Annotating thematic features in English and Spanish: a contrastive corpus-based study

Jorge Arús, Julia Lavid and Lara Moratón

Universidad Complutense de Madrid

Abstract

In this paper we present the preliminary results of an empirical study designed to test contrastive features of the category of Theme in English and Spanish through corpus analysis and manual annotation. Using as our theoretical basis the more general features of the model of thematisation proposed in Lavid, Arús and Zamorano (2010), the study describes the different steps of the methodology used, starting with the selection of the corpus used as 'training suite', followed by the design of the annotation scheme, and ending with a discussion of the results of two annotation experiments carried out so far to test the reproducibility of the annotation scheme. It is expected that the work reported in this paper has a theoretical impact on the area of contrastive corpus studies and serves as the basis for the (semi)-automatic annotation of thematic features in larger bilingual corpora.

1 Introduction

The study of the linguistic category of Theme from a cross-linguistic perspective has attracted the attention of a number of scholars in the functional linguistic community, given its relevance not only for clausal analysis (Rose, 2001) but also for its impact on discourse organisation (Lavid, 1998, 2000a, 2000b, to mention a few).[1] However, to date, there are no studies which investigate the potential of human-coded corpus annotation as a tool to test theoretical aspects of this complex linguistic category in a contrastive manner. As part of a larger research effort within the framework of the CONTRANOT project,[2] in this paper we investigate how certain theoretical issues concerning the semantic definition and the structural delimitation of Theme can be tested through contrastive corpus annotation.

Our methodology is based on a number of steps, borrowed from the standard methodologies used in the computational community, but adapting them for the purpose of testing certain theoretical aspects of the category of Theme in English and Spanish. These steps are explained in detail in the following sections, and are summarised in the Methodology section.

The paper is organised as follows. Section 2 discusses the problems which arise when applying the standard definition of Theme to the analysis of the Spanish clause, and describes the solutions adopted in this study, based on the theoretical model of thematisation proposed in Lavid et al. (2010a). Section 3 outlines the methodology followed in our research, specifically describing the corpus selected for the annotation experiments (3.1), the annotation scheme used for testing Theme contrastively in English and Spanish (3.2) and the two annotation experiments designed to test the reproducibility of the annotation scheme, with a discussion of the cases with the poorest agreement among annotators (3.3). Finally, section 4 provides a summary and some concluding remarks, as well as some pointers to the future.

2 Delimiting Theme in English and Spanish

The standard definition of the category of Theme in the Systemic-Functional Linguistics (hence SFL) literature is the following: 'The Theme is the element which serves as the point of departure of the message; it is that which locates and orients the clause within its context' (Halliday and Matthiessen, 2004:64). As stated in Arús (2007, 2010), this definition represents an important improvement over the one available in the first two editions of *IFG* in that it no longer alludes to the concept of 'aboutness', which at a time led to a big deal of misunderstanding about the nature of this textual category, notably outside systemic circles (see also Fawcett, 2007: 24 on this issue).

A basic example of Theme is provided in (1), where the highlighted *Everybody* functions as point of departure of the clause. The Spanish translation of (1), i.e. (2), in turn shows that the standard definition of Theme is in principle applicable to Spanish, *Todo el mundo* ('everybody') being in this case the thematized constituent.

(1) **Everybody** is having fun

(2) ***Todo el mundo*** *se está divirtiendo*

Although examples (1) and (2) seem to suggest that the standard definition of Theme is valid for both English and Spanish, it is easy to run into examples were language-specific problems arise. This is epitomized by realizations such as (3), where a decision has to be made as to whether the Process *llegaron* is the clausal Theme or if this is a clause with unrealized Subject Theme. This is an extremely important issue,

since Spanish, as the rest of Romance languages but French, is considered a pro-drop language (Arnaiz, 1997: 48), which results in clause-initial pronominal Subjects constantly being left out in unmarked processes. Either option, i.e. unrealized Subject as elided Theme or Process as Theme, is problematic in light of the existing characterization of Theme. If the unrealized Subject is the Theme, 'the element which serves as point of departure', as stated in the standard definition, is not actually present, which is in itself quite paradoxical, if not contradictory. On the other hand, if the Process is taken to be the Theme, then the fact that the elided Subject is in some way present in the verbal inflection -- 3rd person plural *-aron* in (3) --seems to suggest that by considering the Process as Theme we are in fact including the Subject in the Theme. This is not really a problem, since Theme is identified as being realized by 'the first group or phrase that has some function in the experiential structure of the clause' (Halliday and Matthiessen, 2004: 66), and the Process is *one* and only one experiential element. However, the question arises of what to do in cases such as (4), where the 3rd person plural inflection, *–an,* is part of the Finite, and the Finite is not supposed to have thematic status by itself. As Halliday and Matthiessen say, referring to English yes/no questions, 'since that [the Finite operator] is not an element in the experiential structure of the clause, the Theme extends over the following Subject' (2004: 76). We could then think that, as the Finite *Han* in (4) is not an experiential element, the Theme should extend over the Predicator *llegado*. However, the following consideration should be taken into account before making a decision: could the presence of the inflection referring to the Subject suffice to consider the Finite as Theme, or should the Theme be extended to the Predicator?

