interest, investigation, investment, investor, law, level, majority, material, matter, means, measure, notice, origin, panel, paragraph, party, period, person, policy, procedure, process, procurement, producer, product, production, protection, report, request, review, right, service, supplier, territory, trade, treatment, value.

Many of these terms common to specialized collocations also emerge in the "cloud" of terms extracted with Termostat (Figure 5.2), such as *agreement, good, paragraph, entity, service, procurement, request, supplier, party, measure* and *territory*. These comparisons suggest that these terms represent central notions within trade agreements. These terms, the concepts they embody in addition to the specialized collocations they form, should therefore be relevant in LSP courses in the domain of international trade.

The list above includes terms from contract law such as *arbitration, contract, dispute, law, panel, person, procedure* and *process*, or terms from the field of economics, such as *producer, product, production, supplier* and *trade*; also, the geographic area where they are applied, for example with the terms *country* and *territory*.

5.5.2 Examples and frequencies with a particular term and its verbal collocates

A term can take many verbs that enter into a collocational relation with it. For example, Table 5.24 shows the candidate specialized collocations of Spanish term *arancel aduanero* extracted with IMS CWB. This is indeed a

relevant term in FTA texts. Several of the verbal collocates that this term takes are semantically related, such as synonym verbs *incrementar, acelerar* and *aumentar*, *adoptar* and *aplicar* or are antonyms, such as *reducir* and *eliminar*.

Table 5.24: *Candidate specialized collocations of Spanish term* arancel aduanero *extracted with IMS CWB*

Frequency	**Term**
10	incrementar un arancel aduanero
4	aplicar un arancel aduanero
4	aumentar un arancel aduanero
3	incrementar ninguno arancel aduanero
2	adoptar ninguno arancel aduanero
2	reducir suyo arancel aduanero
1	acelerar del arancel aduanero
1	adoptar un arancel aduanero
1	aplicar el arancel aduanero
1	eliminar el arancel aduanero
1	incrementar el arancel aduanero
1	reducir un arancel aduanero

One equivalent English term for the Spanish term *arancel aduanero* is *custom duty*. To match all the results under one query, the lemma *custom* was preferred over the word form *customs*. Table 5.25 presents the verbal collocates for this term in the English subcorpus. Here, we can also see synonym verbs that serve as collocates for the term, such as *apply* and *adopt* or *increase, mantain* and antonym verbs *reduce* and *raise*. Besides, regional differences between European and American English are seen in verbs such as *favour* and *favor*.

Table 5.26 shows the top 20 collocates of Spanish noun *procedimiento* extracted with Xaira, using the Z-score AM, searching one item to the left and one to the right.

Table 5.25: *Candidate specialized collocations of English term* custom duty *extracted with IMS CWB*

Frequency	**Term**
10	increase a custom duty
5	apply a custom duty
3	favour nation custom duty
3	pay any custom duty
3	raise a custom duty
1	adopt any custom duty
1	apply the custom duty
1	favor nation custom duty
1	impose the custom duty
1	increase any custom duty
1	maintain any custom duty
1	reduce a custom duty

Word	**Frequency**	**Z-score**
legislativo	146	241.7
previsto	74	96.3
arbitral	34	74.8
al	212	53.0
conducente	5	46.2
un	212	45.0
el	302	35.0
abreviado	2	32.0
jurisdiccional	8	30.4
simplificado	3	30.3
administrativo	16	30.2
establecido	31	29.8
análogo	2	26.1
ante	20	25.9
siguiente	1	22.6
Contradictorio	1	22.6
Patentado	5	22.0
Expedito	2	18.4
Contemplado	5	17.9

Table 5.26: *20 top frequent collocates of Spanish noun* procedimiento *extracted with Xaira*

5.6 Gold standard of terms in the specialized dictionaries and term bases

By means of some in-house Python scripts, I compared the terms from the gold standard of 441 terms with a set of specialized dictionaries and term bases listed in Section 4.3. The comparison was made to see to what extent terms found in FTA texts are also listed as entries in specialized dictionaries from the subject fields of international trade, economics, accounting, finance, banking, business and enterprise. These lexical resources, which are listed in Section 4.3, comprise, in total, 69,643 terms but once duplicates were removed, 64,521 unique terms remained.

The comparison revealed that 185 of the terms included in the gold standard are included in these lexical resources. This represents 41.9% of the total of 441 terms. In contrast, 253 of the terms from the gold standard are not included in these resources, representing 57.5% of the terms. This information can provide insights into the new FTA terms that could be included in future specialist dictionaries or online WTO glossaries dealing with international trade and the field of FTAs or IATE or other similar terminological resources such as Termportalen, the national terminology portal for Norway.[50] These terms could also be relevant for LSP courses in the field of international trade and courses related to FTAs.

5.7 Analysis

After the extraction of the terms and the candidate specialized collocations was carried out, several observations regarding the extraction can be made. First, three-word terms, such as those formed by the patterns Adjective + Noun + Noun or Noun + Preposition + Noun are less frequent in the corpus than terms formed by a noun or Noun + Adjective. However, the use of morphosyntactic patterns to extract these less frequent complex units from the corpus produces less noise than other more frequent patterns, such as simple terms composed by a noun. Second, the preposition *of*, (*de* in Spanish)

[50] `http://www.terminologi.no/`

should not be discarded from the patterns because it is a frequent lexical item appearing in terminological units, especially in Spanish. Regarding adverbs, only deadjectival adverbs ending in *-ly* or *-mente* in Spanish have been taken into account.

From the observations made on the data extracted from the FTA corpus, it is evidenced that the majority of the verbal collocates of both the terms of the gold standard and the list of candidate terms enter into specialized collocations when the verbal collocates are found at three tokens to the left from the term. This means that, presumably, the term syntactically occupies the object role. The second place is occupied by collocates that co-occur at two tokens to the left from the term in the same role of object.

The following subsections present the types of qualitative analyses that have been carried out. First, the morphosyntactic analysis is presented, followed by semantic, terminological and pragmatic analyses.

5.7.1 Morphosyntactic analysis

Several morphosyntactic patterns constitute specialized collocations. However, the focus here is set on the most frequent patterns among these.

The terms employed in FTA texts exhibit the canonical features of specialized texts. There is a preference for nominalization, expressed linguistically by the frequent occurrence of deverbal nouns. For example, in the gold standard of 441 terms, there are 40 occurrences of terms ending with the suffix *-tion*, corresponding to 9% of the terms, such as *information, application, authentication, legislation, consideration, importation, communication* and *authorization*. The verbs that co-occur with terms and form specialized collocations can also take a morphological realization as deverbal nouns. To illustrate, the specialized collocations formed by the term *measure*, occur 42 times in the corpus in conjunction with these verbal collocates, in order of frequency: *maintain, apply, adopt, mean, impose, enforce* and *execute*. In turn, the deverbal nouns *maintenance, application, adoption, meaning, imposition, enforcement* and *execution* can also form a specialized collocation with the term *measure*. Thus, *to adopt a measure* and *measure adoption* are both terminologically relevant and observable in the context of FTA texts.

The terms found in the gold standard emphasize the most relevant notions related to the implementation of FTAs. The most productive and frequent among the 441 items found in the gold standard of terms are the adjectives *agricultural, commercial, financial* and *public* and the nouns *customs, goods, government, import, information, investment* and *service*. These terms are also quite frequent in the formation of specialized collocations in the FTA corpus.

As evidenced in the data, the collocational relation among two or more lexemes is kept, despite the morphosyntactic processes that they might undergo. Hence, the verbs that co-occur with terms and form specialized collocations exhibit specialized features such as the ones held by the realization of their counterpart deverbal nouns. These deverbal nouns and their verbal realization, which frequently co-occur with terms in the FTA corpus frequently denote processes.

A relevant morphosyntactic pattern that is not taken into account by Termostat is the one conformed by Adverb + Adjective / Past Participle + Noun. Even though it is not very frequent, it is still relevant from a terminological point of view. Its structure serves to synthesize a whole sentence with less words, in harmony with the preference of conciseness which is a relevant feature of terms (Cabré, 1999; Gotti, 2003). Using this pattern, 108 candidate specialized collocations were retrieved from the data. For example, the most frequent is *directly competitive good* with 17 occurrences, followed by *mutually satisfactory solution* with 16 occurrences. Other units that correspond to the same pattern are constituted by *freely usable currency, substantially equivalent trade* and *economically disadvantaged minorities* with 15, 12 and 10 occurrences, respectively.

Some verbs are highly productive in the formation of specialized collocations in FTA texts. The following are representative examples of these verbs: *include, apply, provide, require, use, make, maintain, relate, submit, permit, supply, allow, designate, grant, regard, adopt, affect, establish, authorize, consider, constitute* and *identify*. These verbs are associated with processes. For example, the verb *include* co-occurs with 33 terms from the gold standard, such as *commission, debt, domestic support, entry, establishment, financial institution, financial service, financial service supplier, good*

and *government*. Another productive verb is the verb *adopt* that forms a specialized collocation with 8 terms from the gold standard of terms and with 71 of the candidate terms. The specialized collocation *adopt decision* comes on top with 46 occurrences, followed by *adopt measure* with 24 occurrences. Other frequent collocates of the verb *adopt* are FTA terms *consultation, agreement, customs duty* and *safeguard measure*. The verb *apply* co-occurs as a collocate of 28 terms, many of them being complex terms found in the gold standard of terms, such as *agricultural safeguard measure, commercial presence, covered procurement, customs duty, decision, good, import licensing, measure, preferential tariff treatment, procurement, safeguard measure* and *sanitary or phytosanitary measure*.

Another frequent case is constituted by the verb *require* which collocates with 27 terms from the gold standard such as *business person, collective investment scheme, financial institution, importer, information, insurance, investment, respondent, service* and *service provider*. Next appears the verb *provide*, which collocates with 26 terms such as *information, written, service, information, financial service, telecommunications regulatory body* and *good*.

5.7.2 Semantic analysis of CSCs

According to Gallegos (2003), it is problematic to attribute certain linguistic features exclusively either to general language or to specialized language. They are instead interdependent subsystems of a language.

Gallegos (2003) points out that specialized languages exhibit some features such as the following:

1. specialized languages display a certain functional style;
2. they share a specific semantic field and
3. they are typically assigned to a determined social group.

All of these conditions hold for FTA texts.

Relational adjectives are frequent and relevant lexical items occurring in specialized texts, forming part of terms (Daille, 1999, 2001). This type of adjectives are carriers of a naming function, thus, they are closely related

to terms. For example, 308 different adjectives form specialized collocations with terms with the pattern Adjective + Noun, co-occurring three tokens after the verb. Some relational adjectives in this sample are *administrative, advisory, agricultural, confidential, constitutional, inter-governmental, financial, official, procedural* and *juridical*, where all of them are denominal adjectives, therefore, are closely related to a noun with terminological relevance in FTA texts.

In the case of Spanish, some examples of the most frequent adjectives in the FTA corpus are: *nacional, público, financiero, comercial, internacional, relativo, arancelario, material, regional, competente, contendiente* and *arbitral*.

In the following paragraphs I present some findings based on the FTA data and the features of specialized languages attributable to specialized collocations in common with terms.

An experiment to inquire into the lexical features of specialized collocations was carried out by the combination of three NLP tools, based on the method suggested by Burgos (2014): the combination of Freeling (Padró and Stanilovsky, 2012), Princeton's WordNet (Miller, 1995) and NLTK (Bird et al., 2009).[51] The aim of the experiment was to automatically tag the sense of the lexical items with the senses annotated in Wordnet, as listed in the Appendix, Table 3. This experiment involved 1,589 candidate specialized collocations including the lexical units that co-occur with the gold standard of terms and which had been extracted previously using the program *CWB-Scan-Corpus* included in the IMS CWB toolkit.

