

Generating thematic choices for multilingual text generation*

Julia Lavid

Dpt. English Philology, Faculty of Philology, Universidad Complutense de Madrid, 28040 Madrid, Spain

KEYWORDS: multilingual text generation, pragmatic adequacy, multilingual methodologies

E-MAIL: fling01@sis.ucm.es

FAX NUMBER: +34-1-394-5396

PHONE NUMBER: +34-1-518-5799

* This work is supported by the Commission of the European Communities under the project LRE 062-09 GIST (Generating Instructional Text)

Abstract

The present paper addresses two important issues for generating pragmatically adequate administrative forms in the context of GIST, a multilingual text generation system for the automatic production of administrative forms in English, German and Italian. First, it investigates the influence of socio-cultural factors (such as the social distance and the social role) on the mood structures preferred by each language when requesting action from users. Given a specific mood choice, the influence of pragmatic constraints (such as weight, identifiability, and topicality) is also explored to account for the different thematic realizations found in the corpus. Second, it addresses issues of computational specification for a multilingual generation architecture by using mechanisms exploited in a number of diverse multilingual generation projects.

1 Introduction

The GIST project addresses the development of a multilingual generation system for the automatic production of administrative forms in three different languages: English, German and Italian. In this context, it is part of our task to provide a contrastive study and computational specification of the preferred thematic realization options which each language selects to express the information drawn from a common application domain dealing mainly with pension and family benefits, unemployment and disability allowances.

Careful analysis of the forms has shown that thematic realizations occur in different mood contexts which each language prefers to express the same speech function. For example, when reque-

sting action from users, English forms prefer to thematize the Predicator element (unmarked theme in imperative clauses) while Italian and German prefer to thematize participants functioning as Affected or Agent in the transitivity structure of the clause. Example 1, taken from a bilingual Italian/English form, illustrates this contrast (Themes are underlined):

Example 1:

Italian: Il presente formulario, debitamente compilato, deve essere presentato o inviato al più presto possibile alla Sede provinciale dell' Istituto Nazionale della Previdenza Sociale (I.N.P.S) competente per territorio.

English: Send or take the completed form to your I.N.P.S local office as soon as possible.

These unmarked realizations, very frequent in our multilingual corpus, are a clear indication that thematic realization options can be partially controlled by the preferred mood selections which each language chooses to express the same speech function. For example, in English, unmarked theme realizations will be Subject in the declarative clauses, wh-element in wh-interrogative clauses, and the Predicator in imperative clauses. These realizations are language-dependent, i.e., what is unmarked theme in English in a given mood context is not necessarily the same in Italian and German. However, as will be shown below, thematic realization is not fully determined by mood, but is the result of the interplay of different factors which impose a specific thematic structure on the message.

Being aware of these facts, we have opted for an analysis methodology which will allow us to contrast the language-dependent thematic realizations occurring in specific mood contexts under language-specific interpersonal selections. We have concentrated on the thematic realizations which each language prefers when expressing requests, a speech act which predominates in administrative forms.

The paper is organized as follows. First, we describe the multilingual corpus used for the study and explain the methodology used for its analysis. Second, we suggest how specific socio-cultural selections (such as social distance and social role options) preselect the mood structures where thematic realizations occur. Third, we present the language-specific thematic realizations found in requests, and present possible motivating factors which contribute to control them. These are then captured as inquiries which control the different options in each of the proposed language-specific thematic networks.

2 Corpus Analysis

The GIST corpus of English forms contains approximately 25,000 words and comprises eight forms, all of them issued by the Department of Social Security (DSS). The Italian/German corpus contains 36,248 words and comprises 12 bilingual (Italian/German) and 2 bilingual (Italian/English) forms. The documents are produced in Italian by INPS (Istituto Nazionale Previdenza Sociale) and adapted to include a German translation at PAB (Provincia Autonoma di Bolzano), the local government of the province of Bolzano.