(3) *Llegaron los primeros* ('[they] arrived the first')

(4) *Han llegado los primeros* ('[they] have arrived the first´)

As we can see from the discussion above, establishing the *semantic* delimitation of Theme goes hand in hand with establishing its *structural* delimitation, i.e. its boundaries. One does not need to look at Theme cross-linguistically to run into difficulties concerning the boundaries of Theme. Different systemic authors have provided alternative views on this issue; Berry (1989), for instance, considers the Theme everything that goes before the verb in the clause, whereas Ravelli (1995) speaks of the difficulties to establish the thematic extent. Not even the initial status of Theme

can be taken for granted; Fawcett follows Rose (2001) to claim that 'there is no necessary connection between "coming early in the clause" and "realizing a thematic meaning" -- and it is clear that no linguist -- and especially no functional linguist -- should infer too readily that the set of phenomena that have the characteristic of occurring early in the clause in English have a corresponding generalized meaning at the level of semantics' (Fawcett, 2007: 8).

As discussed elsewhere (e.g. Lavid et al., 2010a; Arús, 2010; Lavid, 2010), concerning the unrealized-Theme vs. Process-as-Theme choice, we opt for the latter, by which we align ourselves with authors such as Taboada (1995) or McCabe and Alonso (2002). An important reason for not considering the unrealized pronoun as Theme, besides the aforementioned paradox it would entail, is the existence in Spanish of examples such as (5), where the Subject *algunos* ('some') follows the verb *Dicen* ('say'), Process ^ Subject being a fairly common sequence in Spanish. This clearly shows that when the Subject is not present there is no reason to assume that if it had been present it would have been in thematic position. It seems sensible, therefore, to assign thematic status to the first realized element, whether Process or participant. When any of these elements is preceded by a Circumstance, the experiential Theme includes both the Circumstance and the first element from nuclear transitivity, as in (6) below, where *Before the meeting* is a Circumstance, and the Theme extends to, and includes *everybody*.

(5) *Dicen algunos* *que es mejor…*

Say some that it is better…

'Some say that it is better…'

(6) **Before the meeting, everybody** was glad to hear the good news

The question remains open, however, as to what to do about the verbal inflection and how to integrate its presence in a theoretically sound model of Theme. As proposed in Lavid et al. (2010a), we have found it necessary to a) create several layers of analysis, and b) break up the Theme.[3] At the most general level of analysis is the *Thematic Field*, which we define as the 'Complex functional zone in clause-initial position serving a variety of clausal and discourse functions' (Lavid et al., 2010a: 299).

This rather general definition allows us to consider, within the Thematic Field, all sort of textual and interpersonal meanings, as well as Circumstances, preceding the first experiential element of nuclear transitivity[4], as well as to defer the identification of the thematic climax, and therefore its end, to a lower stage of analysis.[5] In an unproblematic example such as (6) above, the thematic field would be *Before the meeting, everybody*. In an example starting with textual (*But*) and interpersonal (*surprisingly*) elements, such as (7), the Thematic Field is longer, as shown:

(7)

But, surprisingly, before the meeting everybody	*was glad to hear the news*
Thematic Field	*Rhematic field*

The Thematic Field consists of up to two parts, the *Inner Thematic Field* (ITF), made up by the 'elements from the experiential structure of the clause' (including Circumstances) and an optional *Outer Thematic Field* (OTF), or the 'elements which surround and complete the Inner Thematic Field', i.e. textual and interpersonal ones (Lavid et al., 2010a: 299, 302).[6] The example in (7), above, is reproduced as (8), below, with the specification of ITF and OTF. The inclusion of the Circumstance within the ITF reflects the fact that, although these elements do not belong to nuclear transitivity, they are still part of the experiential structure.

(8)

<table>
<tr><td>But, surprisingly,</td><td>before the meeting everybody</td><td>was glad to hear the news</td></tr>
<tr><td>OTF</td><td>ITF</td><td rowspan="2"></td></tr>
<tr><td colspan="2">Thematic Field</td></tr>
</table>

The OTF consists then of any textual and interpersonal Themes that may precede the ITF. Based on the standard definitions of these two categories in *IFG*, for Lavid et al. (2010a: 209-302) Textual Themes are elements which "are instrumental in the creation of the logical connections in the text, such as linkers, binders, and other textual markers", whereas Interpersonal Themes are "those elements which express the attitude and the evaluation of the speaker with respect to his/her message, including those expressing modality and polarity".

