This is the original version of: Lavid, Julia, et al. «Investigating thematic choices in two newspaper genres : A methodological proposal».
***Choice in language : applications in text analysis*, edited by Gerard O'Grady et al., Equinox Publishing, 2013, pp. 187-214.**
The published version is available at: https://www.equinoxpub.com/home/view-chapter/?id=21928 https://www.equinoxpub.com/home/choice-language/
https://utppublishing.com/doi/10.3138/9781908049551.007 https://utppublishing.com/doi/book/10.3138/9781908049551

Investigating thematic choices in two newspaper genres: A methodological proposal

Julia Lavid, Jorge Arús and Lara Moratón
Universidad Complutense de Madrid

Abstract

In this paper we describe a methodology for the investigation of thematic choices in two newspaper genres (news reports and commentaries) based on a number of corpus analysis and annotation tasks, as currently carried out within the CONTRANOT project.[i] Using categories from a recent SFL-based model of thematisation (Lavid, Arús and Zamorano 2010), we performed a preliminary corpus analysis on a training corpus to extract the relevant thematic features which characterize these two genres. These were used to design an annotation scheme, with a core and an extended tagset, whose reliability was tested through two agreement studies. The final phase of the study outlines the main steps for the semi-automatic annotation of the thematic choices in these two genres with the aim of creating an annotated corpus available to the SFL community.

1. Introduction

Most current corpus-based work in SFL uses an analysis methodology based on the manual extraction of linguistic features from authentic texts, usually on the basis of some pre-established theoretical categories. However, as pointed out by Elke Teich, "the creation, exploration and sharing of SFL descriptions is impeded because of inadequate tools, lack of accountability and diverging terminology" (Teich 2007). Indeed, most SFL corpus analysis work has been produced using inadequate computational tools (e.g. word processors) resulting in formats which are not amenable to computational treatment. The analysis scheme and the procedure are often not included, the terminology used may vary, may not be complete or may be unknown (*ibídem*). The linguistic descriptions resulting from such analysis methodology are often neither accountable nor reproducible by other researchers and impede the sharing of resources within the SFL community.

While this situation was difficult to overcome in the pre-computing age, this is clearly not the case in the current state of linguistic knowledge, where the need for and the importance of computation for better practice is evident in different areas of linguistic enquiry, from language documentation to language description and theory building. This is even more evident in linguistic work centered around corpora where methodological refinement is essential in data analysis.

In this paper we investigate the linguistic phenomenon of thematisation in two newspaper genres using an empirical methodology which combines a preliminary corpus analysis with a number of corpus annotation tasks, as proposed and developed within the CONTRANOT project. Our final aim in this project, part of which is the work reported in this paper, is to contribute to the task of creating, exploring and sharing linguistic resources in English and other languages, as currently done within the SFL community (see Teich 2008). Our most immediate goal in this paper is to illustrate the methodology used in our project with a study of the thematic choices which characterise news reports and commentaries, thus advancing and extending previous corpus research on the correlations between thematic choice and genre in English and other languages (Eiler, 1986; Francis 1989, 1990, 1992; Fries and Francis 1992; Ghadessy 1995; Nwogu and Bloor 1991; Lavid 1998, 2000a, 2000b, 2010; Lavid, Arús and Moratón 2010, forthcoming; *inter alia*).

The paper is structured as follows: section 2 outlines the methodological steps that we propose for investigating thematic choices in news reports and in commentaries; section 3 reports on the results of the preliminary corpus analysis performed on the initial training corpus and section 4 describes the core and the extended tagset of the annotation scheme used in this study. Section 5 explains the annotation procedure used and describes the two agreement studies performed and their results. On the basis of these agreement studies, we describe the procedure for the semi-automatic annotation of the investigated thematic choices in section 6. Finally, section 7 provides a summary of the work reported in this paper and some concluding remarks.

2. Methodological proposal

Our methodological proposal consists of a number of tasks, some of which are outlined in this section, while others are explained in detail in sections 3, 4 and 5. We begin with the first task, the compilation of the training corpus.

