SPECIFYING THE DISCOURSE-SEMANTICS OF GRAMMATICAL THEME FOR MULTILINGUAL TEXT GENERATION: PRELIMINARY FINDINGS[1]

Julia Lavid
Department of English Philology
Faculty of Philology
Universidad Complutense de Madrid
Tfno y fax: +34-1-518-5799
e-mail: fling01@emducms1.sis.ucm.es

ABSTRACT

This paper is an attempt to specify the discourse-semantics of grammatical Theme for multilingual text generation. In order to do so, we use previous empirical investigations, which have specified its higher-level contextual sources of control (Lavid 1994), to propose a discourse-semantic resource -chaining strategy- which will be useful for thematic control at the text-planning level. However, the discourse-semantic specifications provided at this level are insufficient to control the fine-grained distinctions available at the lexicogrammatical level in different languages. In order to bridge this gap a preliminary interface is presented here in the form of a system network which abstracts from the lexicogrammatical options available in different computational grammars and realizes text-building categories.

Keywords: multilingual text generation, textual resources, sentence planning

---

[1] The type of architecture adopted in this paper is based on the 'vertical methodology' (Bateman 1993) used in the Esprit BR DANDELION 6665 project, where the author contributed as team leader of the Madrid site. The work reported in this paper is partly based on her contribution to this project.

1 INTRODUCTION

A notorious problem in many full-scale generators is the currently existing gap -the *generation gap* (Meteer 1992)- between text planning (strategic generation) and lexico-grammatical expression (tactical generation). The output quality of many generators is compromised because the text planner cannot exercise sufficient control of the fine-grained distinctions available in the grammar. This is partly due to a general underestimation of the complexity of the linguistic resources needed for the control of tactical generation. This paper tries to bridge the descriptive gap in the area of textual meaning by exploring the discourse-semantics of grammatical Theme for multilingual text generation.

In the following section we briefly sketch the theoretical framework from which we depart and present the discourse-semantic resources of textual meaning which a text-planner can fruitfully use to exercise control at this level. Section 3 focuses on the discourse-semantics of theme and its (con)textual motivations. Section 4 outlines the options for grammatical theme in English, German and Spanish as specified in existing computational grammars and other computational descriptions. Section 5 proposes a semantic interface in the form of a system network which will function as an interlevel between lexicogrammar and discourse-semantics. This interface can be used as a resource by a sentence planner to exercise control on grammatical Theme.

2 THEORETICAL FRAMEWORK

Our theoretical approach to the study of theme is a functional-stratificational one. More specifically, following systemic-functional linguistic (SFL) theory (cf. Halliday 1985; Martin 1992), we view language as a resource for making meaning, which is both metafunctionally and stratally organized. Thus meaning is metafunctionally organized into ideational, interpersonal and textual functional components which are manifestations of very general uses of language (cf. Matthiessen & Bateman 1991). Meaning is also stratally organized into subsystems forming different levels of abstraction: discourse, semantics, lexicogrammar and phonology/graphology. This stratal organization is based on abstraction (higher strata are abstractions of lower ones) and realization (a stratum is realized by the next stratum directly below). The highest level of abstraction is represented by the *discourse context*, viewed here from a sociosemantic perspective as a connotative semiotic including two communication planes: genre (context of culture) and register (context of situation)(cf. Martin 1992). This is the most abstract level of abstraction at which registers, generic discourse purposes, etc. are specified. The relationship between the higher-level discourse parameters and the lower-level of grammatical abstraction is mediated by a semantic interface -the *discourse semantics* stratum- which generalises across grammatical resources. Table

TOURING FROM CARDIFF

Here we shall concentrate on the fascinating variety of attractions withing a hour' s drive of Cardiff – too numerous to list in full so we offer you a sample:

Immediately to the NORTH of Cardiff lie the industrial valleys of South Wales made famous throughout the world by films such as How Green was my valley. Here small working communities lie hemmed in by green hills. Here the archeology of the industrial revolution abounds in vivid contrast to the breath-taking natural beauty you enjoy when negotiating the steep winding roads that go "over the top" from one valley to the next, such as the road over the Rhigos, from Maerdy to Aberdare. Beyond the valleys lie the mountains of the Brecon Beacons National Park. Here the sheep predominate. Man-made lakes that supply water to the population of South East Wales lie between the mountains. Stopping in the car park at Storey Arms one can relax and enjoy the beautiful surroundings or, more energetically, follow the footpath to the top of 2,907 ft high Penyfan. Visit the mountain centre at Libanus or the interpretive centre in the Garwnant Forest, for an insight into the wildlife of the area.