The analysis involved addressing the following issues:

1. Determining the interpersonal context in which thematic realization variants occur. This involved identifying the social distance and the social role options which characterize each culture.
2. Determining the preferred mood structures which each language uses to realize requests; dividing them into direct and indirect requests depending on the selected mood structure, and establishing the probability of occurrence of one type or another in each language under study.
3. Determining the thematic realization forms selected by each language in the mood structures identified in 2.
4. Determining the pragmatic factors which constrain the selection of a given thematic realization occurring in a specific mood context.

As explained above, this paper concentrates on a specific class of the basic speech function of demanding goods-and-services, namely, on requests, and the preferred thematic realization options which each language chooses in administrative forms. However, the proposed methodology has also been employed for the comparison of preferred thematic realizations of other speech functions which frequently appear in administrative forms.

3 Interpersonal options in administrative forms

Determining the interpersonal context in which thematic realizations occur involved identifying the social distance and the social role options which characterize each culture in the administrative setting where the forms are produced. The English forms have been written according to the DDU (Document Design Unit) guidelines and are the result of extensive user testing to achieve a warm and non-intimidating tone. These forms tend to minimize the social distance and to level out the social roles between the administration and the users, and this is achieved through a number

of devices, among them by the direct expression of requests by means of the imperative mood. In contrast, both the Italian/German forms, produced in a different socio-cultural setting, where the social distance is maximal and the social roles are hierarchic, prefer to use indirect strategies for the realization of requests, thus opting for the declarative mood. Figure 1 illustrates these language-specific preferences

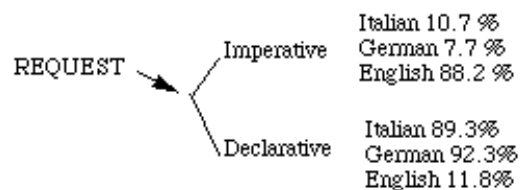


Figure 1: Language-specific preferences in the realization of requests in the GIST corpus

While the English forms prefer the imperative mood in 88.2 % of the cases and the declarative only in 11.8%, the Italian/German forms opt for the declarative in 89.3% and 92.3% of the cases, respectively, selecting the imperative in only 10.7% and 7.7% of the cases. The difference in these proportionalities is statistically significant ($p < 0.001$).

4 From mood structures to thematic realizations

As we explained above, thematic realizations always occur in specific mood contexts, which can be direct or indirect realizations of a given speech function. Table 1 below illustrates the range of thematic realizations selected by each language and their frequency distribution when requesting action from users.

Table 1: Distribution of thematic realizations of direct and indirect requests in the three languages

Italian				German				English			
Indirect		Direct		Indirect		Direct		Indirect		Direct	
S / Agent	19	Predicator	6	S / Agent	20	Predicator	6	S / You	28	Predicator	124
S/ Aff.	51	Cond.Cl.	10	S/ Aff	52	Circumst.	10	S / We	5	Please+P	209
Od/Aff	3	Circumst.	12	Od/Aff	2	Cond. Cl	3	S / NG	6	Do not+P	19
Circumst.	65	Od/Aff	3	Circumst.	53	Inter. T	1	S / They	1	Cond. Cl	89
Finite	24	Temp. Cl	2	S / Es	14			Od/ Aff.	1	Temp. Cl	16
S/ Carrier	3			Cond. Cl	29			Cond.Cl	18	Locat Cl	2
Temp.Cl	5			Temp. Cl	3			Temp. Cl	5	Final Cl	2
Cond.Cl	32			Finite	3			Conc. Cl	1	Conc. Cl	1
Final Cl	1			Textual	7					Circumst.	24
Textual T	5			NF	3						
Neg+ Fin	1			Neg.+NF	1						
Si + Fin	1			Neg +Aj.	1						
Infinitive	48			Neg+ Infi	1						
Da+ Infi	16			Od + Infi.	41						
Neg+Infi.	3			Infinitive	13						
TOTAL	278	TOTAL	33	TOTAL	243	TOTAL	20	TOTAL	65	TOTAL	486
%	89.3	%	10.7	%	92.3	%	7.7	%	11.8	%	88.2

S= Subject; Od= Direct Object; P= Predicator; Aff.= Affected; Circumst.=Circumstance; Temp. Cl= Temp. Clause; Cond.Cl= Conditional Clause; Locat.Cl= Locative Clause; Final Cl= Final Clause; Conc. Cl= Concessive Clause; Textual T= Textual Theme; Inter. T= Interpersonal Theme; Fin.= Finite; NF= NonFinite; Infi= Infinitive.