In the case of nouns, the results of the experiment indicate that most of them correspond to nouns used for acts or actions, such as *claim, custom, decision, duty, enterprise, establishment, investment, procurement, qualification* and *safeguard*. In the second place come nouns tagged as attributes, for example: *ability, agreement, authority, information, jurisdiction, purpose, service* and *value*. In the third place come nouns tagged as related to people, for example: *arbitrator, claimant, importer, investor, mechanism, order, provider, respondent* and *supplier*. In the fourth place in the rank appear

[51] Special thanks are due to Associate Prof. Dr. Diego Burgos for kindly providing the scripts that combine Freeling and NLTK and tags the senses of the specialized collocations.

nouns corresponding to artifacts, such as *aircraft, apparel, body, component, container, document, material, textile, good* and *vehicle*. In the fifth place appear nouns corresponding to locations, namely the countries participating in FTAs.

In the case of verbs, most of them have to do with cognition, communication and change. For verbs that were tagged as corresponding to cognition, there are 75 of these, for example: *accept, allow, approve, assess, assign, associate, base, calculate, choose, classify, conclude, connect, consider, construe, consume, control* and *decide*. Next, there are 68 verbs related to communication, such as *to advertise, advise, agree, annotate, appeal, apply, ask* and *authorize*.

In addition to this, as expected in texts related to international trade such as the texts that make part of the FTA corpus, verbs related to buying, selling or owning are relevant lexical units in this kind of texts. For example, the verbs *to own, store, purchase, finance, trade, earn, furnish, award* and *possess*. These verbs are specialized, and are carriers of relevant semantic information in conjunction with terms. These verbs enter into specialized collocations such as these: *furnish access to information, own financial institution, owe financial institution, store carrier medium, owe financial service supplier, expropriate transfer, calculate value, purchase enterprise, resell good, owe importer, trade product, stock product, trade relevant market, award supplier* and *sell transfer*. Table 5.27 presents the complete list of verbs found in the cognition category, whereas Table 5.28 presents the verbs found in the communication category and Table 5.29 comprises the verbs found in the change category.

Table 3 in the Appendix presents the most relevant WordNet semantic classification in descending order of frequency, extracted for English candidate specialized collocations. On top of these results, we can see the categories mentioned above related to nouns and verbs. The nouns denote actions, attributes, events, artifacts and locations while the verbs are related to cognition, communication and change. This type of lexical units is expected to appear frequently and preeminently in a specialized text. In contrast, at the bottom of the list we can see that verbs of perception, nouns of shape and verbs of emotion only appear once or twice in the data. This is supported

Table 5.27: *WordNet classification of English cognition verbs in candidate specialized collocations*

WordNet category	**Verbs**
Cognition	accept, allow, approve, assess, assign, associate, base, calculate, choose, classify, conclude, connect, consider, construe, consume, control, decide, demonstrate, describe, design, designate, detail, determine, discredit, disregard, distinguish, divide, earmark, elect, ensure, establish, estimate, except, exclude, favor, favour, hear, identify, influence, inspect, intend, interpret, link, maintain, mean, name, offer, pay, prejudice, propose, prove, rate, reason, recognise, recognize, reexamine, refer, reflect, regard, register, reject, relate, rely, reserve, respect, review, schedule, select, show, specify, submit, support, test, understand, verify

Table 5.28: *WordNet classification of English communication verbs in candidate specialized collocations*

WordNet category	**Verbs**
Communication	admit, advertise, advise, agree, annotate, appeal, apply, ask, authorize, avoid, bear, cancel, cause, challenge, circumvent, claim, commit, communicate, confer, contact, declare, define, deny, disclose, dispute, disseminate, distribute, encourage, entitle, execute, force, grant, import, impose, indicate, investigate, invite, mention, negotiate, notify, oblige, permit, precede, present, prevail, programme, prohibit, promote, publicize, publish, reach, record, request, require, revise, say, seek, sell, send, sign, speak, subject, supply, threaten, transmit, undertake, wish, write

Table 5.29: *WordNet classification of English change verbs in candidate specialized collocations*

WordNet category	**Verbs**
Change	abuse, accrue, accumulate, adapt, address, adjust, adopt, advance, amend, appear, become, begin, broaden, change, commercialize, complete, confirm, conform, continue, convert, deepen, delay, deprive, derive, detain, dilute, diminish, distort, disturb, edit, enable, engage, enhance, enter, exchange, exit, facilitate, fail, find, go, implement, improve, include, incorrect, increase, incur, inform, introduce, involve, issue, keep, limit, locate, mark, market, match, measure, modify, nationalize, number, obtain, order, prepare, preserve, privatize, process, provide, qualify, regulate, remove, replace, represent, restrict, result, scramble, settle, shape, start, structure, substantiate, tender, terminate, transpose, wear, withdraw

by Gotti (2003) who signals that lack of emotion is a feature of specialized discourse while Cabré (1999) lists impersonalization as another feature of this type of texts.

Koike (2002) divides verbs that form collocations into functional and lexical and subsequently he subdivides each one of these categories into general and specific verbs. According to Koike, lexical specific verbs collocate with less nouns than general verbs and therefore provide for a stronger semantic link. Once this link is created, several verbs that display "straight" semantics, i.e. its meaning is transparent, become semantically neutralized. Thus, even though, taken in isolation, some verbs are not synonyms among themselves, their meaning becomes synonymous once they collocate with a given noun. For example in the FTA data we have the verbs *apply, adopt* and *impose* which collocate with the term *custom duty*. The same author also holds that abstract nouns tend to combine with the figurative sense of the verb. Therefore, verbs tend more to specialize their straight meaning with abstract nouns than with concrete nouns.

Drouin (1997, 2004) reports how heterogenous corpora comparison and comparison against non-terms are effective approaches to extract terms, especially to detect simple terms. In these papers, he proposes a method that opposes technical data against a reference non-technical data as a way to evidence the terms that are unique to the technical data. Similarly, I carried out an experiment to compare the terms that appear in the FTA corpus with a radically different data set, which would help contrast terms that are highly frequent in FTAs in contrast to general language. To attain this, a component of the OPUS corpus (Tiedemann, 2012), namely, the English-Spanish section of the OpenSubtitles 2011 corpus, was selected. This corpus contains parallel data from aligned movie subtitles from the website OpenSubtitles.org. The aim of the experiment was to find the key words of the FTA corpus, by using the Keywords program from the Word Smith Tools (Scott, 2007) to compare the two different data sets as a way to evidence the most relevant words in the FTA data.

Next, the second part of the experiment consisted of looking for the usual collocates of this top-100 key words, in a window of 5 tokens to the right and to the left from the node term, namely, each of the 100 key words.

The top-100 English key words in the strict order as they were extracted from the data are presented in Table 5.30. Some of these key words refer to the objectives set forth in FTAs, for example, *cooperation, development, provisions, tariff, customs, regulations, obligations, arbitration, dispute, rights, procurement* and *production*. Other terms refer to the type of interchange and business that is regulated through FTAs, for example: *services, goods, trade, investment, protection, financial, information* and *telecommunications*.

FTAs provide and regulate all kinds of services among trade partners. Processing the data with Termostat (Drouin, 2003) reveals that indeed the term *service* is quite relevant in FTA texts because this term enters into 133 terms and specialized collocations. Some of the collocates that the term *service* frequently takes in the FTA corpus are: *foreign, auditing, protection, certification, intermediary, printing, integrated, satellite, data, dental, settlement, cross-border* and *specialized*. In addition to this, FTAs provide multiple types of measures to regulate trade among nations. For the term *measure*, Termostat identifies 68 terms and specialized collocations. Some

of the most relevant examples of specialized collocations including the term *measure* appearing in the FTA corpus are the following: *measure building, anti-dumping measure, policy measure, duty measure, subordinate / actual / definitive / transitional* and *incentive measure*.

Other relevant terms in the top keywords from the FTA corpus refer to geographic areas where FTAs are applicable: *territory, European, international* and *domestic*, while other terms refer to the internal organization of FTA texts, such as *article, chapter, paragraph* and *subparagraph*.

This list of the top-100 English key words also comprises some frequent verbs that frequently emerge in specialized collocations from the FTA corpus, such as *provide, apply, request, include* and *ensure*.

Table 5.30: *Top-100 English keywords with the OpenSubtitles2011 as contrast corpora*

article	apply	ensure	authorities
party	customs	entry	origin
agreement	suppliers	purposes	regulations
services	measure	pursuant	wto
parties	request	commission	activities
chapter	procurement	subject	arbitration
measures	service	conditions	telecommunications
paragraph	application	relevant	administrative
goods	products	obligations	cooperation
provisions	national	means	authority
trade	committee	relating	applicable
member	members	articles	basis
provided	treatment	treaty	procedure
territory	international	domestic	materials
european	subheading	economic	implementation
accordance	panel	persons	originating
procedures	rights	supplier	entities
financial	dispute	consultations	include
including	investment	date	disputing
information	tariff	period	subparagraph
referred	public		agreements
states	requirements	established	protection
council	section	respect	heading
union	technical	related	provision
provide	entity	development	production

5.7.3 Terminological and pragmatic considerations

FTA terms, in conjunction with the terms used in other subject fields, also tend to be concise, nominal and impersonal (Cabré, 1999; Gotti, 2003). The candidate specialized collocations found in the previous paragraphs illustrate this conciseness. For example, the concept of a *directly competitive good* could otherwise be expressed as *a good that is directly competitive*.

Terms composed by the pattern Adjective + Noun + Noun, such as *agricultural export subsidy, agricultural safeguard measure, collective investment scheme, economic integration agreement* and *financial service supplier* are themselves specialized collocations. The reason for this is that the Noun + Noun segment in this combination is already a term. Their termhood can be confirmed by consulting the *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English* (Alcaraz and Castro, 2007), which includes the terms *export subsidy, safeguard measure* and *service supplier* as entries. The adjective in this pattern that co-occurs with a Noun + Noun term modifies it in various ways: morphosyntactically, semantically, pragmatically, terminologically and phraseologically and in this way provides valuable lexical information that contributes to delineate the domain-specificity of FTAs. This emphasizes the role that adjectives play in the semi-automatic extraction of terms and specialized collocations as well as in other NLP tasks.

In addition to this, some terms from the gold standard and the candidate terms extracted with Termostat display a degree of terminological variation. The topic of terminological variation has been the focus of recent research (Freixa, 2003; Suárez, 2004; Freixa, 2006; Fernández, 2011). One example to illustrate it is the Spanish term *tasa arancelaria*, which is realized in the English subcorpus with three equivalents: *customs duty* (with 130 occurrences), *rate of duty* (70 occurrences) and *rate of customs* (47 occurrences). The term *rate of duty* collocates as object of these verbs: *increase, apply* and *raise*. In the case of the term *customs duty*, it collocates with *adopt* and *increase* and the term *rate of customs* collocates with *determine, apply* and *qualify*. Thus, the method proposed in this work to address the semi-automatic extraction of specialized collocations could also be used to detect

the term variants which can be useful both for term harmonization or term variation purposes. To this end, morphosyntactic patterns could be used in conjunction with an anchor list of terms, by means of using a lexical item found in a complex term to match candidate variants. For example, we could include in that anchor list the terms *duty*, *rate* or *customs*, which are highly frequent in FTA texts.