Immediate to the WEST of Cardiff lies the Vale of Glamorgan with its villages, leafy lanes, farmland and heritage coastline. Seaside resorts start just beyond the city boundary with the Edwardian-style resort of Penarth, complete with pebble beach and pier. A few miles west lies the livelier type of resort, Barry Island, with all the joys of the pleasure park and a wide, sandy beach. And beyond lie more beaches-Fontygary, Llantwit Major, Southerndown, Ogmore. In the Vale of Glamorgan you will find a variety of seaside catering for most tastes. Behind those beaches lie unspoilt country with historic villages and delightful pubs. Well within the hour' s drive lies Porthcawl another lively resort complete with funfair.
Just about that hour' s drive away to the west lie some of Britain' s loveliest coastline, the Gower Peninsula, with its wide sandy beaches and great headlands.
To the EAST we have the beautiful wooded Wye Valley, with the remains of a famous abbey on the banks of the river at Tintern, the Roman remains of Caerleon, Caerwent, Raglan Castle, Monmouth and the Forest of Dean.

(The Cardiff Guide, Cardiff City Council)

The text is steered along a spatial line of development (*spatial chaining strategy*) which consists of references to different locations with respect to a central place of observation and to references to spatial locations indicating a trajectory through which the reader travels. The spatial chaining strategy is signalled most of the times by the selection of locative themes (underlined in the text) to mark the reference to a new point in a spatial line: *Immediately to the NORTH of Cardiff; Beyond the valleys; Immediately to the WEST of Cardiff; A few miles west; and beyond; Behind those beaches; Well within the hour's drive; Just about that hour's drive away to the west.*

After an initial empirical analysis of 60 English texts of different discourse types, later extended to 60 Spanish discourses, Lavid (1994, 1997a) found different types of chaining strategies which showed statistically significant correlations between the chaining strategy selected by the writer to organize his/her text and the type of theme selected to signal the selected chaining strategy[2] . Table 2 below illustrates these characteristic correlations:

---

[2] Lavid's categories and analysis methodology have also been applied to the analysis of German and Spanish texts, yielding similar correlations between chaining strategies and types of themes selected (see Villiger 1996; Lavid 97a).

1 below illustrates the relationship between the contextual dimensions of the communicative situation -which together constitute the register- and the categories of the linguistic system. Discourse purposes -belonging to the more abstract generic plane- cut accross the different register dimensions, thus not being represented in the figure.

Table 1: Register and metafunction in relation to discourse semantic and lexicogrammatical systems (adapted from Martin 1992: 403)

REGISTER	DISCOURSE SEMANTICS	LEXICOGRAMMAR
Tenor (Relationship between participants)	Interpersonal Metafunction NEGOTIATION	Clause: Mood; Ellipsis Modalization; Modulation Polarity; Tagging; Vocation
Mode (Language role, medium & channel)	Textual Metafunction IDENTIFICATION CHAINING STRATEGY + FOCUS	Nominal Group: Deixis; Pronouns; Proper Names; etc. Clause: Theme; Voice; Culmination; Internal matter
Field (Experiential domain)	Logical Metafunction CONJUNCTION Experiential Metafunction IDEATION	Clause Complex Logico-Sem. & Interdependency Clause: Transitivity; lexis; group rank experiential grammar;

As the figure illustrates, the discourse semantic stratum is metafunctionally organized into four central discourse systems interfacing between the lexicogrammatical resources and the organisation of context into the register variables of tenor, mode and field. These four central systems are *Negotiation, Identification, Conjunction,* and *Ideation.* To these systems studied in Martin (1992), I propose to add two more resources -*Chaining Strategy* and *Focus*-- which, together with *Identification*, consider the discourse semantics of textual meaning, thus interfacing between the textual grammatical resources of Theme, Voice and Culmination and their higher-level contextual motivations. I will concentrate on *Chaining strategies* and *Theme* in the following sections.

