Contrastive corpus annotation in the CONTRANOT project: Issues and Problems

Julia Lavid, Jorge Arús, Marta Carretero, Lara Moratón and Juan Rafael Zamorano

University Complutense of Madrid (Spain)

1. Introduction

The task of manual (or human-coded) corpus annotation is currently the object of extensive research in the Natural Language Processing (NLP) community for a number of computational applications.[1] NLP researchers have managed to automate the annotation of various kinds of 'low level' linguistic tasks (e.g. tokenisation, lemmatisation, part-of-speech tagging, parsing) with a reasonable degree of accuracy, but they find it difficult to automate the annotation of higher levels of linguistic processing (e.g., semantic, pragmatic, or discourse categories) for use in applications such as Information Extraction (IE), Information Retrieval (IR), Automated Text Summarisation, or Machine Translation, among others. Automatic annotation of semantic, pragmatic or discourse categories requires manual annotation by humans first, to produce a small corpus with high-quality human-coded annotations on which computer algorithms can be trained to be able to build computational systems that 'learn' from human-coded data. As a result of this requirement, an active area of research in the NLP community is now focused on ensuring the quality of the human-coded annotations, i.e., the extent to which the annotation procedures are well designed and 'reliable' (Reidsma and Carletta 2008).

In the Linguistics camp, the topic of manual corpus annotation has not received the same attention as in the NLP community, mainly due to a lack of concern for the reliability of the corpus annotation process: manual corpus annotation is often considered to be the same as traditional corpus analysis by a single linguist (McEnery et al. 2006). However, as explained elsewhere (Hovy and Lavid 2010), and as shown by the work carried out within the framework of

the CONTRANOT project, manual corpus annotation can be considered as a topic of methodological cutting-edge research both for theoretical and applied corpus studies (see Lavid 2012). More specifically, manual corpus annotation can be exploited as a mechanism to test aspects of linguistic categories empirically, and to reveal and reformulate features of complex linguistic categories in a contrastive manner (see Arús, Lavid and Moratón 2012; Lavid, Arús and Moratón 2010a, 2010b; Carretero and Taboada in press; Taboada and Carretero in press; Carretero and Zamorano 2010).

In this paper we focus on the contrastive corpus annotation of certain aspects of the phenomena of Thematisation, on the one hand, and of Modality, on the other, in the framework of the CONTRANOT project, a research effort aimed at the creation and validation of contrastive functional descriptions through corpus analysis and annotation. Our most immediate aim in this paper is to illustrate the issues and challenges researchers have to face when confronted with the task of developing well-designed and reliable annotation procedures for complex linguistic phenomena in a contrastive manner. This includes the presentation of the annotation schemas developed so far and a description of the agreement studies carried out to test the reliability of these annotation schemas. Our final aim is to create high-quality human-coded annotations on which computer algorithms can be trained to automate the corpus annotation of complex linguistic categories such as Thematisation and Modality for computational applications.

The paper is structured as follows: section 2 provides the background for the work by outlining the main annotation tasks and procedures developed in the CONTRANOT project. Section 3 describes the annotation tasks in the area of Thematisation and section 4 in the area of Modality. Section 5 evaluates and discusses the results of the annotations in these two areas. Finally, section 6 summarises the work reported and provides some pointers for the future.

2. Background issues: the CONTRANOT project

The background for the work reported here is the CONTRANOT project, one of whose aims is to produce reliable and consistent human-coded annotations which can serve as quality data for the training of machine-learning algorithms. In order to ensure the reliability and consistency of the human-coded annotations, we have followed a number of steps and procedures which include:

- the selection of the training corpora
- the selection and instantiation of the theoretical categories to be annotated
- the design of annotation schemes
- the performance of agreement studies
- the evaluation of the annotations

In the following sections we will illustrate these steps through the description of the annotation tasks carried out in the area of Thematisation and Modality in English and Spanish.

3. Contrastive annotation of Thematisation

For the contrastive annotation of thematic features we first selected a training corpus consisting of thirty two newspaper texts (16 English and 16 Spanish) equally divided into news reports and commentaries. The selection of this corpus was motivated by our interest in journalistic discourse and by the electronic availability of comparable data in both languages.

The thematic features which were selected for the design of the annotation scheme were based on the recent model of thematisation for Spanish in contrast to English described in Lavid et al. (2010, 294-306). We designed two annotation schemes, one for English and one for Spanish, including both coarse-grained and more fine-grained annotations. Table 1 below displays the English and the Spanish core tagsets which include six coarse-grained tags reflecting the range of

possible thematic types which can occur as part of the Thematic field in English and Spanish declarative clauses, both in news reports and in commentaries. Definitions and realisations of these tags for each language are provided in Appendix 1 (for English) and Appendix 2 (for Spanish) at the end of the paper.

Table 1: Core tagsets for English and Spanish

English	**Spanish**
Thematic Head (TH)	Thematic Head (TH)
PreHead (PH)	PreHead (PH)
Textual Theme (TT)	Textual Theme (TT)
Interpersonal Theme (IT)	Interpersonal Theme (IT)
Predicated Theme (PT)	Predicated Theme (PT)
'There'-type Theme (There-T)	'Hay'-type Theme (Hay-T)

The extended tagset includes more fine-grained subtypes of Thematic Heads, namely, those conflating with experiential roles within the clause (e.g. Actor, Senser, Phenomenon, etc…). Table 2 presents a preliminary extended tagset for Thematic Head realisations, subject to further refinements. Definitions for these tags and examples of realisations are provided in Appendix 1 (for English) and Appendix 2 (for Spanish) at the end of the paper.

Table 2: Preliminary extended tagset for Thematic Head types

Participant Type as Thematic
TH-Actor
TH- Goal
TH- Beneficiary
TH- Senser
TH- Phenomenon
TH- Sayer
TH- Carrier
TH- Token
TH- Value

In order to test the reliability of the core and part of the extended tagset presented for both languages, we performed two agreement studies on a small fragment of our initial training corpus consisting of a total of 143 clause complexes for the English dataset and 79 for the Spanish one. On both datasets, the first agreement study measured inter-annotator agreement on the

identification of thematic spans, while the second measured inter-annotator agreement on the type of label chosen by the annotators on the previously selected spans. We used two types of agreement metrics: the Agreement Metric (AGR) and Kappa (K). For the first task –the identification of thematic spans– we used the Agreement Metric (AGR) rather than Kappa because the annotators could be coding different expressions (markables) in identifying thematic spans. For the second task –the labelling of the thematic types– we used the kappa coefficient (K), which measures agreement when two independent coders are analysing the same element.

The first agreement study focused on the identification thematic markables and consisted of two tasks. In the first task, annotators had to identify the whole Thematic field in each clause complex (the definition and realisations of the Thematic field are provided in Appendix 1 at the end of this paper).

In the second task, annotators were asked to identify the spans realising only specific thematic types from the Thematic field, i.e. those included in the core tagset (definitions and realisations of these tags are provided in Appendix 2 at the end of this paper).

In the first task the agreement between coder (a) and (b) was very high on average (0.97 %) in the annotation of the English dataset. In the second task, agreement was high in the identification of the span expressing the Thematic Head (0.9384%) and the Textual Theme (0.965%), but lower –although still substantial- in the identification of the PreHead (0.787%). By contrast, agreement was only fair (0.375%) in the identification of the Interpersonal Theme, which was labelled as Textual Theme by one of the annotators. This could be due to a performance error, but may indicate that the definitions for these two tags need reformulation or extension so as to make them more clearly distinguishable from each other.

With respect to the Spanish dataset, agreement was high on average in the first task (0.91%), i.e. the identification of the whole Thematic field in each clause span. Disagreement mainly occurred when confronted with a complex Verbal group which included an auxiliary and a lexical part. Here annotators hesitated about including the lexical part or not within the Thematic field. In

the second task agreement was almost perfect in the identification of the Predicated Theme (0.98%), the Textual Theme (0.97%), 'Hay' Theme (0.98%), but slightly lower –although still high- in the identification of the Interpersonal Theme (0.94%), the Thematic Head (0.91%), and the PreHead (0.89%). The few disagreements that occurred in the identification of the Thematic Head were due to discrepancies between annotators as to what part of the verbal morphology realised the Thematic Head, specially when the verb was irregular. For example, when confronted with the form 'es' (3rd person singular of the Present tense 'to be' in Spanish), annotators found it difficult to decide which would be the Head. Also some disagreements occurred in the identification of the Interpersonal Theme, where one annotator included elements which were part of the PreHead. For example, one coder considered '*Quizás*' as an interpersonal element whereas the other has coded it here as Pre-Head. Table 3 below summarises the results of the first agreement study on the English and the Spanish dataset.