1. *Compiling the training corpus*. Our first task in this study was to compile an initial corpus of newspaper texts to create what is known as the 'training corpus', i.e., the data set on which the annotations would be coded. The training corpus for the current study consisted of a total of 901

clause complexes (895 declaratives, 2 interrogatives and 4 imperatives) belonging to two groups of texts collected from published sources in 2008 and 2009. A first group comprised seventeen newspaper commentaries extracted from the Project Syndicate (http://www.project-syndicate.org). The second group was made up by sixteen news reports from the news section of *The Times* online (http://www.timesonline.co.uk). The motivation for compiling this corpus rested mainly in the electronic availability both in English and in other languages of newspaper texts and in our current interest in journalistic discourse, more specifically in newspaper genres. Within newspaper genres, news reports and commentaries offered an interesting contrast in terms of their communicative purposes which made them a good data source for investigating thematic choice.

2. *Instantiating the theory*. The next step was the definition and delimitation of the theoretical categories that would be tested in the corpus annotation task. As explained in Hovy and Lavid (2010):

> instantiating the theory encounters the problem that no theory is ever complete, and few if any are developed to an equal degree for all variants of the phenomena they address. Since theories tend to focus on some phenomena over others, uncertainty arises about exactly which categories to define as tags for annotation, how to define them exactly, and what to do with the residue not covered by the theory.

In order to decide which categories would be used for the annotation of our corpus, we first selected four coarse-grained categories from the recent SL-based model of thematisation developed by Lavid, Arús and Zamorano (2010), and performed a preliminary corpus analysis on the training corpus, adding some conflations and realisations of one the categories (the Thematic Head). This analysis allowed us to generate hypotheses about the behaviour of thematisation in the two genres under study, as described in section 3 below.

3. *Designing the annotation scheme and guidelines*. This task involved 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. The process is a complex one which requires step-wise refinements and modifications during the whole annotation process. On the basis of the preliminary corpus analysis we selected certain thematic features of the two genres under study and created a simplified annotation scheme which was modified and perfected during the annotation process. The annotation scheme is described in section 4.

4. *Performing agreement studies*. In order to test the reliability of the core and the extended tagsets of the annotation scheme, we performed two agreement studies on some fragment of the training corpus and evaluated the results in order to determine the reproducibility of the annotation schemes. The results of these agreement studies are presented in section 5.

5. *Annotating a larger corpus* using tags from a reliable annotation scheme. Once the annotation scheme was validated through the agreement studies, we proceeded with the semi-automatic annotation of the investigated thematic choices using the UAM corpus tool. This final phase is described in section 6 below.

3. Preliminary corpus analysis

As explained in section 2 above, we performed a preliminary corpus analysis on the training corpus. Although this analysis is not carried out in the NLP community, and it may be considered unnecessary for some computational applications, we think that it is a useful preliminary step for deciding which features to include in the core and the extended tagsets of the annotation scheme and to generate hypotheses about the behaviour of a given linguistic phenomenon in specific contexts of use.

Our analysis focused on four general thematic categories from the Inner and the Outer Thematic Fields of the English clause complex, as defined in Lavid, Arús and Zamorano (2010). These were the Thematic Head (TH), the PreHead (H), the Interpersonal Theme (IT) and the Textual Theme (TT). The Thematic Head is the first element with a function in the experiential configuration of the clause; it is central to the unfolding of the text by allowing the tracking of the discourse participants. With respect to the Thematic Head, we made two interesting observations: first, it could conflate with different experiential roles in the transitivity structure of the clause (e.g. with Actor, Senser, Phenomenon, etc…). Second, it was usually realised by different types of Noun Groups (either concrete or abstract) and these groups could be simple or complex. On the basis of these observations we undertook a quantitative analysis on the two newspaper genres, obtaining the following results:

1.- The two newspaper genres choose different types of experiential roles as Thematic Heads in their clause complexes. News reports opt for Sayers as Thematic Heads while commentaries prefer

to choose Carriers, as illustrated in examples (1) and (2), respectively. Differences were also observed in the selection of "There" as Thematic Head (6.15% in reports [3], where it is particularly helpful to introduce new information, versus only 0.52% in commentaries).

