Using bilingual corpora for the construction of contrastive generation grammars: issues and problems

Julia Lavid
Dep. English Philology I
Universidad Complutense de Madrid
28040 Madrid (Spain)
Phone and fax: +34-91-518-5799
e-mail: julavid@filol.ucm.es

This paper reports on the use of corpora for the construction of a computational grammar of Spanish, contrastive with English, in the application context of Multilingual Natural Language Generation (MLG). The theoretical framework for this work is Systemic Functional Linguistics (SFL) and the computational context provided by KPML (Komet Penman Multilingual), an extensive grammar development environment and generation engine that supports large-scale multilingual development (Bateman 1997). The initial phenomena which are being investigated contrastively belong to three different functional regions of the grammar, i.e., particular subareas of the grammar that are concerned with particular areas of meanings. These regions are transitivity (ideational meaning), thematicity (textual meaning) and mood (interpersonal meaning). The present study concentrates on textual meaning (thematicity and focus) as an illustration.

Following what has now established itself as a standard methodology for empirically-based Natural Language Generation (Bateman 1998a, Reiter and Dale 1997), the following steps were carried out: first, a bilingual corpus (English-Spanish) was selected. For Spanish, a sample of spoken texts from the MacroCorpus of the educated linguistic standard of the main cities of the Spanish-speaking world was used, while for English a comparable sample was selected from the British National Corpus Sampler. This was motivated by the need to provide a realistic account of the behaviour of the linguistic phenomena investigated in unplanned and spontaneous contexts of use. The second step was to carry out a contrastive analysis of the phenomena mentioned before. Finally, the results of the analysis were coded up as resources/processes for generation. In the case of Spanish, these had to be created anew. In the case of English, as the KPML already includes an English generation grammar, this last step consisted on checking the coverage of the existing specifications and extending them when could not cover the instances found in the corpus, and adapting them for effective MLG.

Given the nature of the NLG process, which typically converts communicative goals expressed in some internal representation into surface forms, the kind of information that is most readily usable for NLG are statements of mappings from functions to forms. Therefore, the corpus analysis phase for NLG usually includes an explicit, and usually quite lengthy linguistic analysis where the analyst seeks possible realisations of communicative functions, which restricts the size of the corpus that can be realistically considered.

This paper describes the different steps carried out for the generation of the linguistic phenomena mentioned above, discussing the problems encountered during the corpus analysis phase, and the computational representation derived from it, as well as some of the decisions taken to overcome them.

1. Introduction

Natural Language Generation (NLG), the subfield of Artificial Intelligence and Computational Linguistics that investigates the automated production of texts by machine, is typically a process that converts communicative goals expressed in some internal representation into surface forms. As such, it touches upon different linguistic areas of inquiry such as text planning and discourse organisation, lexical semantics, grammatical and lexical choice, and the relationship between all of these. In fact, one of the main concerns of NLG is the construction of of computational accounts of the linguistic system capable of generating texts in one language (monolingual NLG) or in several (multilingual NLG, henceforth MLG). This theoretical task poses unusual demands on the linguist who has to provide explicit and details accounts of how language works and confront the results of the application of his/her theoretical claims in concrete computational systems. A more practical concern is the creation of computational systems capable of producing acceptable text from various sources and for different types of applications: well-known examples include the generation of weather reports from meteorological data (Kittredge et al. 1986), the generation of letters responding to customers (Springer et al. 1991), and other systems applied in areas such as technical documentation and instructional texts (Not and Stock 1994, Paris et al. 1995, Lavid 1995), patent claims (Sheremetyeva et al. 1996), information systems (Bateman and Teich 1995), computer-supported co-operative work (Levine and Mellish 1994), patient health information and education (DiMarco et al. 1995), and medical reports (Li et al. 1986), to mention a few.[1]

While the first generation systems were limited to the random generation of grammatically correct sentences, the field has experienced a very rapid growth over the past ten years, both as a research area bringing a unique perspective on fundamental issues in artificial intelligence, cognitive science, and human-computer interaction, and as an emerging technology capable of partially automating routine document creation and playing an important role in human-computer interfaces. In this sense, as practical systems became more sophisticated, it was necessary to provide a good understanding of the notion of "textuality" and all the factors involved in the creation of different text types. In this context, it was only natural that corpora started to be used as the empirical basis both for the theoretical investigation of textual phenomena, and as part of the requirement analysis phase for NLG systems. As a result, the use of corpora has now been integrated as part of the standard methodology for NLG, both in theoretically-oriented research and in the development of concrete generation systems, to such an extent that computational tools have been developed to support, among other functionalities, the analysis of machine-readable monolingual or multilingual corpora for NLG (Alexa and Rostek 1997).