The ITF, in turn, has an obligatory element, the *Thematic Head*, and one or more thematic *PreHeads* which, although not really optional, as we will see in due course, are not always present. The Thematic Head is defined as the 'first element with a function in the experiential configuration of the clause which is more central to the unfolding of the text by allowing the tracking of the discourse participants' (Lavid et al., 2010a: 299). This excludes Circumstances, so in (8) above, the Thematic Head is *everybody*, whereas the Circumstance *before the meeting* is the PreHead, which is 'typically realised by *Circumstantial* elements which do not exhaust the thematic potential of the clause' (Lavid et al., 2010a: 301).[7] The analysis reflecting the internal structures of the OTF and the ITF is as shown in (9):

(9)

<table>
<tr><td>But,</td><td>surprisingly,</td><td>before the meeting</td><td>everybody</td><td>was glad to hear the news</td></tr>
<tr><td>Textual Theme</td><td>Interpersonal Theme</td><td>PreHead</td><td>Thematic Head</td><td rowspan="3">Rhematic Field</td></tr>
<tr><td colspan="2">OTF</td><td colspan="2">ITF</td></tr>
<tr><td colspan="4">Thematic Field</td></tr>
</table>

Going back to Spanish, example (10) illustrates a process with a very simple thematic structure, consisting only of the Thematic Head, whereas (11) shows a process from our corpus of newspaper commentaries with a complex ITF.

(10)

<table>
<tr><td>El profesor</td><td>comenzó la lección</td></tr>
<tr><td>Thematic Head</td><td rowspan="2">Rhematic field</td></tr>
<tr><td>Thematic Field</td></tr>
</table>

'The teacher started the lesson'

(11)

De hecho	*hasta 1954, cuando construyeron la carretera entre Río de Janeiro y Santos,*	*el principal acceso a Paraty*	*era marítimo*

Adjunct	Circumstance	Participant	
PreHead		Head	
Inner Thematic field			Rhematic field

'In fact, until 1954, when the road from Rio de Janeiro to Santos was built, the main access to Paraty was by sea.'

The key elements in our analysis that explain thematic choices in a language like Spanish are the PreHead and the Thematic Head. Example (12) shows a process with a number of specific traits of the Spanish language. The clause starts with the '*se'* marker typical of middle processes. When this clitic appears in thematic position, it is a PreHead, as it is part of the experiential structure (it is part of the middle process *hallarse* ['find oneself')], yet it cannot qualify for Thematic Head because it is not a participant (see Lavid et al., 2010a: 114-18). Although *se* is part of the verb, the first real element from nuclear transitivity to appear in the process is *halla*, but not all of it has the same thematic force. The part which allows the tracking of participants and contributes to the cohesive unfolding of the text is the inflection *–a*, which textually fulfils a similar role to *(s)he* in English. That is why the non-inflectional part of the verb, i.e. *hall-*, is also a PreHead, whereas the inflection is the Thematic Head.

(12)

Se	*hall-*	*-a*	*ahora ante un proceso de expansión internacional*
'*se*' CLITIC	Find	3PSG. PRES. IND.	
PreHead		Head	
Thematic Field			Rhematic Field

'it finds itself in the face of an international expansion process'

Example (12) quite clearly illustrates the advantage of breaking up the standard ideational Theme into PreHead and Thematic Head, as we can now easily account for the thematic structure of the no longer problematic *han llegado los primeros*, which we now analyze in (13), with the inflection *–an* as Thematic Head and the non inflectional part of the verb preceding it, i.e. the form *h-*, as PreHead. We thus solve the problem of the extent of the Theme when a process starts with a Finite ^ Predicator sequence.

Having identified the Thematic Head as the nuclear element in the ITF, the presence of such an element marks the end of the Theme. Finally, in (12) and (13) we can see why the PreHead, although not an obligatory element, is not always really an optional element: when the verb is in thematic position, as in these two examples, the PreHead has to be present, as it is the non-inflectional part of the verb. In fact, the only PreHeads than can be added to, or removed from, a process without substantial semantic change are certain Circumstances.

(13)

H-	*- an*	*llegado los primeros*
PreHead	Thematic Head	
Thematic Field		Rhematic Field

3 Methodology

To carry out the research here presented, we followed a number of steps:

1. Selecting the corpus. This step involves the compilation of those texts to create what is known as the 'training corpus' on which the annotations will be hand-coded both by human experts and by trained annotators.

2. Delimiting the category to be annotated and its features, as described in section 2 above.

3. Designing the annotation scheme and guidelines. This involves instantiating all or part of the features of the selected theoretical model and developing a core and an extended tagset to be used in the process of annotating the training corpus.

4. Performing annotation experiments on some fragment of the training corpus, in order to determine the feasibility both of the instantiation and the annotator manual.

