LINGUISTIC AND COMPUTATIONAL APPROACHES TO INFORMATION IN DISCOURSE: THEME, FOCUS, GIVEN, AND OTHER DANGEROUS THINGS

Julia Lavid López
Universidad Complutense de Madrid

ABSTRACT

This paper addresses the information communication problem faced by interlocutors during the process of text production and understanding. Different processes govern the flow of information in discourse, such as the process of content selection, the process of information distribution, and the process of thematic organization.These processes have to be encoded by the text producer into a single linear structure to be decoded by the text receiver in the complementary process of discourse comprehension. Different disciplines, such as Linguistics and Artificial Intelligence, have studied these phenomena from different perspectives, thus developing different, but related categories, with overlapping meanings in many occasions, for the study of information structure in discourse. This paper reviews the commonalities and the different uses of categories such as *topic*, *theme*, *given/new*, and *focus* in both disciplines, and provides operational redefinitions for these notions in an attempt to integrate them into a unified account of information structure in discourse which might be useful for both descriptive and computational purposes.

1. INTRODUCTION: THE INFORMATION COMMUNICATION PROBLEM

In order to introduce the wide of issue of information in discourse, I would like to present a very simplified model of discourse production, which, in my view, reflects the information communication problem faced by the speaker during the process of text production.[1] This outline will help understand the different processes that govern the flow of information in discourse:

1.- The speaker has in mind a particular mental representation of some event or other subject matter, which he intends the hearer to create during the complementary process of discourse comprehension (the topic of discourse).
2.- To help the addressee create his mental representation, the speaker creates a text representation (linear sequence of utterances) out of a non-linear structure.
3.- A coherent text representation must meet, among others, the following needs:
   (a) indicating attachment point (i.e., how to connect preceding and following information = *bridge* or *theme*)
   (b) indicating principal inference point (= *focus*)
   (c) indicating overall communicative goal content (= *topic*)

As a result of these needs, the speaker performs a series of processes which can be provisionally described as follows:

a) The process of content selection: the speaker' selection of some subpart of the information contained in the knowledge base during discourse production.
a) The process of information distribution: the speaker's estimates of the listener's familiarity with the subject matter. This process is responsible for the assignment of different textual statuses (Given / New) to elements in discourse.
b) The process of thematic organization: the speaker's attempts to help the listener appreciate some particular point of view toward the information contained in the knowledge base. This process is responsible for the division of messages into thematic and rhematic sections.
c) The process of information presentation: the speaker's adjustment to the syntactic requirements imposed by each language, and the use of cohesive mechanisms to produce smooth text flow.

As we can see, the information communication problem manifests itself through a series of processes which the speaker performs during discourse production in order to convey and negotiate meaning to the addressee during the complementary process of discourse comprehension. This paper concentrates on the resources that both Linguistics and Artificial Intelligence have brought to address the problem, and on how the results obtained in both fields have provided only partial solutions which need to be redefined and operationalized to speed up the investigation of this burning issue in discourse studies. With this aim in mind, the first section provides an overview of the different terms and definitions used in Linguistics and in Artificial Intelligence to extract their commonalities and divergences. The next section provides operational redefinitions for the notions of focus and theme from a discourse perspective, in an attempt to integrate the findings in both research fields. The last section concentrates on the progression of *focus* and *theme* and summarizes the empirical results obtained by recent interdisciplinary studies.

2. OVERVIEW OF TERMS AND DEFINITIONS

In order to account for the some of the processes outlined above, researchers from different fields have used different terms, such as *topic*, *focus*, *theme* and *rheme*, *given* and *new*, with such ranges of overlap and lack of cross-reading that anyone who starts reading the literature experiences a certain amount of confusion. To clarify the existing terminological cocktail, I propose to assign each of the notions studied in both fields to the different processes outlined above, mapping out the functionalities and extracting the commonalities among concepts, so that the linguistic and computational communities have a better understanding of these interrelated phenomena. This pursuit seems to me a useful step towards a fruitful cooperation between both fields. I will start with the notion of *focus*.