In this paper, we report on the use of bilingual corpora for the construction of a computational grammar of Spanish, contrastive with English, in the application context of Multilingual Natural Language Generation (MLG), concentrating on the textual phenomenon of thematicity and its relationship with the related notion of focus as an illustration. Section 2 describes the two corpora selected and the criteria for concentrating on two comparable samples as the empirical basis for computational specifications. Section 3 presents the theoretical framework selected for this study and the issues which must be explored contrastively with respect to the textual phenomena mentioned before. Section 4 describes the corpus analysis phase of the bilingual samples. Section 5 presents a computational specification of the results of the analysis as resources for generation. The specification is based on the notion of functional typology as developed in SFL and implemented in the KPML development environment. Finally, section 6 provides a summary and discusses the implications of these results for corpus-based MLG.

2. Bilingual corpora for NLG

Two electronic corpora were initially targeted as the empirical basis for the contrastive work proposed in this paper. These were the *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* (Samper et al. 1998), a corpus of the educated linguistic standard of the main cities of the Spanish-speaking world, and the British National Corpus. These two corpora were initially selected for two main reasons: they both describe the educated speech of both Spanish and English and contain samples of

[1] For an extensively documented review of the field see Bateman (1998b)

unplanned and spontaneous speech, which are necessary to provide a realistic account of the behaviour of the linguistic phenomena investigated in their contexts of use. In the case of NLG, it is frequent to find computational specifications of linguistic phenomena based on the analysis of specific text types, most of them written monologues, which, though useful for the generation of texts of a specific domain, can only provide a partial view on the phenomena studied. As the purpose of this study is to investigate the behaviour of textual phenomena for a multipurpose contrastive grammar of English and Spanish, the choice of samples from unplanned, spontaneous speech offered a rich and unexplored empirical basis for the study of the textual phenomena mentioned before.

For the study of Spanish, a sample of spoken texts from the Spanish Macrocorpus was selected. This Macrocorpus includes the transliteration of 84 hours of recording by 168 native speakers representative of the educated speech of twelve Hispanic cities, nine of them from SouthAmerica and three of them from the Iberian Peninsula. The recordings are basically unstructured interviews conducted in a conversational style where the interviewer introduces a few questions and topics to stimulate the conversation and to establish some uniformity in the topics discussed by the speakers. In general, the speakers were left free to talk about the topics suggested by the interviewer and to introduce any new topics. The sample selected for this study consists of 10 interviews from the three cities of the Iberian Peninsula contained in the Macrocorpus, i.e. from Madrid, Seville and Las Palmas de Gran Canaria.

For the study of English, a comparable sample was selected from the BNC Sampler Corpus, a subcorpus of the British National Corpus, consisting of approximately one-fiftieth of the whole corpus, viz. two million words. The Sampler Corpus consists of 184 different texts, comprising just under 127,000 sentences and two million words. From this sampler corpus, a subcorpus of spoken texts was chosen for the purposes of the present study, consisting of 10 conversations.

3. The textual resources of thematicity and focus in English and Spanish

The linguistic phenomena selected for illustrating the use of corpora in NLG belong to the textual region of the grammar, and have been the subject of several functional accounts. The theoretical framework for the study of these phenomena is SFL, as this is the theoretical basis for the implementation of the current English generation grammar Nigel. In SFL, the textual clause grammar is composed of two complementary systems, the systems of Theme and Information, characterised as assigning two different kinds of textual prominence to elements of the clause: thematic prominence, in the case of the system of Theme, and prominence as news, in the case of the system of Information. The notion of Focus, however, has not received a comparable attention within SFL, but has been the subject of a number of discourse-oriented linguistic and computational studies (Lavid 2000, McCoy and Cheng, McKeown, among others), and has been treated in Dik´s Functional Grammar (Dik 1978) as a pragmatic function which assigns more salience to some clausal constituent with respect to the contextual (pragmatic) information between language producers and receivers.

