THE TEXTUAL DEVELOPMENT OF NON-STEREOTYPIC CONCEPTS

Karin Haenelt and Michael Könyves-Tóth
Integrated Publication and Information Systems Institute (IPSI)
GMD
Dolivostraße 15, D 6100 Darmstadt, Germany
haenelt@ipsi.darmstadt.gmd.dbp.de
koenyves@ipsi.darmstadt.gmd.dbp.de
tel. ++49/(0)6151/875-811, fax -818

ABSTRACT

In this paper the text theoretical foundation of our text analysis system KONTEXT is described. The basic premise of the KONTEXT model is that new concepts are communicated by using the mechanisms of text constitution. The text model used assumes that the information conveyed in a text and the information describing its contextual organization can be structured into five layers (sentence structure, information on thematic progression, referential structure, conceptual representation of the text and conceptual background knowledge). The text analysis component constructs and traverses the information of these layers under control of the discourse development. In this way, it can incrementally construct a textual view on knowledge, rather than only recognizing precoded concepts.

1 INTRODUCTION

In the field of knowledge-based text analysis it has been regarded as insufficient to analyze a text against the background of static and stereotypic default assumptions for some time (cf. [Hellwig84], [Scha/Bruce/Polanyi87]). By applying this method the pre-coded concepts are invoked again and again during the process of text analysis, regardless of the changes and the new concepts being constituted by the ongoing text. The function of a text, however, is not confined to concept selection as in current knowledge-based applications. In addition, textual mechanisms are used to operate on concepts and to compose them to actual contexts, i.e. to constitute (new) concepts. Textually the contexts are established by the thematic and by the referential structure. Thus, new mechanisms are required which permit the textual organization to control the creation and manipulation of concepts in text processing. In a way, this is to tie linguistic and knowledge-based approaches to text processing together into a single method.

2 THE KONTEXT MODEL

The basic premise of the KONTEXT model is that the relationship of expression and concept are changed during a text and concepts are communicated by using the mechanisms of text constitution. The KONTEXT model is based on the assumption that

- the information conveyed in a text and the information describing its contextual organization can be structured into five layers. They define the sentence structure, information on thematic progression, the referential structure, the conceptual representation of the text and the conceptual background knowledge;
- discourse provides the basic mechanisms by which concepts are constructed. Discourse is defined as sequences of transitions between discourse states and discourse states are defined by the information represented in the layers.

The text analysis component constructs and traverses the information of these layers under control of the discourse development. In this way, it can incrementally construct a textual view on knowledge, rather than only recognizing precoded concepts.
We will now describe the layers of the text repre-

sentation. In the following section we discuss the conception of discourse in more detail.

2.1 LAYERS OF TEXT REPRESENTATION

There are five layers of text representation:

sentence structure
thematic structure
referential structure
view
background knowledge

The lowest layer is the basis for textual communication. It is a formal representation of concepts modeling an open world and serves as *background knowledge*. Since we allow for the construction of new details and concepts, an organization of concepts is provided which supports this task. Our background knowledge differs from traditional knowledge bases in that it does not represent a particular domain model which assigns a predefined and fixed structure to the concepts. It is rather organized around expressions and models their referential potential in terms of concepts. It resembles a meaning dictionary (like e.g. [COBUILD87] which is used as the basic material), where with expressions concepts are constituted and used to explain other concepts. Basically all concepts are of the same rank with respect to an open world. During discourse the concepts are accessed via explicitly modeled perspectives on them [Kunze90] [Melcuk87] depending on the actual textual development (e.g. actual state of contexts, c.f. 2.2 *discourse state*).

The next layer, the *view*, models the subject matter of the text using the concepts which are defined in the background knowledge. The ongoing discourse selects concepts from the background knowledge or the already existing view, reorganizes their structure and (re–)integrates them coherently into the already existing view. The concepts constructed in the view during discourse provide the text specific perspective on the background knowledge.

The layer of the *referential structure* represents reference objects and their relationships. It drops details of the concept definition in accordance with the abstraction level of references in the text, and represents those complexes as units which are explicitly referred to by linguistic means in the text.

The layer of *thematic structure* traces the discourse development. It represents the contextual clustering of reference objects and traces the development of their clustering. This trace represents the progression of themes and the development of focusing. The notion of thematic structure is based on the Prague School approaches to the thematic organization (e.g. [Danes70][Hajicová/Sgall88][Hajicová/Vrbová82]), which we refine by distinguishing the mechanisms involved in terms of the textual function of linguistic means with respect to the different layers of the text representation.
In our model the units of the layer of thematic structure are *contexts*. By context we understand a cluster of reference objects, where within a context the relationship between a reference expression and its reference object is unequivocal. During the ongoing discourse, however, this relationship and the groups of reference objects which are clustered together change. Whether or not linguistic means create new contexts, and which kind of clustering of reference objects they effect, depends on their textual function and on the state of discourse they operate on (examples of this are given below). Contexts are the units of the thematic progression. It is this grouping of reference objects that is referred to by linguistic means immediately, that is changed, resumed, revised and tied up to during discourse. The thematic structure is the result of creating, closing and referring to contexts. The movement of contexts traces the growth of the view.
It should be noted that complex progression types can be constructed. This is due to the ability of predicative expressions to cover several themes by virtue of their arity and due to the textual possibility of changing the structure of a contextually clustered concept by changing the focus when referring to a context. Therefore hierarchical structures as proposed by different approaches to describing the structuring of actual texts are not sufficient to cope with the ability of natural language texts to constitute contextual relations (cf. content

oriented structures: e.g. thematic progression [Danes70] – at least the five forms elaborated are hierarchical –), or discourse segmentations: e.g. discourse constituent units [Polanyi88], context spaces [Reichman85], rhetorical structures [Mann/Thompson88], superstructures and macrostructures [vanDijk83]).

