A specialized parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations

Pedro Patiño

NHH Norwegian School of Economics

Summary
This paper describes the Corpus of Free Trade Agreements (henceforth FTA), a specialized parallel corpus in English and Spanish from Europe and America that is being prepared and aligned with Translation Corpus Aligner 2 (Hofland & Johansson, 1998). The data is taken from Free Trade Agreements officially signed and ratified by several countries and blocks of countries. Once complete, the corpus will contain about 1.3 million words in the English section and 1.5 million words in its Spanish counterpart. One of the aims is to study the specialized collocations that appear in this kind of texts and the terminological value of specialized collocations as carriers of specialized information.

1 Introduction

In the field of terminology, there is an interest to research into the features of specialized texts in the field of economics. Several authors affirm that phraseological data like collocations are not represented in a systematic way in general and specialized dictionaries (Orliac 2004; Moon 2008). The compilation of a parallel corpus made from texts from FTAs would allow to study a kind of specialized text in use and to derive specialized phraseological units and terms that have been so far neglected in specialized dictionaries and term bases. This paper presents the work towards the compilation of a specialized English and Spanish parallel corpus. The data that will be used to create this corpus are the texts of FTAs that have been officially signed and ratified by the parties involved, specifically the agreements involving Colombia, Mexico, Chile, the Caribbean Community, the United States, the European Union, Canada and the European Free Trade Association. These agreements are normally written in English and then translated to another language, in this case into Spanish. In other cases, as the negotiation advances, the teams of free trade experts draft the texts simultaneously, with each team writing in its mother tongue.

These are some examples of collocations that appear in the FTAs under consideration. First, some examples in Spanish are presented: *adoptar medidas tributarias; adoptar medidas de salvaguardia; alcanzar una solución mutuamente satisfactoria; aplicar un arancel aduanero; percibir derechos antidumping*. The following are some examples in English: *apply a customs duty; enforce its environmental laws; arrive at a mutually satisfactory resolution*.

This research is part of a PhD project aimed at investigating the specialized collocations that appear in official FTA texts. Specialized collocations can pose problems even for native speakers who do not know the subject field. Bartsch (2004: 20) addresses the fact 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". Knowledge of the phraseology of FTA texts could be useful for language professionals such as translators,

technical writers, terminologists and specialist lexicographers. Besides, this information can serve to enrich computational lexicons for machine translation and natural language processing. 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". In the frame of this PhD project, in order to be considered as specialized, the candidate collocation has to include a constituent as a noun, adjective or verb that has terminological value. The terminological value will be attested by the occurrence of the candidate term in a specialized dictionary or term base, specialized corpus or by consultation with a field expert. The working definition of collocation that will be used in this paper is the one offered by Bartsch (2004: 76) 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 relevant as it does not exclude collocations that span more than two lexical items neither does it exclude non-adjacent word combinations.

2 Corpus description

So far, the FTA corpus consists of 170 XML files in each language. Once complete, the corpus will contain at least 1.3 million words in the English section and approximately 1.5 million words in its Spanish counterpart. It will also comprise texts from different language varieties as English from the United States and the European Union texts, as well as Spanish from Latin American countries and Spain. Free trade agreements are specialized official documents that set the norms for the trade of goods among two or more parties and thus are a rich repository for terminology and phraseology that is used in different fields of business activity throughout the world. For example, those are some terms that appear in the Spanish section of the corpus: *derechos antidumping, procedimientos judiciales civiles, derecho internacional consuetudinario, arancel aduanero*. These terms are used in law and international trade. Translators, and other language professionals, need clear information on the collocations that are formed by those terms, which are not always available in current general and specialized dictionaries, as there are no clear criteria for the systematic inclusion or exclusion of collocations (Benson 1985; Moon 2008).

Most of the data (85%) has already been gathered and processed and is being aligned using the last version of the software Translation Corpus Aligner 2 (Hofland & Johansson 1998), which allows for the exportation of XML files compliant with the Text Encoding Initiative. The data were downloaded from the web pages of the Foreign Trade Information System of the Organization of American States[1] and the European Union[2]. The original files were downloaded as PDF, HTM and RTF files that were converted to a XML code that is readable by the Translation Corpus Aligner 2 software.