This terminological variation could be explained by different trade professionals or even translators intervening in the writing of a FTA text or by the various language variants represented in the corpus, besides its geographic, time and origin peculiarities.

The comparison between the terms included in the lexical resources composed by specialized dictionaries and term bases from the subject field of international trade, economics, accounting, finance, banking, business and enterprise suggests that there is a high degree of exclusion of terms (57.5%) from FTA texts that are not included in specialist dictionaries. One reason for this might be that some of the FTA texts included in the corpus have been enacted in the last few years and current dictionaries do not include many of the terms from these FTA texts.

Additionally, terms exhibit particular pragmatic features. Cabré (1999, 112) enumerates the following pragmatic factors to differentiate terms from general words:

1. the basic purpose
2. the subject dealt with
3. the users
4. the communicative situations in which both codes are found
5. the types of discourse in which terms or general language words appear.

The terms and specialized collocations that appear in FTA texts are no exception. In the case of the terms that occur in FTA texts, they have a basic purpose, namely to serve as a body of norms for international trade among nations or blocs of nations. Besides, the FTAs deal with specific subjects that regulate such trade. Also, these texts also have specific users, namely

governments, private firms and supranational entities that engage in international trade. Finally, FTA texts also respond to a particular communicative situation and incorporate specific types of discourse in which its terms appear.

In addition to the above factors, FTA texts are restricted to a narrow domain and can be highly frequent within this domain but outside such a domain, they do not occur often or do not occur at all. Thus, the frequency of co-occurrence of two or more lexical items is indicative of its pragmatic features that restrict their co-ocurrence only in association to a particular context. To illustrate this point, I will reuse an example from Section 2.11. In the FTA corpus, the Adjective + Noun + Noun collocation *preferential tariff treatment* appears 70 times in the FTA English subcorpus with roughly 1.5 million words.

Figure 5.4: *Presence of the term* preferential tariff treatment *in Google Books Ngram Viewer (1800-2008)*

In contrast, the Corpus of Contemporary American English (COCA) (Davies, 2009) does not offer any occurrence of *preferential tariff treatment* even though this corpus contains 520 million words. The COCA corpus only has 5 occurrences of *preferential tariff*, all of them extracted from economic newspapers discussing free trade topics. Figure 5.4 illustrates the occurrences

of the term *preferential tariff treatment* in Google Books Ngram Viewer from 1800 to 2008 (Lin et al., 2012). The figure indicates that this term was non-existent throughout most of the 19th century and became popular in the 1940s, coinciding with the GATT agreement and then became even more frequent in the 80s and 90s, when the majority of current FTAs were sanctioned and entered into force. Therefore, the co-occurrence of several lexical items forming a specialized collocation can be associated with a particular domain where it is employed habitually by a professional community. However, outside the context of such professional community belonging to a particular domain, its use is not frequent or even non-existent.

This chapter has offered a description and classification of the most frequent patterns that form specialized collocations that appear in FTA texts. Several considerations were made regarding the morphosyntactic, terminological, phraseological, semantic and pragmatic features that characterize the most frequent morphosyntactic patterns of terms and its verbal collocates that form specialized collocations in the FTA corpus in the English and Spanish data. Both a gold standard of terms and the terms extracted semi-automatically were the basis to perform these analyses.

The following chapter presents a proposal to represent specialized collocations in lexical resources.

CHAPTER 6

Representation of specialized collocations in language resources

6.1 Introduction

This chapter discusses the computational representation of specialized collocations in machine-readable dictionaries (MRDs). Special attention is given to two ISO standards designed for these purposes, namely the Lexical Markup Framework (LMF) and the Terminological Markup Framework (TMF), which were enumerated in subsection 2.6. This chapter discusses the suitability of these ISO standards for encoding linguistic information in computational lexicons, to be able to represent specialized collocations and other phraseological information. Then, a proposal for the computational representation of specialized collocations using one of these standards is made.

6.1.1 The Lexical Markup Framework (LMF)

LMF developers had in mind the idea of designing a metamodel for the creation of two types of computational resources: lexicons designed for NLP and the ones designed for MRD (Francopoulo and George, 2013). According

to its designers (Francopoulo et al., 2006a), its goals are:

- to provide a common model for the creation and use of lexical resources
- to manage the exchange of data between and among these resources, and
- to enable the merging of a large number of individual electronic resources to form extensive global electronic resources.

In its official document, the LMF is defined as "an abstract metamodel that provides a common, standardized framework for the construction of computational lexicons" (ISO, 2008, 5). This standard supports Unicode for the treatment of data in any language. Both the LMF and the TMF are standards built upon the notion of data category registries (DCR) to encode the metadata.

A DCR is defined in the TMF standard as "a set of data category specifications on which any specific TML [Terminology Markup Language] shall rely for creating its own data category set" (ISO, 2001, 8). Its function is to standardize the form in which metadata for extensions or modules are declared. These modules aim to cover several linguistic levels for the treatment of morphology, syntax and semantics. In the case of the LMF, it also handles the representation of equivalent information for translation and MWEs as a means to ensure interoperability among monolingual, bilingual and plurilingual lexical resources.

Because of its modular design, with the LMF a project can be deployed in less time, by using only the relevant modules from the set of available options. This means that there is no need to encode, for instance, the translation information of a lexicon that is only concerned with syntax.

In addition to this, as is usually done in XML-codified data, the data is structured in relation to a Document Type Definition (DTD), which defines the valid data categories according to the project developers. According to Harold and Means (2004), DTDs are written using a formal syntax which explains the exact elements that may appear in the document, their precise location, contents and attributes.

This standard is conceived to work by means of web services, which facilitates performing the queries without the need of downloading huge amounts

of data. Thus, the system displays the data on the screen in a format that is readable by humans or replies to a particular request from a computer program without the need of representing the whole resource.

The LMF standard includes a module for the representation of MWEs, known as NLP Multiword Expression Pattern. In principle, such a module enables the representation of the internal structure of fixed, semi-fixed and flexible lexical units in a computational lexicon (Francopoulo et al., 2006a,b, 2009; Francopoulo and George, 2013), such as the types of MWEs listed in Section 2.8.1. The LMF also includes an extension for dealing with bilingual or multilingual dictionaries, designed to express equivalence relations from the level of sense or transference, which could be used in automatic translation (ISO, 2008).

6.1.2 The Terminological Markup Framework (TMF)

The TMF standard aims to be a metamodel for a Terminological Markup Language (TML) with the objective of providing the infrastructure for the computational representation of terminological data by using XML technology. It is aimed at the standardization of terminological data representation (Romary, 2001).

This standard is built upon the principles of interoperability and blind interchange of data without loss of information. It differs from lexicographical metamodels because it is onomasiological rather than semasiological.

The TMF standard does not specify a separate and detailed module for the representation of phraseological units. Its DTD incorporates some basic data categories for the inclusion of phraseological units, such as *TermType*, which among other tags, takes the attributes *collocation, formula, phrase* and *setPhrase*. Also the data category *terminologicalEntryType* takes, among others, the attributes *collocation, phrase* and *setPhrase* (ISO, 2001). However, this standard only allows to encode MWEs such as specialized collocations as a whole unit and not in a granular way, to be able to specify the individual lexical items that make up a specialized lexical combination. This way, it would be possible to account individually for the lexical items in such a combination, in our case, a term that constitutes the node of the

specialized collocation and the usual collocates that such a node takes. This, in my view, makes the TMF standard unsuitable for the representation of specialized collocations.

6.1.3 The TermBase eXchange (TBX)

The TermBase eXchange (TBX) norm is the ISO standard 30042:2008. It was developed under the Localization Industry Standards Association (LISA). Since LISA ceased to exist in 2011, two identical versions coexist:[52] the original version, released under a Creative Commons license,[53] and the ISO version. Just as the TMF, the TBX also incorporates among its pre-established data categories a *termType* specification that accepts *phraseologicalUnit* as one of its valid attributes.

However, neither the OLIF (Open Lexicon Interchange Format) nor the TBX standards specifically provide a module that specifically includes a component to represent MWEs directly, such as specialized collocations.

As described in Parra et al. (2013), TBX's DTD is extremely flexible. This flexibility simultaneously constitutes an advantage and a disadvantage for the implementation of this standard for terminological and lexical resources. Thus, the user may modify and adapt the DTD to suit his/her needs but at the same time this flexibility could hamper the lossless interchange of information.

The TBX standard was primarily developed for localization and translation. Therefore, it is focused on bilingual or multilingual resources to be used by translators and terminologists but not on the needs of monolingual resources.

Regarding the representation of MWEs such as specialized collocations, the TBX standard does not foresee how to encode these lexical units with NLP tools. This implies that MWEs can only be represented as long strings, not as several elements that make up a phraseological unit. As a consequence, it is not possible to encode the node and the usual collocates that this node may take on a specific domain, which makes the TBX standard less adequate

[52] `http://www.tbxinfo.net`
[53] `http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf`

for the representation of specialized collocations. Thus, it seems to me that an updated version of this standard should allow for granularity to be able to encode data at the token level to be suitable for NLP and specialized phraseology needs.

The following section discusses how specialized collocations can be represented in language resources, using a standard for the development and exchange of computational lexicons.

6.2 Proposal for the representation of specialized collocations in language resources

To be able to represent specialized collocations in a language resource that aims to be reusable and interoperable with other language resources, first of all it would be necessary to mark the node of the collocation as a term and to specify in which specific domains(s) it is used. This implies the incorporation of the relevant semantic and pragmatic information related to the term under consideration.

Second, it should encode the most usual collocates that frequently co-occur with this term, thus forming a specialized collocation with that particular term, and which co-occur in the same domain.

In addition to this, information on syntactic, morphological, pragmatic as well as regional aspects should be encoded to account for the multiple realizations of these units in different varieties of the same language.

The following text presents an example of XML code based on the LMF standard for the bilingual representation from the level of sense, using as an example the English collocation *"preferential tariff treatment"*:[54]

```
<?xml version=‘‘1.0’’ encoding=‘‘UTF-8’’?>
<LexicalResource dtdVersion=‘‘16’’>
<GlobalInformation>
<feat att=‘‘label’’ val=‘‘Representation of a specialized collocation’’/>
<feat att=‘‘comment’’ val=‘‘English specialized
        collocation preferential tariff treatment’’/>
```