Table 3: Results of first agreement study for English and Spanish datasets

	English	Spanish
Task 1: identification of Thematic field	AGR = 0.97 %	AGR = 0.91%
Task 2: Identification of thematic spans realising core tags	AGR for TH = 0.9384 AGR for TT = 0.965 AGR for PH = 0.787 AGR for IT = 0.375	AGR for TH = 0.91 AGR for TT = 0.97 AGR for PH = 0.89 AGR for IT = 0.94

The second agreement study focused on the labelling of markables and consisted of two tasks. In the first task, annotators had to label the thematic markables which had been agreed on in the previous agreement study. For this task the agreed thematic spans were highlighted in the coding sheet so that coders could carry out the classification task on the same span. We also included some 'red herrings' in this task, i.e., highlighted items which did not correspond to any of the thematic types of the core tagset and asked the coders to classify those as 'none' with the aim of checking the annotator's knowledge of the different types. In the second task annotators

had to choose the tags from the extended tagset for Thematic Head types, corresponding to different experiential roles conflating with Thematic Heads, as specified in Table 2 above.

In the first task, agreement was almost perfect (kappa=0.915) in the English dataset. In the second task, agreement was substantial (kappa=0.875), with disagreement occurring only in fifteen cases, probably due to the inherent difficulty in disambiguating experiential roles conflating with Thematic Heads. The dividing line between material and metaphorical relational processes proved to be particularly problematic. Within relational processes, the differentiation between attributive and identifying, as well as the directionality of identifying processes, were also important sources of disagreement.

In the Spanish dataset agreement was substantial in the first task of labelling the thematic markables which had been agreed on in the previous agreement study (kappa =0.839), which indicates that thematic labelling using the core tagset was easier once the thematic spans were previously identified. By contrast, agreement was only moderate in the second task (kappa=0.475), probably for the same reason as in the English dataset: the inherent difficulty in disambiguating experiential roles conflating with Thematic Heads. Confusion between Thematic Heads as Carriers in attributive processes and Tokens in identifying processes were the most frequent sources of disagreement. Table 4 below summarises the results of the second agreement study on the English and the Spanish dataset.

Table 4: Results of second agreement study for English and Spanish datasets

	English	Spanish
Task 1: labeling of core tags	kappa = 0.915	kappa = 0.839
Task 2: labeling of Thematic Head types	Kappa = 0.875	kappa = 0.475

4. Contrastive annotation of Modality features

For the annotation of modality features we designed an annotation scheme focused on three main types of modality and their realisation through a limited number of modal verbs both in English

and in Spanish. The three modality types were those which are most commonly distinguished in the literature, namely, epistemic, deontic and dynamic (Hermerén 1978; Palmer 1990; Perkins 1983; Silva-Corvalán 1995; Nuyts 2001; Wärnsby 2006; Collins 2009). Definitions and illustrative examples for these three types are provided in Appendix 3 at the end of the paper. The reason for the choice of a limited number of expressions as point of departure lies in the difficulty to determine all the realisations of each category in the present state of the definitions of the categories. This difficulty, which became obvious in previous attempts to annotate complete texts on the basis of these categories, explains why monographs on modality are most often based on concrete expressions and hardly ever aim to give exhaustive accounts of the different realisations of each type. The modal verbs selected as markables were the following:

- The English verbs *must* and *have to* (the latter both in the present and past tenses), and their Spanish equivalents *deber (de)*[2] and *tener que* (the latter both in the present and past tenses to enable comparison with English *have to*).
- The English verbs *can* and *may* and their respective past forms *could* and *might*, as well as their Spanish equivalent *poder* (both in the present and in the past tenses).

The reasons for choosing these verbs are various: first, these verbs are highly frequent in the spoken and written varieties of English and Spanish; second, they are highly comparable, being translatable in both languages even if each expression has its own idiosyncrasies; third, modal verbs are typical realisations of modality, both in English and in Spanish, as is attested by the copious bibliography devoted to them; fourth, each of these expressions has different modal meanings which are far from easy to distinguish in many cases. This poses a challenge for the design of the annotation system of modality, since decisions will have to be made concerning the classification of unclear occurrences. Predictably, these decisions will be applicable to the annotation of other expressions of modality and will contribute to future theoretical descriptions of the types of modality. The tagset used for the annotation is specified in Table 5.

Table 5: Tagset for modality in English and Spanish

TYPE OF MODALITY	SHORT DEFINITION	REALISATIONS IN ENGLISH	REALISATIONS IN SPANISH
Epistemic (EPIST)	Degrees of probability	*could, may, might, must*	*deber (de), poder, tener que*
Deontic (DEONT)	Degrees of obligation, permission and prohibition	*can, could, have to, may, might, must*	*deber (de), poder, tener que*
Dynamic (DYNA)	Inevitability, tendency, ability, natural possibility and impossibility	*can, could, have to, must*	*deber (de), poder, tener que*

In order to test the reliability of the tagset, we performed two agreement studies on an initial training corpus consisting of a total of 480 tokens (40 markables for each of the modal verbs mentioned above, including both English and Spanish), randomly extracted from the British National Corpus (BNC) (http://corpus.byu.edu/bnc/) for English and the Corpus del Español (CdE) (http://www.corpusdelespanol.org) for Spanish. In the agreement studies each markable was classified as expressing epistemic [EPIST], deontic [DEON] or dynamic [DYNA] modality by two annotators. The purpose of the first agreement study was to measure the level of agreement in the assignment of one of the three possible types of modality by two different annotators (inter-annotator agreement) with the aim of detecting the areas of lower levels of agreement and proposing guidelines for the future annotation of the cases that fall into these areas. The second study, which was carried out six months after the first study, consists of the annotation of exactly the same tokens by the same annotators after having discussed problematic cases between annotators and reached a consensus as to how to annotate those cases on the basis of specific contexts of use.

The agreement levels were measured using two indicators: the percentage or proportion of examples that showed agreement in the annotation, and the Kappa coefficient. The overall results obtained in the first experiment are summarized in Table 6 below. These values indicate that in general terms the definitions of the three modality types (epistemic, deontic, dynamic) yield acceptable levels of agreement between annotators, but there is still some margin for improvement.

Table 6: Inter-annotator agreement in the identification of modal meanings (first experiment)

NUMBER OF EXAMPLES ANNOTATED	PROPORTION OF AGREEMENT	KAPPA VALUE
480	0.75	0.64

The cases of disagreement were classified into three groups:

- Dynamic-epistemic discrepancy. This occurred when one of the annotators tagged an expression as indicating dynamic modality while the other opted for epistemic modality.
- Dynamic-deontic discrepancy. This occurred when one of the annotators chose dynamic modality whereas the other chose deontic modality.
- Epistemic-deontic discrepancy. The disagreement occurred between epistemic and deontic modality.

Figure 1 summarizes the frequency of occurrence for each of the three types of disagreement.

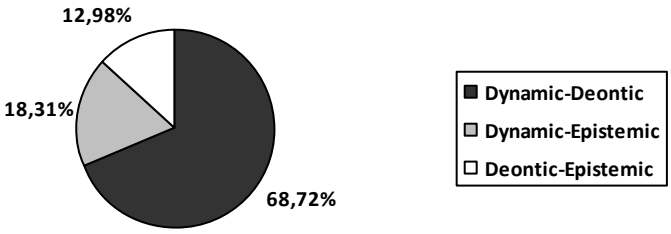

Figure 1. Distribution of causes of disagreement between annotators

The results suggest that dynamic modality is the least well defined of the three types, since the confusion between this and the other two types of modality accounts for about 87% of all disagreement cases. A more detailed discussion of the three types of disagreement and their causes is given below. The results offered up to now concern the three types of modality (dynamic,

epistemic, deontic) from a global perspective. In other words, they provide valuable information about the precision and replicability of the definitions for each type of modality when used to classify a significant number of examples by two different annotators.

However, since modal verbs differ markedly in their polysemous nature, it is also worth considering the individual results obtained for each verb in both languages: these will offer an insight into the difficulties posed by the semantics of the each modal verb. Tables 7 and 8 below show the agreement level obtained in the first agreement study for each of the English and the Spanish modals, respectively.

