Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 249 – 256

Procedia Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Creation of a high-quality, register-diversified parallel (English-Spanish) corpus for linguistic and computational investigations

Julia Lavid[a], Jorge Arús[a], Bernard DeClerck[b], Veronique Hoste[c] *

[a]Dpt. English Philology I, Universidad Complutense, Spain
[b]Dpt. English, University of Ghent, [c]Dpt. Translation, Interpreting and Communication, University of Ghent, Belgium

Abstract

This paper outlines current work on the construction of a high-quality, richly-annotated and register-diversified parallel corpus for the English-Spanish language pair, as currently carried out within the framework of the MULTINOT project. The corpus consists of original and translated texts in both directions and is designed as a multifunctional resource to be used in a number of disciplines such as corpus-based contrastive linguistic and translation studies, machine translation, computer-assisted translation, computer-assisted language learning and terminology extraction. The paper describes the structure of the corpus –which includes four subcorpora: English originals (EO) and Spanish originals (SO), English translations (Etrans) and Spanish translations (Strans)-, the registers selected for inclusion in the corpus, and the methodology used to guarantee the quality of the processing steps to enrich the corpus with linguistic information at different levels.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Corpus creation, corpus annotation, English, Spanish.

1. Introduction

In spite of the increasing need for high-quality and richly-annotated corpora in different languages in the Natural Language Processing (NLP) community and the need for linguistically-interpreted parallel corpora in translation studies, it is difficult to find an integrated multifunctional resource for the English-Spanish pair whose features -in terms of quality of preprocessing, register diversity and multidimensional annotation- can satisfy the needs of a

* Corresponding author.
E-mail address: lavid@filol.ucm.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.443

diverse group of users and disciplines.

From the NLP perspective, existing parallel corpora for the English-Spanish pair are not balanced in their register composition and translation directions and are not subject to high-quality preprocessing and annotation procedures, as it is necessary for NLP tasks such as Named Entity Recognition (NER) or Terminology Extraction (TE).

From the linguistic and translation perspectives, existing corpora for the English-Spanish pair do not reflect the complexity of linguistic knowledge we are used to dealing with in linguistic theory. Linguistic research questions are usually complex, often involving constraints and interactions between different linguistic categories or levels of linguistic description. Simple research questions can be answered on the basis of raw corpora or even with the help of an automatic part-of-speech tagging system (see Lavid 2008; Lavid et al. 2010), but when investigating more challenging interactions and relations, it is necessary to count on resources with multiple levels of annotation which allow the extraction of features at different levels.

In this paper we report on the creation of a high-quality, register-diversified parallel and medium-sized corpus (half a million words) for the English-Spanish pair, consisting of originals and translated texts in both directions and enriched with linguistic annotations which can be exploited in a number of linguistic, applied and computational contexts. The creation of such a corpus is currently being carried out in the framework of the MULTINOT project, a research effort jointly developed between two European research groups (FUNCAP at Universidad Complutense and LT3 at Ghent University), with international expertise in contrastive, corpus-based linguistic and computational investigations.[†]

The paper is structured as follows: section 2 is an introduction to the MULTINOT project and its aims; section 3 describes the methodology used for the development of the corpus, focusing on its design and structure; section 4 elaborates on the different corpus preprocessing stages undertaken so far; section 5 outlines the manual annotation tasks foreseen for the last phase of the project; section 6 ends with some concluding remarks.

2. The MULTINOT project and its aims

The main aim of the MULTINOT project is to create a parallel English-Spanish corpus which is balanced – in terms of register diversity and translation directions – and whose design and enrichment with linguistic annotations focuses on quality rather than on quantity. In pursuing this aim, we have inspired ourselves in the design of other bilingual parallel corpora such as the CroCo corpus, for the English-German pair (Hansen Schirra et al. 2012, 2006; Culo et al. 2008), and the Dutch Parallel Corpus (DPC) for the Dutch-English, Dutch-French pair (Paulussen et al. 2013).

The second aim of the project is to offer the scientific community a multifunctional resource which can be used by a variety of potential users and in a number of theoretical and applied contexts. For example, linguists working on contrastive and corpus-based analysis, translators in need of bilingual parallel texts in both directions, translation trainers as a resource for computer-assisted translation, language teachers and computational linguists developing NLP applications.