[54] This XML code is based on a proposal of LMF implementation made by Gil Francopoulo, available here as of September 4, 2016: `http://www.tagmatica.fr/lmf/FrenchLMFTestSuites2.xml`

```
<feat att=‘‘languageCoding’’ val=‘‘ISO 639-3’’/>
</GlobalInformation>
<Lexicon>
        <feat att=‘‘language’’ val=‘‘eng’’/>
<LexicalEntry mwePattern=‘‘AdjNN’’>
  <feat att=‘‘partOfSpeech’’ val=‘‘SpecCol’’/>
<Lemma>
  <feat att=‘‘writtenForm’’ val=‘‘preferential tariff treatment’’/>
</Lemma>
<ListOfComponents>
        <Component entry=‘‘E1’’/>
        <Component entry=‘‘E2’’/>
        <Component entry=‘‘E3’’/>
</ListOfComponents>
</LexicalEntry>
<LexicalEntry id=‘‘E1’’><feat att=‘‘partOfSpeech’’ val=‘‘adj’’/>
<Lemma><feat att=‘‘writtenForm’’ val=‘‘preferential’’/></Lemma>
</LexicalEntry>
<LexicalEntry id=‘‘E2’’><feat att=‘‘partOfSpeech’’ val=‘‘noun’’/>
<Lemma><feat att=‘‘writtenForm’’ val=‘‘tariff’’/></Lemma>
</LexicalEntry>
<LexicalEntry id=‘‘E3’’><feat att=‘‘partOfSpeech’’ val=‘‘noun’’/>
<Lemma><feat att=‘‘writtenForm’’ val=‘‘treatment’’/></Lemma>
</LexicalEntry>
<!-- Code for bilingual information -->
<SenseAxis id=‘‘SA1’’ senses=‘‘eng.preferential tariff treatment1
                        esp.trato arancelario preferencial1’’>
        <SenseAxisRelation targets=‘‘SA1’’>
                <feat att=‘‘label’’ val=‘‘SpecCol’’/>
        </SenseAxisRelation>
</SenseAxis>

<!-- Specialized collocation with the pattern Adj+N+N -->
<MWEPattern id=‘‘AdjNN’’>
<MWENode>
        <feat att=‘‘syntacticConstituent’’ val=‘‘SpecCol’’/>
<MWELex>
        <feat att=‘‘rank’’ val=‘‘1’’/>
        <feat att=‘‘graphicalSeparator’’ val=‘‘space’’/>
        <feat att=‘‘grammaticalNumber’’ val=‘‘singular’’/>
</MWELex>
<MWELex>
        <feat att=‘‘rank’’ val=‘‘2’’/>
        <feat att=‘‘graphicalSeparator’’ val=‘‘space’’/>
        <feat att=‘‘grammaticalNumber’’ val=‘‘singular’’/>
</MWELex>
<MWELex>
        <feat att=‘‘rank’’ val=‘‘3’’/>
        <feat att=‘‘graphicalSeparator’’ val=‘‘space’’/>
```

```
56          <feat att=‘‘grammaticalNumber’’ val=‘‘singular’’/>
57  </MWELex>
58  </MWENode>
59  </MWEPattern>
60  </Lexicon>
61  </LexicalResource>
```

The section marked with the comment

```
<!-- Code for bilingual information -->
```

on the line 31 of the code, can also be codified in LMF for the bilingual representation of equivalence from the level of transference, assuming that elsewhere in the code there is a reference to the ID of the TransferAxis:

```
1  <TransferAxis
2          id=‘‘SpecCol1’’
3          syntacticBehaviours=‘‘eng.preferential tariff treatment1
4          esp.trato arancelario preferencial1’’>
5  </TransferAxis>
```

However, this section of the standard could be modified to possibly facilitate a more direct implementation, possibly by taking other norms as a basis, such as the TMF.

The LMF standard introduces other data categories for the representation of dictionaries or terminological databases, such as *SourceLanguage* and *TargetLanguage*, which could even be used to express the equivalence relations in a more accessible code for users without a background in computer science.

6.3 Application

The analysis of this information to develop a representation metamodel can be useful for the constitution of multilingual term bases and ontologies, for corpus-based and corpus-driven term and collocation extraction and for ontology-based domain recognition of text.

The constitution of this kind of language resources is listed among the objectives of current or recent research projects in the field of language and technologies such as the EU-funded projects CLARIN, Common Language Re-

sources and Technology Infrastructure[55], META-SHARE, Multilingual Europe Technology Alliance[56], CLARA, Common Language Resources and their Applications[57] and national initiatives such as the project CLARINO, Common Language Resources and Technology Infrastructure Norway.[58]

6.4 Implementation and final remarks

Depending on the preferences, skills or technical aspects determined by the developers, there are several programming languages that would enable experts to perform automatic data treatment. This way, it could be processed into XML compliant code. Some available choices are Python's *ElementTree* module (Bird et al., 2009), Perl's *XML::Parser* module or the XSLT language, which is oriented toward the transformation of XML code into other formats or their representation on a web browser. Examples of data processing include the extraction of a lexicon section, importation or exportation of data and the conversion to other formats such as CSV, RTF, HTML or PDF. Some of these formats are designed to be read by humans (Tanguy and Hathout, 2007).

ISO standards designed for the standardization of language resources, such as the LMF and the TMF, deployed in XML format, offer a platform for the encoding of computational lexicons that is applicable in NLP applications, such as lexicography, terminology, computer assisted translation and machine translation, and also for the creation of electronic dictionaries for human users. Today, there is no single standard that is embraced by the industry and research communities. Nevertheless, some initiatives continue to be developed in projects that are aimed at the creation of reusable, interoperable, polytheoretical, multifunctional and interchangeable language resources without any data loss (Calzolari et al., 2013).

It is yet unknown whether standards such as the LMF or the TMF will be adopted by the worldwide terminology community as a standard to encode

[55] `http://clarin.eu/`
[56] `http:/www.meta-share.eu`
[57] `http://clara.b.uib.no/`
[58] `http://https://clarin.b.uib.no/`

lexical and terminological information, but they are certainly likely candidates. Also, the question remains as to whether commercial and open source translation and terminology management software packages will implement the option of being able to read, write and interchange data using these standards. The definition and adoption of these standards would be highly desirable for terminology and other language resources, both in the industry as well as in academia. Certainly, much effort has been carried out by several projects and it could be optimized and put to good use for the coming years and decades.

The final chapter presents the conclusions of this study, its limitations and perspectives for future work.

CHAPTER 7

Conclusions

The structure of the present chapter is the following. First, I assess the attainment of the hypotheses and objectives set forth at the beginning of the thesis, by using examples excerpted from the FTA corpus. Then, I continue with the contributions of this work. Next, I present the limitations of the present work and the lines for future research.

7.1 Testing of hypotheses

This section is aimed at the validation of the hypotheses set forth in Section 1.3 using the method described in Chapter 4, and the corpora described in Section 4.2.2. The hypotheses set forth at the beginning of the thesis are repeated in the following subsections for convenience.

7.1.1 First hypothesis

Specialized collocations contribute to delineating domain-specificity in a similar way as do the terms used in such a domain. Therefore, specialized collocations are part of specialized language. In the following discussion, I argue that the first hypothesis is supported.

The experiments described in Section 5.7.2 were carried out to assess the first hypothesis. The terms that are used in a specialized context are vital information for the specific subject matter being treated. Thus, they provide crucial information to delineate a domain-specificity. Whether the field in question is medicine, chemistry, biology or economics, each domain will have a preference for the usage of a particular terminological inventory that is unique or most commonly used in such a genre. That is why several terminology-aware NLP applications are designed to take into account the notion of termhood of certain lexical units. This implies that if the terms of a domain could be identified automatically or semi-automatically, then a system could also identify the domain to which the text belongs.

The words that enter into a collocational relation with terms may help to disambiguate the subject field in which the term is typically used. Let us take as an example the term *good* which in isolation is ambiguous. *Good* can be an adjective as in *keep up the good work*. Besides, it can be a noun as in *teachers can be a strong force for good* or it can also be an adverb as in *the team is doing good this year*.[59] The verbal collocate *to trade* enters into a collocation with the term *good* which is highly frequent in FTA texts. This specialized collocation occurs 14 times when the verb *to trade* is found at position -2 from the term *good*. Therefore, a system for NLP could incorporate linguistic rules and statistical information to disambiguate its lexical category and also to identify the domain where the term is being used. A query of *trade a good* in Google Books[60] indicates that it is highly frequent in texts from the field of economics. The string "trade a good" can also occur in counter-examples as in *The possibility of profit makes trade a good activity*. In this case, a linguistic rule could indicate that if a verb occurs before *trade*, then *good* should be tagged as a noun, and it contributes to identifying a domain, while the definite article before *good* helps to disambiguate it as an adjective.

Other terms and their collocates evidence that specialized collocations contribute to delineate a domain-specificity, such as *maintain / adopt / apply measure, submit claim, apply taxation measure* and *determine tariff classi-*

[59] Examples taken from the online Merriam-Webster dictionary `http://www.merriam-webster.com`

[60] `http://books.google.com`

fication. All of these examples are frequent in FTA texts or in texts where FTA-related issues are discussed, such as economics newspapers. In other words, these facts provide enough support to validate this hypothesis.

7.1.2 Second hypothesis

Collocations may be unpredictable and require idiomatic specialist knowledge.

As pointed out in the literature, there is an arbitrary factor in the formation of collocations. This implies that these units are unpredictable if based only on the syntactic and semantic rules of the language (Benson, 1985; Zuluaga, 2002; Seretan, 2011). This means that the preference of one particular noun, verb, adjective or adverb to co-occur with a term over other lexical options is unpredictable if based on syntax alone. Thus, even native speakers of a language might have problems producing the right combination of a specialized lexeme with a noun, verb, adjective or adverb (Bartsch, 2004; L'Homme, 2006). The specialized collocations formed in FTA texts confirm that also in this domain, only experts in international trade are able to produce the right combination of terms with other lexemes from the open categories, namely, verbs, nouns, adjectives and adverbs.

As an example, let us take the specialized collocation formed by a verb and a term with the pattern Adjective + Noun, such as *provide judicial authority*. This specialized collocation presents a frequency of 22 occurrences in the English subcorpus. The verbal collocate *to provide* is the base for the deverbal noun *provision* which in turn is a frequent term in FTA texts. The verb *to provide* usually collocates with the term *judicial authority* while other near-synonyms of this verb do not enter into such a collocation. For example *deliver, feed, give, hand, hand over, furnish* and *supply*.[61] Thus, specialist knowledge from the field of FTAs is necessary to account for the right combination of a term with other lexical units to attain accuracy and the adequate combination of words.

According to the above, the second hypothesis is also validated by the findings.

[61] Synonyms obtained from `http://www.merriam-webster.com`

7.1.3 Third hypothesis

The attribute of domain-specificity of specialized collocations is activated by some linguistic features of the constituents. The identification of these features can be useful to further describe the domain-specificity of phraseological units and also to represent specialized collocations for the creation of language resources.

I hold that this hypothesis is validated as will be explained in the following paragraphs. According to the definition of specialized collocation offered in Section 2.14, the linguistic constituents of specialized collocations are a simple or a complex term plus the lexical words that co-occur with it, in a direct syntactic relation with the term.

In the case of other nouns or adjectives that co-occur with terms, these are also complex terms from a morphosyntactic point of view, such as *preferential tariff treatment*, where *tariff treatment* is also a term in the field of international trade. The same applies to the Spanish term *procedimiento legislativo*, 'legislative procedure', which collocates with the verb *adoptar*. The same Spanish term also co-occurs with two adjectives that modify the type of procedure: *procedimiento legislativo especial*, 'special legislative procedure', and *procedimiento legislativo ordinario*, 'ordinary legislative procedure'.

Verbs and deverbal nouns play a definitive role in the definition of the linguistic features of specialized collocations. I agree with Estopà (1999) who argues that deverbal nouns form specialized lexical combinations in specialized texts. For example, in the FTA corpus, the term *provision* and the verb *to provide* enter into a specialized collocation with the term *judicial authority*. Other examples are *supply financial service* and *apply rate of duty*.

Though morphosyntactic patterns alone can be powerful enough to retrieve hundreds and thousands of candidate specialized collocations, there is still the issue of noise, because some of the verbs are not tagged correctly by the TreeTagger. Some of the candidate specialized collocations retrieved in this way are non-relevant. However, the use of linguistic and, more specifically, terminological knowledge expressed by means of a list of "seed" terms (Baroni and Bernardini, 2004; Burgos, 2014) in combination with the morphosyntactic patterns provides a substantial improvement over querying the

corpus merely with morphosyntactic patterns. Therefore, based on the above discussion, I consider that this hypothesis is supported.