In view of the need to provide form-function mappings of these phenomena which can be used by a MLG system, a corpus-based analysis for generation purposes must explore at least the following issues contrastively:

1. Do both English and Spanish grammars have a Theme and a Focus function ?
2. How do English and Spanish realise the Theme and the Focus functions - e.g.: sequence, inflection, adposition, intonation?
3. Are there marked and unmarked Theme and Focus selections in both languages, and to what extent do they depend on other systems (e.g. Mood, Voice)?
4. Are there resources in English and Spanish to combine thematicity and focus?

With respect to the first question, different linguistic studies acknowledge the existence of these two functions in different languages (see Caffarel et al. in preparation; Dik et al. 1981), so this issue will not be further explored here. The rest of the issues, however, are central for a functional characterisation of the textual phenomena selected for this study, and their corpus-based empirical study is the basis for the specification of resources required for MLG. Therefore, the next section will describe the corpus analysis carried out for this study and the problems encountered when attempting to investigate the issues mentioned above.

4. Contrastive corpus analysis for MLG: problems and decisions

As in the area of discourse analysis, the type of information that NLG needs from corpus investigations is one which basically consists of statements of mappings from functions to forms. Therefore, when investigating specific discourse phenomena, a NLG system requires as a first step a specification of the mappings from functions to forms with the purpose of duplicating the text analysed. In this sense, the corpus analysis phase raises some problematic issues which must be considered by NLG practitioners. One of these issues is the size of the corpus: if, at is usually the case, the analyst has to intervene looking for possible realisations of communicative functions, the use of large corpora becomes impractical. Also, it is not possible to mark-up large quantities of texts according to functional categories if they cannot be recognised automatically on the basis of tagged texts or syntactic analysis.

In view of these problems, the following decisions had to be taken to investigate the issues mentioned above:

1.- With respect to the first issue, i.e., the ways in which English and Spanish realise the Theme and the Focus functions, the corpus analysis was based on the following assumptions, based on previous linguistic studies:
a) The Theme function was recognised as realised by clause initial position in both languages, as several linguistic studies have demonstrated (Lavid in press; Taboada 2000).
b) The Focus function was recognised on the basis of non-prosodic realisations, such as word order patterns, focus markers and characteristic constructions, since the BNC corpus does not include prosodic annotation. [2]

2.- With respect to the second issue, i.e., the existence and realisations of marked and unmarked theme and focus selections in both languages, the corpus analysis was based on the following assumptions:
a) marked and unmarked themes were recognised in specific mood contexts, such as declarative, interrogative and imperative options. Absolute themes were recognised by the presence of a pause and a comma separating them from the rest of the predication.
b) unmarked focus was assumed to coincide with the last lexical element of the clause.[3]

3.- With respect to the third issue, i.e., the existence and variation of resources which combine thematicity and focus in both languages, the corpus analysis concentrated on the so-called cleft and pseudo-cleft sentences as realisation strategies used by both English and Spanish to combine the Theme and the Focus functions. These were semi-automatically distinguished in both corpora on the basis of their characteristic form.

5. Towards a computational specification for MLG

The second step in what has now established itself as a standard methodology for empirically-based NLG is the codification of the corpus analysis results as resources/processes for generation. The current implementation of the English generation grammar contained in the KPML development environment already includes a computational specification for the Theme function in English as a textual region of the grammar. However, as will be shown in the following sections, it became necessary to modify the existing specification to account for instances found in the corpus and to ensure maximal sharing of resources for contrastive generation. For Spanish a new specification was created on the basis of the contrastive corpus analysis. It should be noted, however, that it is not the purpose of this study to provide a full specification

[2] Considering that Focus in English is predominantly realised by marked prosodic prominence (Martínez-Caro 1999), this apparent limitation of the corpus-analysis phase must be overcome by explicitly representing this realisation in the computational specification for English.