As explained above, English prefers the direct expression of requests by means of the imperative mood (88.2%). This fact is reflected in the preferred thematic realization options selected: typically, English administrative forms select the interpersonal modal “please” followed by the Predicator element in an imperative clause type. This option occurs in 209 clauses from a total of 486, thus constituting the 43 % of all direct thematic realizations in the English corpus. Indirect realizations are much more unfrequent (11.8%) than direct ones, with a predominance of personal forms; among them, the second person pronoun “you” functioning as Subject in declarative clauses is the most widely used when referring to the form-filler, and the first person plural pronoun “we” when referring to the source of the form. By contrast, German and Italian prefer indirect strategies for the expression of requests due to culture-specific requirements where the social distance and the social role relationships between the administration and the users is maximal and hierarchic. This fact is reflected in their preferred thematic realizations. Leaving aside the circumstantial elements, the conditional clauses, and the

realizations by means of an infinitive – with a high number of occurrences in both corpora for reasons which we will explain below, both languages tend to thematize the Affected (Patient) element as Subject in independent declarative clauses, thus creating a passive construction. It is well-known that passive constructions contribute to create a distant, impersonal impression on the reader by not encoding agents as subjects, thus obscuring this ‘natural’ expected mapping. The influence of these interpersonal factors is even more apparent in those (highly frequent) cases where the thematized Affected element is a nominalization, where the action is hidden out as a noun and there is no mention of the participant involved. Also, when the Subject is the Agent, all thematic realizations are expressed by Nominal Groups, rather than by personal pronouns referring to the addressee which contributes to an impersonal and distant tone.

The high percentage of infinitives both in Italian (some of them preceded by “da”) and their translation into German (most of them preceded by the Direct Object) is, again, an indication of the tendency to use impersonal non-finite forms at the

beginning of sections where the writer expects an immediate action from the reader and does not need to provide him with details about the filling of the form.

Other significant realizations captured on table 1 can be explained by the influence of varied factors, among which we have isolated two: *weight* and *topicality*. These two factors combined explain the high percentages of conditional clauses and circumstantial elements in thematic position.

Topicality is the tendency to present in first initial position those elements which are 'topical' in the sense of shared by the speaker and the hearer in the discourse context. Careful analysis of the initial conditionals in the three corpora has shown that, in general, they function as a premise, as material that the addressee is expected to take as given (see Lavid & Taboada, 1994). The high percentage of circumstantial elements, most of which express condition realized by prepositional groups in Italian and in German, is due to the same factor. Most of these circumstantial elements pick up information which has been mentioned before in the discourse, thus preserving topic continuity and contributing to smooth information flow.

Weight is the principle according to which long, "heavy" constituents tend to come late in the clause (Leech, 1983: 65). In the case of Italian, this results in the thematization of the Finite element of the Verbal Group, while the long Subject is delayed to the end of the clause. This factor accounts for 24 occurrences in the Italian forms (11.3 %). The German translations use the rather infrequent construction introduced by the particle "Es" to imitate the Italian structure in 7.4 % per cent of the cases.

5 Representation

As illustrated in the previous sections, the combination of pragmatic factors such as weight and topicality, on the one hand, and the social distance and social role selections preferred by each culture, on the other, control language-dependent thematic realizations in our multilingual corpus.

Using the chooser-and-inquiry framework developed by Mann and others (Mann, 1983) and explained elsewhere (Matthiessen & Bateman, 1991), we propose to represent these factors as inquiries (semantic choices) associated with features in each of the language-specific thematic networks¹ which we present in Figure 2, next page.