2.1 THE NOTION OF FOCUS IN LINGUISTICS

The term *focus* has been used for two different and opposite notions in Linguistics and in Artificial Intelligence (Hajicova 1987:311). In Linguistics, the term was first introduced by Halliday (1967) and Chomsky (1971), and later used by different linguists (Sgall 1979, Sgall et al. 1973) to specify that part of the sentence that conveys some irrecoverable information predicating something about the >given=, recoverable, contextually bound part, that is the *topic*, the *theme*, or *the ground* of the sentence (Halliday 1967; Chomsky 1971; Sgall 1979; Sgall et al. 1973; Vallduví 1994). This dichotomy of topic-focus was also called *topic-comment* (Bloomfield 1935; Hockett 1958) and *presupposition-focus* (Jackendoff 1972), depending on the authors. For Hockett (1958: 201), *topic* is what the speaker announces in a sentence, before proceeding to say something about it in the *comment*. Similarly, Jackendoff (1972:239-278) distinguishes *presupposition* («the information in the sentence that is assumed by the speaker to be shared by him and the hearer») from its complementary term *focus*, the latter determining stress assignment and consequently "pitch accent." The *topic* (*presupposition*) of a sentence would be roughly equivalent to *given* or recoverable information, while the term *focus* would refer to *new*.

The consideration of focus as locus of the new information that a sentence conveys was further validated by Quirk and Halliday, who accorded a central role to intonation in the structuring of information. According to Quirk et al. (1972: 94): "The focus, signalled by the intonation nucleus, indicates where the new information lies." For Halliday (1967) the information *focus* is where the main burden of the message lies, and this choice is realized by the assignment of prominence in the tone-unit. The focus of the information is that which is presumed to be New, which is made tonic; and in the unmarked case this will be realised on the last, non-anaphorical lexical item in the tone group. For example, in (1) below, the item "LAND" is the information focus of the utterance, and, thus, carries nuclear stress:

(1) // in a / far-away / LAND//

Although this is the *focus*, it does not define the extent of the *new*, which will depend upon how much of the text repeats or renews what has been previously said or assumed (by the speaker) to be known by the listener.

2.2 THE THEME/ RHEME STRUCTURE AND THE GIVEN/NEW DISTINCTIONS

In general, the notion of *givennness* has been defined in terms of different parameters such as predictability (Halliday 1967), saliency (Chafe 1976), shared knowledge

(Haviland and Clark 1974; Clark and Haviland 1977), assumed familiarity (Prince 1981), and recoverability (Geluykens 1991). More specifically, in the works of the Prague School of Linguistics (Firbas 1974), the notions of *given/new* information have been treated jointly with the *theme/rheme* structure (Firbas 1974). So, the notion of *theme* included two types of information: a) information which is known or *given* in the situation, and b) information from which the speaker proceeds. Chafe (1970: 210-211) adopted a similar position to that of the Prague School, claiming that known or "old" information is "shared information" which serves as a kind of starting point for the message.

In this sense, linguists like Halliday (1967) and Quirk et. al. (1972) established a distinction between these two notions: information from which the speaker proceeds, "which serves as the point of departure of the message" (Halliday 1967: 12; 1985: 39) is called *theme* (also defined as "what the clause is going to be about») which, together with *rheme*, structures the clause as a message. *Theme* in English is realized through initial position in the sentence, but this is just a language-particular realization. Independently from this type of structuring, Halliday studies the distribution of information into material which the speaker presents as *given* (in the sense of predictable or recoverable), and *new* (not-recoverable). Therefore, the functions of of *given* and *new* are not the same as those of *theme* and *rheme*:

> The *theme* is what I, the speaker, choose to take as my point of departure. The *given* is what you, the listener, already know about or have accessible to you. *Theme-rheme* is speaker-oriented, while *given-new* is listener-oriented (Halliday 1985: 278).

In spite of this distinction, even those linguists who advocate for the separation of these notions (e.g. Halliday 1985: 278), recognize the existence of a semantic correlation between information structure and thematic structure, such that, in most of the cases the speaker will choose the *theme* from within what is *given*, and locate the *new* within the *rheme*. Divergent cases can be found, however, i.e., cases where the above correlation is reversed as in (2), where the *theme* is *new* information whereas the material in the *rheme* is *given*; or cases where both the *theme* and the *rheme* of the message are *given*, as in (3):

(2) [No one has ever seen donkeys fly]
A young boy from Lanzarote recently saw one
Theme *Rheme*
New *Given*

(3) [Bob helps Mary's brother a lot]
He helps him too.
Theme *Rheme*
Given

2.3 The notion of focus in artificial intelligence

If we now turn to the field of Artificial Intelligence (AI) to see how these related notions have been dealt with, we will find that researchers concerned with the structure of discourse have used the term *focus* to refer to that element of the sentence on which the discourse participants center their attention as the discourse unfolds (Sidner 1983). *Focusing* is then the active process on the part of the speaker and the listener by which they concentrate their attention on some subset of their knowledge.