The FTA texts comprise the Spanish and English versions of the agreements signed by these countries or blocks of countries, as shown in Table 1:

Countries	English words	Percentage	Year
Canada - Peru	69930	6.02	2008
CARICOM (Caribbean Community) – Dominican Republic	9458	0.81	1998
CARIFORUM – European Union	51483	4.43	2008
Chile – Australia	64841	5.58	2008

[1] http://www.sice.oas.org/agreements_e.asp
[2] http://eu-lex.europa.eu/JOHtml.do?uri=OJ:C:2010:083:SOM:es:HTML

Chile – EFTA	16671	1.43	2003
Chile - European Union	34381	2.96	2002
Chile – United States	86112	7.41	2003
Colombia – EFTA	69569	5.99	2008
European Union	133237	11.47	1992 / 2007
Free Trade Area of the Americas (draft)	179747	15.47	2003
Mexico – EFTA	14862	1.28	2000
Colombia - United States	160091	13.78	2006
Colombia - European Union	ND	ND	2010
World Trade Organization	88548	7.62	1994
NAFTA	182990	15.75	1992
Total	1161920	100	

Table 1. Free Trade Agreements included in the corpus

As a next step in the research, it is necessary to perform a semi-automatic extraction of a list of candidate collocations for the different subsets of the corpus in English and Spanish. To attain this end, a preliminary list of 45 terms will be used to research the lexical collocations that these terms produce. The list of terms was drawn from the section of definitions of every agreement appearing typically in the first part of each one of the agreements.

Table 2 shows the top 20 collocates of Spanish noun "*procedimiento*" extracted with Xaira[3], searching one item to the left and one to the right.

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 2. 20 top frequent collocates of Spanish noun "procedimiento" extracted with Xaira

Later, the occurrence of these candidate collocations will be studied in the different subsets of the corpus in the two languages under study. Then, these collocations will be compared with general and specialized corpora and dictionaries. To contrast the FTA collocations that are found, five reference corpora (three of them annotated) and several specialized dictionaries of economics will be used:

[3] http://www.oucs.ox.ac.uk/rts/xaria/

- English Corpora: the Corpus of Contemporary American English (Davies 2009), with 410 million words and the British National Corpus[4], with 100 million words.

- Spanish Corpora: Corpus Tècnic de l'IULA (Bach et al. 1997), Corpus de referencia del español actual (CREA-RAE)[5] with 200 million words, and Corpus del Español with 100 million words (Davies 2002).

- Bilingual specialized dictionaries: SICE-OAS online Dictionary of Trade Terms[6], Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English (Alcaraz & Castro 2007), Dictionary of International Business Terms (Capela & Hartman 2000), the Routledge Spanish Dictionary of Business, Commerce and Finance (1998).

- Specialized English dictionaries: the Routledge Dictionary of Economics. Second edition (Rutherford 2002) and The Dictionary of International Business Terms (Shim et al. 1998).

3 Conclusions

Parallel corpora are valuable language resources to study language in context and as a rich repository of terminology and phraseology. Language professionals, like translators, technical writers, LSP instructors and learners, terminologists and lexicographers can benefit from the exploitation of this type of corpus. The constitution of a language resource such as a parallel corpus from specialized texts from FTA's can be useful for the study of specialized collocations that are used in texts of free trade agreements.

Acknowledgements

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA).

References

Bach, Carme /Saurí, Roser /Vivaldi, Jordi /Cabré, María Teresa (1997) *El Corpus de l'IULA: descripció. Papers de l'IULA. Sèrie Informes, 17*. Barcelona: IIULA, Universitat Pompeu Fabra.

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

Benson, M. (1985) Collocations and idioms. In: Ilson, Robert (ed.) *Dictionaries, lexicography and language learning*. Oxford: Pergamon Press. 61-68.

Davies, Mark (2002) Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Actas del Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* .Valladolid: SEPLN. 21-27.

Davies, Mark (2009) The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics* 14. 159-190.

Hofland, Knut /Johansson, Stig (1998) The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In: Johansson, Stig /Oksefjell, Signe (eds) *Corpora and Cross-linguistic research. Theory, Method, and Case Studies.* Amsterdam/Atlanta: Rodopi. 87-100.

[4] http://www.natcorp.ox.a.uk
[5] http://www.corpus.rae.es/creanet.html
[6] http://www.sice.oas.org/dictionary_Dictio_e.asp

L'Homme, Marie Claude (2009) A methodology for describing collocations in a specialised dictionary. In: Nielsen, Sandro /Tarp, Sven (eds) *Lexicography in the 21st century*. Amsterdam: John Benjamins. 237-256.

Moon, R. (2008). Dictionaries and collocation. In: Granger, Sylviane /Meunier, Fanny (eds) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins. 313-336.

Orliac, Brigitte (2004) Automatisation du repérage et de l'encodage des collocations en langue de spécialité. Doctoral dissertation. Montreal: University of Montreal.

Dictionaries

Alcaraz, Enrique /Castro, José (2007) *Diccionario de comercio internacional: importación y exportación: inglés-español, Spanish-English*. Barcelona: Ariel.

Capela, John /Hartman, Stephen (2000) *Dictionary of International Business Terms.* 2nd Ed. Hauppauge, NY: Barron's Educational Series.

Routledge (1998) *Spanish Dictionary of Business, Commerce and Finance*. CD-ROM. London/New York: Routledge Software.

Rutherford, Donald (2002) *Routledge Dictionary of Economics.* London/New York: Routledge.

Shim, Jae K. /Siegel, Joel G. /Levine, Marc H. (1998). *The Dictionary of International Business Terms.* Chicago: Glenlake Publishing Company.