(1) [Thematic Head/Sayer:] Mr Tilmant said the bank had the trust of its customers and had not seen a large outflow of funds (Report 7)

(2) [Thematic Head/Carrier:] The main expectations are for a reduction of nuclear armaments (Comment. 2)

(3) *Later* [Thematic Head:] there were unconfirmed reports that six people were still alive in the rubble of a building. (Report 10)

2.- News reports and commentaries also seem to differ in the types of nouns selected as Thematic Heads. While news reports typically select concrete nouns, referring to individuals, groups of people or institutions, as illustrated by (4), commentaries prefer abstract nouns, as shown in (5). As we see in (4), although Noun groups in reports are shorter, they are often clarified through the use of appositions which clearly identify the referent. Conversely, the higher complexity of Noun groups in commentaries gives an impression of academic, formal discourse and of a more elaborated style than that of news reports.

(4) [Thematic Head:] Dominique Strauss-Kam, the French head of the International Monetary Fund, escaped dismissal for a one-night stand with a subordinate today, but was denounced by board members for a "serious error of judgment". (Report 1)

(5) As a result, [Thematic Head:] its ability to maintain services – and the military capacity to respond to any maneuver by the Khartoum government aimed against the peace agreement – is seriously compromised. (Commentary 1)

3.- Both newspaper genres show a low frequency of interpersonal Themes, which points to a preference for the use of linguistic means other than Theme for expressing interpersonal meanings. A cursory analysis reveals the use of alternative resources such as verbal modality and evaluative lexis, as illustrated in (6).

(6) Reaching out to the SCO would certainly seem to support NATO's stated objectives (Commentary 5)

4.- As for textual Themes, differences were found pointing to the different textual structures which characterize these two genres. In news reports textual Themes are not frequent, since the textual organization relies on paragraphing. Each paragraph reports a finding or a comment by the writer.

By contrast, textual Themes are a fundamental tool for writers of commentaries, who systematically rely on textual Themes to scaffold the text's argumentative structure, and to signal logico-semantic relations between complex ideas, as illustrated by (7).

(7) [Textual Theme:] *Thus, for example,* [Thematic Head:] at the SCO summit in August, Russia did not get the support of other members regarding the Georgia conflict.

As shown by the results described above, the preliminary corpus analysis on the initial sample reveals interesting differences in the thematic choices characterizing these two newspaper genres. These thematic choices served as the theoretical basis for the creation of the core and the extended tagsets of the annotation scheme in the subsequent corpus annotation phase, as explained below.

4. Annotation scheme

The thematic features extracted in the corpus analysis phase served as the theoretical basis for the design of a preliminary annotation scheme, which includes both coarse- and more fine-grained annotations of some of the features. The coarse-grained annotations are specified in a core tagset and the more fine-grained ones in an extended tagset. Our preliminary core tagset includes four tags, reflecting the range of possible thematic types which can occur as part of the Thematic field in English declarative clauses, both in news reports and in commentaries. Definitions and realisations of these tags are provided in Appendix 2 at the end of the paper. The four tags of the core tagset are the following:

1. Thematic Head (TH)
2. PreHead (PH)
3. Textual Theme (TT)
4. Interpersonal Theme (IT)

The extended tagset includes more fine-grained subtypes of some of the tags contained in the core tagset. These tags reflect the more fine-grained thematic choices investigated in the preliminary corpus analysis phase, namely, the conflation of the Thematic Head with certain experiential roles (e.g. Actor, Senser, Phenomenon, etc…), the semantic nature (concrete or abstract) and the complexity (simple or complex) of the Noun Group chosen for its realization. Table 1 presents a preliminary extended tagset for Thematic Head realisations, subject to further refinements.