The resulting corpus -the MULTINOT corpus- is a half-a-million-word sentence-aligned, and linguistically-annotated parallel corpus for the language pair English-Spanish. As the MULTINOT corpus is bidirectional, it can be used as comparable corpus to study contrastive differences between original texts in English and Spanish, between translations in both directions, and to study differences between translated versus non-translated texts in both languages.

[†] The MULTINOT project is financed by the Spanish Ministry of Economy and Competitiveness under project grant FFI2012-32201. The authors of this paper –all members of the FUNCAP (http://www.campusmoncloa.es/en/groups/funcap/), and the LT3 team (www.lt3.ugent.be) research groups- gratefully acknowledge the support provided by the Spanish authorities.

The MULTINOT corpus distinguishes itself from other parallel corpora by having a balanced composition (both in terms of registers and translation directions) and by focusing on quality rather than quantity. Thus, during the data collection phase, we made sure that the text samples were extracted from published online materials provided by publishing houses, press, government, corporate enterprises, European institutions, and other organizations under the 'fair use' agreement. Also, during data processing we also focused on corpus quality by manually correcting text samples at different processing stages such as sentence splitting, alignment and part-of-speech tagging. Furthermore, interannotator agreement is also foreseen for the manual annotation phase of several higher level features (modality, thematisation, discourse markers), which will be undertaken in the last phase of the project.

In order to enable corpus users to query the corpus selecting the texts that fulfill their specific needs, each sample has an accompanying metadata file including text-related and translation-related information. The whole corpus will be released in XML format, and made available through a password-protected online interface to be requested from corpus compilers.

3. Methodology

The design principles of the MULTINOT corpus have been based on three main principles: quality, diversity and balance, as explained below.

3.1. Quality

Given the intended multifunctionality of the corpus, it must contain high-quality representative data, so a quality control system has been developed for each step in compiling, annotating and aligning the corpus. We use two main forms of quality control in the MULTINOT corpus:

- Manual verification: this is performed by qualified linguists with native and near-native language proficiency. Manual validation of each processing step will be guaranteed for minimally 50% of the whole corpus.
- Spot checking: a spot-checking module is being developed on the basis of an error analysis of the manually verified data.
- Automatic control procedures: this refers to the automatic comparison of the output from different alignment programs.

3.2. Diversity

Given the variety of potential users of the MULTINOT corpus, we aimed at compiling a wide variety of registers from the written mode, following typologies used in other parallel corpus projects, such as the DPC corpus (Paulussen et al. 2013) and the CroCo Corpus (Hansen Schirra et al. 2012). Therefore, we opted for five main general text types or domains, which could be then subdivided into more delicate categories. The five macro-registers or domains were: literature, journalism, instructive texts, administrative texts and external communication. These five main types represent high-level macro-registers or domains and each contain several basic-level categories, as specified below:

- novels and short stories (from the fiction literature domain)
- essays and expository popular science texts (from the nonfiction domain)
- news reporting articles (from the journalistic domain)
- manuals and legal documents from webpages (from the instructive domain)
- official speeches and proceedings of parliamentary debates (from the administrative domain)
- annual reports and letters of self-presentation of companies, promotion and advertising brochures, and scientific texts (from the domain of external communication)

3.3. Balance

Translation direction is an important balancing criterion and for this purpose we set a target figure of 50,000 words per translation direction. Also, in order to preserve the balance in terms of providers, the combination of register and translation direction comes from at least three different providers. It should be noted, however, that it was very difficult to obtain enough data for certain registers in some translation directions. For example, expository

popular science texts are hardly translated from Spanish into English and it is also difficult to find scientific texts translated from Spanish into English. Therefore, we have been forced to make exceptions to the global design for those cases in which it was not possible to find material in both translation directions, or when it was difficult to find information on the translation direction. At the time when this paper is being written the number of words in the MULTINOT corpus is specified in Table 1 below:

Table 1: Word count distribution of different registers in MULTINOT

MacroRegister	**Sub-register**	**Source=> Target**	**English**	**Spanish**	**Total**
Literature	Novels	EN=>ES	24886 orig	26927 trans	51813
		ES=>EN	27939 tran	26672 orig	54611
	Short stories	EN=>ES	2186 orig	2088 trans	4274
		ES=>EN	1175 orig	1197 trans	2372
	Essays	EN=>ES	27382 orig	27235 trans	54617
		ES=>EN	32517 trans	30362 orig	62879
Journalism	News reporting articles (popsci)	EN=>ES	10658 orig	9753 trans	20411
		ES=>EN	24579 orig	23730 trans	48309
Administrative	Official speeches	EN=>ES	25373 orig	27112 trans	52485
		ES=>EN			
	Proceedings of debates	EN=>ES	25620 orig	26450 trans	52070
		ES=>EN	27390 orig	26320 trans	53710
External communication	Promotion/advertising brochures	EN=>ES	23695 orig	27790 trans	51485
		ES=>EN	25761 orig	25367 trans	51128
	Self-presentation documents	EN=>ES			
		ES=>EN	31188 orig	27326 trans	58514
	Scientific texts	EN=>ES	42580 orig	45047 trans	87627
		ES=>EN	24322 orig	23430 trans	47742
	Legal procedures	EN=>ES	36688 orig	38640 trans	75328
		ES=>EN	23984 orig	22185 trans	46169

4. Corpus preprocessing stages

The data compiled from different sources come in different formats and have to be normalized, preprocessed through the LeTs preprocessing pipeline (Van de Kauter et al. 2013) and aligned. We describe each of these processes in detail below.

4.1. Text normalization

Text normalization steps include the conversion of texts to .txt format, assigning documents a unique standardized name, normalizing the character encoding to the Unicode standard UTF8, cleaning the data and content removal (e.g.: tables, indexes, headers and footers, images).

4.2. LeTs Preprocessing pipeline

Once the texts are normalised they are further processed through the LeTs preprocessing pipeline (Van de Kauter et al. 2013). The LeTs Proprocessing toolkit accepts input files in various character encoding formats and first converts these files to UTF-8 encoding. The texts are then split into sentences and tokenized by rule-based methods

based on the Sentence Splitter and Tokenizer developed for the Europarl Parallel Corpus (Koehn 2005) and for TreeTagger (Schmid 1994, Schmid 1995). Subsquent steps of the LeTs pipeline include Part-of-Speech Tagging, Lemmatization and Named Entity Recognition. Figure 1 below illustrates the architecture of the LeTs Preprocess pipeline which will be used in the MULTINOT Project.

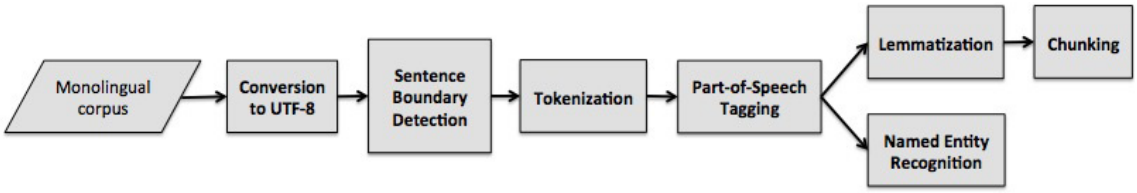

Figure 1: Architecture of the LeTs Preprocess pipeline (after Van de Kauter et al. 2013)

For English, tools such as the English tagger with tags from the Penn treebank tag set are being evaluated. For Spanish, we will use the CorNLP tagger for Spanish (Manning et al. 2014).

4.3. Alignment

In MULTINOT we will basically focus on sentential alignment, i.e., those cases where each sentence of the source language text is connected with the equivalent sentence or sentences of the target language text. The alignment links which are legitimate in MULTINOT are the following:

- 1:1 (one sentence in a source language is aligned with one sentence in a target language)
- 1: many (one sentence in a source language is aligned with two or more sentences in a target language)
- many:1 (two or more sentences in a source language are aligned with one sentence in a target language)
- many : many (two or more sentences in a source language are aligned with two or many sentences in a target language)
- 0 : 1 (no alignment links for a sentence in a target language)
- 1 : 0 (no alignment links for a sentence in a source language)

The sentence alignment process requires a previous task of paragraph alignment for the normal functioning of one of the aligners foreseen for the MULTINOT Project and used previously in the DPC corpus project: the Vanilla aligner (Danielsson and Ridings, 1997). The Vanilla aligner is an implementation of the Gale and Church algorithm (1993), and requires prior alignment of paragraphs to reduce the search space. We will, therefore, perform paragraph alignment manually with ParaConc (Barlow, 2002). It is also foreseen the use of a second aligner, the Microsoft Bilingual sentence aligner (Moore, 2002), which uses word correspondences - generated by a word translation model (IBM Translation Model 1) - to improve the initial alignment based on sentence length.