[3] According to Halliday (1994), unmarked focus falls on the last lexical element of the tonic group, which in unmarked circumstances coincides with the clause.

of the behaviour of the textual phenomena of thematicity and focus in both languages, but rather to discuss and illustrate with some examples some problems and issues raised by corpus-based contrastive MLG specifications. The following sections, therefore, will illustrate some of these problems and the solutions suggested in the context of a MLG architecture based on functional typology.

5.1. Thematic resources for English and Spanish: a functional-typological characterisation

This section presents a partial specification of the thematic resources available in English and Spanish on the basis of the contrastive corpus-based analysis. More specifically, the area of theme markedness will be discussed in detail, as there exist some interesting commonalities and differences which must be accounted for when generating English-Spanish thematic variants. The approach rests on the notion of functional typologies, as pursued in systemic-functional linguistics (cf. Halliday 1978), later developed for MLG purposes (see Bateman et al. 1999). Functional typologies are constructed by means of classification hierarchies called system networks (Halliday 1966), where each disjunction, or grammatical system is seen as a point of abstract functional choice, capturing those minimal points of alternation offered by a linguistic representation level. This representation of the so-called paradigmatic axis of linguistic description is complemented with the corresponding syntagmatic realisation, i.e. the structural expression of the choices made in the paradigmatic axis and associated with individual grammatical features. This task is carried out by the so-called realisation statements which set allowable constraints on configurations of syntactic constituents such as linear precedence, immediate dominance, or "unification" of functional constituents. This type of approach, where the functional coherence of the paradigmatic organisation is preferred over generalisations concerning possible structures, has been found to provide effective multilingual linguistic descriptions and computational representations which maximise the factoring out of generalisations across languages (cf Bateman et al. 1991, 1999).

Therefore, following this approach, and on the basis of empirical corpus-analysis, a specification was created to account for Theme markedness in Spanish, as shown in Figure 1 below:

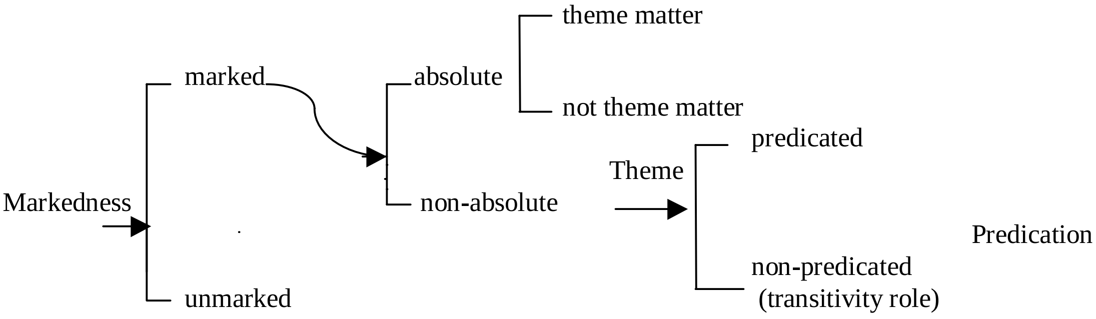

Figure 1: Spanish Theme markedness systems

As the system network shows, the primary distinction in Spanish is between marked and unmarked themes, and within marked themes between absolute or non-absolute themes. Absolute themes do not map onto any transitivity or interpersonal functions, and are normally separated from the transitivity /interpersonal structure of the clause by a comma in writing and by a pause in speech. They can be of two types: representing a theme matter or not. If non-absolute, they can be predicated or non-predicated. When predicated, they are normally realised by pseudo-cleft sentences, as will discussed in detail below. If they are not predicated, thematic status may be assigned to any role within the transitivity structure of the clause (participants, processes or circumstances).