Table 1: Preliminary extended tagset for Thematic Head types

Participant Type as Thematic Head
TH-Actor
TH- Goal
TH- Beneficiary
TH- Senser
TH- Phenomenon
TH- Sayer
TH- Carrier
TH- Token
TH- Value
Semantic nature of NG
TH- Concrete
TH- Abstract
Complexity of NG
TH- Simple
TH- Complex

The PreHead category in the core tagset was also specified as a simple Circumstance (PH-Circumstance), realized by Adverbial or Prepositional Groups, as (PH-CCL), when realized by a dependent clause, or as a Finite (PH-Fin), since these the main choices. Similarly, the Textual Theme was specified with three possible tags: Linker (TT-Link), Binder (TT- Bind) and Correlative (TT- Cor). The Interpersonal Theme was further divided into Vocative (IT- Voc), Comment Adjunct (IT- Com), and Modal Adjunct (IT- Mod) (see Appendix 2).

In section 5 below we will explain how we tested the reliability of the core and part of the extended tagset presented here, using inter-annotator agreement measures.

5. Agreement studies and annotation procedure

Agreement studies (also called reliability studies) are common in the NLP community where the quality of the annotations is essential for the success of the annotation project. As explained in Hovy and Lavid (2010: 23):

> It is taken as axiomatic that any annotation must be performed by at least two, and usually more, people acting independently, so that their tagging decisions can be compared; if they do not agree with enough reliability then the whole project is taken to be ill-defined or too difficult … The underlying premise of annotation is that if people cannot agree enough, then either the theory is wrong (or badly stated or instantiated), or the annotation process itself is flawed. In any case, training of computer algorithms is impossible on inconsistent input.

In the Linguistics community, and within the CONTRANOT project, agreement studies are used to test hypotheses about the behaviour of linguistic categories empirically (Hovy and Lavid 2010). More specifically, agreement studies are designed to test the reliability of the tags included in the annotation scheme. In the current study we performed two agreement studies on a small fragment of the training corpus consisting of a total of 143 clause complexes. The first 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 operation is based on the difference between the actual agreement and the expected agreement by chance. The K value ranges from 0 (the agreement is no other than the expected by chance) to 1 (there is total agreement), as shown in table 2 below:

Table 2: Interpretation of Kappa (from Viera and Garret 2005: 362)

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

The annotation procedure was the following: two annotators (or coders), #1 and #2, were asked to analyse the thematic features of six texts individually, both having studied Lavid et al.'s model of thematisation in depth and internalised the definitions for the core and the extended tagset contained in the annotation scheme and guidelines.[ii] The lead researcher of the project, the first author of this paper, managed the annotators and organized regular meetings with them. The annotators were given coding sheets with instructions, each corresponding to a different type of task. In each agreement study there were two tasks: in agreement study 1 the two tasks focused on the identification of thematic spans. In agreement study 2 the two tasks focused on the labelling of the agreed thematic spans. These are explained in detail below.

Agreement study 1

This study consisted of two tasks focused on the identification of markables.

The first task was aimed at the identification of the whole set of potential thematic markables. This was carried out by asking the two coders to identify the Thematic field of each clause complex in each text. The definition and realisations of the Thematic field are provided in Appendix 1 at the end of this paper.

The second task focused on the identification of spans realising only specific thematic types from the Thematic field. Here the two coders were asked to identify in each clause complex the spans realising the Thematic Head (TH), the PreHead (PH), the IT (Interpersonal Theme), and the Textual Theme (TT). The definitions and realisations of these tags are provided in Appendix 2 at the end of this paper.

Agreement study 2

This study consisted of two tasks focused on the labelling of markables.

The first task focused on the labelling of the thematic markables agreed 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.

The second task was the labelling of the Thematic Heads as one of the possible subtypes specified in the extended tagset, i.e, as Actor, Goal, Beneficiary, Senser, Phenomenon, Sayer, Verbiage, Token, Value, Carrier, Attribute, Process, and the "There" element.

To sum up, the process consisted of two rounds in which the annotators worked individually and agreement results were measured. In the first round coders were asked to identify thematic spans, and in the second one to label specific thematic types, with an intermediate phase in which results were discussed and a consensus was reached with respect to the spans realising the thematic markables. The results of the two agreement studies are presented in the following sections.

5.1. Results of Agreement study 1

The first agreement study focused on the identification of two types of thematic spans.

In the first task, annotators had to identify the whole Thematic field in each clause span. The results of the first task are graphically presented in table 3 below.