5. Manual Annotation Tasks

Manual annotation tasks for higher-level discourse categories of relevance for English and Spanish comparative and translation studies are foreseen for the last phase of the project. On the basis of previous studies focusing on the creation of reliable annotation schemas for phenomena such as thematisation (Lavid et al. 2013), modality (Zamorano and Carretero 2010) and appraisal (Taboada and Grieve 2004), carried out by members of the FUNCAP team, we will develop a representation standard which allows the alignment not only of raw text sequences but also of annotated translation units, thus facilitating the exploitation of the different parts of the MULTINOT corpus. For illustration purposes, Tables 2 and 3 below show the core tagsets for English and Spanish, respectively, based on the theoretical model proposed by Lavid (see Lavid et al 2010, chapter 5).

Table 2. Core Tagset for Theme types in English (after Arús, Lavid and Moratón 2012)

Annotation Layer	**Thematic field**	**Description**
Unit		Main clause
Core annotation scheme		
Tags:	TH (Thematic Head)	First nuclear constituent (Participant or Process, not Circumstantial) in main clause, or 'There' in Existential clauses.
	PreHead	Any Circumstantial element and/or Finite element preceding the Thematic Head.
	Textual Theme	Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders and other textual markers.
	Interpersonal Theme	Elements which express the attitude and the evaluation of the speaker with respect to his/her message, such as Vocatives and Modal Adjuncts, including mood and comment adjuncts

Table 3. Core Tagset for Theme Types in Spanish (after Arús, Lavid and Moratón 2012)

Annotation Layer	**Thematic field**	**Description**
Unit		Main clause
Core annotation scheme		
Tags:	TH (Thematic Head)	First nuclear element (not circumstantial) in main clause, realized by either lexical or morphological means.
	PreHead	Elements preceding the Thematic Head, including: Circumstantials, pronominal 'se', lexical part of Verbal Group.
	Textual Theme	Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders and other textual markers.
	Interpersonal Theme	Elements which express the attitude and the evaluation of the speaker with respect to his/her message.

For the manual annotation of the bilingual texts, the Ellogon annotation platform will be used. This is an annotation tool which concentrates characteristics for bilingual annotation of parallel and comparable texts that cannot be found altogether in a single tool (Petasis and Soumari 2012). A screenshot of the annotation tool can be seen in fig. 2. below:

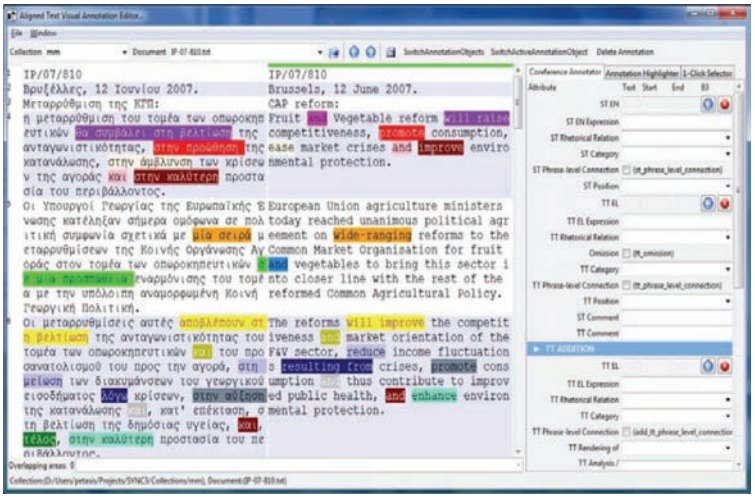

Fig. 2. Screenshot of the Ellogon annotation tool (Petasis and Soumari 2012).

As shown in Figure 2 above, aligned documents are displayed one next to the other, aligned at sentence level. Segments having the same color between texts in the two languages in fig. 2 denote that they belong to the same annotation (group of features). Clicking on any of them enables the editing/modification of the relevant annotation, through the inputs on the right side of the tool.

6. Concluding remarks

In this paper we have described current work on the construction of the MULTINOT corpus, a bilingual English-Spanish parallel textual database consisting of original and translated texts in both directions of the translation. The corpus mainly differs from other existing parallel corpora in the following aspects:

- Quality control: in order to guarantee corpus quality, a considerable part of the MULTINOT corpus is being checked manually at different levels, including sentence splitting, alignment and linguistic annotation.
- Level of annotation: the MULTINOT corpus is aligned, tagged on part-of-speech level and lemmatized. These are carried out automatically with existing processing tools. The annotation of more complex discourse features is foreseen through the use of manual annotation procedures.
- Balanced composition: the MULTINOT corpus contains texts from a wide range of registers and domains.