7.2 Attainment of objectives

According to the objectives set forth under Section 1.4, this study was aimed at determining how specialized collocations contribute to delineating the domain-specificity of English and Spanish FTA texts.

The lexical units that co-occur with FTA terms shed light on the domain-specificity of international trade. As suggested by the findings, specialized collocations transmit valuable information in relation to the terms which they include. This information can aid several NLP tasks listed under Section 2.9.2, besides lexicography and terminography.

The experiments carried out with Freeling, NLTK and Termnet, described in Section 5.7.2, evidenced how the verbs and the FTA terms that co-occur frequently with these verbs are carriers of specialized meaning related to FTA-related activities, such as buying, selling and trading goods and services. Because of the arbitrary nature of the lexical items that enter into specialized collocations with terms, it seems to me that it is mandatory to build large enough corpus data from which professional users can obtain information on the distribution of words and their lexical preferences with other words.

An applied objective of this work was aimed at assessing the applicability of linguistic annotation schemes for the representation of specialized collocations in term bases and computational lexicons. In Chapter 6 it was argued that even though several standards have been published by the ISO and other initiatives, not all of them are suitable for the computational representation of MWEs such as specialized collocations. Some of these standards such as MARTIF were designed from an onomasiological rather than a semasiological approach and this makes them unsuitable for representing MWEs. Other standards such as TBX are quite flexible and do not include a detailed model for representing MWEs. It was found that especially the LMF standard offers reasonable suitability for the computational representation of specialized collocations. However, there is room for improvement, since some aspects are underspecified in these standards, such as the code

for the bilingual representation of equivalence from the level of transference, as described under Section 6.2. This underspecification can create data loss when merging, importing or exporting data among lexical or terminological language resources.

The specific objective set forth in Section 1.4 was to perform a linguistic classification, description and comparison of FTA specialized collocations that appear in a corpus of English and Spanish from American and European FTA texts. Chapter 5 offered a description of the most frequent morphosyntactic patterns that participate in the formation of specialized collocations in English and Spanish and aimed to constitute a contribution for the development of a semi-automatic system for the extraction of specialized collocations from a tagged corpus. The same chapter also documented the most frequent terms and their usual verbal collocates appearing in the FTA corpus. These frequent terms found in FTA data could be used as seed terms to improve the extraction of terms and specialized collocations.

This objective was partially attained. In my view, performing a detailed cross-language comparison on the linguistic behavior of specialized collocations that takes into account the language varieties of American and European English and Spanish would merit special attention. Thus, the study of the contrastive aspect of this objective was not feasible within the framework of this project and remains as future work. However, a prototype of some scripts aimed at the semi-automatic extraction of English and Spanish specialized collocations was developed. These scripts are a starting point to develop a better extraction system, by using morphosyntactic, semantic, terminological and statistical information.

The comparison of the characteristics of specialized collocations found in FTA texts with general and specialized English and Spanish corpora, dictionaries and term bases indicates that specialized collocations found in FTA texts are highly frequent in such texts and to a much lesser degree appear in economics textbooks or newspapers but are virtually non-existent in general texts. However, the same morphosyntactic patterns that were used to extract terms and specialized collocations in FTA texts are useful in other text types, both general and specialized.

7.3 Contributions and applications of this research

This thesis has documented the most relevant terms that appear in English and Spanish FTA texts, their usual length and internal composition. Additionally, it has documented which are the most productive morphosyntactic patterns that can be used to extract these terms semi-automatically. It has also provided evidence for the most frequent specialized lexical combinations that involve a verb and a term found in these texts and the most frequent position where verbs can be found in relation to terms, which can be useful to improve the extraction of specialized collocations. This work has also documented the FTA terms that are not yet included in specialist dictionaries from the fields of international trade and economics, which can be useful to improve lexical resources in the field of FTA and international trade. This thesis also provides a relevant methodology to carry out corpus linguistics work with specialized parallel corpora, that can be applied to other text genres.

The experiments described under Section 5.7.2 involving semantic tagging and a combination of several tools seems to be an effective way to study a corpus to derive important linguistic information. It seems desirable to perform further and deeper experiments involving more data and to contrast FTA texts with texts of economic and legal nature, such as EUR-Lex, the European Commission's Directorate-General Translation Memory (DGT-TM), and the Europarl corpus, obtained from European Parliament proceedings (Tiedemann, 2012).

Furthermore, this thesis has provided a proposal for the computational representation of MWEs such as specialized collocations for the lexical, terminological and phraseological enrichment of lexical resources, by using the LMF standard.

Several observations from FTA corpus data and the literature review on the topic of the collocations that appear in specialized texts provide a basis for several concluding remarks. First, the study of specialized collocations using a corpus-based and corpus-driven approach requires an interdisciplinary approach, as described in Section 1.2. Current language resources such as

dictionaries and term bases do not systematically include the usual collocates that co-occur with terms in a specific domain, which appear almost exclusively in specialized texts as suggested by the data. Therefore, it would be desirable to include these lexical units in specialized language resources. Besides, the inclusion of specialized collocations in lexicons, glossaries and term bases could contribute to improving translation quality, regardless of whether it is done by a human using computer assisted translation tools or by a machine translation system that is supported by a statistical or a phrase-based translation engine.

Another remark that can be made regarding specialized collocations is that they are restricted to a subject field and have a regular tendency to maintain lexical stability among the constituents of the collocation, as suggested by the data presented in Section 5.7.3.

This work has amply illustrated that corpus linguistics tools and techniques provide efficient resources for the retrieval of these specialized collocations, which are not currently offered readily and systematically in general or specialized dictionaries.

This research on specialized collocations can be useful for NLP applications for the exploitation of language resources, such as in the fields of terminology, terminography, specialized lexicography and machine translation (Gillam et al., 2002). In addition to this, it can also be used to determine how to merge and harmonize language resources without loss of information.

The lexical combination between terms and other lexical units such as verbs, adjectives, adverbs and other nouns is relevant information that should be taken into account by LSP teachers and learners. Therefore, the information of how these words combine with others in a specialized setting can also serve for the teaching of LSP and specialized translation.

Specialized collocations can also contribute to the interpretation and production of natural sounding text (McCarthy, 2006), not only in general but also in specialized domains. Besides, since the same term can be used in different domains, with different senses, the collocates of that term can be useful for the automatic identification of a topic.

7.3.1 Specialized collocations in specialized dictionaries

Currently, specialized dictionaries are published in a paper version while others also have an online counterpart while still others are still published on a CD/DVD format. Yet others are encoded as machine-readable lexicons meant for NLP applications. Currently, none of the above types of dictionary customarily provides the usual collocates for the terms included in specialized dictionaries. Specialized lexicography and terminography projects would greatly profit should word repertoires that include lexical units such as specialized collocations be developed. The specialized collocations identified by the method and the tools proposed in this work can help enrich lexical resources in the field of macroeconomics and international trade. In fact, Pustejovsky (1998) claimed that in the future it would be difficult to carry out serious linguistics and NLP research without the help of adequate language resources such as electronic dictionaries and computational lexicographic resources.

7.3.2 Collocation extraction

The approach for collocation extraction employed for this research could be used for further work on the topic. A team made by linguists, terminologists and computer scientists could use the patterns suggested in Chapter 5, to develop an improved version of a collocation extraction tool aimed at the semi-automatic identification of collocations found in specialized corpora, not only in the field of FTAs, but also in related domains such as legal and economic texts as well as in medical and scientific texts. Such a system could benefit from the findings of this thesis, regarding the lexical, semantic and morphosyntactic distribution and patterns that form specialized collocations in FTA texts. Also, the experiments suggest that the extraction could be greatly improved by means of the use of a list of seed terms. These seed terms can be taken from the gold standard of terms constituted for this project or the candidate terms extracted semi-automatically with Termostat.

7.3.3 Specialized translation

The use of the adequate collocations is necessary for the transmission of a specialized message and a qualified translator is well aware of this, as pointed out by Fontenelle (1994)

> It is therefore important that students should be aware of such collocations and able to use them adequately when translating a text into a foreign language, since they are going to be judged by their ability to manipulate these ready-made chunks of language.

This work has provided a list of central terms and the lexical items that form specialized collocations with these terms and which are relevant for translation purposes in domains related to international trade. This knowledge is relevant for translation instructors and students as well as translation professionals when dealing with texts from the field of international trade or economics-related topics.

7.4 Future work

The nature of this project sets time constraints for its development. Even though several topic and phenomena are relevant and merit attention in future work, they were deliberately omitted in this study.

In order to improve the extraction of terms and specialized collocations, some further steps could be taken. The identification of the morphological and semantic features of the verbs, adjectives and adverbs that co-occur with terms are a relevant aspect that merits special attention to improve the extraction of specialized collocations from a tagged corpus.

Therefore, a corpus enriched with morphological and semantic annotations could also provide further insights and would be highly desirable to further study specialized collocations. Subsequently, this information could be used to represent specialized collocations in lexical resources such as computational lexicons for several NLP tasks that would benefit from this phraseological information.

It seems relevant to further study the semantic features of the nouns, verbs and other lexical units that collocate with terms in the FTA corpus

because they can shed light on these legal texts that bear a normative status.

Additionally, further comparative cross-language studies could be performed to see how specialized collocations behave across English and Spanish by the use of parallel corpora.

It seems relevant to carry out a future study on the terminological and denominative variation of specialized collocations occurring in English and Spanish FTA texts both from the European and American variants of both languages, regarding aspects such as its lexical, terminological, phraseological, morphosyntactic and semantic variation.

Once it was manually cleaned, the list of candidate terms extracted with Termostat (Drouin, 2003) from the FTA texts, left 10,430 candidate terms in Spanish and 6,285 in English. This indicates that for Spanish there is much variation, possibly because different teams of translators or technical writers of FTA texts introduced new term variants. For example, since the FTAA subsection of the corpus is a draft version of a free trade agreement, it includes stylistic differences between brackets as well as term variants that the teams of negotiators have suggested and that could be used to compare polysemy and other semantic aspects present in FTA texts. The FTAA could also be interesting to carry out future studies on terminological variation across FTAs or to find synonyms or build an ontology of FTA texts. Also, as new agreements are being signed and ratified and others are amended periodically, they could provide data for a study on the variation of terms and specialized collocations over the last decades.

Several studies in terminological and denominative variation have been carried out before and could provide valuable insights to carry out a future study on the term variation that is present in FTA texts. Some of these studies have been published by Freixa (2003, 2006); Suárez (2004) and Fernández (2011).

Also, a study could be performed on the lexical, terminological and phraseological variation from a historical perspective, starting from the predecessors of modern FTAs, such as the GATT and GAT agreements from 1947 to modern FTAs involving many nations throughout the world, influenced by historical, political, social, economical and cultural factors.

Given the nature of FTA texts, it would also be interesting to perform

future studies on the formulaic language of FTA texts and the lexical bundles used in this kind of texts (Biber et al., 1999; Cortes, 2004). Furthermore, it seems pertinent to research the role of specialized collocations in the growing field of sentiment analysis. Another possible line of future research has to do with the initialisms, understood as linguistic units of lexical reduction (Giraldo, 2008) that co-occur with specialized collocations in FTA texts. Initialisms have a nominal value and therefore in specialized texts they constitute terms that co-occur with other lexical units.

This thesis aims to be a contribution to understanding the role of specialized collocations in specialized texts and how these lexical units provide valuable information regarding the terms that are part of specialized collocations. The lines of future work mentioned above offer interesting and challenging endeavors to continue the research in the field of specialized phraseology and terminology.