5. Measuring the results of the annotations by comparing the degree of agreement between the annotators' decisions. This step also involves deciding which measures are appropriate, how they should be applied, and determining the level of agreement which will be considered satisfactory. Here our goal is to identify all relevant phenomena, test the theory instantiation, and hence validate the underlying theory, investigating the cases of poor agreement.

3.1 Selecting the corpus

Our training corpus consists of a selection of comparable newspaper texts in English and Spanish belonging to two genres -news reports and commentaries. The selection of these texts as our training corpus is motivated by two main factors: 1) their online availability (Hauser, 2001: 291); 2) the fact that they were subject to a previous contrastive analysis (Lavid, Arús and Moratón, 2010) using categories from the model of thematisation that we want to test in this study. Both reports and commentaries were extracted from online press editions (British press for English, press from Spain for Spanish). The total number of clauses and words are specified in tables 1 and 2:

Table 1 The English corpus

ENGLISH	**# CLAUSES**	**# WORDS**
REPORTS	325	8743
COMMENTARIES	576	13583
TOTAL	**901**	**22326**

Table 2 The Spanish corpus

SPANISH	**# CLAUSES**	**# WORDS**
REPORTS	111	3338
COMMENTARIES	111	3453
TOTAL	**222**	**6791**

The research presented in Lavid et al., (2010b) shows that these two genres do show different thematic features in their textual unfolding. These differences concern a) the experiential roles selected as Thematic Head -- e.g. much higher presence of Sayer as Theme in news reports vs. relative preference for Carrier as Head in news commentaries; b) the semantic nature of the nominals realizing the Thematic Head characterizing each genre -- tendency for concrete nouns referring to individuals, group of people or institutions in news reports vs. higher frequency of abstract nouns in news commentaries; and c) the internal structure of the Nominal Groups as Thematic Heads -- more complex in commentaries than in reports (see Lavid et al., 2010b: 90 for

interpretations of these contrasts). In the present work, however, our main concern is not the thematic contrasts between these two newspaper genres but, as established in the introduction, the delimitation of thematic categories for later validation.

3.2 Designing the annotation scheme

Our next step in this study was to design the annotation scheme and guidelines which would allow us to test the theoretical model of thematisation, partially described in section 1 above.[8] An annotation scheme is, according to Leech "the document describing and explaining the scheme of analysis employed for the annotations' (Leech, 1997: 6). The main challenges in the creation of our annotation scheme were: a) the identification and definition of the thematic categories to include as part of the core annotation scheme; and b), the creation of an extended annotation scheme that would help annotators overcome the difficulties posed by the shortcomings of the core annotation scheme.

Tables 3 and 4 show the core tagsets for English and Spanish, respectively, based on the theoretical model outlined in section 1. It may seem *prima facie* surprising that, after all our efforts to find and define thematic categories that would cover the notion of Theme cross-linguistically, we finally decided to create a different tagset for each language. The reason is that the annotation scheme is going to be the reference for the annotators and, once each thematic category had been identified and defined, we found it would be more useful for the annotators to be provided with additional information, notably concerning specific realizations. Once realizations come into the fore, language-specific tagsets are unavoidable. Thus, in table 3 we can see a reference to existential *there* in English as Thematic Head, whereas the tagset in table 4 refers to the Spanish clitic *se*.

Table 3 Core Tagset for Theme types in English

Annotation Layer	Thematic field	Description
Unit		Main clause
Core annotation scheme		
Tags:	TH (Thematic Head)	First nuclear constituent (Participant or Process, not Circumstantial) in main clause,

		or “There” in Existential clauses.
	PreHead	Any Circumstantial element and/or Finite element preceding the Thematic Head.
	Textual Theme	Elements which are instrumental in the creation of the logical connections in the text, such as linkers, binders and other textual markers.
	Interpersonal Theme	Elements which express the attitude and the evaluation of the speaker with respect to his/her message, such as Vocatives and Modal Adjuncts, including mood and comment adjuncts

Table 4 Core Tagset for Theme Types in Spanish

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

On the basis of these core tagsets, we designed two annotation experiments to test the feasibility of the proposed annotation schemes. Given the intrinsic complexity of tagging semantic categories -- as opposed, for instance, to POS (Part-Of-Speech) tagging, we did not expect the core annotation schemes for Theme to be final. The annotation experiments were conceived as a probe to find out what was lacking in the core schemes with a view to the design of the extended schemes.