For example, in a sentence like "I got a really pretty turtle this weekend," the *focus* would be the turtle, because it is very likely that the participants will continue to talk about it in the next sentences (Sidner 1983:116). In general, most of the work on *focus* describes it as a working construct for tracking discourse referents, for example for the resolution of anaphora, to aid the understanding of text for a particular task (Grosz and Sidner 1985; Reichman 1981; Reichman-Adar 1984; Hobbs 1979; Linde 1979; Cohen 1987; Carberry 1983). In these approaches, *focus* appears as a constructive category, which, even though is identifiable with a particular discourse referent, it arises in the context of a discourse.

This is the reason why, unlike what is common practice in linguistics, the studies on *focus* in AI-oriented research have not paid much attention to its syntactic or phonological realization. Being a discourse category, researchers have not been concerned with identifying it with a particular sentence constituent. Why? Because it would be misleading to draw a parallellism and claiming that the *focus* —in the AI sense— of a discourse is the same as the *theme* (or topic) or the *rheme* (or comment) of a sentence. However, as the intuitive notion of *focus* and the definition of theme or topic involve the notion of "aboutness," they have been frequently treated as synonymous. The fact that they may coincide in their realization adds more confusion to the issue.

In the following example, the *focus* of (5) —in the AI sense— refers to the computer, since the computer is one of the items "just introduced" and the utterer of (4) focuses his/her attention on it. In terms of linguistic analysis, however, the pronoun 'it', referring to the computer, belongs to the *topic* or *theme* rather than to the *focus* of (5).

(4) John switched off the computer

(5) It had been on nearly all the day

The problem lies in the identification of *topic* or *theme* (defined as "what the sentence is about») with a sentence constituent, which, as long as we keep to typical declarative clauses in isolation, coincides with the Subject and appears in first initial position. However, sentences in isolation do not have *topics* per se. In example (4), is the sentence about *John*, about the *computer*, or about the *switching of the computer*? Without the following clause (5) which establishes the context of (4), it is delusive to assign a topic to (4). Therefore the notion of "aboutness," captured by the intuitive notion of *topic*, is only applicable in the wider context of a text in which a sentence is inserted.

This requirement was also observed by Schank (1977) who claimed that sentences out of context cannot be said to have a topic, because the topic arises only out of the interaction of adjacent sentences by the process of intersection. Therefore, once the notion of *topic* is removed from the domain of the sentence and redefined as a discourse category which describes what a text, or part of a text is about, it appears to refer to the same intuitive notion captured by *focus* in Artificial Intelligence.

To summarize, it appears that the concepts of *focus*, *given-new*, and *theme-rheme* describe related, albeit different phenomena, since they are the result of different processes.

Focus in Artificial Intelligence is a constructive category resulting from the process of content selection by which the discourse participants center their attention on some subpart of the information contained in the knowledge base. In Linguistics, however, *focus* is a sentence-level category, which captures contextually non-bound (irrecoverable) information predicating something about the *given,* recoverable, contextually bound part, that is the *topic* of the sentence. Therefore, it is the result of the process of information distribution which reflects the speaker's estimates of the listener's familiarity with the subject matter.

Theme is that element from which the clause "departs," according to Halliday, and, as such, it is the result of a process by which the speaker presents his material from a particular perspective or point of view. However, in the Prague School treatment of the Theme-Rheme structure, those elements within the sentence that carry the lowest degrees of communicative dynamism (i.e. which are 'given') are thematic, versus those rhematic elements which carry the communication forward and are, which convey new information.

Topic has been traditionally treated as a category which refers to the recoverable, contextually-bound part of the sentence, thus distinguishing a dichotomy of *topic* and *focus*, also called *topic-comment*, depending on the authors, as was explained above. However, as a discourse category, the notion of *topic* roughly captures the same notion studied under the label of *focus* in computational linguistics (Givón 1984: 137). Table 1 below summarizes these points:

ARTIFICIAL INTELLIGENCE	LINGUISTICS	DEFINITION
Focus	Topic	recoverable information
	Theme1	perspectival departure
	Theme2 (= Given)	low communicative dynamism
	Focus	irrecoverable information

Table 1. Summary of terms and definitions

3. TOWARDS AN INTEGRATING ACCOUNT: THEME AND FOCUS REDEFINED

As we can see, there is no single uniform account of these interrelated phenomena. An interesting question now arises. Why is it that Artificial Intelligence researchers have found it necessary for their work to define and use the notion of *focus*, while linguists have used varied terms for several interrelated phenomena?