3 THEME, CHAINING STRATEGIES AND DISCOURSE TYPES

According to Lavid (1994), *chaining strategies* are "devices used by the writer to steer his/her text along a specific line of development or frame with the purpose of achieving a maximally profitable text organization, in view of the discourse purpose and the subject matter." Chaining strategies include temporal and spatial developments, general to specific, sequential, through-argument or counter-argument patterns, rhetorical patterns centering around participants, or points of contrast, enumeration, elaboration, etc... The textual role of *Theme* in these developments is to act as a signpost for the reader, as a guide along a specific line of development selected by the writer. Example (1) below illustrates this guiding function of Theme in discourse:

Table 2: Chaining strategies and thematic selection

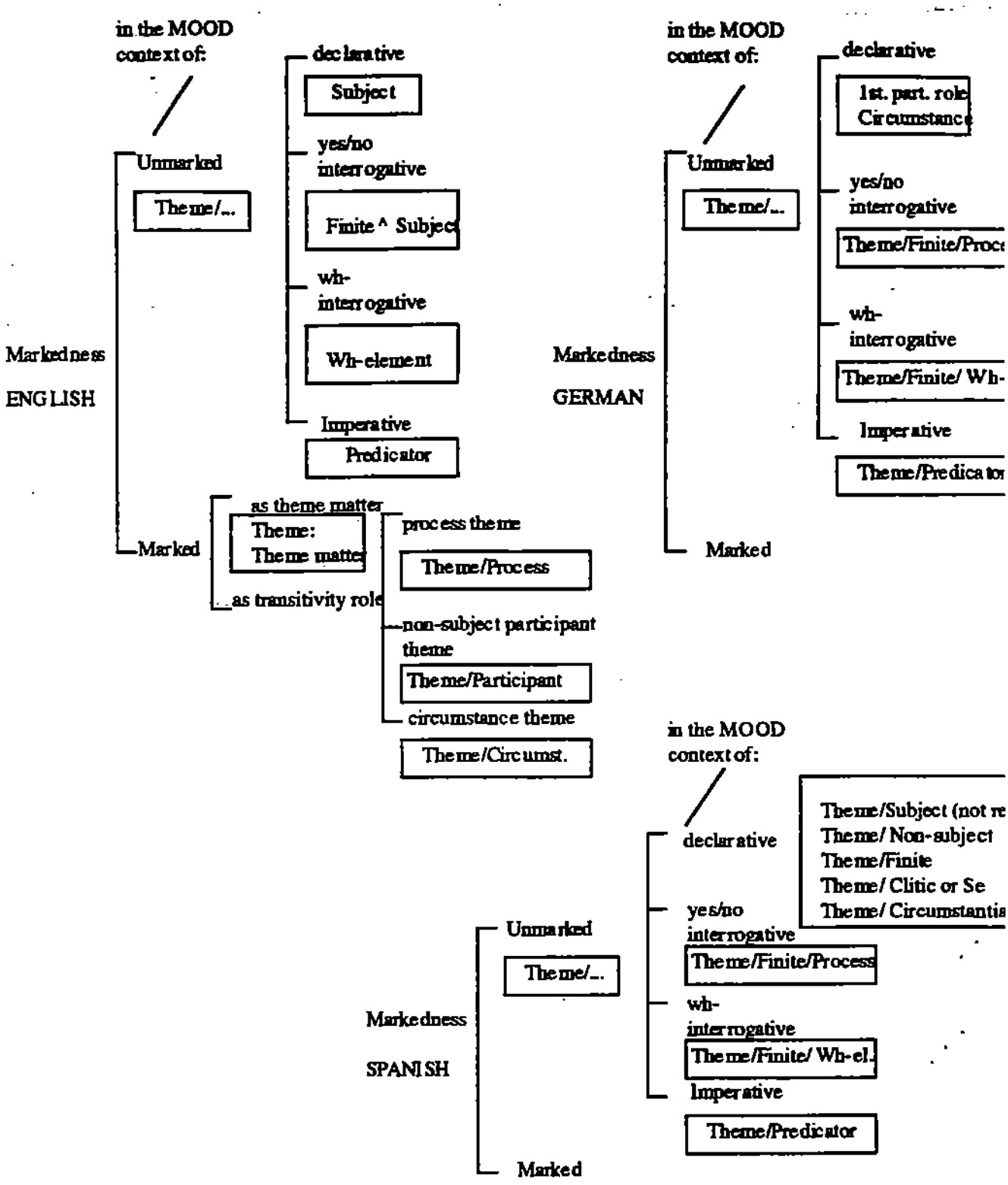
CHAINING STRATEGY	THEMATIC SELECTION
Characterization	Participant
Spatial	Location
Temporal	Time
Sequential	Time Process
Through- or counter-argument	Text-builder

