Annotating complex linguistic features in bilingual corpora: The case of MULTINOT

Julia Lavid

Department of English
Philology I,
Faculty of Linguistics
Universidad Complutense
E-mail:
julavid@filol.ucm.es

Abstract

In spite of the current need in the computational community for digital corpora in different languages with complex linguistic annotations going beyond morphosyntactic features, there is not much work within the Digital Humanities community dedicated to this task. In this paper I describe recent work on the development of a bilingual (English-Spanish) corpus consisting of original comparable and parallel texts from a variety of genres and annotated with complex linguistic features such as modality and evidentiality, metadiscourse markers, and thematisation, as carried out within the framework of the MULTINOT project (Lavid et al. 2015).

1 Introduction

Digital corpora annotated with complex linguistic information (i.e. semantic, pragmatic and discourse features) are fundamental for training and testing algorithms in the Natural Language Processing (NLP) community, and they are essential as gold standards for testing the performance of human language technology. In the Corpus Linguistics community, the annotation of texts adds value to a corpus in terms of reusability, stability and reproducibility. Moreover, corpus annotation has a tremendous potential as a topic of methodological cutting-edge research both for theoretical and applied corpus studies (Lavid 2012, Lavid et al. 2014).

However, 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 diverse group of users and disciplines.

In this paper I describe a number of issues and problems which have emerged in the annotation of complex linguistic features (i.e. semantic, pragmatic and discourse features) in the bilingual MULTINOT corpus, a high-quality, register-diversified parallel and medium-sized corpus (one 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 has been carried out in the framework of the MULTINOT project, a research effort jointly developed between two European research groups (FUNCAP at Universidad Complutense Madrid and LT3 at Ghent University) with international expertise in contrastive, corpus-based linguistic and computational investigations.[1]

The paper is organised as follows: section 2 describes the annotation tasks carried out within the project; section 3 discusses the main issues and problems emerged during the manual annotation tasks described in section 2; section 4 ends with some concluding remarks.

2 Annotation tasks in MULTINOT

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, 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 the focus was on corpus quality by manually correcting text samples at different processing stages such as sentence splitting, alignment and part-of-speech tagging. Furthermore, inter-annotator agreement studies have been performed for the empirical validation of complex linguistic features, such as modality, thematisation, and discourse markers. These are described in detail below:

2.1 Annotating modal meaning

As explained elsewhere, the annotation of modal meaning is a complex task (see Lavid et al. 2016a, 2016b). The difficulties derive not only from the practicalities of the annotation process, but also from the subtle distinctions which emerge in the modal domain. The complexity increases when dealing with more than one language, given the language-specific features that have to be considered in the annotation process. As

[1] The MULTINOT project is financed by the Spanish Ministry of Economy and Competitiveness under project grant FFI2012-32201. As principal investigator of the project, I gratefully acknowledge the support provided by the funding authorities.

explained elsewhere (see Lavid et al. 2016 a), for the annotation of modal meanings in English and Spanish we first designed a core tagset, consisting of four basic types of modal meanings, and an extended tagset, capturing the different subtypes. The four basic tags are [EP] (epistemic), [DE] (deontic), [DY] (dynamic) and [VO] (volitional). These are elaborated by more refined tags which capture more specific modal meanings. Thus, for example, 'epistemic' meanings, which express a qualification of the truth of a proposition (Boye, 2012), are divided into two main subtypes: a) 'evidential' meanings, defined in terms of the notion of source of information, evidence, or epistemic justification; and b) 'non-evidential' meanings, referring to the degree of certainty or epistemic support.[2]

More specifically, for example, 'evidential' meanings can be subdivided depending on the source of the evidence that the speaker has or claims to have at his/her disposal, for or against the truth. These sources can be:

1. *perceptual* [PE], referring to non-linguistic sources obtained through the senses, as in (1) and (2) below:
    (1) Instead, scientists say, the mountains of Oahu are *actually* dissolving from within.
    (2) En realidad, según los resultados de una nueva investigación, esas montañas, derivadas de dos volcanes *aparentemente* extintos, se están disolviendo desde dentro

2. *cognitive* [COG], referring to evidence coming from knowledge by someone different from the speaker/ writer, including thoughts, beliefs and apprehension, as in (3) and (4) below:
    (3) Martin Rees, Britain's astronomer royal, *believes* that there are many universes, possibly an infinite number (EO_EXPE_001)
    (4) Pablo *sospecha* que si contara uno por uno los ladrillos que dibuja a mano alzada sobre la fachada se encontraría en cada boceto con idéntica cantidad. (SO_FICTION_017)