A partial answer to this question lies in the level of description on which researchers have grounded their definitions. The definitions provided in Linguistics for notions such as *theme, given, topic,* and *focus* always apply on the level of the sentence, in spite of the attempts to resort to discourse to identify some of the notions. By contrast, the notion of *focus* in Artificial Intelligence has always been a discourse category, “a computational account of one of the ways speakers structure communication over several clauses in a discourse.” (Sidner 1983: 127).

In view of this difficulty, operational redefinitions for these concepts are offered below, in an attempt to integrate the findings in both research fields and make them useful for a theory of discourse structure. In order to do so, a discourse perspective will be adopted in order to allow us to extract their functionalities from their behaviour in real texts.

3.1. TOWARDS AN OPERATIONAL REDEFINITION OF FOCUS

Probably the main reason why *focus* has been studied in Artificial Intelligence and Computational Linguistics is because one cannot simply string together sentences arbitrarily and expect smooth text flow. Therefore, research efforts in AI with respect to *focus* have concentrated upon the design of a set of rules which dictate the legal or permissible focus moves or transitions as the discourse unfolds (McKeown 1985; Grosz and Sidner 1985; Brennan, Friedman and Pollard 1987; Lambert and Carberry 1991). These transitions or focus-shift rules are a reflection of the speaker’s preferences in the presentation of information to ensure coherent discourse.

Thus, for example, McKeown claims that in (6) and (7) the discourse focuses on a single entity (the balloon), since the speaker wants to convey information about several of its properties. As she points out, these properties cannot simply be presented in any order, since in most random cases the text will be judged not smoothly developing. She claims that a speaker will group together properties that are in some way related to each other, making (6) more connected than (7):

(6) The balloon was red and white striped. It had a silver circle at the top to reflect heat. Because this balloon was designed to carry men, it had to be large. In fact, it was larger than any balloon John had ever seen.

(7) The balloon was red and white striped. Because this balloon was designed to carry men, it had to be large. It had a silver circle at the top to reflect heat. In fact, it was larger than any balloon John had ever seen.

More recently, researchers in Computational Linguistics have developed the so-called *centering* theory (Grosz, Joshi and Weinstein 1995) to model the the local component of attentional state, i.e., the *focus* or *center* of attention at any given point in discourse. The *center transition rules* in this theory reflect our intuitions as to how to link a number of utterances together in a coherent local segment of discourse.

In the area of text generation, McCoy and Cheng introduced the notion of a *discourse focus tree* as a mechanism for controlling focus shifts in discourse (1990). Their hypothesis is that a focus tree is constructed and traversed by the participants as the discourse progresses, one node being visited at a time. Each node in the tree is something which is being talked about in the discourse and points to an entity from the knowledge base. The type of the currently visited node —*object, attribute, setting, action*, and *event*— determines what entities from the knowledge base are in focus, and generates expectations about what may be said next in a discourse, which must generally come from the highlighted set of knowledge. A change in the focus of attention corresponds to changing the currently visited node. Table 2 below illustrates these focus nodes categories and their focus-shift candidates.

NODE TYPE	FOCUS-SHIFT CANDIDATES
OBJECT	attributes of the object; actions the object plays a prominent role in
ATTRIBUTE	objects which have the attribute; more specific attribute
SETTING	objects involved in the setting; actions which typically occur in the setting
ACTION	actor, object, etc... of the action; any participant role
EVENT	actions which can be grouped together into the event

Table 2. Focus-shift candidates for selected node types
(adapted from McCoy and Cheng 1990)

The use of different categories of *focus* nodes causes the highlighting of a particular set of knowledge base entities, thus constraining what can be said next as the discourse progresses. In (6) above, the first focus node would be the physical attributes of the balloon, more specifically its color and shape. The expectation would be that the next attributes that are mentioned would probably be a "subclass" of this inferred attribute, thus favouring version (6) over (7) since it follows a recognised pattern of inferencing. That is, given a specific concept in *focus*, the listener's expectation would be that the next sentence would continue to add information that is semantically related to the previous one. That is, concepts cannot simply follow one another without some logical-semantic relationship; they have to be connected via recognisable patterns of inference. Thus, for example, if the concept in question is an attribute (say, an adjective), the expectation would be that the next attributes that are mentioned would be a "subclass" of the one in focus.