3.3 Annotation experiments

In order to test the annotation scheme, two experiments were conducted. The first one used three human experts with substantial amount of training (namely two linguists and one PhD student in linguistics) who tagged thirty-four clauses from a randomly-chosen English text from our corpus, and twenty clauses from an English text, also randomly selected, for all the categories in the core annotation scheme, i.e. Thematic Head, PreHead, Textual Theme and Interpersonal Theme.[9] The main purpose of this experiment was to measure the reproducibility of our annotation scheme and, eventually, its stability by having the same annotators tag the same texts. The annotators were given the texts already divided into clause complexes, as shown in table 5, below. In our research on Theme, we decided to start by looking at this textual category only in the clause complex, to make things easier. Once the thematic categories under scrutiny are stable enough, we will move to the (simple) clause, the 'central processing unit in the lexicogrammar' (Halliday and Matthiessen, 2004: 10). Table 5 shows an extract from the English text annotated in this study. The clauses shown are already tagged by the three annotators for PreHead -- in the experiment, of course, none of the annotators could see the others' responses; what we see in table 5 is a compilation of the three annotators' responses. The annotators were asked to select and to insert in the column under PreHead the chunk of text they thought would correspond to such a thematic element.

Table 5 Annotation exercise on PreHead in English

#	**TEXT**	**Annotator 1**	**Annotator 2**	**Annotator 3**
		PreH	PreH	PreH

1	**The most immediate result of the arrest warrant issued for Sudanese President Omar Hassan al-Bashir by the International Criminal Court last month was the expulsion of most aid agencies from the country.**			
2	**But this global focus on Sudan's Darfur region, though justified, has overshadowed an even more vital issue: sustaining the quest for a broader peace in all of Sudan.**			
3	**What is most needed now is to build an international consensus on a strategy to implement fully the 2005 Comprehensive Peace Agreement (CPA) for Sudan.**			What i-
4	**The CPA ended Africa's longest civil war, which had left behind over two million dead.**			
5	**That agreement not only contains benchmarks that should lead to self-determination for Sudan's South;**			
6	**it also spells out a democratization process in Sudan itself.**			
7	**After all, the oppressive nature of the regime in Khartoum is at the root of the many conflicts that have torn the country apart.**			
8	If the government in Khartoum persists in undermining the reform process and derailing the referendum on self-determination promised for the South in January 2011, **a return to full-scale civil war, with calamitous consequences for the peoples of Sudan and the entire region, is a real possibility.**	If the government in Khartoum … January 2011	If the government in Khartoum … January 2011	If the government in Khartoum … January 2011

After the annotation exercise, agreements and disagreements were quantified, as shown in table 6, where the figures indicate the number of clauses on which the annotators agree, and then statistically measured to find out about the significance, or not, of the agreement among the three experts. For this, we used *Krippendorff's alpha***,** 'a reliability coefficient developed to measure the agreement between observers, coders, judges, raters, or measuring instruments, [which] emerged in the field of content analysis but is widely applicable wherever two or more methods of processing data are applied to the same set of objects, units of analysis, or items and the question is how much they agree' (Krippendorff, 2007: 1).

Table 6 Annotator agreement for thematic categories (experts)

	English (34 clauses)		
Category	**3 annotators agree**	**2 annotators agree**	**No agreement**
Thematic Head	29	4	1
PreHead	30	3	1
Textual Theme	27	7	0

	Spanish (20 clauses)		
Thematic Head	15	4	1
PreHead	16	3	1
Interpersonal Theme	19	1	0
Textual Theme	20	0	0

A look at table 6 reveals that total agreement was the norm, although never, except for Textual Theme in Spanish, without flaw. In fact, some categories show high degrees of unreliability, mostly if we consider that for such a small number of sentences there are occasions when the three annotators disagreed. Krippendorff's alfa confirms this: in a scale from 0 (least reliable agreement) to 1 (most reliable agreement), all categories, except for Textual Theme scored closer to 0 than to 1. This pointed to a need for reworking the definitions in our annotation scheme, but, before doing so, we wanted a) a corroboration of the results obtained in this experiment, and b) to answer the question of whether the significant amount of training effort used in the first experiment could be reduced, i.e. whether the use of the annotation scheme as a guide in the annotation process makes linguistic training unnecessary.

With this twofold purpose, we conducted a second experiment. On this occasion, the annotators were 18 Spanish students of English Linguistics with no prior annotation training and provided only with minimal instruction on Theme. The texts on which they worked contained a roughly similar number of clauses to those in the experiment with experts, and the annotation activities were the same as before. Table 7 shows the results of their annotations. To make the table reader-friendly, we have grouped the results into 6 columns, ranging from total agreement ('all agree') to no agreement ('none agree'), with partial agreements expressed by the in-between columns. These indicate agreement among two to five students, six to nine, ten to thirteen and fourteen to seventeen. As in table 6, above, the figures in the cells indicate the number of sentences on which annotators (i.e. students) agree.