Table 7: Initial inter-annotator agreement (English verbs)

MODAL VERB	PROPORTION OF AGREEMENT	KAPPA VALUE
must	0.93	0.85
may	0.95	0.84
might	0.95	0.48
had to	0.75	0.55
can	0.72	0.49
could	0.68	0.49
have to	0.65	0.35

Table 8: Initial inter-annotator agreement (Spanish verbs).

MODAL VERB	PROPORTION OF AGREEMENT	KAPPA VALUE
deber	0.85	0.69
poder [past]	0.75	0.47
tener que [present]	0.62	0.43
tener que [past]	0.62	0.30
poder [present]	0.50	0.28

As shown in Tables 7 and 8, the highest proportion of agreement occurred in the annotation of the modal verbs *must*, *may* and *might* in English, and of deber in Spanish. The rest of verbs produced very similar results ranging from 65 % to 75 % of agreement. However the Kappa value for these verbs shows more variation. This is because the Kappa formula is sensitive to the number of choices actually made during the process of annotation, reflecting the fact that some of the tags did not pose a problem for the annotators. However the Kappa value for these verbs

shows more variation. This is because the Kappa formula assigns a different weight to the each individual case of disagreement depending on the overall results produced by the annotators. For example, one case of disagreement will have a more serious impact on the resulting Kappa value if the annotators had to choose between two possible tags than if they had to choose between three tags. Likewise, one case of disagreement will affect the final Kappa value more negatively when the annotation for the rest of the cases is dominated by one of the tags than when this annotation is more varied between the possible tags. All this explains why one obtains the same Kappa value for *can* and *could*, even though the proportion of agreement is not the same. This reflects the fact that the range of problematic decisions is slightly wider in the case of *could* than in the case of *can*, and consequently one needs fewer cases of disagreement with *can* to have a negative impact on the Kappa value. Something similar could be claimed about *may* and *might*. In spite of the fact that they yielded the same proportion of agreement (95% of examples), the Kappa value varies considerably because *may* apparently presents more challenging decisions than *might*. That explains why one obtains the same Kappa value for *can* and *could*, even though the proportion of agreement is not the same. This means that the range of problematic decisions is slightly wider in the case of *could* than in the case of *can*. Something similar could be claimed about *may* and *might*. In spite of the fact that they yielded the same proportion of agreement (95% of examples), the Kappa value varies considerably because *may* presents more challenging decisions than *might*.

In several cases, the similarity in level of agreement between equivalent modals in the two languages is noticeable. This is the case of *must* and *deber (de)*, *tener que* and *have to* (both in the present and in the past), and, to a lesser extent, *poder* and *can / could*. These similarities indicate that the areas of disagreement in English and Spanish are similar from the conceptual point of view.

In the following paragraphs we will describe the main areas of disagreement detected in the first agreement study in specific contexts of use. Given the inherent ambiguity of certain uses,

annotators reached a consensus as to what tag to attach to each verb on the basis of a number of factors, as described below:

1) Disagreement between deontic and dynamic modality:

-Context A. Disagreement between strong dynamic and deontic modality with *have to* and *tener que* (present and past), due to the difficulty to distinguish between necessity that stems from social laws (deontic) and obligation that stems from natural laws (dynamic). The problematic examples typically refer to actions whose (non-) performance is voluntary (in this respect, they can be considered as deontic) but inevitably connected with certain results (and this inevitability connects it with dynamic modality). We consider that, in these cases, the factor of common sense or reasonableness is predominant and the decision was made to annotate them as deontic.

(1) y resulta que el muchacho estaba cayendo al mar, y entonces una de las normas de seguridad es que cuando uno ve que está cayendo al mar *tiene que* soltar todo el equipo, de manera de no hacer mucho peso.

"and it turns up that the boy was falling into the sea, and then one of the safety rules is that when someone sees that they are falling to sea, they must get rid of all the equipment, so that the weight is not much."

- Context B. Disagreement between dynamic and deontic modality with *can* and *poder* in pragmatic uses of these verbs that have evolved into a conventionalised construction to issue commands in a polite way ("Can you pass me the salt?"). The difficulty involved in these cases lies in that the modality is originally dynamic, but the illocutionary force of a directive speech act (command, request, suggestion…) evokes deontic modality. In order to keep the criterion of annotating the tokens according to the semantic meaning rather than the pragmatic force, the decision was made to annotate them as dynamic.

- Context C. Cases of *can* and *poder* with mental processes ("we can think, one can expect") or by material processes of analysing or classifying ("X can be classified/divided/analysed"), etc. to tentatively introduce their opinion in discourse. These examples have a component of adequacy to reality (dynamic) and another of reasonability (deontic). The agreement was reached to give priority to the second component, so that these cases were annotated as deontic.
- Context D. Cases of unspecified (im)possibility with *can/could* or *poder*. In some examples it is hard to tell if the meaning of (im)possibility stems from social or natural factors, as defined by Perkins; in certain cases, as in (2), language users could well leave the source of the (im)possibility unspecified in order to save face. In these contexts, the annotation will depend ultimately on the social or natural source of the modality. The modality would be deontic in (2), since the possibility was due to social reasons (the money was given by persons or institutions):

(2) La música contemporánea española está de enhorabuena: se mantiene el Festival de Alicante. El INAEM *pudo* encontrar los dinerillos necesarios, aunque no los suficientes para llevar a su Orquesta Nacional, que andaba deprimida por no prestar sus servicios en ningún festival reputado.

"Contemporary Spanish music is to be congratulated: Alicante's festival will still be held. INAEM was able to gather the necessary funds, although not the sufficient funds for hiring its National Orchestra, which was undergoing depression due to the absence of performances in any reputed festival."

Disagreement between epistemic and dynamic modality

-Context E. Possibility in generic statements. The verbs *can/poder* are often used in generic statements, as in (3), to indicate that, whenever certain circumstances occur, there is potential for the event to take place, i.e., that nature does not prevent it from occurring (dynamic modality), and also that for each time that these circumstances occur, there

is a probability for the event to occur (epistemic modality). Consequently, these cases have been considered as instances of merger between dynamic and epistemic modality, so that both labels [EPIST + DYNA] should be chosen.

(3) An inadequate diet, as well as large amounts of sugar, *can* also lead to craving, which then results in some very unpleasant symptoms: nervousness and anxiety palpitations headaches dizziness and fainting weight gain.

- Context F. Indeterminacy between epistemic and dynamic impossibility. The distinction between negative epistemic and dynamic modality of high degree is not always clear-cut. Impossibility is clearly dynamic when it is derived from direct evidence or from uncontested knowledge. However, other cases are not so clear, in that the sp/wr does not seem to be totally certain about the impossibility (4). In order to set up a uniform criterion for all cases of impossibility except the social (deontic), we decided to annotate examples of this kind as dynamic.

(4) No one pursued it. North once told Secord that he had gone so far as to mention to the President that the Ayatollah was helping the contras. This was not quite a true story, he admitted, just a joke to raise Secord's morale. He had made the remark, or something like it, ‘to the back of Admiral Poindexter’ as they came out of a meeting with Reagan, and Reagan *could* not possibly have heard it. But North would doubtless have liked him to.

- Context G. Indeterminacy between epistemic and dynamic possibility in questions with *can* and *could*. These questions, an example of which is (5), may be interpreted as asking about naturally possible reasons (dynamic) or as speculative questions whose answer neither the sp/wr addresser nor the addressee knew (epistemic). In order to unify the criteria, the decision was made to annotate such questions as dynamic.

(5) But Mill considers what, what reasons *could* there possibly be for having this two stage process.

2) Disagreement between epistemic and deontic modality:

- Context H. This disagreement was found sporadically, mainly with the verbs *could/poder* [past], in cases in which it was not clear from the context whether the verb meant uncertainty or else possibility of a social kind. The epistemic and deontic meanings could be considered as neutralised, in the sense that they are overriden by the pragmatic force of a polite suggestion (6). The decision was made to analyse these cases as epistemic.

(6) The Group did not support the informal suggestions put forward by Staff Side Whitley that a) there should be a "Whitley" representative on the Group -- it felt that this function *could* be adequately provided by more frequent, regular and open meetings of the Computer Users Group.

The second experiment was carried out six months later using the same examples as in the first experiment, except that the verbs *may* and *might* were not included and additional examples were introduced of the rest of the verbs, so that the total of 480 remained the same. This was done to focus on the modals which had proved problematic in the first experiment and on which consensus had been reached, as described above. The overall inter-annotator results obtained are summarized in Table 9 below. Tables 10 and 11 specify the agreement level obtained for each of the English and the Spanish modals, respectively.

Table 9: Inter-annotator agreement in the identification of modal meanings (second experiment).