It is important to point out that the construction of the discourse focus tree is a joint enterprise undertaken by the discourse participants. That is, while the speaker adds or changes the currently visited node, the hearer must try and make appropriate changes in his/her model of the tree based on what has been said. Therefore, the determination of the *focus* of attention at any given point in the discourse, and the specification as to where the focus may progress to, requires a process of inferencing on the part of the participants. For instance, if the discourse focuses on an object, it may next progress to talk about attributes of that object. This progression would cause an attribute node to be grown in the tree. For this reason, and in an attempt to integrate this notion of *focus* with the one proposed in Linguistics, we propose the following provisional definition:

> **DEF:** *Focus* is the locus, within each message, of principal inferential effort. It is that/those concept(s) most relevant for processing on which the speaker wants the hearer to spend the most thought.

If we define an inference as the mental process of creating a new concept by applying some rule(s) of deduction to an input concept, then inferential effort will be the work involved in:

(a) finding rules of deduction,
(b) testing whether their requirements match the current input (concepts) at hand,
(c) if so, applying the rules, and
(d) instantiating the resulting concepts as new input(s) for further inference.

The speaker determines the *focus* at a given point in discourse by identifying the concept(s) contained in it which are most central to the mental processing (the inferences) required by the listener to understand the message, and to follow a coherent progression of information. By signalling these central concepts as the *focus*, the speaker is saying: "Concentrate your thoughts around this. This is where I want your thinking to start."

The hearer/reader will concentrate on those unit(s) that are of most interest to his/her immediate goals. However, these may not be the ones best suited to the current purposes of the discourse. When the speaker's and hearer/reader's deduction concept chains diverge, then they have increasing trouble communicating. But if the hearer/reader spends his /her principal inferential effort on the same input unit concepts as the speaker does, then the hearer will be, in general, well prepared for what the speaker is about to continue with, to the extent that the hearer's deduction rules are informed about the domain of discourse, of course. If the topic is totally new to the hearer, then he/she will not be able to "pre-think" in the direction the speaker is headed.

According to this redefinition, *focus* appears as a dynamic process of applying deduction rules and prioritizing inference, quite in tune with the centering framework as developed by different authors (Joshi and Kuhn 1979; Joshi and Weinstein 1981; Grosz, Joshi and Weinstein 1995) where focusing functions to limit the inferences required for the understanding utterances in a discourse. If the *focus* is the locus of principal inferential effort, the concept(s) in focus at a given time in a discourse will reduce other alternative inferential possibilities.

3.2 TOWARDS AN OPERATIONAL REDEFINITION OF THEME

As we have seen above, the notion of *theme* was traditionally treated together with givenness by the Prague School, who linked them with the idea of Communicative Dynamism. Halliday and Quirk et al. advocated for a separation of these notions, and, accordingly, *theme* was defined by Halliday as that "element which serves as point of departure of the message; that with which the clause is concerned" (1985:

39). This double-sided definition is, in my view, the origin of the confusion with other related notions such as *topic* and *focus* in the AI sense. For, as has been recently pointed out, the point of departure of the message is not necessarily what that message is about (Downing 1992).

In the first place, the notion of "aboutness" is artificial if applied to sentences in isolation. Secondly, even if we admitted the existence of some clausal element(s) as those about which is being talked about, the parallel identification of Theme with the initial element of the clause as a realisation of the notion of "point of departure" would lead us to accept that elements such as "well," "frankly," or "Mr. Jones" in (11) below are what the message is about.

(11) Well, frankly, Mr. Jones, in Rome I had a really great time

THEME.................................. RHEME
Textual Interpersonal......... Ideational............

Neither the textual, nor the interpersonal themes are even remotely concerned with the definition of theme as what the clause is about. While this way of conceptualizing Theme is useful when dealing with sentences out of context, it is doubtful whether it would account adequately for the progression of thematic information from one sentence to another in naturally occurring text data. Rather, we should strive for a characterization of theme as a functional notion associated with information communicated in discourse. Accordingly, I provisionally propose the following redefinition of Theme:

> DEF: Theme is that element that informs the listener as discourse unfolds how to relate the incoming information to what is already known.

This definition helps illustrate why Theme usually co-occurs with Given, and why it is the "point of departure" of the clause. It fulfills a guiding function in the listener's journey through the series of New information portions. This is why several linguists investigating the function of Theme in discourse have found out that a text's flow of information (what is called its "method of development») is strongly related to the material in the themes of the component clauses (Fries 1983:135).

Having clarified to some degree the notions *focus*, *theme-rheme*, and *given-new*, we turn next to some of the most interesting aspects of these notions: their behaviour in text from one sentence to the next. This is the topic of the next subsection.

3.3 THE PROGRESSION OF FOCUS AND THEME

As explained above (Section 3.1), research efforts in Artificial Intelligence and Computational Linguistics with respect to *focus* have concentrated upon the design of a set of rules which dictate the legal or permissible focus moves during text generation. These so-called focus-shift rules are a reflection of the speaker's preferences in the presentation of information to ensure coherent discourse.