Table 3: Inter-annotator Agreement: Spans expressing Thematic Field

	a	b	agr(a‖b)	agr(b‖a)	average
Text1	A	B	0.96	0.93	
Text2	A	B	0.973	1	
Text3	A	B	0.906	1	
Text4	A	B	1	1	
Text5	A	B	1	1	
Text6	A	B	0.92	1	
Average					0.97408

As shown by the figures in table 3 above, the agreement between coder (a) and (b) was very high on average (0.97 %).

In the second task the annotators were asked to identify the spans realising thematic categories of the core tagset in our annotation scheme. The results for each of the thematic categories are collectively presented in table 4 below.

Table 4: Inter-annotator Agreement: Spans expressing TH, PH, IT, and TT.

	TH	PH	IT	TT
Average	0.9384	0.787	0.375	0.965

As illustrated in Table 4 above, agreement was high in the identification of the span expressing the Thematic Head and the Textual Theme, but lower –although still substantial- in the identification of the PreHead (0.787%). By contrast, agreement was rather low (0.375%) in the identification of the Interpersonal Theme. As shown in table 5, below, the main reason for the lower agreement in the identification of Interpersonal Theme was the labeling, by one of the annotators, of textual Themes as interpersonal Themes. This error, however, did not happen at the time labelling segments (see agreement study 2, below), which suggests a punctual performance error by coder 1. It is worth considering, anyway, whether the definitions for TT and IT may need some sort of reformulation or extension so as to make them more clearly distinguishable from each other. Before doing so, however, further tests will have to be conducted to check whether the confusion was simply a performance error, in which case there might be no need for such reformulation, or whether the problem persists.

Table 5. Textual Themes annotated as Interpersonal Themes by Coder 1

	Text 2: The Vanishing Bomb		
Clause #	Text Clause	Coder 1	Coder 2
19	For example, although the United Nations Mission in Sudan is supposed to monitor implementation of the CPA, Darfur has practically monopolized its attention.	For example	-
	Text 3: The Limits of Energy Innovation		
Clause #	Text Clause	Coder 1	Coder 2
30	For example, if 20% of the world's electricity were to be generated by wind turbines, then, considering their inherently low load factor of about 25% (compared to 75% for thermal stations using steam turbines), we would need to install new capacity of some 1.25 TW in these machines.	For example	-

5.2. Results of Agreement study 2

The second agreement study focused on the labelling of markables. As in the previous study, we designed two tasks:

In the first task, annotators had to label the thematic markables which had been agreed on in the previous agreement study. As said above, coders were requested to classify the highlighted thematic spans on the coding sheet, including those spans which did not correspond to any of the thematic types of the core tagset. The results of this task are graphically presented in table 6 below. The numbers 1 through 5 correspond to the five labels coders were to chose from, i.e. TH, PH, IT, TT and 'none', respectively (note that coder 1 did not mark any span as IT, hence the absence of category 4 in the vertical column. As we can see, the overall Kappa value is quite high, at 0.915, indicating almost perfect agreement.

Table 6. Inter-annotator agreement: Labeling of thematic spans

Contingency table Coder 1 * Coder 2

		Coder 2					
		1	2	3	4	5	Total
Coder 1	1	18	0	0	0	0	18
	2	1	85	0	0	5	91
	3	0	0	19	1	0	20
	5	0	0	0	0	14	14
Total		19	85	19	1	19	143

Symmetric measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Agreement measure	Kappa	,915	,031	17,355	,000
Number of valid cases		143			

The second task focused on the labeling of the Thematic Heads in the clause complexes. 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 1 above. As shown in table 7 below, the Kappa value is rather high, at 0.875, and agreement is therefore considered to be substantial. Disagreement occurred in 15 cases, probably due to the inherent difficulty in disambiguating experiential roles conflating with Thematic Heads. As seen in table 8 below, which

refers to the text with the higher number of disagreements, these often reflected different interpretation of process types for non-clear-cut cases, which automatically involved the assignation of different participant roles. 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.