Bibliography

Aguado de Cea, G. (2007). A multiperspective approach to specialized phraseology: Internet as a reference corpus for phraseology. In Posteguillo, S., Esteve, M. J., and Gea-Valor, M. L., editors, *The Texture of Internet: Netlinguistics in Progress*, pages 182–207. Cambridge Scholars Publishing, Newcastle.

Aguilar-Amat Castillo, A. (1994). Colocaciones en un corpus: detección y aplicaciones. In *Lenguajes naturales y lenguajes formales: Actas del X Congreso de Lenguajes Naturales y Lenguajes Formales*, pages 327–334, Sevilla. Promociones y Publicaciones Universitarias, PPU.

Aijmer, K. (2008). Parallel and comparable corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, pages 275–292. Walter de Gruyter, Berlin/New York.

Alcaraz, E. and Castro, J. (2007). *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English*. Ariel, Barcelona.

Alexander, R. J. (1999). The recent english-language register of economics and its present importance for world commerce and trade in the late 20th century. In Hoffmann, L., Kalverkämper, H., and Wiegand, H. E., editors, *Fachsprachen Languages for Special Purposes: Ein internationales Hand-*

buch zur Fachsprachenforschung und Terminologiewissenschaft, volume 2, pages 1466–1472. Walter de Gruyter, Berlin / New York.

Amsler, R. (1982). Computational lexicology: a research program. In *AFIPS National Computer Conference*, pages 657–663, Houston. ACM.

APEC Study Centre, M. U. (2001). *An Australia-United States Free Trade Agreement - Issues and Implications*. Commonwealth of Australia, Canberra.

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT journal*, 47(1):56–63.

Baker, M. (2011). *In other words: A coursebook on translation*. Routledge, London/New York.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, United States.

Bally, C. (1932). *Linguistique générale et linguistique française*. Francke, Bern.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2004*, pages 1313–1316, Istanbul, Turkey. ELRA.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, Tübingen.

Benson, M. (1985). Collocations and idioms. In *Dictionaries, lexicography and language learning*, pages 61–68. Pergamon Press Oxford, Oxford.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography 3*, 3(1):23–35.

Benson, M., Benson, E., and Ilson, R. F. (1986). *The BBI combinatory dictionary of English: a guide to word combinations.* John Benjamins, Amsterdam/Philadelphia.

Benson, M., Benson, E., and Ilson, R. F. (2010). *The BBI combinatory dictionary of English: a guide to word combinations, third edition.* John Benjamins, Amsterdam/Philadelphia.

Bevilacqua, C. R. (2004). *Unidades fraseológicas especializadas eventivas: descripción y reglas de formación en el ámbito de la energía solar.* PhD thesis, Instituto Universitario De Lingüística Aplicada, IULA, Universidad Pompeu Fabra. Barcelona.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman Grammar of Spoken and Written English.* Longman, Harlow, UK.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python.* O'reilly, Sebastopol, United States.

Bossé-Andrieu, J. and Mareschal, G. (1998a). Trois aspects de la combinatoire collocationnelle. *TTR: traduction, terminologie, rédaction*, 11(1):157–171.

Bossé-Andrieu, J. and Mareschal, G. (1998b). Valeur sémantique du verbe dans les collocations verbales spécialisées. *TTR: traduction, terminologie, rédaction*, 11(1):173–197.

Budin, G. (1990). Terminological analysis of LSP phraseology. *Terminology Science and Research. Journal of the International Institute for Terminological Research (IITF)*, 1 (1-2):64–69.

Burgos, D. (2014). *Towards an Image-Term Co-occurrence Model for Multilingual Terminology Alignment and Cross-Language Image Indexing.* PhD thesis, Universitat Pompeu Fabra, Barcelona.

Cabré, M. T. (1993). *La terminología. Teoría, métodos, aplicaciones.* Antártida, Barcelona.

Cabré, M. T. (1999). *La terminología. Representación y comunicación. Una teoría de base comunicativa y otros artículos.* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

Calzolari, N. (1994). Issues for Lexicon Building. In Zampolli, A., Nicoletta, C., and Palmer, M., editors, *Linguistica Computazionale Vol. IX-X. Current Issues in Computational Linguistics: In Honour of Don Walker*, pages 267–281. Giardini Editori e Stampatori, Pisa.

Calzolari, N., Lenci, A., and Zampolli, A. (2001). International standards for multilingual resource sharing: the isle computational lexicon working group. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources - Volume 15*, STAR '01, pages 71–78, Stroudsburg, United States. Association for Computational Linguistics.

Calzolari, N., Monachini, M., and Soria, C. (2013). Lmf – historical context and perspectives. In Francopoulo, G., editor, *LMF Lexical Markup Framework*, pages 1–18. ISTE Ltd / John Wiley Sons, Inc., London / Hoboken, United States.

Cartagena, N. (1998). Acerca de la variabilidad de los términos sintagmáticos en textos españoles especializados. In Wotjak, G., editor, *Estudios de fraseología y fraseografía del español actual*, pages 281–296. Vervuert / Iberoamericana, Frankfurt am Main / Madrid.

Casares, J. (1992). *Introducción a la lexicografía moderna, 3a edición.* CSIC, Madrid.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of RIAO'1988. International Conference on User-Oriented Content-Based Text and Image Handling*, pages 609–624, Cambridge, United States.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94 3rd Conference on Computational Lexicography and Text Research*, pages 23–32, Budapest.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Clas, A. (1994). Collocations et langues de spécialité. *Meta: Journal des traducteurs*, 39(4):576–580.

Cohen, B. (1986). *Lexique de cooccurrents: Bourse, conjoncture économique*. Linguatech, Brossard, Canada.

Cop, M. (1990). The function of collocations in dictionaries. *BudaLEX '88 proceedings*, 1990:35–46.

Cop, M. (1991). Collocations in the bilingual dictionary. In Hausmann, F. J., Reichmann, O., Wiegand, H. E., and Zgusta, L., editors, *Wörterbücher: ein internationales Handbuch zur Lexikographie*, volume 3, pages 2775–2778. Walter de Gruyter, Berlin/New York.

Corpas, G. (2003). *Diez años de investigación en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos*. Iberoamericana, Vervuert, Madrid/Frankfurt.

Corpas, G. and Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents. In Beeby, A., Inés, P. R., and Sánchez-Gijón, P., editors, *Corpus Use and Translating*, volume 82, pages 75–107. John Benjamins Publishing, Amsterdam/Philadelphia.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4):397–423.

Cowie, A. P. (1986). Collocational dictionaries - a comparative view. In *Fourth Joint Anglo-Soviet Seminar*, pages 61–69, London. British Council.

Coxhead, A. (2007). Acquiring academic and disciplinary vocabulary. In Hyland, K. and Shaw, P., editors, *The Routledge Handbook of English for Academic Purposes*, pages 177–190. Routledge, London/New York.

Crump, L. (2007). Bilateral negotiations in a multilateral world: Implications for the wto and global trade policy development. In Crump, L. and Maswood, S. J., editors, *Developing countries and global trade negotiations*, pages 166–248. Routledge, London/New York.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press, Cambridge/New York.

Crystal, D. (2008). *A Dictionary of linguistics and phonetics*. John Wiley & Sons, Malden, United States / Oxford.

Cunningham, H. and Bontcheva, K. (2006). Computational language systems: Architectures. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 733–752. Elsevier, London.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. PhD thesis, Université de Paris 7, Paris.

Daille, B. (1999). Identification des adjectifs relationnels en corpus. In *Actes de TALN, Traitement automatique du langage naturel*, pages 105–114.

Daille, B. (2001). Qualitative terminology extraction, identifying relational adjectives. In Bourigault, D., Jacquemin, C., and L'Homme, M.-C., editors, *Recent advances in computational terminology*, volume 2. John Benjamins, Amsterdam.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14:159–90.

De Groc, C. (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 497–498. IEEE Computer Society.

Drouin, P. (1997). Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme. *Meta: Journal des traducteurs*, 42(1):45–54.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Drouin, P. (2004). Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal*.

Estopà, R. (1999). *Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)*. PhD thesis, Universitat Pompeu Fabra, Barcelona.

Evert, S. (2004). *The statistics of word cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung. Stuttgart.

Evert, S. (2005). Empirical research on association measures: The UCS toolkit. Technical report, University of Louvain, Louvain-la-Neuve, Belgium.

Evert, S. (2009). Corpora and collocations. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*, pages 1212–1248. Walter de Gruyter, Berlin.

Fernández, S. (2011). *Variación terminológica y cognición: factores cognitivos en la denominación del concepto especializado*. PhD thesis, Universitat Pompeu Fabra, Barcelona.

Fernández, P. (2008). Las colocaciones en el lenguaje jurídico. In Mellado Blanco, C., editor, *Colocaciones y fraseología en los diccionarios*, volume 44, pages 69–84. Peter Lang Pub Inc, Frankfurt.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press, London.

Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In Connor, U. and Upton, T. A., editors, *Discourse in the professions. Perspectives from corpus linguistics*, pages 11–33. John Benjamins, Amsterdam.

Fontenelle, T. (1994). Towards the construction of a collocational database for translation students. *Meta: Journal des traducteurs*, 39(1):47–56.

Foo, J. (2011). Exploring termhood using language models. In *Proceedings of the NODALIDA 2011 workshop. CHAT 2011: Creation, Harmonization and Application of Terminology Resources*, pages 32–35, Riga, Latvia. Northern European Association for Language Technology Proceedings Series, Vol. 12.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006a). Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70. 10.1007/s10579-008-9077-5.

Francopoulo, G., Declerck, T., Monachini, M., and Romary, L. (2006b). The relevance of standards for research infrastructures. In *International Conference on Language Resources and Evaluation - LREC 2006*, Genoa, Italy. elra. LIRICS.

Francopoulo, G. and George, M. (2013). Model description. In Francopoulo, G., editor, *LMF Lexical Markup Framework*, pages 19–40. ISTE Ltd / John Wiley Sons, Inc., London / Hoboken, United States.

Freixa, J. (2003). *La variació terminològica: anàlisi de la variació denominativa en textos de diferent grau despecialització de làrea de medi ambient*. PhD thesis, Universitat de Barcelona.

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. *Terminology*, 12(1):51–77.

Galinski, C. (1990). Terminology and phraseology. *Terminology Science and Research*, 1:70–86.

Gallegos, A. (2003). *Nominalización y registro técnico. Algunas relaciones entre morfopragmática, tradiciones discursivas y desarrollo de la lengua en español*. PhD thesis, Albert-Ludwigs-Universität Freiburg, Freiburg.

Gamero, S. (2001). *La traducción de textos técnicos: descripción y análisis de textos (alemán-español)*. Ariel, Barcelona.

Gaussier, E. and Langé, J.-M. (1994). Some methods for the extraction of bilingual terminology. In *International Conference on New Methods in Language Processing (NeMLaP)*, pages 224–228.

Gelbukh, A. and Kolesnikova, O. (2013). *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer-Verlag, Berlin/Heidelberg.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., and Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage Publications Limited, London.

Gillam, L., Ahmad, K., Dalby, D., and Cox, C. (2002). Knowledge exchange and terminology interchange: The role of standards. In *Proceedings of Translating and the Computer 24*.

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

Gledhill, C. J. (2000). *Collocations in science writing*. Gunter Narr Verlag, Tübingen.

Godfrey, J. and Zampolli, A. (1997). Language resources. In Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Varile, G., Zampolli, A., Cole, R.,

and Zue, V., editors, *Survey of the State of the Art in Human Language Technology*, pages 381–384. Cambridge University Press, Cambridge.