With respect to English, the existing specification was found not to account for all the language instances found in the corpus. Therefore, on the basis of the corpus-based analysis and to provide a maximally effective specification for the purposes of MLG, the following specification was created for English:

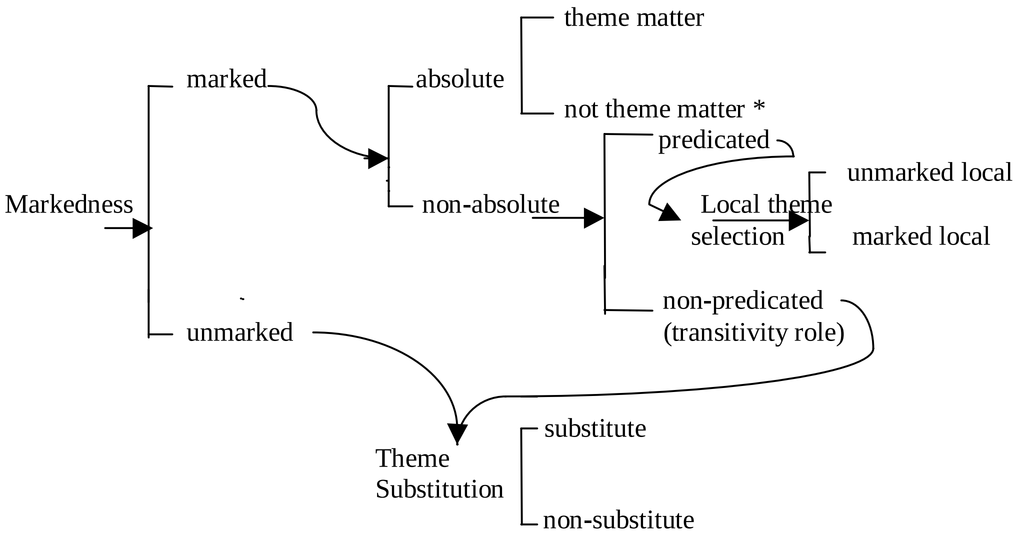

Figure 2. English Theme markedness systems

As can be seen, an attempt has been made at capturing the comnonalities with the Spanish specification while maintaining the English-specific integrity and difference. Thus, while English shares with Spanish the primary distinction between marked and unmarked themes, previously existing specifications (see Matthiessen 1995: 540) did not include a distinction between absolute and non-absolute themes. However, in order to enforce maximal sharing of resources across languages, this system was also included in the English network. In this sense, it should be noted that up to this point both languages are similar, but differ as the features become more delicate: while absolute themes in Spanish may refer to a theme matter or to any other transitivity role coreferential with a constituent within the predication, English has a gap in this common paradigm (indicated by an asterisk in the English network) since it only has the option of a theme matter as absolute theme. Also, English includes more delicate options within the systems of Theme predication, as it includes the possibility of having a local theme selection, and includes a system called Theme substitution, which is not available in the Spanish paradigm. In this new specification, predicated themes have been considered as marked in the sense that they map the Focus constituent, and, in many instances, the New element, onto the Theme (see Halliday 1994: 59 on predicated Themes). Contrastive examples extracted from the corpora are discussed in detail below.

With respect to the syntagmatic realisations of the options presented in the networks above, both languages present divergent grammaticalisations of their common paradigmatic potential. Table 1 below illustrates the different features of the systems of both languages, together with their divergent realisations. These are specified as sets of constraints which specify the syntactic and lexical properties of the linguistic units being generated. For example, the feature 'theme substitute' has the realisation statement +Pro-theme, which means 'insert a Pro-theme' and 'Pro-theme / Subject' which means 'conflate or unify with the grammatical function Subject'. The rest of the realisation specifies the ordering of constituents in relation to one another: thus 'Rheme ^ Theme' means order the Rheme before the Theme.

feature	**English realisation**	**Spanish realisation**
System: Markedness		
marked		
absolute		
theme matter	’As to/as for/... + NG, + Pred.	’en cuanto a + NG’, + Pred.
not theme matter		Left Dislocation + Predication
non-absolute		
predicated	’It + be + X..that/who’	Relative ^ Copula ^ NP NP ^ Copula ^ Relative Copula ^ NP ^ Relative
unmarked local		
marked local	’NP it + be+ that’	
non-predicated	Part, Pro. or Circ. non-Subj.	Part, Pro. or Circ. non-Subj.
unmarked		
non-substitute		
substitute	+Pro-theme/ Subject ^ Rheme ^Theme	

Table 1: Syntagmatic realisations of theme markedness options in English and Spanish

Some of the divergent realisations in both languages are the following:

1. <u>Absolute Themes</u>

Absolute themes refer to constituents which are not integrated within the transitivity/interpersonal structure of the clause, and are, therefore, marked in speech with a pause, and with a comma in writing. As shown in Figure 1 above, Spanish has two options when choosing absolute themes: as theme matter and as not-theme matter. Example (1) below illustrates the choice of a theme matter in Spanish, an instance extracted from the Spanish Macrocorpus:

(1) *En cuanto a los libros que me gusta leer,* tú sabes que a mí me gusta leer todo (SE-10)
As for the books that me likes to read, you know that me likes to read everything
’As for the books that I enjoy reading, you know that I enjoy reading everything’

A contrastive example in English is illustrated by example (2) below, extracted from the BNC sampler corpus:

(2) *As for the past,* I have adopted the doctrine of anamnesis (AEA)

When the absolute theme is not a theme matter, in Spanish it may be coreferential with an element of the structure of the clause, as example (3) illustrates:

(3) ¿Y *el ciudadano, el madrileño madrileño,* se *le* nota un cambio.... (MA-10)
And the citizen, the Madrid Madrid (one), to him notice (3-sing) one change
’ And the citizen, the Madrid citizen, can one notice a change in him...?’

In (3), the constituent *’el ciudadano’* is not integrated in the transitivity structure of the clause, though it is coreferential with an element functioning as beneficiary in the clause, expressed by the clitic *’le’* (italics in the example).

English, however, lacks this option in its paradigm, and this is indicated by an asterisk in the system network in Figure 2.

2. <u>Predicated Themes</u>

Both English and Spanish include predicated themes as a choice between presenting a Theme with a special feature of identification or without that special feature. In fact, this choice is a textual strategy which combines the Theme and the Focus functions into one single realisation, which, depending on the language, may be a cleft, a pseudo-cleft, an identifying clause or a combination of the three. Thus, English typically uses clefts with the form (’it + be + X + that ...) to realise predicated themes (italics) and to mark focus constituents (boldface) as in examples (4) and (5) extracted from the BNC sample:

(4) *It was* ***only a relatively small Arab army*** that arrived in Egypt (JXL)

(5) *It is* ***not the death of the body*** that is important to us, it ’s **the soul** (NHE3)

Textually speaking, this construction serves two textual functions:

a) to identify the Theme in the transitivity role it serves in the clause: by using the identifying type of clause we are achieving a textual distribution of meaning with the added feature of identification. For example, in (4) ’the thing that arrived in Egypt’ is identified as ’a relatively small Arab army’. Similarly, in (5) the ’the thing that is important to us’ is identified as *not the death of the body* but as *the soul.* According to this, the segment ’it’s not the death of the body’ would be the Theme of the clause, and ’that is important to us’ would be the Rheme.

b) to define a specific constituent as having the Focus function, which, in many cases, is presented as contrastive to another alternative. Example (5) is a case of what Dik would call counterpresuppositional Focus, more specifically Substitute Focus. According to his definition, "the information presented is opposed to other, similar information which the Speaker presupposes to be entertained by the Addressee" (Dik 1989: 282). In (5), the speaker rejects the information which he/she presupposes to be entertained by the Addressee (*the death of the body*) and corrects it with information which the speaker considers to be correct (*the soul*).

English also uses pseudo-cleft and identifying clauses to combine the Theme and Focus functions. Example (6) below illustrates a pseudo-cleft construction where the Theme is realised by a nominalising clause introduced by ’what’, and the rest of the clause serves as a Rheme. This construction also serves to mark the focus constituent as the one following the copular verb:

(6) *what we were able to do on one occasion* was er to raise enough money (FYJ)

Spanish, by contrast, tends to use only pseudo-cleft clauses to combine the Theme and Focus functions. The range of possible syntagmatic realisations in Spanish is more varied than in English, and includes three main types:

1. Constructions where the relative clause appears in initial position, followed by the copular verb and the nominal group functioning as an attribute. Exampe (7), extracted from the Spanish Macrocorpus, illustrates this first type:

(7) *Lo que yo necesito [...] es* **un poco más del estilo árabe**... (SE-01)
What I need is a bit more of the style Arabian
’What I need is a bit more of the Arabian style’