3. *communicative* [COM], referring to evidence coming from linguistic messages, as in (5) and (6):
    (5) *According to* the researchers' estimates, the net effect is that Oahu will continue to grow for as long as 1.5 million years. (EO_EXPE_005)
    (6) *Según* las estimaciones de los investigadores, el resultado final es que Oahu continuará creciendo hasta dentro de un millón y medio de años. (STRANS_EXPE_005)

The groupings and the subtypes presented below in Figure 1 below are the result of a number of preliminary annotation experiments during which the tags were elaborated and refined until a consensus was reached on the basic and the secondary meanings. The tagsets are hierarchical, allowing the annotator to choose the coarser tags from the core tagset (EP, DE, DY, VO), when in doubt about the more fine-grained subtypes from the extended tagset. For example, if the annotator is uncertain about whether a markable is 'possibility' or

[2] These are treated in detail in Lavid et al. (2016b).

'probability', s/he can simply tag it as 'epistemic' [EP] and 'non-evidential' [NEV]. The abbreviated form of each tag is given in capital letters in brackets next to the full form.

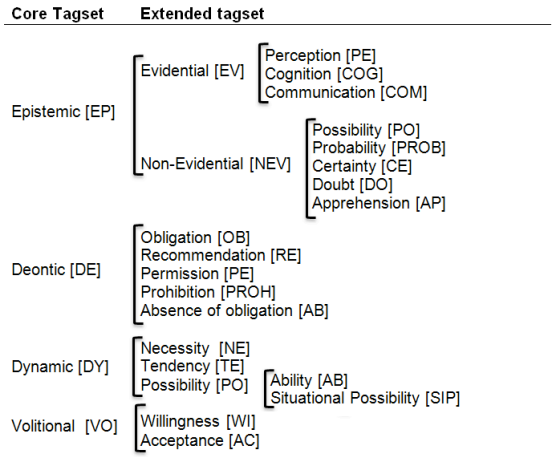

Figure 1: Core and extended tagsets for modal meanings in English and Spanish (after Lavid et al. 2016a)

In spite of the difficulties in distinguishing between these categories, the annotation experiments performed on two datasets (one containing two hundred English sentences and the other two hundred Spanish sentences, all of them extracted from original texts of the MULTINOT corpus) yielded a good degree of agreement between annotators, as shown in tables 1 and 2 below.

	Number of agreements (%)	Cohen's kappa coefficient
Epistemic/Evidential	84.00%	0.792
Deontic modality	80.00%	0.837
Dynamic modality	94.00%	0.868
Volitional modality	86.00%	0.787

TABLE 1 Inter-annotator agreement within modal subtypes in English

	Number of agreements (%)	Cohen's kappa coefficient
Epistemic/Evidential	76.00%	0.703
Deontic modality	84.00%	0.880
Dynamic modality	96.00%	0.875
Volitional modality	88.00%	0.816

TABLE 2 Inter-annotator agreement within modal subtypes in Spanish

This indicates that the proposed tagsets for different modal meanings, i.e., epistemic, deontic, dynamic and volitional, are reliable and consistent and can be used for the large-scale annotation of the bilingual corpus. However, the annotation experiment also revealed difficult cases and problems in the annotation process, which will be discussed in section 3. The large-scale annotation of the bilingual texts of the MULTINOT corpus is currently being carried out by two independent annotators using a pre-annotated list of triggers on aligned versions of the bilingual texts in Excel, as shown in Figure 2 below.

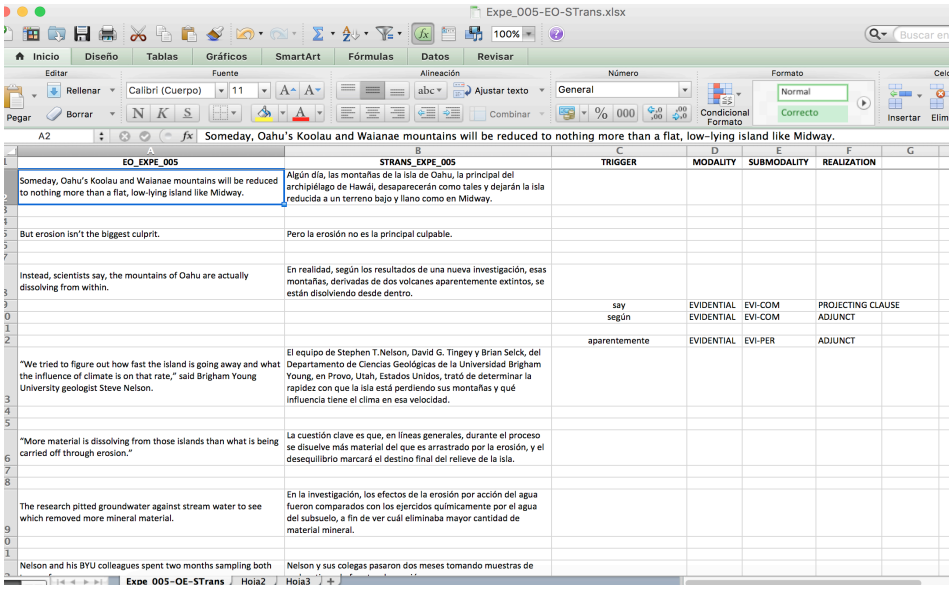

Figure 2: Screenshot of bilingual annotation of aligned original and translated text.