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

Gouadec, D. (1990). *Terminologie: constitution des données*. Afnor, Paris.

Gozdz-Roszkowski, S. (2011). *Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study*. Lodz Studies in Language. Peter Lang, Bern.

Graham, D. (2004). Free trade: Myth, reality and alternatives.

Greaves, C. and Warren, M. (2010). What can a corpus tell us about multiword units? In O'Keeffe, A. and McCarthy, M., editors, *The Routledge Handbook of Corpus Linguistics*, pages 212–226. Routledge, London/New York.

Gries, S. T. (2013). 50-something years of work on collocations: what is or should be next.... *International Journal of Corpus Linguistics*, 18(1):137–166.

Grishman, R. and Calzolari, N. (1997). Lexicons. In Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Varile, G., Zampolli, A., Cole, R., and Zue, V., editors, *Survey of the State of the Art in Human Language Technology*, pages 392–395. Cambridge University Press, Cambridge.

Groom, N. W. (2007). *Phraseology and epistemology in humanities writing: a corpus-driven study*. PhD thesis, University of Birmingham.

Halliday, M. (1961). *On grammar*, volume 1. Continuum, London.

Halliday, M. A. K. (2004). Lexicology. In Teubert, W., Yallop, C., Cermäkova, A., and Halliday, M. A. K., editors, *Lexicology and Corpus Linguistics*, pages 1–22. Continuum, London.

Hanks, P. (2003). Lexicography. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 48–69. Oxford University Press, Oxford.

Harold, E. R. and Means, W. S. (2004). *XML in a Nutshell*. O'Reilly, Sebastopol, United States.

Harris, Z. (1968). *Mathematical structures of language*. Interscience Publishers John Wiley and Sons, New York.

Heid, U. (1999). Extracting terminologically relevant collocations from german technical texts. In *5th International Congress on Terminology and Knowledge Engineering, TKE '99*, pages 241–255.

Heid, U. (2001). Collocations in sublanguage texts: Extraction from corpora. *Handbook of terminology management: Application-oriented terminology management*, 2:788–808.

Heid, U. and Freibott, G. (1991). Collocations dans une base de données terminologique et lexicale. *Meta: Journal des traducteurs*, 36(1):77–91.

Heid, U. and Weller, M. (2008). Tools for collocation extraction: Preferences for active vs. passive. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Hofland, K. and Johansson, S. (1998). The translation corpus aligner: A program for automatic alignment of parallel texts. In Johansson, S. and Oksefjell, S., editors, *Corpora and Cross-linguistic research. Theory, Method, and Case Studies*, pages 87–100. Rodopi, Amsterdam/Atlanta.

Hunston, S. (2006). Corpus linguistics. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 234–248. Elsevier, London.

Ide, N. and Véronis, J. (1994). Multext: Multilingual text tools and corpora. In *Proceedings of the 15th conference on Computational linguistics, Volume 1*, pages 588–592. Association for Computational Linguistics.

Irwin, D. A. (2009). *Free trade under fire*. Princeton University Press, Princeton, United States.

ISO (2001). Computer applications in terminology–terminological markup framework (TMF),ISO/DIS 16642, ISO/TC 37/SC 3/WG.

ISO (2008). Language resource management - Lexical markup framework (LMF), ISO 24613:2008, ISO/TC 37/SC 4 N453 (N330 Rev.16).

ISO (2009). The International Standard ISO 704 Terminology work - Principles and methods.

Jacquemin, C., Klavans, J. L., and Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–31. Association for Computational Linguistics.

Johansson, S. (2007). *Seeing through multilingual corpora: on the use of corpora in contrastive studies*, volume 26. John Benjamins Publishing Co.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

Kjær, A. L. (1990). Phraseology research. state of the art. *Terminology, Science and Research*, 1:3–20.

Kjær, A. L. (2007). Phrasemes in legal texts. In *Phraseology: An International Handbook of Contemporary Research*, volume 1, pages 506–516. Walter de Gruyter, Berlin/New York.

Koester, A. (2010). Building small specialised corpora. In O'Keeffe, A. and McCarthy, M., editors, *The Routledge Handbook of Corpus Linguistics*, pages 66–79. Routledge, London.

Koike, K. (2001). *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico*. Universidad de Alcalá de Henares, Alcalá de Henares, Spain.

Koike, K. (2002). Comportamientos semánticos en las colocaciones léxicas. *LEA: Lingüística española actual*, 24(1):5–24.

Krishnamurthy, R. (2006). Collocations. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 596–600. Elsevier, London.

Kristiansen, M. (2004). *The Multidisciplinary Nature of the Social Sciences. Investigating Disciplinary Autonomy in Organisational Behaviour by means of Terminological Analysis*. PhD thesis, Universitetet i Bergen, NHH Norwegian School of Economics, Bergen.

Lerat, P. (1989). Les fondements théoriques de la terminologie. *La banque des mots, Revue de terminologie française*, pages 51–62.

L'Homme, M. (2001). Combinaisons lexicales spécialisées: Regroupement des mots clés par classes conceptuelles. In Daille, B. and Williams, G., editors, *Journées d'étude de l'ATALA. La collocation. Rapport de recherche, Nantes: Institut de recherche en informatique de Nantes*, pages 19–22.

L'Homme, M. (2002). What can verbs and adjectives tell us about terms? In *Terminology and Knowledge Engineering, TKE 2002. Proceedings*, pages 28–30, Nancy, France.

L'Homme, M. and Bertrand, C. (2000). Specialized lexical combinations: Should they be described as collocations or in terms of selectional restrictions? In *9th Euralex International Congress*, pages 497–506, Stuttgart. Stuttgart University.

L'Homme, M. and Leroyer, P. (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology*, 15(2):258–283.

L'Homme, M. C. (1998). Caractérisation des combinaisons lexicales specialisees par rapport aux collocations de langue generale. In Fontenelle, T., editor, *Proceedings EURALEX '98*, pages 513–522. Universite de Liege.

L'Homme, M. C. (2006). The processing of terms in dictionaries: New models and techniques. A state of the art. *Terminology*, 12(2):181–188.

L'Homme, M. C. (2009). A methodology for describing collocations in a specialised dictionary. In Nielsen, S. and Tarp, S., editors, *Lexicography in the 21st century*, pages 237–256. John Benjamins, Amsterdam.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. Association for Computational Linguistics.

Litkowski, K. (2006). Computational lexicons and dictionaries. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, pages 753–759. Elsevier, London.

Ljubešic, N., Vintar, Š., and Fišer, D. (2012). Multi-word term extraction from comparable corpora by combining contextual and constituent clues. In *The 5th Workshop on Building and Using Comparable Corpora*, page 143.

Lorente, M. (2002a). Altres elements lèxics. In Solà, J., editor, *Gramàtica del català contemporani (Gcc). Volum I*, pages 831–888. Empúries, Barcelona.

Lorente, M. (2002b). Terminología y fraseología especializada: del léxico a la sintaxis. In Guerrero, G. and Pérez, L. F., editors, *Panorama actual de la terminología*, pages 159–180. Comares, Colección Interlingua, 30, Granada.

Lyse, G. I. (2011). *Translation-based Word Sense Disambiguation.* PhD thesis, The University of Bergen, Bergen.

Lyse, G. I. and Andersen, G. (2012). Collocations and statistical analysis of n-grams. In Andersen, G., editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 79–109. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Maks, I., Tiberius, C., and van Veenendaal, R. (2008). Standardising Bilingual Lexical Resources According to the Lexicon Markup Framework. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1723–1727, Marrakech, Morocco. European Language Resources Association (ELRA).

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press, Cambridge, United States.

Martin, W. (1992). Remarks on collocations in sublanguages. *Terminologie et traduction*, 2(3):157–164.

Matsumoto, Y. (2003). Lexical Knowledge Acquisition. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 395–413. The Oxford University Press, Oxford.

Maurer-Stroh, P. (2004). *Towards a Bilingual Adjective-Noun Collocation Dictionary of English and German*. PhD thesis, Universität Klagenfurt, Klagenfurt, Austria.

McCarthy, D. (2006). Lexical acquisition. In Brown, K., editor, *Encyclopedia of Language and Linguistics*, volume 7, pages 61–68. Elsevier, London.

McEnery, T. (2003). Corpus linguistics. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 448–463. Oxford University Press, Oxford.

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

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh.

McEnery, T. and Xiao, Z. (2007). Parallel and comparable corpora - the state of play. In Kawaguchi, Y., Takagaki, T., Tomimori, N., and Tsuruga, Y., editors, *Corpus-based perspectives in linguistics*, volume 6. John Benjamins Publishing Company, Amsterdam/Philadelphia.

McKeown, K. R. and Radev, D. R. (2000). Collocations. In Moisl, H., Somers, H., and Dale, R., editors, *A Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker, New York.

Mel'čuk, I. (1998). Collocations and lexical functions. In Cowie, A. P., editor, *Phraseology: Theory, Analysis, and Applications*. Clarendon, Oxford/NewYork.

Meynard, I. (2000). *Internet: répertoire bilingue de combinaisons lexicales spécialisées: français-anglais*. Linguatech, Brossard, Canada.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moliner, M. (1966). *Diccionario de uso del español*. Gredos, Madrid.

Moon, B. E. (2000a). Free trade area (FTA). In Jones, R. J. B., editor, *Routledge Encyclopedia of International Political Economy*, pages 574–575. Routledge, London/New York.

Moon, B. E. (2000b). Trade agreements. In Jones, R. J. B., editor, *Routledge Encyclopedia of International Political Economy*, pages 1570–1571. Routledge, London/New York.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Clarendon Press, Oxford, UK.

Moon, R. (2008). Dictionaries and collocation. In Granger, S. and Meunier, F., editors, *Phraseology: An Interdisciplinary Perspective*, pages 313–336. John Benjamins, Amsterdam and Philadelphia.

Moreno, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios de Lingüística Española (ELiEs)*, 9(1):1.

Munday, J. (2016). *Introducing translation studies: Theories and applications*. Routledge, London/New York.

Méndez, B. (2008). Las colocaciones en la prosa académica médica: Análisis contrastivo inglés-español. In Mellado Blanco, C., editor, *Colocaciones y fraseología en los diccionarios*, volume 44, pages 117–130. Peter Lang, Frankfurt.

Narlikar, A. (2005). *The World Trade Organization: a very short introduction*. Oxford University Press, Oxford.

Newmark, P. (1988). *A textbook of translation*. Prentice Hall, London.

Nicolaides, P. (2000). Free trade area (FTA). In Jones, R. J. B., editor, *Routledge Encyclopedia of International Political Economy*, pages 575–576. Routledge, London/New York.

Nugues, P. M. (2006). *An Introduction to Language Processing with Perl and Prolog: an outline of theories, implementation, and application with special consideration of English, French, and German*. Springer-Verlag.

Oakes, M. (2009). Preprocessing multilingual corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin/New York.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Orliac, B. (2004). *Automatisation du repérage et de l'encodage des collocations en langue de spécialité*. PhD thesis, University of Montreal, Montreal.

Orliac, B. (2008). Extracting specialized collocations using lexical functions. In Altenberg, B. and Granger, S., editors, *Phraseology: An Interdisciplinary Perspective*, pages 377–390. John Benjamins, Amsterdam/Philadelphia.