2. Constructions where the nominal group is in initial position , the verb in medial position and the relative clause at the end. Example (8) illustrates this type:

(8) ***Eso*** es lo que pienso de la familia (SE-04)
That is what think (1-sing.) of the family
’That is what I think of the family’

3. Constructions with the verb in initial position followed by the nominal group and the relative clause. Example (9) illustrates this type:

(9) *pero ha sido precisamente* ***el avance de la anestesia*** [...] la que ha podido hacer que se [...] pudiera hacer una serie de operaciones (SE-05)

But was (3-sing) precisely the advance of anaesthesia that could allow [...] that could be made certain operations
'But it was precisely the advance of anaesthesia what allowed certain operations to be done'

3. Local Theme selection

In English when theme predication is selected, and it is realised by means of a construction of the type *it + be + ...[[that ...]],* that clause opens up the potential for an additional thematic contrast - a contrast local to that clause. In this case, the Theme local to the identifying clause may be unmarked (as it is in most of the cases) or marked. If it is marked, it will coincide with the Complement / Identifier of the identifying clause and will be given thematic prominence in initial position of the clause. No examples have been found in the BNC sample, probably due to the fact that it is a small one. However, as other studies have proved its existence with empirical evidence, we include the example provided by Matthiessen as an illustration:

(10) There we fell and *my leg* it was that broke (Matthiessen 1995: 567)

4. Substitute Theme

In English, when the Theme is unmarked and not predicated, there is a choice whether to re-introduce the Theme/Subject at the end of the clause so as to give it a "thematic culmination". With substitute Theme, the Theme/Subject is a pronominal nominal group and the substitute Theme is typically a full lexical nominal group, though it is also possible to find examples with *this*, *one*, etc. In examples (11) and (12) below, the Theme/Subjects are realised by the demonstrative 'that', and the lexical items 'piano' and 'day release' are the substitute Themes, which serve as a textual reprise of the thematic referents:

(11) *That* 's good, that 's good for you, *the piano* you just had there (kp8)
(12) *That* 's what you want to be after, *day release*.

6. Summary and concluding remarks

As the contrastive examples above have illustrated, a detailed corpus analysis in search for form-function mappings of linguistic phenomena is an indispensable step as a basis for empirically-based computational specifications for MLG. The study of the phenomenon of thematicity and its relationship with focus in both languages, though partial and purely illustrative, has served to show how a functional analysis of selected corpus samples may shed light on the paradigmatic features and their syntagmatic realisations of those phenomena, which are the basis for computational specifications for MLG architectures based on functional typologies.

However, the requirements of NLG, similar in many respects to the goals of discourse analysis, raise important problems for corpus analysis: the need for semantic and contextual (pragmatic) analysis of the selected corpora, which cannot always be recognised automatically and so cannot be extracted from large-scale corpora without an analyst's intervention, makes, in most cases, the use of large corpora and form-based corpus tools unsuitable for NLG purposes. Given the close relation between discourse analysis and NLG, it can be hoped that new tools and specialised environments are developed to extract more from current electronic corpora than what can actually be obtained from form-based quantitative corpus analysis tools.

References

Alexa, M and Rostek, L 1996 *Computer-assisted corpus-based text analysis with TATOE*, Technical Report, German National Research Center for Information Technology, Institute for Integrated Information and Publication Systems, Darmstadt, Germany.

Bateman, J 1997 Enabling technology for multilingual natural language generation: the KPML development environment. *Journal of Natural Language Engineering* 3 (1): 1-41.

Bateman, J 1998a Using corpora for uncovering text organization. *IV-V Jornades de corpus lingüístics*. Barcelona, Institut Universitari de Lingüística Aplicada, Universidad Pompeu Fabra.

Bateman, J 1998b Automated discourse generation. In Kent A and Hall, C.M (eds), *Encyclopedia of Library and Information Science*, Vol 62, New York, Marcel Dekker, pp 1-54.

Bateman J, Teich E 1995 Selective information presentation in an integrated publication system: an application of genre-driven text generation. *Information Processing and Management* 31(5): 753-767.