NUMBER OF EXAMPLES ANNOTATED	PROPORTION OF AGREEMENT	KAPPA VALUE
480	0.79	0.66

Table 10: Inter-annotator agreement after six months (English verbs).

MODAL VERB	PROPORTION OF AGREEMENT	KAPPA VALUE

must	0.95	0.89 (+0.04)
can	0.90	0.85 (+0.36)
had to	0.78	0.59 (+0.04)
could	0.70	0.55 (+0.06)
have to	0.75	0.49 (+0.14)

Table 11: Inter-annotator agreement after six months (Spanish verbs).

MODAL VERB	PROPORTION OF AGREEMENT	KAPPA VALUE
deber	0.95	0.88 (+0.19)
poder [present]	0.70	0.57 (+0.29)
poder [past]	0.80	0.54 (+0.07)
tener que [present]	0.62	0.43 (+0.00)
tener que [past]	0.68	0.35 (+0.05)

The general proportion of agreement does not seem to change in a very significant way (there is not much difference between Table 9 and Table 5), but Tables 10 and 11 show that there are variations depending on the individual verbs. It can be seen that the level of agreement increases significantly for *can* and *poder* (present). There is also a modest increase for *deber (de)* and *have to* (present). For the rest of the verbs it ranges from (+0.07) to (+0.04) except for *tener que* (present), which shows exactly the same proportion as in the first experiment. The reasons for this distribution of the differences in the increase of agreement among individual modals between the first and the second experiment are discussed in Section 5.

5. Evaluation of the annotations

The annotation tasks described above are currently being evaluated by the authors of this paper in terms of their impact for theory formation and redefinition in the areas of Thematisation and Modality in English and Spanish.

In the area of Thematisation, the agreement studies reveal interesting aspects of the annotation of the thematic features included in the annotation schemas for English and Spanish. The task of identifying spans proved to be more difficult than the labeling one for annotators, and more for the Spanish dataset than for the English one. This is probably due to the complex morphology of the Spanish verbal group. This complex morphology was also a source of disagreement in the labelling task, which caused annotators to hesitate between thematic features such as PreHead or Thematic Head. However, when reading the definitions again and discussing problems of identifying and segmenting the thematic spans with annotators, consensus was achieved. Lower agreement in the second study concerning the labelling of Thematic Head types both in English and in Spanish points to inherent difficulties in disambiguating different types of processes in both languages, and can only be improved through consistent training and practice with the annotators.

In the area of modality, the annotation tasks revealed that the level of inter-annotator agreement was already acceptable in the first experiment, but that a higher level could be achieved by analysing the cases of disagreement. The study of these cases revealed that the main problem for agreement lies in the difficulty to distinguish between deontic and dynamic modality in certain cases of *can, could / poder* and of *have to / tener que*, and that the difficulty affected the present and past tenses of these verbs. The results of the second experiment show a considerably higher proportion of agreement in the annotation of *can* and *poder* in the present tense. This increase in agreement proves the effectiveness of the annotation decisions proposed for the contexts in which these verbs occur, namely B, C, D, E, F and G. These decisions might well be the basis for the creation of subcategories for the three main types of modality. Some of these subcategories could be: for dynamic modality, “dynamic possibility for directives” (Context B) or “possibility in speculative questions” (Context G); for deontic modality, “reasonable expectation” or “reasonable classification” (both from Context C); for epistemic modality, “weak probability as a polite suggestion” (Context H). Another subcategory would be “generic possibility”, which is to be

annotated as [EPIST + DYN]. These labels could be used as annotation options along with more usual subcategories (which would be applied to non-problematic cases), such as "weak probability", "strong probability" or "certainty" for epistemic modality, "obligation" or "permission" for deontic modality and "ability" or "inevitability" for dynamic modality.

However, the annotation decisions have been less effective in the area of necessity: the description of Context A and its associated annotation decision has to be revised. The improvement in the agreement level observed between the first and the second experiment has also been modest for the past forms of modal expressions, which indicates that a more in-depth analysis of the interaction between tense and modality and its influence on inter-annotator agreement has to be carried out.

From the theoretical point of view, the annotation of these modal auxiliaries in a large number of texts on the basis of these categories and expressions, together with the inclusion of more expressions of all the types of modality, will lead to a broader and deeper characterisation of the types of modality. In this characterisation, the kinds of contexts that create problems for the annotation are to be treated as boundary cases between types of modality.

6. Summary and concluding remarks

In an attempt to illustrate the challenges researchers have to face when confronted with the task of developing well-designed and reliable annotation procedures for complex linguistic phenomena in a contrastive manner, the current paper has focused on a number of annotation tasks in the area of Thematisation and Modality in English and Spanish. In the area of Thematisation, we have described the annotation schemas designed so far, including a core and an extended tagset and how we tested their reliability through two different agreement studies, one focusing on the identification of thematic spans and the other on the labelling of the previously identified spans.

The results of the agreement studies indicate that the annotation of the Spanish dataset proved more challenging than the English one, mainly due to the morphological features of the Spanish verbal group. They also showed that, in general, segmentation of spans (markables) can be more problematic than actual labelling, though in both cases consensus could be reached among annotators after discussing problematic cases.

In the area of Modality we focused on three main types of modality and their realisations through a limited set of modal verbs in both languages, and described two annotation experiments. A first inter-annotator agreement test showed that the modalities that led to disagreement most often were the deontic and the dynamic, especially their realisations by *can, could / poder* and of *have to / tener que*. The most problematic contexts in which these expressions occurred were detected, and a number of decisions were adopted for the annotation of the occurrences in these kinds of contexts. A second experiment was carried out, which proved that the annotation decisions increased agreement levels in problematic uses of *can, could / poder*, especially in the present tense. However, the results were less satisfactory for the area of necessity realised by *have to / tener que*, which indicates that the context-detection and the annotation decisions will have to be revised. These decisions might well be the basis for the creation of new subtypes of modality, an interesting task for theory formation in this complex linguistic area.

As shown by the work reported in this paper, contrastive corpus annotation of complex linguistics categories such as Thematisation and Modality can be used as a tool to validate linguistic theories through the design of reliable annotation schemes which can be tested through agreement studies. Although reaching acceptable levels of agreement is necessary to ensure the quality of the annotated data, the investigation of cases of poor agreement gives the researcher the chance to refine problematic aspects of existing theories and to create new subcategories which were not previously proposed. This is the focus of our current work within the CONTRANOT project. Future work will extend the results of this research towards specific applications such as

the training of machine learning systems on large amounts of annotated data at high enough agreement.

Notes

[1] The CONTRANOT project is financed by the Spanish Ministry of Science and Innovation under the I+D Research Projects Programme (FFI2008-03384). As team leader (Julia Lavid) and members of the research team (Jorge Arús, Marta Carretero, Lara Moratón and Juan Rafael Zamorano), we gratefully acknowledge the financial support provided by the Spanish Ministry of Science and Innovation for the work reported in this paper.

[2] Spanish grammars have traditionally stated that, in correct usage, *deber* + infinitive expresses obligation and necessity (which belongs to deontic and dynamic modality), while *deber de* + infinitive expresses deduction (which belongs to epistemic modality). However, language users tend to ignore this difference and employ the two periphrases interchangeably for both meanings. Given this situation, we have opted for including the two periphrases in our analysis and annotating the actual meanings as they occur in the examples retrieved.

[3] Cf. *Se **me** cayó*, where *se* is part of the Verbal Group *caerse* and is, therefore, pre-Head (it fulfils no participant role)

References

Arús, Jorge, Julia Lavid, and Lara Moratón. 2012. "Annotating Thematic Features in English and Spanish: A Contrastive Corpus-based Study." *Linguistics and the Human Sciences* 6: 173–192.

Carretero, Marta, and Juan Rafael Zamorano-Mansilla. 2010. "Annotating English and Spanish corpora for the categories of epistemic and deontic modality." Paper presented at the 4th International Conference on Modality in English. Madrid, Universidad Complutense, 9-11 September.

Carretero, Marta, and Maite Taboada. In press. "The Annotation of Appraisal: How Attitude and Epistemic Modality Overlap in English and Spanish Consumer Reviews." In *Thinking Modally: English and Contrastive Studies on Modality*, ed. by Juan Rafael Zamorano-

Mansilla, E. Domínguez-Romero, C. Maíz-Arévalo, and M. V. Martín de la Rosa. Bern: Peter Lang.

Coates, Jennifer. 1983. *The Semantics of the Modal Auxiliaries*. London: Croom Helm.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20: 37–46.