These characteristic correlations can be fruitfully used by a text planner to control the discourse-semantics of textual meaning in a text, i.e., to specify the kind of chaining strategy and thematic selection which will characterize a given text. However, a more fine-grained type of specification is necessary to control theme at the lexicogrammatical level, as shown by the grammatical theme options available in several computational grammars (i.e. the Nigel component of the Penman generation system, the grammar fragment of the Komet-Penman generator). In section 5 an attempt is made to provide such a specification by presenting presenting a semantic interface which mediates between the discourse-semantics and the lexico-grammatical strata. But before presenting such interface, let us consider the grammatical systems available for Theme in English, German and Spanish.

4 THE GRAMMAR OF THEME FOR ENGLISH, GERMAN AND SPANISH

This section presents the grammatical Theme options for English, German and Spanish, as specified in different computational resources. The specification provided for English is a fragment of the very large computational grammar of English described in (Matthiessen 1995). The specification for German is also an adaptation of the current implementation provided in the Komet-Penman grammar (see Ramm et al. 1995). The specification provided for Spanish is part of ongoing work for a systemic-functional computational grammar of Spanish (Lavid 1997b). Figure 1 illustrates these systems:

Figure 1: Grammatical Theme systems for English, German and Spanish



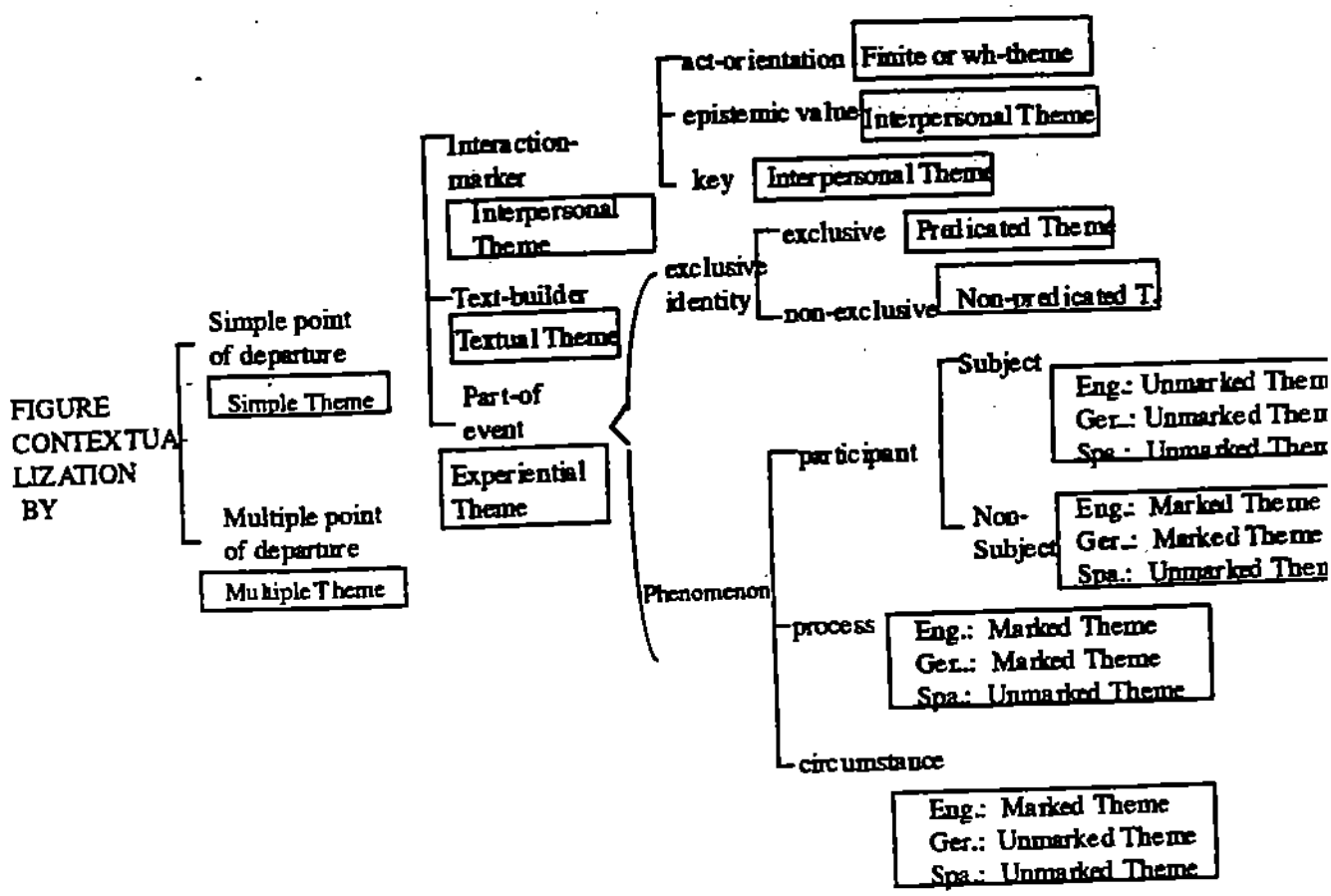
As the network illustrates, the range of unmarked options for theme selection in declarative sentences in Spanish is much wider than it is in English or in German. In English, for example, theme is unmarked only when it coincides with the subject of the clause, all the other options being marked. By contrast, in Spanish both the subject (when not recoverable)[3] , the Finite element, clitics, circumstantials and any non-subject participant of the clause will be considered as unmarked theme. This is probably due to the fact that Spanish has a relatively free word order in comparison with English.