Bateman J, Matthiessen C, Nanri K , Zeng L 1991 The Re-use of linguistic resources across languages in multilingual generation components. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Sydney, pp 966-971.

Bateman J, Matthiessen C, and Zeng, L 1999 Multilingual Natural Language Generation for Multilingual Software: a Functional Linguistic Approach. *Applied Artificial Intelligence* 13: 607-639.

Caffarel A, Martin J R, and Matthiessen C in preparation *Language typology: a functional perspective*.

Dik S C 1978 Functional Grammar. Dordrecht, Foris.

Dik S C, Hoffmann M, de Jong JR, Djiang S, Stroomer H, and de Vries L 1981 On the typology of focus phenomena. In Hoekstra et al. (eds) *Perspectives on functional grammar*. Dordrecht, Foris.

DiMarco C, Hirst G, Wanner L, Wilkinson J 1995 Healthdoc: customizing patient information and health education by medical condition and personal characteristics. In *Proceedings of the Workshop on Patient Education*, Glasgow.

Halliday M A K 1966 Some notes on "deep grammar". *Journal of Linguistics* 2 (1): 57-67.

Halliday M A K 1978 *Language as social semiotic*. London, Edward Arnold.

Halliday M A K 1994 *An introduction to functional grammar*. London, Edward Arnold.

Kittredge R, Polguère A, Goldberg E 1986 Synthesizing weather reports from formatted data. In *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, Germany, pp 563-565.

Lavid J 1995 From interpersonal option to thematic realisation in multilingual instructions. In Kittredge R (ed) *Working notes of the IJCAI-95 workshop on text generation*. Montreal.

Lavid J 2000 Theme, focus, given and other dangerous things: linguistic and computational approaches to information in discourse. *Revista Canaria de Estudios Ingleses :*

Lavid J in press La noción gramatical de tema en un contexto multilingüe: una perspectiva funcional-tipológica. In *Treinta años de la Sociedad Española de Lingüística*. Madrid, Gredos.

Levine J, Mellish C 1994 Corect: Combining CSCW with natural language generation for collaborative requirements capture. In *Proceedings of International Joint Conference on Artificial Intelligence*, Montreal, Canada, pp 1398-1404.

Li P, Evens M, Hier D 1986 Generating medical case reports with the linguistic string parser. In *Proceedings of the 5th National Conference on Artificial Intelligence*, Philadelphia, pp 1069-1073.

Martínez-Caro E 1999 *Gramática del discurso: foco y énfasis en inglés y en español*. Barcelona, Promociones y Publicaciones Universitarias.

McCoy K, Cheng J 1990. Focus of attention: constraining what can be said next. In Paris et al. (eds) *Natural language generation in artificial intelligence and computational linguistics*. Kluwer, Dordrecht, 1990.

McKeown, K 1995 *Text generation: Using discourse strategies and focus constraints to generate natural language text.* Cambridge, Cambridge University Press.

Matthiessen C 1995 *Lexicogrammatical cartography: english systems*. Tokyo, International Language Sciences Publishers.

Not E, Stock, O 1994 Automatic generation of instructions for citizens in a multilingual community. In *Proceedings of the European Language Engineering Convention*, Paris.

Paris C, Van der Linden K, Fisher M, Hartley A, Pemberton L, Power R, Scott D 1995 A support tool for writing multilingual instructions. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal, pp1398-1404.

Reiter E, Dale R 1997 Building applied natural language generation. *Journal of Natural Language Engineering* 3(1): 57-87.

Samper JA, Hernández Cabrera CA, Troya Déniz M 1998 *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Universidad de las Palmas de Gran Canaria.

Sheremetyeva S, Nirenburg S, Nirenburg I 1996 Generating patent claims from interactive input. In *Proceedings of the 8th International Workshop on Natural Language Generation*, Herstmonceaux, England, pp 61-70.

Springer S, Buta P, Wolf T 1991 Automatic letter composition for customer service. In Smith R, Scott C (eds) *Innovative applications of artificial intelligence* 3. Menlo Park, Ca AAAI Press.

Taboada M 2000 *Collaborating through talk*: *the interactive construction of task-oriented dialogue in English and Spanish.* Unpublished PhD thesis, Universidad Complutense de Madrid.