Collins, Peter. 2009. *Modals and Quasi-modals in English*. Amsterdam: Rodopi.

Hermerén, Lars. 1978. *On Modality in English: the Study of the Semantics of the Modals*. Lund: Gleerup.

Hovy, Eduard, and Julia Lavid. 2010. "Towards a Science of Corpus Annotation: A New Methodological Challenges for Corpus Linguistics." *International Journal of Translation* 22 (1): 13–36.

Lavid, Julia. 2012. "Corpus Analysis and Annotation in CONTRANOT: Linguistic and Methodological Challenges." In *Encoding the Past, Decoding the Future: Corpora in the 21st Century*, ed. by Isabel Moskowich, and Begoña Crespo, 205–220. Cambridge: Cambridge Scholars.

Lavid, Julia, Jorge Arús, and Juan Rafael Zamorano-Mansilla. 2010. *Systemic Functional Grammar of Spanish: A Contrastive Study with English*. London: Continuum.

Lavid, Julia, Jorge Arús, and Lara Moratón. 2010a. "Towards an Annotated English-Spanish Corpus with SFL-based Textual Features." Paper presented at the 37th International Systemic-Functional Congress. Vancouver, Canada.

Lavid, Julia, Jorge Arús, and Lara Moratón. 2010b. "Investigating Thematic Meaning in English and Spanish: A Methodological Proposal." Paper presented at the 22nd European Systemic-Functional Linguistics Conference and Workshop. University of Primorska, (Koper, Eslovenia). To be published in G. O'Grady et al. (eds.). *Choice in Language: Applications in Text Analysis*. London: Equinox.

McEnery, Anthony, R. Xiao, and Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York: Routledge.

Nuyts, Jan. 2001. *Epistemic Modality, Language and Conceptualisation: A Cognitive-Pragmatic Perspective*. Amsterdam: John Benjamins.

Palmer, Frank R. 1990. *Modality and the English Modals*. London and New York: Longman.

Perkins, Michael R. 1983. *Modal Expressions in English*. London: Frances Pinter.

Reidsma, Dennis, and Jean Carletta. 2008. "Reliability Measurement without Limits." *Computational Linguistics* 34 (3): 319–326.

Silva-Corvalán, Carmen. 1995. "Contextual Conditions for the Interpretation of 'Poder' and 'Deber' in Spanish." In *Modality in Grammar and Discourse*, ed. by Joan Bybee, and Suzanne Fleischman, 67–105. Amsterdam: John Benjamins.

Taboada, Maite and Marta Carretero. In press. "Labelling Evaluative Language in English and Spanish: The Case of Attitude in Consumer Reviews." In *Contrastive Discourse Analysis: Functional and Corpus Perspectives*, ed. by Maite Taboada, Susana Doval, and Elsa González Álvarez. London: Equinox.

Wärnsby, Anna. 2006. *(De)coding Modality. The Case of Must, May, Måste and Kan*. (Lund Studies in English, 113). Lund: Lund University.

Appendix 1:

<u>Definition of Thematic Field</u> (English)

Thematic Field: Initiating clause span of varying length up to and including the first nuclear constituent [FNC] in main clause (in bold in the examples), or one of the following:

- Predicated Theme construction [PT]

- "There" in Existential clauses.

Examples of Thematic field (in bold) ending in [FNC]:

(1) **[FNC:] The cat** is on the mat

(2) **[FNC:] Eating** is vital

(3) **[FNC:] That he refused to do it** worried me

(4) **[FNC:] Of great importance** is the fact that the liver remains unharmed.

(5) **On the table [FNC:] stood** a lamp

(6) **But, surprisingly, before the meeting [FNC:] everybody** was glad to hear the news.

(7) **In my opinion, Real Madrid, [FNC:] their players** have been holding up a banner

Examples of Predicated Theme Construction [PT] and "There" in Existential clauses:

(8) **In fact [PT:] It is love** that makes the world go round

(9) **When I arrived, [THERE:] there** were three people waiting for the bus

English Core Tagset for Theme categories (Declarative clauses)

This core tagset includes a list of possible elements which may appear in the Thematic field in English declarative clauses and their lexicogrammatical (morphosyntactic) realisations. These elements are:

1. Thematic Head
2. Pre-head
3. Predicated Theme
4. 'There' type constructions
5. Textual Theme
6. Interpersonal Theme

Definitions and realisations are provided for each of these elements below:

1. Thematic Head (TH)

The **Thematic Head** is defined as the first nuclear constituent (not circumstantial) element in the clause. This can be a *Participant* or a *Process*. When the Thematic Head is a *Participant*, it can be realised as:

- a Noun Group. (e.g. ***The cat*** is on the mat; ***Peter*** is at home; ***She*** saw him yesterday)

- an Adverb (e.g. ***Tomorrow*** is a holiday)

- a Non-finite Clause (e.g. ***Eating*** is vital ; ***To live*** is to die)

- a Nominal That-Clause (e.g. ***That he refused to do it*** worried me)

- a Thematic Equative (e.g. ***What I want*** is you)

When the Thematic Head is a *Process*, it is realised as a verbal form, preceded by a Pre-Head element, such as for example a Circumstance (e.g. *On the table **stood** a lamp*).

2. Pre-Head (PH)

The **Pre-Head** element is any circumstantial and/or finite element preceding the Thematic Head. This includes the following realisations (in bold face):

- Adverbial Groups (e.g. [PH-Circ:] ***Afterwards*** there will be another meeting)

- Prepositional Phrases (e.g. [PH-Circ:] ***On your right*** you can see the Royal Palace)

- Circumstantial clauses (e.g. [PH-CCL:] ***After dropping her off***, he continued his trip)

- Finite verbal forms, i.e. auxiliaries, preceding the lexical verb: (e.g. [PH-Finite:] ***Should*** *you decide to leave the country, please let me know.* ***Had*** *I known you were so near, I would have flown to meet you*)

3. Predicated Theme

This is a construction that consists of two parts: (1) an initial thematic segment consisting of 'It' + BE followed by the element in Focus; and (2) a rhematic segment realised by a relative-like clause, as shown below:

(e.g.: ***It is you*** *who are to blame*)

4. "There" type construction

E.g.: **There** were three people waiting for the bus

5. Textual Theme (TT)

Textual themes are elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders or correlatives. These include:

- Linkers (paratactic nexus) (e.g. [TT-Link:] ***And*** *don't tell me you didn't know;* ***but*** *let's change the topic*)
- Binders (hypotactic nexus) (e.g. [TT-Bind:] ***However****, the situation now is different;* ***now*** *we needed to promote the event,* ***secondly****, you should go to a doctor)*.
- Correlatives: (not only…but; either…or) (e.g. [TT-Cor:] ***Not only*** *didn't he call but also forgot completely about us;* ***either*** *you're with us* ***or*** *you're against us.)*

6. Interpersonal Theme (IT)

These are elements which express the attitude and the evaluation of the speaker with respect to his/her message. These include:

- Vocatives, i.e., any item used to address (e.g. [IT- Voc:] ***Tom!*** *This is a nice surprise;* ***Sir,*** *could you follow me, please?)*
- Comment Adjuncts (e.g. [IT- Com:] ***Surprisingly*** *he didn't mention anything;* ***understandably,*** *he kept a low profile)*
- Modal Adjuncts (e.g. [IT- Mod:] ***Probably*** *that's the only lesson we learned;* ***Surely*** *you didn't do that!)*

Extended Tagset for English (Thematic Head Types)

The definitions for Participant types are based on Halliday and Matthiessen (2004) IFG and Martin, Matthiessen and Painter (1997) *Working with Functional Grammar*. All examples include the defined participant in thematic position.