The English theme markedness options are the ones used in NIGEL, the systemic-functional grammar component implemented in the PENMAN System (Matthiessen, 1993). The German options are an adaptation of the account presented by Steiner and Ramm (forthcoming). The Italian options are the result of our own ongoing account

for this language. Interestingly, as the network indicates, marked thematic options in English are no longer dependent on mood but on transitivity. In Italian and in German, marked thematic options are not restricted to a limited number of transitivity roles which are fronted in the structure of the clause, but include a wide variety of clause ordering combinations which mark its whole structure as marked. As an example of how choices could be controlled by a specification of different factors from the communicative situation, we propose the following inquiries which control Italian unmarked thematic choices in the declarative mood context :

Choice:	UNMARKED as either SUBJECT or FINITE or CIRCUMSTANCE
Inquiries (-id):	Is the Subject identifiable, and is there more than one participant in the process?
Then:	conflate THEME with Subject
Inquiry (-id):	Is the Subject unidentifiable, or too long, or is there only one participant in the process?
Then:	conflate THEME with Finite
Inquiry (-id)	Is the current topic a time or space description of the current situation?
	conflate THEME with CIRCUMSTANCE
Inquiry (-id)	Is the current topic a participant recoverable from the co(n)textual situation?
	conflate THEME with CLITIC element

We hope this to be sufficient as an illustration of how different contextual factors represented as inquiries can constrain thematic choices in the markedness options of each of the language-specific thematic possibilities. While these factors are common to the three languages, their interaction with thematic realizations is language-specific.

5 Summary and Conclusion

This paper has employed an intensive corpus analysis to identify the different thematic realizations which three different languages (German, Italian, English) select when requesting action from users in administrative forms. The analysis has shown that thematic realizations are partially determined by the preferred mood structures which each language selects on the basis of culturally-conditioned interpersonal factors such as the social distance and the social role between the interactants. For the purposes of multilingual text generation in the application context of the GIST system, we have proposed to represent the language-specific mood structures which realize direct and indirect requests as options within a system with percentages annotating the preferred selections. These