Thus, McKeown (1985, p.67) worked with three graded notions: (a) the immediate focus of a sentence (current focus, CF), (b) the potential focus list (PFL), which includes the elements of the sentence that are potential candidates for a change in focus, and (c) a focus stack (a stack of past immediate foci). She then described the following three focus shift rules:

1. change focus to member of previous PFL if possible
2. maintain focus if possible
3. return to topic of previous discussion; more precisely, choose the CF from the focus stack

Together, these rules specify the patterns of focus shift one encounters in paragraphs of simple English. They are necessary for smooth text flow; without them, generator systems produce text that seems bizarre, as example (9) above. More recently, the so-called *centering theory* has also proceeded along similar lines in an attempt at specifying the rules for focus/center movement in discourse to achieve coherence (Grosz and Sidner 1997).

In Linguistics, Danes' work shows that the organization of information in whole texts, as opposed to just sentences, is determined by Thematic Progression (TP). This is defined as "the choice and ordering of utterance themes, their mutual concatenation and hierarchy, as well as their relationship to the hyperthemes of the superior text units (such as paragraph or chapter) to the whole of text, and to the situation" (1974:114). In this sense, Danes criticizes Halliday's claim that "thematization is independent of what has gone before," because this would make thematization irrelevant with respect to the organization of the text, thus contradicting our intuitive expectations that the progression of the presentation of subject-matter must necessarily be governed by some regularities (Halliday 1967). Following this argument, he identifies different patterns of Thematic Progression which may occur in text: the *simple linear progression*, the *constant progression*, the *derived hyperthematic progression* and the *splitting progression*. These can be diagrammatically represented as follows:

1) *Simple linear progression*: an item from the rheme of the first clause becomes the theme of the subsequent clause:
Rh(x); ->Th(x) + Rh(y); Th(y)...

2) *Constant progression*: an item in the theme of the first clause is also selected as the theme of the following clause:
Th(x) + Rh(x); Th(x) + Rh(y)...

3) *Derived hyperthematic progression*: the particular themes in subsequent clauses are derived from a "hypertheme»:
T=[Hypertheme]; Th(1) + Rh(1); Th(2) + Rh(2); Th(3) + Rh(3)...

4) *Splitting progression***:** the rheme of the first clause is split into two items, each in turn being then taken as a theme element in the subsequent clauses:
Th(x) +Rh(x) (=Ri + Rii); Th(y) + Ri; Th(w) + Rii

A moment's reflection will shed light as to what researchers in both fields have been aiming at. Whereas in AI the emphasis has been laid on providing guidelines as to how to construct coherent text and how to track participants in discourse, in Linguistics research has concentrated on describing the ways in which information progresses from one sentence to the next in texts.

In both cases, the most burning issue, namely, the enquiry into the principles of information selection and distribution, is still a matter of empirical investigation. As Danes exhorted with reference to theme selection and progression:

> We must not be content with a statement that certain sentence elements convey the known information (in contrast to others conveying the new one), but we ought to find out the principles exactly according to which this information and not another portion of the mass of known information has been selected. In other words, we must enquire into the principles of thematic choice and thematic progression. (1974: 112)

In this sense, several empirical studies offer some promising results on the issue. With respect to the process of thematization, Fries (1995: 10) hypothesized a relationship between theme selection and genre type, and different authors explored this relationship in different genres/registers (Berry 1989; Bäcklund 1990; Francis 1990; Ghadessy 1995; Wang 1991). Other studies investigated the relationship between text types, registers and/or genres and thematic progression patterns (Virtanen 1993; Downing and Lavid 1998), some of them discovering statistically significant correlations between contextual factors from the communicative situation and the thematization process (Lavid 1998, forthcoming).

With respect to focus, recent research has also investigated and empirically validated the influence of contextual factors such as the *discourse purpose*, the *text type* and the *subject-matter* of the discourse, among others, as determining constraints which contribute to the focus selection process (Lavid 1994).

As it usually happens between related disciplines, it takes time and research effort to transfer and successfully apply the results obtained in one field to another. This is certainly the case with the transfer of empirical results from the theoretical field of Linguistics to the practical applications developed by computational systems, most of which are constrained by time and cost-effectiveness limitations. Nonetheless, the fact that some of the above mentioned empirical studies (Lavid 1998, forthcoming) have been successfully used in computational prototypes and applications in the framework of interdisciplinary collaborations, proves that the dialogue between both disciplines is not only fruitful and desirable, but absolutely necessary for the advancement of knowledge of this topic.