1. TH-Actor: the participant doing the deed in a material processes, as in

[TH-Actor:] **Peter** went home

[TH-Actor:] **Mary** received the letter

[TH-Actor:] **John** gave Mary a kiss

2. TH-Goal: the participant impacted by a doing in a material process, as in [TH-Goal:] **Mary** was kissed by Peter, [TH-Goal:] **The letter** was put in the mail or [Goal:] **The bathrooms** are cleaned hourly
3. TH-Beneficiary: the participant benefiting (positively or negatively) from the doing in a material process, as in [TH-Beneficiary:] **Mary** was given a letter, [TH-Beneficiary:] **He** was granted a scholarship or [TH-Beneficiary:] **They** were inflicted a crushing defeat
4. TH-Range (or Scope): the participant that construes the domain over which the process takes places, as in [TH-Range:] **That mountain** is climbed mostly on its northern side, or construes the process itself, either in general or in specific terms, as in [TH-Range:] **Showers** should be taken in the morning.
5. TH-Senser: the participant sensing in a mental process, as in [TH-Senser:] **She** likes ice-cream, [TH-Senser:] **I** can't see the light, [TH-Senser:] **She** knows a lot of stories, [TH-Senser:] **They** prefer to stay.
6. TH-Phenomenon: the participant being sensed in a mental process, as in [TH-Phenomenon:] **he** is hated everywhere, [TH-Phenomenon:] **Deer** can be seen crossing the fields, [TH-Phenomenon:] **That**'s well known by everybody or [TH-Phenomenon:] **That ring** is very much coveted
7. TH-Sayer: the participant saying, telling, stating, informing, asking, threatening, suggesting and so on in a verbal process, as in [TH-Sayer:] **She** never tells the truth, [TH-Sayer:] **They** ordered me to leave or [TH-Sayer:] **She** threatened to kill herself
8. TH-Verbiage: the content of saying in a verbal clause, when expressed as a nominal group, as in [TH-Verbiage:] **That story** has been told many times, [TH-Verbiage:] **Questions** will be asked, or [TH-Verbiage:] **That word** was never uttered by me.
9. TH-Receiver: the addressee of a speech interaction in a verbal process, as in [TH-Receiver:] **I** was told to leave at once, [TH-Receiver:] **The kids** were told a story or [TH-Receiver:] **She** was asked her name.

10. TH-Token: the participant representing the expression, symbol, form, name, function, position or actor in an identifying relational process. Identifying relational processes are reversible and the Token tends to appear in the first position with respect to the Value, as in: [TH-Token:] **Mary** is the best, [TH-Token:] **Green** means "go" or [TH-Token:] **She** played the leading role. The Token is also the participant that tends to go first in possessive and circumstantial identifying relational processes, as [TH-Token:] **They** own the house or [TH-Token:] **Tomorrow** is January the 1st, respectively.
11. TH-Value: the participant representing the content, symbolised thing, meaning, referent, filler, holder of position or role in an identifying relational process. Identifying relational processes are reversible and the Value appears in initial position when the process is reversed, as in [TH-Value:] **The best one** is Mary, [TH-Value:] **"Go"** is symbolised by green or [TH-Value:] **The leading role** was played by her. The Value is also the participant that goes first in possessive and circumstantial identifying relational processes when these are reversed, as in [TH-Value:] **The house** is owned by them or [TH-Value:] **January the 1st** is tomorrow, respectively.
12. TH-Carrier: the participant to which an Attribute is assigned in an attributive relational process, whether intensive, possessive or circumstantial. These relational processes are not easily reversed. Examples: **She** is quite wise in general, [TH-Carrier:] **I** have a guitar or [TH-Carrier:] **The movie** is about a multimillionaire
13. TH-Attribute: what is assigned to the Carrier in an attributive relational process, whether intensive, possessive or circumstantial. As attributive processes are not easily reversed, Attributes are not found in thematic position except in exclamations such as How [TH-Attribute:] **clever** she is!
14. "There": The starting element in an existential process. It is not a participant. Examples: [TH-there:] **There** is a hair in my soup or there are many people here

15. TH-Process: a whole process, whether material, mental, verbal, relational or existential: [TH-Process:] **Gone are** the days when my heart was young.

Appendix 2:

Definition of Thematic Field (Spanish)

Clause-initial material which goes from the beginning of the clause complex up to and including the first nuclear experiential constituent [FNC] in the main clause. The FNC can be realised by either lexical or morphological means. Examples of Thematic field (in bold) ending in [FNC]:

(1) **[FNC:]** ***El gato*** *está en la alfombra*

(2) **[FNC:]** ***Cansadísimo*** *llegué ayer a casa*

(3) **[FNC:]** ***Se*** *está muy bien aquí*

(4) **[FNC:]** ***Se*** *lo di ayer* (*le*-allomorph)

(5) ***Se*** **[FNC:]** ***me*** cayó (*se* is here part of the Verbal Group *caerse* and is, therefore, pre-Head [it fulfils no participant role]).

(6) ***En realidad,*** **[FNC:]** ***corriendo*** no se consigue nada

(7) **[FNC:]** ***que me digas eso*** significa que no me has entendido

(8) ***Teng-*****[FNC:]** ***o*** frío (verbal base is pre-Head; both together, ITF)

(9) **[FNC:]** ***Ten*** cuidado!

(10) ***Las vacaciones,*** **[FNC:]** ***todo el mundo*** sueña con ellas

(11) **[FNC:]** ***Lo que necesitas*** es amor (Thematic Equative)

Spanish Core Tagset for Theme categories (Declarative clauses)

This core tagset includes a list of possible elements which may appear in the Thematic field in Spanish declarative clauses and their lexicogrammatical (morphosyntactic) realisations. These elements are:

1. Thematic Head

2. Pre-head

3. Predicated Theme

4. 'Hay' type constructions

5. Textual Theme

6. Interpersonal Theme

1. Thematic Head (TH)

The Thematic Head is defined as the first nuclear (not circumstantial) element in the main clause. This can be a *Participant* or a *Process*. If the Thematic Head is a *Participant*, it can be encoded through independent lexical and grammatical forms or through verbal prefixes or suffixes. Examples of lexical and grammatical realisations are the following:

- a Noun Group functioning as Subject (e.g. ***El gato*** *está en la alfombra*)

- A Noun Group functioning as Complement followed by its corresponding clitic (e.g. ***A la niña la*** *llamaron Ana*)

- an Adverbial Group functioning as Subject (e.g. ***Mañana*** *será otro día*)

- The **Se** clitic when impersonal (e.g. ***Se*** *está muy bien aquí*), passive (e.g. ***Se*** *venden libros*), reflexive (e.g. ***Se*** *lavaron los pies*), reciprocal (***Se*** *insultaron sin piedad*), or *le*-allomorph (e.g. ***Se*** *lo dio ayer*).[3]

- Non-finite clause (e.g. ***Corriendo y estresándose*** *no se consigue nada*; ***nadar a braza todo el rato*** *me aburre*)

- Nominal clause (e.g. ***Que me digas eso*** *significa que no me has entendido*)

- A 'lo que' nominalised clause (e.g. ***Lo que queremos*** *es saber el título*)

Examples of realisations through verbal prefixes or affixes (clitics) are given below:

- A verbal suffix or inflection indicating the person and number of the *Participant-Subject*. This is a very common kind of TH in Spanish, where the verbal inflection realises the Head, whereas the preceding lexical part of the verb realises the PreHead as in, e.g., *Aprendi-**ó***

*pronto; Est-**oy** cansada, Llegam-**os** enseguida ('[we] arrived quickly'), H-**an** encontrado petróleo ('they've found oil'), Est-**áis** haciendo mucho ruido ('you're making a lot of noise')*

- A verbal affix (or clitic) indicating the person and number of the *Participant-Complement* (e.g. ***La** convenció su mirada*)

When the TH is a *Process*, it is realised as Verbal Group (command)

E.g., ***Ten** cuidado!*

2. Pre-Head (PH)

The Pre-Head is any element preceding the Thematic Head, including, circumstantials, middle *se* and the lexical part of the Process. This includes the following realisations (in bold face):

Circumstantial realisations:

- Adverbial Groups (e.g. [PH-Circ:] ***Mañana** nos vemos*)
- Prepositional Phrases (e.g. [PH-Circ:] ***En tal caso**, será mejor no hacer nada*)
- Circumstantial clauses (e.g. [PH-Circ:] ***Sin mediar palabra**, le dio una bofetada*)

Middle marking (*me, te, se,* etc.) realisations:

- Personal pronouns (not reflexive but morphologically identical to these); (e.g. [PH-middle marking:] ***Se** me cayó, ¿**te** convences?, **Nos** fuimos pronto*)

Lexical part of Process (Verbal Group minus inflectional ending) realisations:

- Predicator minus inflectional ending (e.g. [PH-lexical part:] ***Teng**-o frío*; ***v**-i a María con su novio*
- Finite minus inflectional ending (e.g. [PH-lexical part:] ***H**-e comido demasiado; **Esta**-mos hartos*

3. Predicated Theme (PT)

A construction that consists of two parts: (1) an initial thematic segment consisting of the copular verb followed by the element in Focus; and (2) a rhematic segment realised by a relative-like clause, as shown below:

"***Fue Fermín*** el que me dejó triste"

4. 'Hay' Type construction

This is a type of construction which occurs in existential clauses. It is realised by the "Hay" element and its temporal variants *"había", "hubo", "habrá"* (e.g. ***Había*** *tres chicas esperando en la puerta*) followed by the element which is presented, called the Existent. The starting element in an existential process. It is variable with respect to tense but invariable with respect to person. The final inflection does therefore not serve to track participants, so the whole unit is Theme. Examples:

[TH-existential:] **Hay** un pelo en mi sopa ('there is a hair in my soup')

[TH-existential:]**Hay** mucha gente aquí ('there are many people here')

5. Textual Theme (TT)

Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders or correlatives. These include:

- Linkers (paratactic nexus) (e.g. [TT-Link:] *¿**O** te crees más listo que los demás?; **pero**, bueno, vamos a dejarlo*).
- Binders (hypotactic nexus) (e.g. [TT-Bind:] ***Además,*** *tú no sabes nada de mí;* ***por lo tanto,*** *nos vimos obligados a cerrar*).
- Correlatives: (*no solo…sino que; o…o*) (e.g. [TT-Cor:] ***No solo*** *nos toma por tontos* ***sino que*** *además se cree que somos idiotas*).