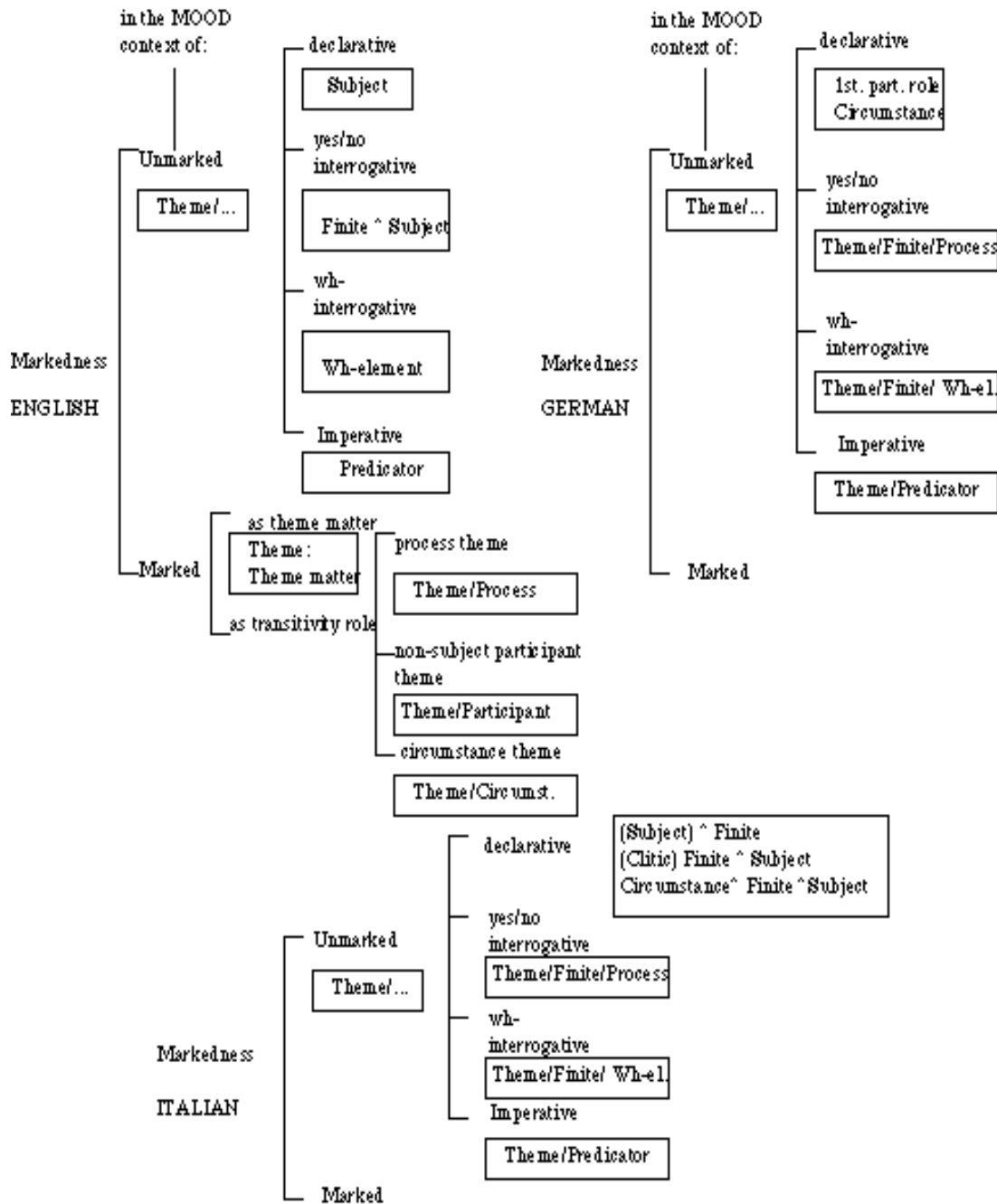


Figure 2: Theme markedness in English, German and Italian.

mood structures constitute the context where thematic realizations can be further specified by representing motivating factors from the communicative context as inquiries which control choices for each of the three language-specific thematic networks.

This paper, therefore, has presented a useful methodology for the generation of thematic choices in requests which can be fruitfully applied to other speech acts, such as questions and statements,

which are also common in administrative forms. In fact, the current text structurer of the GIST system is using the proposed specification for the multilingual generation of mood structures and thematic choices in these three types of speech acts, thus validating the applicability of this method in a practical computational scenario.

Note

1. The use of the system network and its translation into (typed) feature structures and of the inquiry interface is well understood in computational systemic-functional accounts, being a common representational means in PEN-MAN-style generators used in a number of diverse multilingual generation projects. These include: DRAFTER (British EPSRC Project J19221) (Paris et al., 1995), KOMET (GMD/IPSI) (Bateman and Teich, forthcoming), TECHDOC (Rösner and Stede, 1991), DANDELION (EP6665), and GIST (European LRE Project 062-009) (Not and Stock, 1994).

References

- Halliday, M.A.K. (1976). *System and Function in Language*. Oxford University Press, London. edited by G.R. Kress.
- Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. Edward Arnold, London.
- Lavid, J. and Taboada, M.T. (1994) *Specification of Evaluation Criteria for Text Quality*. GIST Internal Report INT-12. Universidad Complutense de Madrid, September, 1994.
- Leech, G.N. (1983) *Principles of Pragmatics*. Longman, London.
- Matthiessen, C. M.I.M. (1992) *Lexicogrammatical cartography: English systems*. Technical report, University of Sydney, Linguistics Department, 1992. Ongoing expanding draft.
- Matthiessen, C. M.I.M. and Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York.
- Mann, W. C. (1983). *An overview of the the Penman text generation system*. Technical Report ISI/RR-83-114, USC/Information Sciences Institute, Marina del Rey, CA, 1983.
- Steiner, E. and Ramm, W. (forthcoming). On Theme as a grammatical notion for German. *Functions of Language*.