2.2. Annotating metadiscourse markers

For the annotation of metadiscourse markers, an annotation scheme was designed on the basis of Hyland's distinction between interactive and interactional markers, supplemented with more specific ones from the area of epistemicity and evidentiality. Interactive (textual) markers are concerned with ways of organising discourse to anticipate readers' knowledge and concerned with ways of organising discourse to anticipate readers' knowledge and include *transitions, frame markers, endophoric markers, evidentials* and *code glosses.* The initial tagset for interactive markers is graphically displayed in Table 3:

TEXTUAL MARKERS	**FUNCTION** Enable the writer to manage the information flow so as to provide his preferred interpretations	**ENGLISH**	**SPANISH**
Transitions	Express semantic relation between main clauses.	In addition, but, thus, and, when, etc…	Además, y, pero, cuando, etc..
Frame Markers	Indicate text boundaries or elements of schematic text structure.	Finally, first, second, to conclude, etc…	Primero, finalmente, para terminar, etc…
Endophoric Markers	Refer to information in other parts of the text	See figure X, in section Y, here, etc..	Véase, en la sección, ahí, en eso
Evidentials	Refer to sources of information from other texts/people	X states, (Y, 2010), According to X	X dice que, según Y,
Code glosses	Help readers grasp functions of ideational material.	namely, such as, in other words, e.g., parenthesis, punctuation devices e	A saber, en otras palabras, paréntesis, puntuación

TABLE 3 Initial tagset for interactive markers in English and Spanish

Interactional (interpersonal) markers focus on the participants of the interaction and "seek to display the writer's persona and a tenor consistent with the norms of the disciplinary community" (Hyland & Tse 2004, 139). These include *hedges, boosters, attitude markers, engagement markers and self-mention markers.* The initial tagset for interactional markers is graphically displayed in Table 4:

INTERPER-SONAL MARKERS		**FUNCTiON**	**ENG.**	**SPA.**
Stance Markers [SM]	Hedges	withhold writer's full commitment to proposition	Might, perhaps, possibly	*Quizás, posible…, podría ser, etc…*
	Boosters	emphasize force or the writer's Incertainty	It is clear that, in fact, definitely, etc…	*Está claro que, de hecho, definitiva-mente*
	Attitude	writer's appraisal	Unfortunatel y,surprisingly	*Desgracia-damente*
	Self-mention	writer's personalisation strategy	I was a poll clerk	*Yo me encuentro entristecido*
Engagement markers [EM]	Questions	call attention to uncertainties	What to do?	*¿Quién lo supera?*
	Inclusive 1sp.pl	pull the reader along with the writer's argument	We need to act , Let's wait	*Pensemos, date cuenta, etc…*
	Indefinite, 2nd P.Pr	recognise the presence of their readers	you cannot seek	*No sé, ustedes..*
	Directives	Direct reader to particular action	Tell that to my parents	
	Deontic Modals		We need to act fast	*Es necesario que, hay que..*
	Asides	Seek complicity with reader		-eso sí-

TABLE 4 Initial tagset for interactional markers in English and Spanish

The annotation experiments were performed on eighteen comparable texts randomly selected from the bilingual MULTINOT corpus and divided into news reports, editorials and letters to the editor. Except for some disagreements in the annotation of engagement markers, the experiments showed that the categories used are valid and can be fruitfully used to characterize the three journalistic genres studied.

3 Issues and problems

In spite of the agreement rates obtained in the annotation experiments, a number of issues and problems emerged during the annotations which deserve to be analysed and discussed.

In the area of modality, the difficulties encountered in the annotation experiment can be divided into two main types:

1. Cases where there was a degree of overlap in the modal meanings expressed by the triggers, such as modal auxiliaries (*can, may, might, must*, etc.) in English and their counterparts in Spanish (*poder, deber, tener que*), as well as with some related adjectives (*possible*). These items are polysemous, i.e., they tend to express more than one modal meaning (*must* : obligation, necessity, prohibition; *can*: permission, ability, situational possibility, prohibition), and this can give rise to potential disagreement between annotators.
2. Cases where annotators disagree on the modal nature of the triggers. This group includes mostly lexical verbs, adjectives and nouns such as *prohibit*, *necessary* or *obligation* which have a meaning that is closely related to one of the modality types. The annotation experiments showed that the real challenge with these triggers is deciding whether they express a modal meaning or not. This is because these words are little or not grammaticalized at all, and while some of their uses are equivalent to modal constructions, other uses are clearly non-modal.