It is expected that the final corpus will be a useful and multifunctional resource for a wide range of applications in different research fields such as contrastive linguistics and translation studies, language teachers and language learners, human translators, and NLP developers.

Acknowledgements

The MULTINOT corpus is carried out within the MULTINOT project, which is funded by the Spanish Ministry of Economy and Competitiveness (MINECO).

References

Arús, J., Lavid J. and Moratón L. (2012). Annotating thematic features in English and Spanish: a contrastive corpus-based study. *Linguistics and the Human Sciences, 6.1.3*, 173 - 192.

Barlow, M. (2002) ParaConc: Concordance software for multilingual parallel corpora, in *Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research*., Las Palmas, Spain, pp. 20–24.

Culo, O., Hansen-Schirra, S, Neuman, S and Vela, M. (2008). Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. In *Proceedings of the LREC 2008 Workshop "Building and Using Comparable Corpora"*, Marrakech, Morocco.

Danielsson, P. and D. Ridings (1997) Practical presentation of a "vanilla" aligner, in *Proceedings of the TELRI Workshop on Alignment and*

Exploitation of Texts, Ljubljana.
Gale, W. A. and K. W. Church (1993) 'A program for aligning sentences in bilingual corpora'. *Computational Linguistics*, 19(1), pp. 75–102.
Hansen-Schirra, S, Neumann, S. and Steiner, E. (2012). *Cross-linguistic corpora for the study of translations – insights from the language pair English-German*. Berlin: de Gruyter.
Hansen-Schirra, S, Neumann, S. and Vela, M. (2006). Multidimensional Annotation and Alignment in an English-German Translation Corpus. In *Proceedings of the workshop on NLPXML-2006*. Italy.
Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation, *Proceedings of the tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86.
Kunz, K. and E. Steiner (2012). Towards a comparison of cohesive reference in English and German: system and texts. In M. Taboada, S. Doval Suarez and E. González Álvarez (eds.) *Contrastive Discourse Analysis: Functional and Corpus Perspectives.* London: Equinox.
Lavid, J. (2008). Contrastes: an online English-Spanish textual database for contrastive and translation learning. In Barbara Lewandowska-Tomaszczyk (ed.) *Corpus Linguistics, Computer tools and Applications: State of the Art*. Frankfurt: Peter Lang, 431-443.
Lavid, J. Arús, J. and JR Zamorano (2010). Designing and exploiting a small online English-Spanish parallel corpus for language teaching purposes. In *Corpus-based Approaches to English Language Teaching*. London: Continuum, 138-148.

Lavid, J., Arús, J., Carretero, M., Moratón, L. and JR Zamorano (2014) Contrastive corpus annotation in the CONTRANOT Project: Issues and problems. In Gómez González, María de los Ángeles, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García and Angela Downing (eds.). *The Functional Perspective on Language and Discourse: Applications and implicatio*ns. Amsterdam: John Benjamins, pp 57–86. DOI: 10.1075/pbns.247.04lav
Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.
Moore, R. C. (2002) Fast and accurate sentence alignment of bilingual corpora, in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, Tiburon, California, pp. 135–244.
Petasis, G. and Tsoumari, M. (2012). A New Annotation Tool for Aligned Bilingual Corpora. In Sojka, Petr, Horák, Alevs, Kopevcek, Ivan and Pala, Karel (eds.). *Text, Speech and Dialogue*, Springer Berlin Heidelberg., 01/2012: pages 95-104. ISBN: 9783642327896
Paulussen, H., L. Macken, W. Vandeweghe, and P. Desmet (2013). Dutch Parallel Corpus: a balanced parallel corpus for Dutch-English and Dutch-French, *Essential Speech and Language Technology for Dutch* pp. 185–199, Springer.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, pp. 44–49.
Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland, pp. 47–50.
Taboada, M. and J. Grieve (2004) Analyzing Appraisal Automatically. *American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text*. pp.158-161, Stanford. March 2004. AAAI Technical Report SS-04-07.
Van de Kauter, M., Coorman, J., Lefever, B., Desmet, B., Macken, L. And V. Hoste (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* 3 (2013) pp. 103-120.
Zamorano-Mansilla, JR and M. Carretero *An annotation scheme for dynamic modality in English and Spanish*. Linguistics and the Human Sciences, Vol 6, No 1-3 (2010), 297-320.