6. TOWARDS A SEMANTIC INTERFACE FOR GRAMMATICAL THEME

The grammatical specifications for Theme presented in the preceding sections provide us with Theme options of different languages at the lexicogrammatical level. These options cannot be controlled directly by the output of a text planner, which, as explained in section 3 above, would just tell us whether the theme selected as signal of a given chaining strategy is locative, temporal, or topical. In order to control the thematic options available in the grammar, we need an interface which mediates between the grammatical system and its discourse semantics and which recognizes the types of information which the discourse resources will preselect. Such a system will function as an *interlevel* (cf. Halliday et al. 1964) between lexico-grammar and discourse semantics abstracting from lexicogrammatical and realizing text building categories. Figure 2 below presents a preliminary specification of such a system.

[3] When the subject is recoverable from the co(n) text, its presence in the clause is a marked choice on the part of the speaker.

Figure 2: Theme Semantics



Using the previous characterization of Theme as that element of the message which serves as its point of departure, which allows the speaker to manipulate the *local context* of the clause, we can specify that point of departure as *simple*, i.e., embracing only one of *element-of-process*, *text-builder*, *interaction-marker*, *part-of-event*, or it may be *multiple*, i.e. embracing one of several-parts-of-event, other-multiple-contextualization. The grammatical realization in the first case is *simple Theme*, in the second case *multiple Theme*. When the point of departure is simple, it can be further classified metafunctionally into the following options:

- Interaction-marker. The lexicogrammatical realization of this is interpersonal theme.
- Text builder. The lexicogrammatical realization of this would be a textual theme.
- Part of event. The realization would be an experiential theme.