Table 7. Inter-annotator agreement: Thematic Head types (Experiential roles)

Contingency table Coder 1 * Coder 2

		Coder2												Total
		1	2	5	6	7	9	10	11	12	13	14	15	1
Coder1	1	39	1	1	0	1	0	4	0	1	0	0	0	47
	2	0	12	0	0	0	0	0	0	0	0	0	0	12
	5	0	0	6	0	0	0	1	0	0	0	0	0	7
	6	0	0	0	1	0	0	0	0	0	0	0	0	1
	7	2	0	0	0	9	0	0	0	0	0	0	0	11
	9	0	0	0	0	0	1	0	0	1	0	0	0	2
	10	1	0	0	0	0	0	8	0	1	0	0	0	10
	11	0	0	0	0	0	0	1	5	0	0	0	0	6
	12	0	0	0	0	0	0	0	0	24	0	0	0	24
	13	0	0	0	0	0	0	0	0	0	6	0	0	6
	14	0	0	0	0	0	0	0	0	0	0	6	0	6
	15	0	0	0	0	0	0	0	0	0	0	0	11	11
Total		42	13	7	1	10	1	14	5	27	6	6	11	143

Symmetric measures		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement	Kappa	,875	,031	26,282	,000
N of Valid Cases		143			

Table 7. Disagreement in the assignation of experiential roles to thematic Heads

	Text 1. The Bigger Issue in Sudan		
Clause #	**Clause**	**Coder 1**	**Coder 2**
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.	Value	Token
6	After all, the oppressive nature of the regime in Khartoum is at the root of the many conflicts that have torn the country apart.	Actor	Token
7	If the government in Khartoum persists in undermining	Token	Carrier

	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.		
18	The Government of Southern Sudan suffers from serious financial constraints, owing to unrealistic assumptions about its oil revenues.	Receiver	Carrier
28	China, a close ally of the government in Khartoum, is now carefully weighing its oil interests and its strategic concerns in the South.	Actor	Senser

6. Steps in the semi-automatic annotation of thematic features

On the basis of the agreement studies described in the previous section, we undertook the final phase of the study, namely, the semi-automatic annotation of the investigated thematic choices. For this purpose we used a well-known tool in the SFL community, the UAM corpus tool, available at <http://www.wagsoft.com/CorpusTool/>. This tool is specifically designed to support SFL-based annotation, which makes it very useful for the kind of annotation carried out in our study (see O'Donnell 2008 for details).

Although the annotation with the UAM corpus tool is still done by a human annotator and not automatically by a computer programme, there is a difference with purely manual annotation: here the program offers annotators a number of labels from which to choose. These labels are automatically generated by the tool based on the schemes previously created. The procedure for the creation of schemes for thematic annotation and how this enabled the semi-automatic annotation of the texts in our corpus was as follows:

1. A first step involved the creation of a system network which captured the whole set of potential thematic elements in the English clause complex, i.e., the Thematic field. Being SFL-oriented, the UAM Corpus Tool allows the creation of system networks, called *schemes* in this tool, by means of a user-friendly application, where the user adds subsystems to the general network and features to each system. Figure 1 shows a screenshot from the application in use, whereas figure 2 reproduces the final product, i.e. the system network for Thematic Field.

[FIGURE 1 NEAR HERE]

[FIGURE 2 NEAR HERE]

The created annotation scheme is therefore the resource feeding the text-segment annotation template. The features included in the scheme will appear in the template in the same sequence as in the scheme. This can be seen in figure 3, below, where the highlighted thematic segment has already been annotated: the features in the 'Assigned' box – i.e. 'thematic-field', 'outer-thematic-field', 'textual-theme', 'linkers', 'no-interpersonal theme', 'inner-thematic-field' and 'no-prehead' – can be traced in the system network in figure 2, above, from left to right and from top to bottom.