Ostojska-Asensio, M. (2002). *Diccionario de Términos Jurídicos, Español-Inglés English-Spanish*. Espasa Calpe, Madrid.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Palmer, H. E. and Hornby, A. S. (1933). *Second interim report on English collocations*. Institute for Research in English Teaching, Tokyo.

Parra, C., Smørdal Losnegaard, G., Lyse Samdal, G. I., and Patiño, P. (2013). Representing multiword expressions in lexical andterminological resources: An analysis for natural languageprocessing purposes. In *Electronic Lexicography in the 21st century:thinking outside the paper*, Tallinn, Estonia. Institute of the Estonian Language and Trojina, Institute for Applied Slovene Studies.

Patiño, P. (2010). Extracción semiautomática de locuciones especializadas de economía en español. *Lenguaje*, 38(1):235–255.

Patiño, P. (2013). FTA Corpus: a parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations. In *The many facets of corpus linguistics in Bergen - in honour of Knut Hofland. Bergen Language and Linguistics Studies*, volume 3, pages 81–92. University of Bergen, Bergen, Norway.

Pavel, S. (1993a). Neology and phraseology as terminology-in-the-making. In Sonneveld, H. B. and Loening, K. L., editors, *Terminology: applications in interdisciplinary communication*, page 21–34. John Benjamins, Amsterdam.

Pavel, S. (1993b). Vers une méthode de recherche phraséologique en langue de spécialité. *Actualité terminologique/Terminology Update*, 26(2):9–13.

Pearson, J. (1998). *Terms in Context*. John Benjamins, Amsterdam.

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658. Association for Computational Linguistics.

Picht, H. (1987). Terms and their LSP environment - LSP phraseology. *Meta : journal des traducteurs / Meta: Translators' Journal*, 32(2):149–155.

Picht, H. (1990a). LSP phraseology from the terminological point of view. *Terminology Science and Research. Journal of the International Institute for Terminological Research (IITF)*, 1(1-2):33–48.

Picht, H. (1990b). A study of LSP phraseological elements in spanish technical texts. *Journal of the International Institute for Terminological Research (IITF)*, 1(1-2):49–58.

Pustejovsky, J. (1998). *The generative lexicon*. The MIT Press, Cambridge, United States.

Quiroz, G. (2008). *Los sintagmas nominales extensos especializados en inglés y en español: descripción y clasificación en un corpus de genoma*. PhD thesis, Universitat Pompeu Fabra.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer, Cham, Switzerland / Heidelberg.

Romary, L. (2001). An abstract model for the representation of multilingual terminological data: TMF - Terminological Markup Framework. In *TAMA 2001*, Antwerp, Belgium.

Routledge (1998). *Routledge Spanish Dictionary of Business, Commerce and Finance*. Routledge Software, London/New York.

Russell, J. (2005). *Dictionary of Banking and Finance*. A and C Black, London.

Saeed, J. (2003). *Semantics*. Blackwell Publishing, Malden, United States.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 189–206. Springer.

Schäfer, R. (2015). Processing and querying large web corpora with the cow14 architecture. In *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)*, pages 1–8, Lancaster. Institut für Deutsche Sprache, Mannheim and Institute for Corpus Linguistics and Text Technology, Vienna.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Schott, J. J. (2005). Does the ftaa have a future? *Institute for International Economics*, 44.

Scott, M. (2007). *Oxford WordSmith Tools 4.0*. Oxford University Press, Oxford.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, Dordrecht.

Seretan, V. (2013). On collocations and their interaction with parsing and translation. *Informatics*, 1(1):11–31.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Sinclair, J., Jones, S., and Daley, R. (1970). *English Collocation Studies: The OSTI Report*. Continuum, London.

Spang-Hanssen, H. (1983). Kommunikation og fagsprog [Communication and LSP]. *Språk i Norden*, 22:25–37.

Stevenson, M. and Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

Stubbs, M. (2002). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell Publishing.

Stubbs, M. (2009). Technology and phraseology. with notes on the history of corpus linguistics. In Römer, S. U. and Schulze, R., editors, *Exploring the lexis-grammar interface*, pages 15–31. John Benjamins, Amsterdam.

Suárez, M. (2004). *Análisis contrastivo de la variación denominativa en textos especializados: del texto original al texto meta*. PhD thesis, Universitat Pompeu Fabra.

Tanguy, L. and Hathout, N. (2007). *Perl pour les linguistes*. Lavoisier, Paris.

Tecedor, M. (1998). Consideraciones lingüístico-pragmáticas acerca del trasvase de las expresiones fijas del lenguaje taurino al código general. In *Estudios de fraseología y fraseografía del español actual*, pages 129–154. Iberoamericana.

Thomas, P. (1993). Choosing Headwords from Language-for-Special-Purposes LSP Collocations for Entry into Terminology Data Bank (Term Bank). In Sonneveld, H. B. and Loening, K. L., editors, *Terminology: Applications in interdisciplinary communication*, pages 43–68. John Benjamins.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Tognini-Bonelli, E. (2002). Between phraseology and terminology in the language of economics. In Nuccorini, S., editor, *Phrases and Phraseology – Data and Descriptions*, pages 65–83. Peter Lang.

Toury, G. (1995). *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam/Philadelphia.

Van Eijck, J. and Unger, C. (2010). *Computational semantics with functional programming*. Cambridge University Press, Cambridge.

van Sterkenburg, P. (2003). 'the' dictionary: Definition and history. In van Sterkenburg, P., editor, *A Practical Guide to Lexicography*, volume 6, pages 3–25. John Benjamins Publishing, Amsterdam/Philadelphia.

Wanner, L., Bohnet, B., Mark, G., and Vidal, V. (2007). The first steps towards the automatic compilation of specialized collocation dictionaries. In Ibekwe-SanJuan, F., Condamines, A., and Cabré, M. T., editors, *Application-Driven Terminology Engineering*, pages 127–161. John Benjamins, Amsterdam.

Wilcock, G. (2009). *Introduction to Linguistic Annotation and Text Analytics*. Morgan & Claypool, San Rafael, United States.

Wilks, Y., Fass, D., Guo, C.-m., McDonald, J. E., Plate, T., and Slator, B. M. (2008). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of the 12th Conference on Computational Linguistics*, pages 750–755, Morristown, New Jersey. Association for Computational Linguistics.

WTO (2015). *Understanding the WTO*. World Trade Organization, Information and External Relations Division, Geneva.

Wynne, M. (2009). Searching and concordancing. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: An International Handbook*. Berlin/New York: Mouton de Gruyter.

Zampolli, A. (1997). The PAROLE project. In *The general context of the European actions for Language Resources. Proceedings of the Second European Seminar: Language Applications for a Multilingual Europe*, pages 185–210, Kaunas, Lithuania. TELRI.

Zuluaga, A. (1998). Análisis y traducción de unidades fraseológicas desautomatizadas. *Lingüística y Literatura, Universidad de Antioquia*, 34/35:203–220.

Zuluaga, A. (2002). Los "enlaces frecuentes" de María Moliner. observaciones sobre las llamadas colocaciones. *Philologie im Netz*, 22:56–74.

Šarcevic, S. (2000). Creativity in legal translation: how much is too much? In *Translation in context: selected contributions from the EST Congress, Granada, 1998*, volume 39, page 281. John Benjamins Publishing Company.

Appendices

APPENDIX A

Appendix

Table 1: *English Morphosyntactic patterns used by Termostat and their frequencies*

Pattern	**Freq**	**Pattern**	**Freq**
Adj N	2105	N N N N N	3
N N	1165	N N Prep N N	3
N	1073	N N Prep N N N	3
N Prep N	685	N Prep Adj Adj N N	3
Adj N N	234	N Prep N Prep Adj N	3
N Prep Adj N	211	Adj Adj N Prep Adj N	2
Adj Adj N	170	Adj Adj N Prep N	2
Adj Coord_Conjunction Adj N	164	Adj N N Prep N	2
N N N	123	N N Adj N	2
Adj N Prep N	93	N N Prep Adj Adj N	2
N Prep N N	59	N Prep Adj N N N	2
N Prep Adj N N	33	Adj Adj N N N	1
Adj N Prep Adj N	30	Adj Adj N N N Prep N Prep N	1
N Adj N	23	Adj Adj N Prep N N	1
N N Prep N	20	Adj N N Adj N	1
Adj Adj N N	17	Adj N N Prep Adj N	1
Adj N Prep N N	15	Adj N Prep Adj Adj N	1
Adj N N N	14	Adj N Prep N Prep Adj Adj N	1
N Prep Adj Adj N	12	Adj N Prep N Prep N	1
Adj N Prep Adj N N	9	N Adj Adj N Prep Adj N	1
N N N N	6	N N Adj Adj N	1
N N Prep Adj N	6	N N N Adj N	1
Adj N Adj N	5	N N N N Prep Adj N	1
N Prep N N N	5	N N N N Prep N	1
N Prep N Prep N	5	N Prep Adj N Prep N	1
Adj Adj Adj N	4	N Prep N N Prep N	1
N Prep Adj N Prep Adj N	4	N Prep N Prep Adj Adj N	1

Table 2: *Spanish Morphosyntactic patterns used by Termostat and their frequencies*

Pattern	**Freq**
N Prep N	3253
N Adj	3093
N	1463
N Prep N Adj	1312
N Adj Adj	591
N N	525
N Adj Coord_Conj Adj	179
N Adj Adj Adj	16
N N N	4
N N N Adj	0
N V N	0
N V N Adj	0

Table 3: *Relevant categories in WordNet classification for English candidate specialized collocations*

Freq.	**WordNet category**	**Meaning**
829	noun.act	nouns denoting acts or actions
499	noun.attribute	nouns denoting attributes of people and objects
478	verb.cognition	verbs of thinking, judging, analyzing, doubting
436	verb.communic.	verbs of telling, asking, ordering, singing
429	verb.change	verbs of size, temperature change, intensifying, etc.
422	adj.all	all adjective clusters
231	noun.person	nouns denoting people
226	noun.event	nouns denoting natural events
219	noun.artifact	nouns denoting man-made objects
149	noun.location	nouns denoting spatial position
141	noun.Tops	unique beginner for nouns
129	verb.possession	verbs of buying, selling, owning
102	noun.cognition	nouns denoting cognitive processes and contents
96	adj.pert	relational adjectives (pertainyms)
91	noun.plant	nouns denoting plants
91	verb.body	verbs of grooming, dressing and bodily care
83	noun.communic.	nouns denoting communicative processes and contents
82	verb.creation	verbs of sewing, baking, painting, performing
77	verb.consumption	verbs of eating and drinking
69	verb.social	verbs of political and social activities and events
66	noun.time	nouns denoting time and temporal relations
61	noun.substance	nouns denoting substances
54	verb.competition	verbs of fighting, athletic activities
47	adv.all	all adverbs
42	noun.group	nouns denoting groupings of people or objects
38	verb.contact	verbs of touching, hitting, tying, digging
37	noun.animal	nouns denoting animals
23	verb.stative	verbs of being, having, spatial relations
22	noun.body	nouns denoting body parts
18	noun.possession	nouns denoting possession and transfer of possession
7	noun.object	nouns denoting natural objects (not man-made)
6	noun.process	nouns denoting natural processes
5	noun.quantity	nouns denoting quantities and units of measure
4	noun.food	nouns denoting foods and drinks
4	noun.state	nouns denoting stable states of affairs
3	verb.motion	verbs of walking, flying, swimming
2	noun.relation	nouns denoting relations between people, things or ideas
2	verb.perception	verbs of seeing, hearing, feeling
1	noun.shape	nouns denoting two and three dimensional shapes
1	verb.emotion	verbs of feeling