If we move one step further in delicacy, the semantic choices under *part-of event* include *exclusive identity* and *phenomenon*. When an *exclusive identity* is assigned to the point of departure, the grammatical realization is a *predicated theme*. If the choice is *non-exclusive*, then the realization is *non-predicated theme*. When a *phenomenon* is chosen as a point of departure, the major option

are *participant, process* and *circumstance*. Under the semantic type *participant*, the options are *1st participant subject* or *participant non-subject* .The grammatical realization of making the 1st participant subject a Theme is an *unmarked theme* in English, German and Spanish. The realization of making a non-subject participant the Theme is a *marked theme* in English and German but *unmarked* theme in Spanish. When the phenomenon is a *process* rather than a participant the grammatical realization is a *marked theme* in English and German but *unmarked theme* in Spanish. When the phenomenon is a *circumstance* of some kind (location, time, matter, accompaniment, manner, role, etc..) the grammatical realization is a *marked theme* in English, but *unmarked theme* in German and in Spanish.

As we can see, the choices made at this semantic level are language-independent while their realizations at the grammatical level are language-dependent. Therefore, the introduction of this semantic interface has important advantages for multilingual generation architectures which recognize the complexity of the different types of information necessary to generate texts in different languages. As the network illustrates, the categories used here are functional-semantic ones, abstracted from various interrelated areas of the grammar so that they can interface between grammatical categories and their (con)textual motivations. This seems to be a necessary stratum in order to avoid the methodological gap between high-level sources of control and lexicogrammatical choices.

5 CONCLUDING REMARKS

In this paper we have attempted to reduce the descriptive gap in the area of textual meaning as specified in existing computational grammars of English, German and the one under construction for Spanish by providing a preliminary specification of the discourse-semantics of grammatical theme in English, German and Spanish . This preliminary specification is, in our view, a necessary step for exercising adequate control in this area, and an urgent task for the pragmatically-motivated generation of different textual variants which express the same propositional content. Further work, both linguistic and computational, will have to be performed to extend and refine the results of this initial investigation. A crucial step, however, is to adopt a generation architecture based on a stratified approach to language where descriptive responsibilities are distributed across several levels.

REFERENCES

(Halliday et al. 1964) M.A.K. Halliday, A. McIntosh, and Peter Strevens. *The linguistic sciences and language teaching*. Longman, London.

(Halliday 1985) M.A.K. Halliday. *An introduction to functional grammar*. London: Edward Arnold, 1985.

(Lavid 1994) Julia Lavid. Theme, discourse topic and information structuring. Technical report, Universidad Complutense de Madrid, Madrid, 1994. (ESPRIT BR Project Dandelion, EP6665; Deliverable R1.2.2.b.)

(Lavid 1997a) The relevance of corpus-based research for contrastive linguistic and computational studies: thematization as an example. Paper presented at the *Workshop on Corpora in Semantic and Pragmatic Research*. Instituto de Lingüística Aplicada. Universidad Pompeu Fabra. Barcelona, 12 Abril 1997.

(Lavid 1997b) Building textual resources for multilingual text generation. Working paper.

(Martin 1992) James R. Martin. *English Text: System and Structure*. John Benjamins, Amsterdam.

(Matthiessen 1995) C. M.I.M. Matthiessen. *Lexicogrammatical cartography: English systems*. International Language Sciences Publishers. Tokyo.

(Matthiessen and Bateman 1991) C.M.I.M. Matthiessen and John Bateman. *Text generation and systemic functional linguistics: experiences from English and Japanese*. Frances Pinter, London.

(Meteer 1992) M.W. Meteer. *Expressibility and the problem of efficient text planning*. Pinter, London.

(Ramm et al. 1995) W. Ramm, A. Rothkegel, E. Steiner, and C. Villiger. *Discourse Grammar for German*. Deliverable R2.3.2 of WP 2 'Grammar Integration', ESPRIT Basic Research Project 6665 DANDELION, University of Saarland, Saarbrücken, October 1995.

(Villiger, C. 1996) Claudia Villiger. Theme, discourse topic and information structuring in German texts. Universität Saarbrücken.(ESPRIT BR Project Dandelion, EP6665; Deliverable R1.2.2.b.)