2. After creating the annotation scheme, the next logical step was to upload our corpus to the tool so as to run some preliminary annotation tests. Given the semi-automatic nature of the annotation, it was important to make sure that annotators would be given the right labels at the right stage. In fact, what the testing revealed was that the scheme in figure 2 was only suitable for the annotation of whole thematic fields, where the annotator specifies the existence or not of an OTF, a PH, etc., as shown in figure 3, where the whole Thematic Field '*Moreover, Sudanese security forces*' has been annotated. The scheme is not suitable, however, for the annotation of the specific components of the thematic field, i.e. TT, IT, PH and TH, where, once one of them and its dependant features have been selected, the annotation for that component should end. With annotation based on the scheme in figure 2, however, if one chooses, for instance, "textual-theme: linker", the tool will then offer choices for "interpersonal-theme", for "prehead" and, finally, for "thematic-head". This is shown in figure 4, where the textual Theme *Moreover* has already been annotated as such (see annotation in the 'assigned' box, leftmost), and the annotator is still faced with further choices to make, this time for interpersonal Theme (see labels in 'interpersonal Theme' box, center).

[FIGURE 3 NEAR HERE]

[FIGURE 4 NEAR HERE]

3. The scheme shown in figure 2 is therefore valid for the correct description of the systems of English Theme for representational purposes, as well as for the annotation of the whole Thematic Field, with specification of its complexity but without the possibility to segment and tag its internal components. To overcome this annotation problem, it was necessary to create a second annotation layer with a scheme – see figure 5, below – where the features TT, IT, PH and TH are presented as alternative rather than parallel. This is reflected by the different kind of brackets used in each system: braces for parallel features, square brackets for alternative ones. The new scheme cannot be used for representational purposes, as the relations within the network are not the real ones, but it

now allows the independent annotation of each of the thematic components, as illustrated in figure 6, where once the corresponding labels for the annotation of *Moreover* have been selected and duly assigned, no more choices are given to the annotator for that segment, because they are not needed.

[FIGURE 5 NEAR HERE]

[FIGURE 6 NEAR HERE]

4. Once the annotation layers for Thematic Field and its components were created, the rest of the annotation layers to account for the features in the extended tagset in table 1 (section 3, above) were created and likewise tested. The result is the one shown in figure 7, where five different annotation layers can be differentiated – i.e. English-thematic-field, thematic constituents, thematic-head-participant, semantic-nature-of-NG and complexity-of-NG – ready to annotate our corpus for each of those parameters. This is the phase we are in at the present moment: the texts constituting our training corpus are being annotated in the manner here specified. This will be followed by the annotation of a larger amount of texts. At any stage in the annotation process, statistics concerning the annotated texts can be obtained (see 'Statistics' tag in figure 7, below). This will help us validate – and, if necessary, adjust – the results obtained in the manual annotation phase given that automatic data mining is a more reliable way of tackling statistical tasks than manual scrutiny, which tends to be not only fastidious but also error-prone.[iii]

[FIGURE 7 NEAR HERE]

7. Summary and concluding remarks

In an attempt to contribute to current efforts within the SFL community oriented to the task of creating, exploring and sharing linguistic resources in English and other languages in more comprehensive and effective ways, in this paper we have presented a methodology for the creation of quality (reliable) data in the area of thematisation in English. By quality data we refer to a corpus annotated with thematic features where the annotations have been tested experimentally to ensure their reproducibility. We have shown how it is possible to create such a corpus following a number of methodological steps which include the compilation of a training corpus, the performance of a preliminary corpus analysis to extract relevant choices to be included as tags of an annotation scheme, and the validation of the annotation scheme through agreement studies. The final phase of

the current study has focused on illustrating a procedure for the semi-automatic annotation of the investigated thematic choices. On the basis of such reliable annotations and provided we get enough data, it should be possible to develop machine-learning algorithms for the automatic annotation of a larger corpus with the thematic features investigated in this paper. We expect that such a corpus would be a useful resource not only for the SFL community, but also of potential interest for NLP tasks such as information extraction and text classification.

Works Cited

Counsell, S., Loizou, G. and Najjar, R. (2006) Quality of manual data collection in Java software: an empirical investigation. *Empirical Software Engineering* 12.3: 275--293.

Eiler, M. (1986) Thematic distribution as a heuristic for written discourse function. *Functional Approaches to Writing, Research Perspectives*. B.Couture (ed.). Norwood, New Jersey: Ablex. 49-68.