6. Interpersonal Theme (IT)

These are elements which express the attitude and the evaluation of the speaker with respect to his/her message. These include:

- Vocatives, i.e., any item used to address (e.g. [IT- Voc:] ***¡Profesor!*** *¿puedo hablar un momento con usted?;* ***tío,*** *esto es la bomba*)
- Comment Adjuncts (e.g. [IT- Com:] ***desgraciadamente*** *no podremos acudir a tu fiesta*)
- Modal Adjuncts (e.g. [IT- Mod:] ***Tal vez*** *esté en su casa;* ***seguramente*** *no lo vio*)

Extended Tagset for Spanish (Thematic Head Types)

As for English, the definitions for Participant types are based on Halliday and Matthiessen (2004) *IFG* and Martin, Matthiessen and Painter (1997) *Working with Functional Grammar*. All examples include the defined participant in thematic position.

1. TH-Actor: the participant doing the deed in a material processes, as in:

[TH-Actor:] **El niño** tiene anginas ('the kid has tonsillitis')

[TH-Actor:] **Los tres** recibieron un premio ('the three of them were given [lit. received] an award')

[TH-Actor:] **La madre** le dio un beso al niño ('the mother gave the kid a kiss')

2. TH-Goal: the participant impacted by a doing in a material process, as in:

[TH-Goal:] **El premio** lo entregó el president de la academia ('the award was delivered by the president of the academy')

[TH-Goal:] **Las paredes** han de pintarse cada dos años ('the walls are to be painted every other year')

[TH-Goal:] **El cadáver** fue encontrado en la orilla del río ('the corpse was found on the river bank')

3. TH-Beneficiary: the participant benefiting (positively or negatively) from the doing in a material process, as in:

[TH-Beneficiary:] **A mí** me dieron tres entradas ('I was given three tickets')

[TH-Beneficiary:] **Me** han concedido una beca ('I've been granted a scholarship')

[TH-Beneficiary:] **A este perro** le han arrancado una oreja ('this dog has had an ear torn off')

4. TH-Range (or Scope): the participant that construes the domain over which the process takes places, as in:

[TH-Range:] **Esa pared** fue escalada por primera vez en 1915 ('that wall was first climbed in 1915')

or construes the process itself, either in general or in specific terms, as in:

[TH-Range:] **Esta música** es tocada en funciones sociales ('this music is played in social events').

5. TH-Senser: the participant sensing in a mental process, as in:

[TH-Senser:] **A mi hermano** le encanta este osito (‘my brother loves this teddy bear’)

[TH-Senser:] **Los daltónicos** no pueden percibir algunos colores (color blind people cannot make out some colors’)

[TH-Senser:] **Mi abuelo** se sabe un montón de historias (‘my grandfather knows a lot of stories’)

[TH-Senser:] **Algunos** prefirieron quedarse (‘some of them preferred to stay’)

6. TH-Phenomenon: the participant being sensed in a mental process, as in:

[TH-Phenomenon:] **Estas galletas** no me gustan mucho (‘I don’t really like these cookies’)

[TH-Phenomenon:] **Este tipo de flor** se puede ver en primavera (‘this kind of flower can be seen in the Springtime’)

[TH-Phenomenon:] **Esto** es sabido de todos (‘this is known by everyone’)

7. TH-Sayer: the participant saying, telling, stating, informing, asking, threatening, suggesting and so on in a verbal process, as in:

[TH-Sayer:] **Los borrachos** siempre dicen la verdad (‘drunken people always tell the truth’)

[TH-Sayer:] **El juez** ordenó evacuar la sala (‘the judge ordered to create the room’)

[TH-Sayer:] **La mujer** amenazó con tirarse a la vía (‘the woman threatened with throwing herself onto the rail track’)

8. TH-Verbiage: the content of saying in a verbal clause, when expressed as a nominal group, as in:

[TH-Verbiage:] **Esa historia** ya me la han contado (‘I’ve already been told that story’)

[TH-Verbiage:] **Eso** no me lo dices otra vez (you’re not telling me that again’)

9. TH-Receiver: the addressee of a speech interaction in a verbal process, as in:

[TH-Receiver:] **Me** dijeron que me olvidara de ello (‘I was told to forget about that’)

[TH-Receiver:] Siempre **nos** cuentas historias (you’re always telling us stories’)

[TH-Receiver:] **Le** preguntaron su nombre (‘she was asked her name’)

10. TH-Token: the participant representing the expression, symbol, form, name, function, position or actor in an identifying relational process. Identifying relational processes are reversible and the Token tends to appear in the first position with respect to the Value, as in:

[TH-Token:] **Juan** es el responsable ('John is the responsible one [the person in charge])'

[TH-Token:] **El color verde** significa "adelante" ('green color means "go")

[TH-Token:] **Mi hermano** interpretó el papel principal. ('my brother played the leading role')

The Token is also the participant that tends to go first in possessive and circumstantial identifying relational processes, as:

[TH-Token:] **Alguien** debe tener el dinero ('someone must have the money')

[TH-Token:] **Mañana** es el primero de enero ('tomorrow is January 1st')

11. TH-Value: the participant representing the content, symbolised thing, meaning, referent, filler, holder of position or role in an identifying relational process. Identifying relational processes are reversible and the Value appears in initial position when the process is reversed, as in:

[TH-Value:] **El responsable** es Juan ('the responsable one [the person in charge] is John')

[TH-Value:] **"Adelante"** se representa mediante el color verde ("go" is represented by means of the green color')

[TH-Value:] **El papel principal** lo interpretó mi hermano ('the leading role was played by my brother')

The Value is also the participant that goes first in possessive and circumstantial identifying relational processes when these are reversed, as in:

[TH-Value:] **El dinero** lo debe tener alguien (lit. 'the mone, someone must have it')

[TH-Value:] **El primero de enero** es mañana ('the 1st of January is tomorrow')

12. TH-Carrier: the participant to which an Attribute is assigned in an attributive relational process, whether intensive, possessive or circumstantial. These relational processes are not easily reversed. Examples:

[TH-Carrier:] **Este niño** es bastante bueno, en general ('this kid is quite good, in general')

[TH-Carrier:] **Mi madre** tiene dos guitarras ('my mother has two guitars')

[TH-Carrier:] **La película** trata de un niño abandonado ('the movie concerns an abandoned kid')

13. TH-Attribute: what is assigned to the Carrier in an attributive relational process, whether intensive, possessive or circumstantial. As attributive processes are not easily reversed, Attributes are not found in thematic position except in exclamations such as: ¡Qué [TH-Attribute:] **inteligente** es! ('How intelligent she is')

In other constructions where the Attribute appears in thematic position is it of the Absolute theme type, and in that case it does not fulfil a transitivity role:

[TH-absolute:] **Inteligente**, no lo es mucho ('intelligent, she's not really')

Appendix 3:

Definition of Modality

The concept of modality chosen for this annotation system, perhaps the most widely used in the literature on English modality (see references in 3.2.), is built around the logical notions of possibility and necessity, and the main modal categories distinguished are the epistemic, deontic and dynamic. The main types of modality share a number of common semantic features, among which we may signal: a) the expression of an attitude, normally that of the speaker/writer at the speech moment, towards the state of affairs communicated; b) modality commonly expresses non-factuality, that is, the utterance is neither true nor false, or the action has not been performed yet at the speech time; c) the modal meanings can be described in terms of scales, along the possibility-necessity axis.