Table 7 Annotator agreement for thematic categories (students)

	English (33 clauses)					
Category	**all agree**	**14-17**	**10-13**	**6-9**	**2-5**	**None agree**
Thematic Head	6	20	4	3	0	0
PreHead	18	6	9	0	0	0
Interpersonal Theme	13	10	10	0	0	0
Theme Textual	4	9	16	4	0	0
	Spanish (16 clauses)					
Thematic Head	2	8	3	3	0	0
PreHead	3	6	5	2	0	0
Interpersonal Theme	8	6	0	2	0	0
Textual Theme	1	1	8	6	0	0

The range of agreement is much more varied here than in table 7, above, given the substantially higher number of annotators, and their lack of thematic training resulted in lower reliability coefficients for the tested categories: Note that only in one of the categories, i.e. English PreHead (18/33), was there total agreement in more than 50% of the clauses, and in all but two categories, i.e. English PreHead and Interpersonal Theme, we find clauses on which only ≤ 50% of students agree, notably 6/16 clauses in Spanish Textual Theme. Also note that it makes sense that we found no case of 2-5 or no agreement, because the limited number of participants in each clause virtually precludes the combinatory possibilities necessary for such disagreement among 18 students.

An interesting fact was that several students seemed to confuse PreHead and Textual Theme, as illustrated by (14), where Textual Theme *but* was repeatedly analyzed as PreHead, and, conversely, by (15), where PreHead *If…2011* was on several occasions chosen as Textual Theme. The same phenomenon was observed in Spanish.

This could very well be pointing to a terminological overload, as the Textual Theme, the same as the Interpersonal Theme, occurs in Pre-ITF position.

(14) **But** this global focus on Sudan's Darfur region, though justified, has overshadowed an even more vital issue: sustaining the quest for a broader peace in all of Sudan**.**

(15) **If the government in Khartoum persists in undermining the reform process and derailing the referendum on self-determination promised for the South in January 2011,** a return to full-scale civil war, with calamitous consequences for the peoples of Sudan and the entire region, is a real possibility**.**

Another interesting finding was that students occasionally -- and this seems to be a problem derived from the lack of training -- were at a loss concerning where they should be looking for the thematic elements. Such was the case with (16), where several students had problems to discern *For it should be pointed out* as the main clause, with textual Theme *For* and Thematic Head *it*.

(16) For it should be pointed out that when the great earthquake occurred, the city's name was not La Antigua Guatemala but Santiago de los Caballeros de Goathemala, graced with the title of 'Very Noble and Very Loyal City', it was one of the oldest and richest metropolises in Spanish America, and as the capital of a Captaincy General, it was the most important city between Lima and Mexico City.

The findings just mentioned have allowed us to draw some preliminary conclusions concerning not only the design of the annotation scheme but also the design of the activities carried out to test the scheme. The task of annotating examples taken directly from the corpus may be too hard for non-trained annotators, which means that at the beginning this task should be simplified as much as possible, e.g. by highlighting the main clause or even by using contrived or selected examples and, once the thematic categories seem reproducible and stable, moving on to the corpus as it is. The terminological overload evinced by the students' confusion between PreHead and OTF makes us think that future experiments with students should avoid testing all the categories in the same session. Devoting one session to testing Thematic Head and

PreHeaD and another one to textual and interpersonal Theme, i.e. the OTF, will most likely improve the rate of agreement.

In any case, it was clear, from the experiment with expert annotators, that the categories in the annotation scheme needed redefining or extending. In that light, we created an extended annotation scheme where the thematic categories contained not only a definition but also a specification of the possible realizations for each one of them in English and in Spanish. Table 8 below shows the extended tagset for Thematic Head with illustrative examples.

Table 8 Extended tagset for Thematic Head

Thematic Head:

(English)

a) First nuclear constituent (participant or Process) in main clause.

Realizations:

Participant:

Noun Group

E.g. ***the cat*** *is on the mat;* ***Peter*** *is at home;* ***she*** *saw him yesterday*

Clause (non-finite)

E.g. ***Eating*** *is vital ;* ***to live*** *is to die*

Clause (*that*-)

E.g. ***that he refused to do it*** *worried me*

Process:

Verbal Group (command)

E.g. ***Eat*** *your soup!*

Verbal Group (finite, preceded by circumstantial pre-Head)

E.g. *On the table* ***stood*** *a lamp*

b) *There* in existential clauses.

Realization:

There

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

(Spanish)

First nuclear element (participant or process) in main clause, realized by either lexical or morphological means.