Francis, G. (1989) Thematic Selection and Distribution in Written Discourse. *Word* 40: 201-222.

Francis, G. (1990) Theme in the Daily Press. *Occasional Papers in Systemic Linguistics* 4: 51-87.

Fries, P. H., Francis G. (1992) Exploring Theme: Problems for Research. *Occasional Papers in Systemic Linguistics* 6: 45-59.

Ghadessy. M. (1995) Thematic Development and its relationship to register and genres. In M. Ghadessy, ed. Thematic Development in English Text. London: Pinter. 129-146.

Hovy, E., Lavid, J. (2010) Towards a science of corpus annotation: a new methodological challenge for Corpus Linguistics. *International Journal of Translation*. 22.1: 13-36.

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., Arús, J. and Moratón, L. (2010) Signalling genre through Theme: the case of news reports and commentaries. Proceedings of MAD10. 82--92.

Lavid, J., Arús, J. and Moratón L. (forthcoming) Signalling genre through Theme: The case of news reports and commentaries. *Discours. Special issue on Multidisciplinary Perspectives on Signalling Text Organisation*.

Lavid J., Arús J. and Zamorano J. R. (2010) *Systemic-functional Grammar of Spanish: a Contrastive Study with English.* London: Continuum.

Nwogu K., Bloor T. (1991) Thematic Progression in Professional and Popular Medical Texts. *Functional and Systemic Linguistics: Approaches and Uses*. E. Ventola (ed.) Mouton de Gruyter. 369-384.

O'Donnell, M. (2008) Demonstration of the UAM CorpusTool for text and image annotation. *Proceedings of the ACL-08:HLT Demo Session (CompanionVolume)* 13-16. Columbus, Ohio: Association for Computational Linguistics. Retrieved on 15 December 2011 from http://www.aclweb.org/anthology-new/P/P08/P08-4004.pdf

Teich, E. (2007) SFL and linguistic computing: An analysis and some recommendations. Plenary talk at the *34th International Systemic Functional Congress*, Odense, Denmark.

Teich, E. (2008) IRSFL: An initiative for a Repository of SFL Resources. Presentation at the *35th International Systemic Functional Congress*, 21-25 July, 2008, Sydney, Australia.

APPENDIX 1: Definition of Thematic Field

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 unequal relevance** is …..
(5) **On the table [FNC:] stood** a lamp
(6) **But, surprisingly, before the meeting [FNC:] everybody** was glad to hear the news.
(7) **[FNC:] What I want** is you

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

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

(9) **In fact [PT:] It is love** that makes the world go round
(10) **When I arrived**, **[THERE:] there** were three people waiting for the bus

APPENDIX 2: English Core Tagset for Theme categories (Declarative clauses)

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,* a *Process, an Absolute Theme,* a *Thematic Equative, a Predicated Theme* or the "*There*" element in existential clauses. When the Thematic Head is a *Participant*, it can be realized as:

- a Noun Group. (e.g. ***The cat*** *is on the mat;* ***Peter*** *is at home;* ***She*** *saw him yesterday)*
- an Adverbial Group (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)*
- an Absolute Theme (e.g. "***Real Madrid, their players*** *have been holding up a banner*")
- a Thematic Equative (e.g. ***What you need*** *is love*)
- a Predicated Theme (e.g.: ***It is you*** *who are to blame)*

When the Thematic Head is a *Process*, it is realized as a verbal form, preceded by a Pre-Head element, such as for example a Circumstance (e.g. *On the table* ***stood*** *a lamp) or an auxiliary.* When the clause is existential, the Thematic Head is realized by the "There" element (e.g. ***There*** *were three people waiting for the bus)*

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. 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:] ***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.)*

4. 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!*)

APPENDIX 3

Extended Tagset (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, [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, symbolized 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 symbolized 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: 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.

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

[ii] The two annotators were two members of our research group at UCM, namely Dr. M. Juan Rafael Zamorano and Dr. Marta Carretero.

[iii] See Counsell, Loizou and Najjar (2006) on the advantages of automatic over manual data collection. For these authors, although manual data collection can be more accurate than expected and sometimes unavoidable, automatic data collection is usually preferable.