Core Tagset for Modality categories

The examples include of the modal expressions studied in this paper, i.e. a subset of the modal auxiliaries in English, and modal periphrases in Spanish.

1. EPISTEMIC. Concerns degrees of probability, i.e. the speaker/writer's estimation of the chances that a state of affairs has for being or becoming true. Perkins (1983) characterizes it as possibility and necessity derived from rational laws.

English examples:

-COULD:

[EPIST] It seemed undesirable to use force against the Yugoslavs at the moment, but incidents **could** occur or be provoked, and clear instructions should be issued soon as to whether Alexander should order Eighth Army to close the Austrian frontier to the Yugoslavs and eject them from Carinthia, which would of course mean by force.

-MAY:

[EPIST] of one's own past (Fabian 1983). The temporality of the object **may** also contribute to our sense of identity by evoking the past of our own society

[EPIST] While the exact size of this latter group is unclear, Peter Clark has estimated that as many as one fifth of the population of Kent regularly stayed away from church in the later sixteenth century, and the situation **may** well have been worse in the peripheral 'dark corners of the land'

-MIGHT:

[EPIST] an idea. It's no use my telling you what it is -- she **might** not agree, and then it would only be a waste of time. "

[EPIST] Alexander made all the estates of Scotland bind themselves by oath to acknowledge the Maid of Norway as his heir, failing any children Alexander **might** have in the future.

-MUST:

[EPIST] Oh goodo Right, well, better be off then, bye! But like She **must** be really picky if she's doing that!

Spanish examples:

-DEBER (DE):

[EPIST] Visto así, el primer local del mercado, hoy llamado Ramón Castilla, **debe** haber sido muy hermoso.

-PODER:

[EPIST] Enc. - Exacto. Inf. - Es decir que lo reciba allá después. Enc. - No, y creo que **puede** ser interesante para él tomar contacto y hacer parte de un reportaje, etcétera, porque él quiere combinación de reportajes mutuos, no sé.

[EPIST] No concibiendo qué es lo que los marinos ingleses **podían** hacer a tal hora en casa del Vizir, y sobre todo en estado de ebriedad, permanecí petrificado en la plaza.

-TENER QUE:

[EPIST] es una especialidad tan concreta como, como pueda ser un idioma, pues, no sé, **tiene que** ser interesante...

[EPIST] Por lo tanto, si así lo quisiera y sin tener que soportar ningún reclamo de conciencia, podría dar inmediatamente la espalda y alejarme de allí, seguro de que todo **tenía que** continuar tal como está, y sin embargo, extremando una concesión innecesaria, sólo para que no se fuera a decir que tuve demasiada prisa y porque después de todo me daba igual permanecer en esta esquina que en cualquier otro sitio

2. DEONTIC. This modality is characterized by Perkins (1983) in terms of possibility and necessity derived from social or institutional laws. It concerns obligation, recommendation, permission and prohibition.

English examples:

-CAN:

[DEONT] Where a party adducing documentary evidence has access to the original document it should be produced. The court **can** accept a copy if satisfied that the original document has been lost or destroyed.

-COULD:

[DEONT] The appeal was decided on the comparatively short point that although undue influence had been exercised by the husband upon the wife, it had not been shown that the transactions were to the manifest disadvantage of the wife, in which circumstances the court held that the wife **could** not be relieved from the effect of the charges on her home.

-HAVE TO / HAD TO:

[DEONT] They said they'd send the visas to Tam, but what bothers me is how they can do that. The visa **has to** be stamped in your passport, and they can't do that while you've got it.

[DEONT] In the forests of Chippenham and Melksham, Dean, Feckenham, Peak and Windsor the warden had also the custody of royal manors in the forest, and he **had to** see that they were properly stocked and managed.

-MAY:

[DEONT] A solicitor employed by a non-lawyer **may** not carry out professional work for any person other than his employer (ie working directly with the employer's clients is not permitted) but **may** act for a company or other organisation controlled by the employer or over which the employer has substantial control or for a company in the same group as the employer or which controls the employer.

-MIGHT:

[DEONT] In my misery. Sun, my father, moon, my mother, You **might** look at my face Where the tears of blood run down.

[DEONT] I'm all for women's lib. " " **Might** have bloody known. **Might** have bloody known you would be. Bloody typical, if you ask me

-MUST:

[DEONT] The justification for this is that established Government is necessary for the existence of society and therefore its safety against violent overthrow **must** be secured.

Spanish examples:

-DEBER (DE):

[DEONT] No confundas las cosas. El pasado, todas las experiencias cuentan siempre. Y lo que no funciona **debe** obviarse sin otorgarle beneficio a la indecisión.

-PODER:

[DEONT] El señor dijo por fin: - Muchas gracias. **Puede** retirarse. Lo felicito

[DEONT] Muchos eran llevados a América en barcos de otras naciones europeas. La población esclavizada realizaba una gran variedad de actividades; **podían** trabajar como vigilantes, artesanos, pastores, granjeros, porteadores, mineros o sirvientes.

-TENER QUE:

[DEONT] El gobierno Clinton **tiene que** hacer frente a las críticas por el llamado escándalo Whitewater, polémica que cuestiona el papel de Clinton y su mujer en la quiebra de una empresa inmobiliaria en Arkansas.

[DEONT] La tía Julia **tenía que** presentar su partida de nacimiento y la sentencia de divorcio legalizada por los Ministerios de Relaciones Exteriores de Bolivia y del Perú.

3. DYNAMIC. Dynamic modality is described by Perkins (1983) as possibility and necessity derived from natural laws (i.e. those of physics, chemistry, biology, etc.). Dynamic modality includes the meanings of tendency, ability, natural possibility and natural impossibility.

English examples:

-CAN:

[DYNA] There's even a fireplace which **can** be installed on a wall with no flue: a container of special fuel is lit under "coals", and gives you up to three hours of warmth and flames with no mess, no de-ashing, and no chimney is needed.

-COULD:

[DYNA] They took Shanti out in her push chair and amused her as much as they **could**.

-HAVE TO / HAD TO:

[DYNA] Any resistance or reluctance by the scion to take everything, perhaps because it is getting some of what it needs from its own roots, and the stock **has to** start looking for ways to get

rid of the unused energy, and that means making its own top growth, which takes the form of suckers or " briars ".

[DYNA] Esmerelda jumped up and down and told me to hurry up and make the kite fly. I took a last look round, then only **had to** kick the top edge of the kite up a little for it to take the wind and lift.

-MUST:

[DYNA] However, plants designed to flower in December and January, such as Cineraria, Gloxinia and Primulas, **must** be fed and watered regularly as soon as buds begin to develop.

Spanish examples:

-DEBER (DE):

[DYNA] Para que sea eficaz, este equipo protector **debe** ser adecuado y mantenerse en buenas condiciones.

-PODER:

[DYNA] Una vez fijadas las fechas radiométricas de evolución y extinción de un fósil guía, éste **puede** usarse para determinar la edad de cualquier estrato rocoso en que aparezca.

[DYNA] Fui a su casa de Bagur, en la Costa Brava, y le dije que sí, absolutamente, por supuesto. Pero no se **pudo** hacer porque su enfermedad se agravó enseguida.

-TENER QUE:

[DYNA] En primer lugar, yo creo que todo hombre, cuando habla de la mujer, **tiene que** pensar en su madre. Yo creo que ahí... ahí se ter... ahí ccmienza y se termina toda duda al respecto, ¿verdad?

[DYNA] El gran físico y astrónomo toscano desató su fantasía sin darse cuenta del alcance sorprendente de esa vaga intuición acerca del papel que iba a jugar el magnetismo. La Ciencia **tuvo que** madurar otros dos siglos y medio hasta que en 1864 el físico escocés James Clerk Maxwell enunció su teoría sobre las ondas electromagnéticas

4. **Merger epistemic-dynamic** (generic statements with CAN and PODER):

English example:

-[EPIST + DYNA] I like a picture to be fairly full as if it is too empty it **can** become boring to look at after a few months, but it is very easy to get carried away with all your lovely pressed flowers and try to fit them all into one picture.

Spanish example:

-[EPIST + DYNA] Oratorio, composición musical de gran desarrollo para voces e instrumentos, de naturaleza dramática o contemplativa y generalmente sobre un tema religioso. Si bien el libreto **puede** contener incidentes dramáticos, como en la ópera, los oratorios suelen interpretarse en concierto, sin escenarios ni vestuario o atuendo especial.