4 CONCLUSION

This paper has tried to describe the way in which the information communication problem is a burning issue for the study of discourse. Presenting some of the processes through which the problem manifests itself, it has concentrated on the resources that both Linguistics and Artificial Intelligence have brought to address it. Arguing for the need to clarify a series of related notions —such as *theme-rheme*, *given-new*, *topic* and *focus*— which have been misleadingly defined in Linguistics and insufficiently treated in Artificial Intelligence, the paper proposes an approach which separates these notions, showing their interrelationships and their function within texts. It also shows how Artificial Intelligence and Linguistics have described the development of information by providing focus-shift rules/ centering rules (in the former case) and thematic progression patterns (in the latter). However, describing the focus-shift or the thematic progression rules is one thing, but providing motivation for these shifts or progression patterns is another. This requires a theory of linguistic context that rejects the treatment of these phenomena as sentence-level ones and points to the role of the context in the determination of how the information is going to be distributed along a sequence of sentences. Recent studies which correlate thematization and focus selection with several factors from the communicative context seem to be pointing in the right direction.

Note

[1] For more complete descriptions see Van Dijk and Kintsch (1983), and Tomlin (1987).

Works Cited

Bäcklund, I. "Theme in English Telephone Conversations." Paper delivered at the *17th International Systemic Congress*, Stirling, Scotland, June 1990.

Berry, M. "Thematic Options and Success in Writing." *Language and Literature-Theory and Practice: A Tribute to Walter Grauberg*. Ed. C. Butler, R. and J. Cardwell. Nottingham: U of Nottingham, 1989.

Brennan, S., M.W. Friedman, and C.J. Pollard. "A Centering Approach to Pronouns." *Proceedings of the 25th Meeting of the Association for Computational Linguistics*, 1987. 155-162.

Bloomfield, L. *Language* London: Allen and Unwin, 1935.

Carberry, S. "Tracking User Goals in an Information-seeking Environment." *Proceedings of the National Conference on Artificial Intelligence* (AAAI-83), Washington, D.C., August 1983, 59-63.

Cohen, R. "Analyzing the Structure of Argumentative Discourse." *Computational Linguistics Journal* 13.1-2 (1987): 11-24.

Chafe, W. "Givenness, Contrastiveness, Definiteness, Subjects, Topics and Points of View." *Subject and Topic*. Ed. C.N. Li. Academic Press, New York, 1976.

Chomsky, N. "Deep Structure, Surface Structure and Semantic Interpretation." *Semantics: Interdisciplinary Reader in Philosophy, Linguistics and Psychology.* Ed. D. Steinberg and L.A. Jakobovits. Cambridge, Cambridge UP, 1971.

Danes, F. "Functional Sentence Perspective and the Organisation of the Text." *Papers on Functional Sentence Perspective*. Ed. F. Danes. *Papers on Functional Sentence Perspective*. The Hague: Mouton, 1974.

Dijk, T.A. Van, and W. Kintsch. *Strategies of Discourse Comprehension.* New York: Academic Press, 1983.

Downing, A. “An Alternative Approach to Theme: A Systemic-functional Perspective.” *Word* 42. 2 (1991): 119-143.

— and J. Lavid. “Information Progression Strategies in Administrative Forms: A Cross-linguistic Study.” *Linguistic Choice Across Genres: Variation in Spoken and Written English*. Current Issues in Linguistic Theory, 158. Ed. Sánchez-Macarro, A. and R. Carter. Amsterdam: John Benjamins, 1998. 99-116.

Firbas, J. “Some Aspects of the Czechoslovak Approach to the Problems of FSP.” *Papers on Functional Sentence Perspective*. Ed. F. Danes. Academia, Prague, 1974.

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

Fries, P.H. “On the Status of Theme in English: Arguments from Discourse.” *Micro and Macro Connexity of Texts*. Ed. Petöfi and Sozer. Hamburg: Buske Verlag, 1981.

— “A Personal View of Theme.” *Thematic Development in Texts*. Ed. M. Ghadessy London: Pinter, 1995. 1-19.

Geluykens, R. “Information Flow in English Conversation: A New Approach to the Given-New Distinction.” *Functional and Systemic Linguistics: Approaches and Methods*. Ed. E. Ventola Mouton de Gruyter, 1991.

Ghadessy, M. “Thematic Development and Its Relationship to Register and Genres.” *Thematic Development in Texts*. Ed. M. Ghadessy. London: Pinter, 1995. 129-146.