The annotation of metadiscourse markers also yielded problematic cases, but, in general, proved to be easier than the annotation of modal meanings. Interestingly, the annotation results revealed a number of tendencies in the use of metadiscourse markers in the three journalistic genres, which are summarised below:

a) Interactive (textual) markers are the most frequent metadiscourse markers in all three journalistic genres in comparison with Interpersonal markers.

b) Interactional (interpersonal) markers (both Stance and Engagement) appear only in Editorials and Letters to the Editor, but not in News Reports. This is probably due to the fact that news reports must be 'impartial' and 'objective' and avoid – or at least minimize – showing their interpersonal involvement in the text's construction.

c) Transition Markers are the most frequent of all interactive (textual) markers in all three journalistic genres.

d) Evidentials are more frequent in News reports and Editorials in comparison

with Letters to the Editor, probably due to the tendency to rely on other sources for attribution of information.
e) Transition markers are used differently depending on the text's communicative purpose: higher frequency of Temporal in News Reports in English and Spanish in comparison with Editorials and Letters; High frequency of adversative markers in English Editorials vs Additive in Spanish, and similar frequency of Cause, Concession, Adversative markers in Letters to the editor.

4 Conclusion

This paper has summarised and discussed work on the manual annotation of complex linguistic categories such as modality and metadiscourse markers in the framework of the MULTINOT project. The annotation of these categories in the bilingual corpus is important not only from a contrastive and translational point of view, but also for its computational relevance in the NLP community, where modally-annotated corpora are an indispensable resource for training systems to automatically interpret modality. Likewise, the annotation of discourse markers is a very useful task for the understanding of text coherence in the context of NLP applications such as Automated Text Generation and in a number of application contexts such as Contrastive linguistics and Translation studies, Computer-Aided and Machine Translation.

References

[1] Boye, Kasper. 2012. Epistemic Meaning: A Crosslinguistic and Functional-Cognitive Study. Empirical Approaches to Language Typology 43. De Gruyter Mouton: Berlin and Boston.

[2] Hyland, K. (2004). Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* , 13, 133-151. http://dx.doi.org/10.1016/j.jslw.2004.02.001

[3] Lavid, Julia (2012). Corpus Annotation in CONTRANOT: Linguistic and Methodological Challenges. In Isabel Moskowitz and Begoña Crespo (eds.*) Encoding the Past: Decoding the Future: Corpora in the 21st Century.* Cambridge Scholars, 205-220. ISBN: 1-44383581-1

[4] Lavid, Julia, Marta Carretero, Jorge Arús Hita, Lara Moratón and Juan Rafael Zamorano-Mansilla (2014): Contrastive corpus annotation in the CONTRANOT project: Issues and problems. In María de los Ángeles Gómez González, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García and Angela Downing (eds.). *The Functional Perspective on Grammar and Discourse*. Amsterdam: John Benjamins, 57-86.

[5] Lavid, J. and L. Moratón (2016) Contrastive annotation of interpersonal

discourse markers in English and Spanish journalistic texts. Paper presented at the *International IWODA Conference*, September 2016. University of Santiago de Compostela (Spain).

[6] Lavid, Julia, Carretero, Marta and Zamorano, Juan Rafael. (2016a). Contrastive Annotation of Epistemicity in the MULTINOT Project: Preliminary Steps, in Harry Bunt (ed.). *Proceedings of the LREC 2016 Workshop ISA-12 – 12th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, 28 May 2016 – Portorož, Slovenia (2016a), 81-88.

[7] Lavid, Julia, Carretero, Marta and Zamorano, Juan Rafael. (2016b). A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT, in *LiLT (Linguistic Issues in Language Technology), Volume 14, Special Issue on Modality in Natural Language Understanding*. ISSN: 1945-3604. Available online at http://csli-lilt.stanford.edu/ojs/index.php/LiLT/index, CSLI Publications, Stanford University, 1-33.

[8] Lavid, Julia, Arús, Jorge, DeClerck, B and Hoste, Veronique (2015). Creation of a high quality, register-diversified parallel corpus for linguistic and computational investigations. In *Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics* (CILC2015). Procedia - Social and Behavioral Sciences, Volume 198, 24 July 2015, Pages 249–256