Realizations:

Participant:

Noun Group

E.g. ***El gato*** *está en la alfombra;* ***Pedro*** *está en casa;* ***yo*** *sí la ví;* ***Le*** *pareció poco apropiado*

Se clitic (impersonal, passive, reflexive, reciprocal, *le*-allomorph)

E.g. ***Se*** *está muy bien aquí,* ***Se*** *venden libros*; ***Se*** *insultaron sin piedad;* ***se*** *lo di ayer*

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

Clause (non-finite)

E.g. ***Corriendo*** *no se consigue nada ;* ***nada****r me aburre*

Clause (*que*)

E.g. ***que me digas eso*** *significa que no me has entendido*

Verbal inflection (verbal base is pre-Head; both together, ITF)

E.g. *Teng-****o*** *frío; ayer v-****i*** *a María*

Process:

Verbal Group (command)

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

Once new experiments are performed and the annotation scheme proves to be reliable, we will embark on an annotation campaign by members of the research group using the tags of that scheme. The annotation campaign will use coding software, i.e. two coding tools being presently tested so as to see which lends itself better to the annotation of such an abstract concept as Theme -- as well as the other linguistic categories covered by CONTRANOT: the UAM's Corpus Tool (O'Donnell, 2010), a potentially very useful tool for the kind of annotation carried out by our group, as it is specifically designed to support SFL-based annotation; and the UCAT Coding Tool (University of

Pittsburg, 2010), which offers powerful resources for automatic annotation and data analysis.

Although the annotation campaign will take place once the annotation schemes are reliable, this does not mean that annotating the texts will be an easy task, yet the reproducibility of the scheme and the linguistic experience of the annotators should yield satisfactory results. Even less easy will be an eventual further step which currently falls beyond the scope of CONTRANOT but which has to be contemplated as a natural follow-up to the work being done: the development of a semi-automatic annotation program for Theme, i.e. annotation automatically done by a tagging program, which is then manually edited by experts. At this point it is not realistic to think of fully automatic tagging for Theme systems. As Matthiessen points out 'probably for the majority of [linguistic] systems -- it is not yet possible to carry out automatic analysis: computational analysis tools cannot yet cope with the combination of rich analysis and a flow of registerially unrestricted text' (2006: 141). Even semi-automatic annotation should be expected to require a good deal of editing, given the complexity of this category.

4 Summary and concluding remarks

In this paper we have explored how certain theoretical issues concerning the semantic definition and the structural delimitation of Theme can be tested through contrastive corpus annotation. Using as our theoretical basis the model of thematisation proposed in Lavid et al. (2010a), we have described a methodology to test the proposed categories empirically. This methodology, borrowed from the standard ones used in the computational community, consists of a number of steps, starting with the selection of the corpus used as 'training suite', followed by the design of the annotation scheme, with a specification of the core and the extended tagset, and the two annotation experiments carried out so far to test the reproducibility of the annotation scheme.

As pending tasks, we have referred to the need for further tests and evaluation of the extended annotation scheme for a final tune-up, all this with a view to performing an annotation campaign by members of the CONTRANOT project. This campaign has been presented as the final stage of our ongoing project but by no means as the end of our work on annotation for Theme, as future projects should see us engaging in: a) the

expansion of the annotation campaign to other genres; and b) the development of (semi)automatic annotation systems for Theme.

Before embarking on the development of computerized annotation systems, we need to adapt the linguistic descriptions in our annotation scheme to a machine-readable format that allows (semi)automatic tagging. As said above, there is a long way between the creation of a stable and reproducible annotation scheme and the development of a semi-automatic -let alone automatic- annotation system. The challenge is up for grabs and we expect to rise to it sooner than later.

References

Arnaiz, A. R. (1997) An overview of the main word order characteristics of Romance. In A. Siewierska (ed.) *Constituent order in the languages of Europe*. 47--73. Berlin: Mouton de Gruyter.

Arús, J. (2010) On Theme in English and Spanish: a comparative study. In E. Swain (ed.) *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses*. 23--48. Trieste: EUT.

Arús, J. (2007) On the aboutness of Theme. In M. Losada, P. Ron, S. Hernández and J. Casanova (eds.) *Proceedings of the 30th International AEDEAN Conference* (CD-ROM).

Berry, M. (1989) Thematic Options and Success in Writing. In C. Butler, R. Cardwell and J. Cardwell (eds.) *Language and Literature: Theory and Practice. A Tribute to Walter Grauberg*. 62--80. Nottingham: University of Nottingham.

Fawcett, R. (2007) The many types of "Theme" in English: their semantic systems and their functional syntax. Retrieved on 10 June 2010 from http://www.cardiff.ac.uk/chri/researchpapers/humanities/papers1-10/4Fawcett.pdf

Halliday, M.A.K., and C.M.I.M. Matthiessen (2004) *Introduction to Functional Grammar*. London: Arnold.

Hausser, R. (2001) *Foundations of Computational Linguistics*. Berlin: Springer.

Krippendorff, K. (2007) Computing Krippendorff's Alpha-Reliability. Retrieved on 21 March 2010 from www.asc.upenn.edu/usr/krippendorff/webreliability.doc

Lavid, J. (2010) Contrasting choices in clause-initial position in English and Spanish: a corpus-based analysis. In E. Swain (ed.) *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses*. 49--68. Trieste: EUT.