Givon, T. *Syntax: A Functional-typological Introduction*. Vol. 1. Amsterdam: John Benjamins, 1984.

Grosz, B. and C. Sidner. “Discourse Structure and the Proper Treatment of Interruptions.” *Proceedings of the 1985 Joint Conference on Artificial Intelligence*, IJCAI85, Los Angeles, Ca., August 1985.

—.A. Joshi and S.Weinstein. “Centering: a Framework for Modelling the Local Coherence of Discourse.” *Computational Linguistics* 21. 2 (1995): 203-225

Hajicova, E. “Focussing: A Meeting Point of Linguistics and Artificial Intelligence.” *Artificial Intelligence* 14 (1987): 311-321.

Halliday, M.A.K. “Notes on Transitivity and Theme in English.” *Journal of Linguistics* 3 (1967): 37-81.

— *Language as Social Semiotic: The Social Interpretation of Meaning*. London: Edward Arnold, 1978.

— *An Introduction to Functional Grammar*. London: Edward Arnold, 1985.

Hobbs, J. “Coherence and Coreference.” *Cognitive Science* 3 (1979): 67-90.

Hockett, C.F. Hockett. *A Course in Modern Linguistics*. New York: Macmillan, 1958.

Jackendoff, R.S. *Semantic Interpretation in Generative Grammar*. Cambridge : MIT, 1972.

Joshi, A. and S. Kuhn. “Centered Logic: The Role of Entity Centered Sentence Representation in Natural Language Inferencing.” *Proceedings of the International Joint Conference on Artificial Intelligence.* 1979. 435-439.

— and S. Weinstein. “Control of Inference: Role of Some Aspects of Discourse Structure-centering.” *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981. 385-387.

Lambert, L., and S. Carberry. "A Tripartite Plan-based Model of Dialogue." *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991. 47-54.

Linde, C. "Focus of Attention and the Choice of Pronouns in Discourse." *Syntax and Semantics*, Vol. 12. Ed. T. Givón. New York: Academic Press, 1979.

Lavid, J. *Theme, Discourse Topic and Information Structuring*. Deliverable R1.2.2b of WP1.2.2. Esprit Basic Research Project 6665 DANDELION, Universidad Complutense de Madrid, Madrid, October 1994.

— "The Relevance of Corpus-based Research for Contrastive Linguistic and Computational Studies: Thematization as an Example." *IV-V Jornades de corpus lingüístics: els corpus en la recerca semántica y pragmática*. Barcelona: Institut Universitari de Lingüística Aplicada, Universidad Pompeu Fabra, 1998. 117-139.

— "Contextual Constraints on Thematization: An Empirical Study." *Formal Aspects of Context*. Kluwer Academic Publishers, forthcoming.

McCoy, K., and J. Cheng. "Focus of Attention: Constraining What Can Be Said Next." *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Ed. W. Swartout Paris and W. Mann. Kluwer, Dordrecht, 1990.

McKeown, K. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge: Cambridge UP, 1985.

Prince, E. "Toward a Taxonomy of Given-new Information." *Radical Pragmatics*. Ed. Cole. New York: Academic Press, 1981.

Quirk, R.S., G. Greenbaum, G. Leech and J. Svartvik. *A Grammar of Contemporary English*. London: Longman, 1972.

Reichman, R. "Conversational Coherency." *Cognitive Science* 2 (1981): 283-327.

Reichman-Adar, R. "Extended Person-machine Interface." *Artificial Intelligence* 22.2 (1984): 157-218.

Sgall, P. "Towards a Definition of Focus and Topic." *Prague Bulletin of Mathematical Linguistics* 31 (1979): 3-26.

— E. Hajicova and J. Panevova. *Topic, Focus, and Generative Semantics*. Kronberg/Taunus, 1973.

Sidner, C.L. "Focusing and Discourse." *Discourse Processes* 6 (1983): 107-130.

Tomlin. R.S. "Linguistic Reflection of Cognitive Events." *Coherence and Grounding in Discourse*. Ed. R. Tomlin. Amsterdam: John Benjamins, 1987. 455-480.

Vallduví, E. "The Dynamics of Information Packaging." *Integrating Information Structure into Constraint-based and Categorial Approaches.* Ed. Elisabeth Engdahl. (DYANA-2 Report R1.3.B) Amsterdam: ILLC, 1994.

Virtanen, T. *Discourse Functions of Adverbial Placement in English*. Abo Akademi, 1992.

Wang L. Analysis of Thematic Variations in Buried Child. Paper delivered at the *First Biennial Conference on Discourse*. Hangzhou Peoples Republic of China. June 1991.