Lavid, J. (2000a). Contextual constraints on thematisation in written discourse: an empirical study. In P. Bonzon, M. Cavalcanti and R. Nossum (eds.). *Formal Aspects of Context*. 37--47. Dordrecht/Boston/London: Kluwer Academic Publishers.

Lavid, J. (2000b). Text types, chaining strategies and Theme in a multilingual corpus: a cross-linguistic comparison for text generation. In J. Bregazzi, A. Downing, D. López and J. Neff (eds.). *Estudios de Filología Inglesa: Homenaje a Jack White.* 107--121. Madrid: Editorial Complutense.

Lavid, J. (1998). The relevance of corpus-based research for contrastive linguistics and computational studies: thematisation as an example. In M.T. Turell and E. Vallduví (eds.). *IV i V Jornades de corpus lingüistics (1996-1997): els corpus en la recerca semàntica i pragmàtica*. 117--40. Barcelona: Publicaciones del Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra.

Lavid, J., J. Arús and J.R Zamorano (2010 [Lavid et al., 2010a]) *Systemic-Functional Grammar of Spanish: a Contrastive Account with English*. London: Continuum.

Lavid, J., J. Arús and L. Moratón (2010 [Lavid et al., 2010b]) Signalling genre through Theme: the case of news reports and commentaries. In L-M. Ho-Dac, ed. *Proceedings of the 8th MAD: Signalling Text Organisation*. 82-92. Moissac (France): University of Toulousse. Available at http://w3.workshop-mad2010.univ-tlse2.fr/MAD_files/papers/LavidArusMoraton.pdf

Leech, Geoffrey (1997) Introducing corpus annotation. In R. Garside, G. Leech and A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. 1--19. London: Longman.

Matthiessen, C.M.I.M. (1995) *Lexicogrammatical Cartography: English Systems*. Tokyo: International Language Science Publishers.

Matthiessen, Christian (2006) *Frequency Profiles of some Basic Grammar Systems*. In G. Thomson and S. Hunston (eds.) *System and Corpus: Exploring Connections*. 103--42. London: Equinox.

McCabe, A. and I. Alonso (2001) Theme, transitivity and cognitive representation in Spanish and English written texts. In *CLAC* 7/2001. Retrieved on 10 February 2009 from www.ucm.es/info/circulo/no7/mccabe.htm

O'Donnell, M. (2010) *UAM Corpus Tool*. Available at http://www.wagsoft.com/CorpusTool/

Ravelli, L. J. (1995) A dynamic perspective: implications for metafunctional interaction and an understanding of Theme. In R. Hasan and P. H. Fries (eds.) *On Subject and Theme*. 187--234. Amsterdam and Philadelphia: Benjamins.

Rose, D. (2001) Some variation in Theme across languages. *Functions of Language* 8.1: 109--45.

Taboada, M. (1995) *Theme Markedness in English and Spanish: A Systemic-Functional Approach*. Retrieved on 24 September 2010 from www.sfu.ca/~mtaboada/docs/taboada-theme-markedness.pdf

University of Pittsburg (2010) *UCAT Coding Tool*. Available at http://cat.ucsur.pitt.edu/

[1] We would like to thank the anonymous reviewers of this paper for their insightful comments on an earlier draft. Any remaining mistakes are the authors' own.

2 The CONTRANOT project is financed by the Spanish Ministry of Science and Innovation under the I+D Research Projects Programme (reference number FFI2008-03384). Julia Lavid, as team leader, Jorge Arús, as member of the research team, and Lara Moratón, as doctoral student, gratefully acknowledge the support provided by Spanish Ministry and also the BSCH-UCM grant awarded for the work reported in this paper.

3 In our description, we propose using the model of thematisation put forward for Spanish by Lavid, Arús and Zamorano (2010), since it offers a wide range of options to cover most of the features displayed by the phenomenon of thematisation both in English and in Spanish. Note, however, that not all the thematic features described in Lavid et al. (2010) are tested in the work reported in this paper. Our aim is to test some general features first, and then proceed with more specific ones.

4 Following Matthiessen (1995), we take nuclear transitivity as concerning processes and participants and circumstantial transitivity as concerning circumstances.

5 Once all the categories of analysis are defined, the actual thematic annotation will proceed in the opposite direction, as is to be expected, i.e. from the most specific to the most general units.

6 *Surround* is not to be taken literally here, as the OTF actually precedes the ITF.

7 Pending further research, we are for the time being treating all Circumstances as not exhausting the Thematic potential. We are aware, however, that some Circumstances seem to have a more relevant status in the unfolding of text, such as, for instance, with such attitude in (i) with such attitude, you'll achieve nothing.

8 For a complete description of the model upon which this study is based see chapter 5 of Lavid et al. (2010a).

9 Interpersonal Theme was tagged only on the Spanish text due to an error in the administration of the experiment. This could not be redressed, as we found out about it later, when the annotators had already been exposed to the improved annotation scheme, which would have invalidated the results.