The *sentence structure* describes the linguistic means used in the text to express the information encoded in the lower layers.

Our representation models structural relationships of text constitution principles. The background knowledge provides concepts for the constitution of the semantic text representation (view). The concepts constructed in the view during discourse provide the text specific perspective on the background knowledge.
Referential structure and thematic structure each cluster structures of the lower layers. Reference objects group conceptual definitions into units which can be referred to by ensuing linguistic expressions. The sequence of thematizing defines a clustering of reference objects into contexts.
Whilst the lower layers contain more static information which is independent of the actual sequence of the textual presentation, the dynamic of discourse, i.e. the growth of the view during the ongoing discourse, is represented in the layers of thematic structure and sentence structure.
The modeling allows for a text driven control of operations on the knowledge base and on the view, because the manipulations of the lower layers depend on the interpretation of the upper layer phenomena.
We define the types of manipulations necessary in terms of the contribution linguistic means make to the layers of the text representation. The definitions are placed in a text lexicon (cf. the example given below).

2.2 DISCOURSE

By *discourse* we understand a sequence of state transitions which is guided by the interpretation of linguistic means. It models textual access to concepts: A text does not communicate concepts at once. It rather guides sequential access and operations on knowledge that produce a particular *view* on the concepts of the background knowledge.

A *discourse state* is defined by the actual state of all the five layers of the text representation, which renders the actual state of the view and the actual access structure to view and background knowledge. While the view grows during the analysis, only a small segment of it is in the focus of attention at one state, and the objects which are referred to by linguistic expressions may change state by state. A discourse state provides the immediate context to which ensuing linguistic means can refer directly, and also previous contexts.

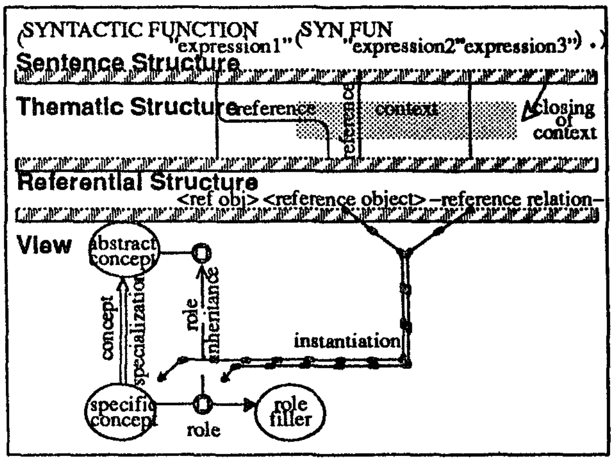
The *transition of a discourse state* is the effect of the interpretation of a linguistic expression. It is determined by the textual function of linguistic means. Modeling the operational semantics of linguistic means within the framework outlined leads to our text lexica.
Differences of the view of two discourse states which are produced by a discourse state transition can be regarded as the semantic contribution of a linguistic expression. But it is important to note that this contribution is not only determined by the isolated expression, and that therefore analysis does not involve a static mapping from a textual expression to some semantic representation or vice versa. The contribution rather depends on the actual state of the preceding discourse on which the expression operates. Note also that there are expressions whose interpretation does not contribute to the growth of the view. In an actual text they rather are used in order to manipulate the thematic organization (e.g. redirections).

3 EXAMPLE

With a small example we illustrate how the KONTEXT model works. We show how a reference object and a concept corresponding to a referential expression is created, and how the relationship between expression and concept is changed during the discourse. From a sample text we take the following sentence and show that discourse state transitions already occur while interpreting this sentence textually:
"The electronic dictionaries that are the goal of EDR will be dictionaries of computers, by computers, and for computers."
We provide a selection of three discourse states showing view and access structure after the interpretation of "The electronic dictionaries" (figure

1), after "that are the goal of EDR" (figure 2), and after "will be dictionaries of computers, by computers, and for computers." (figure 3). Each figure then is explained by describing the textual function of the linguistic means concerned, i.e. by describing how they operate on previous discourse states and what their contribution to the layers of the text representation is. These definitions are placed in a text lexicon. Because we want to draw the attention to the nature of textual functions of linguistic means and to the possibility to distinguish and to describe these functions with respect to the layers of the text representation, we confine ourselves to demonstrating this by discussing only those readings which lead to a solution in our example.

The sentence structure used is the structure the PLAIN grammar [Hellwig80] attributes to a sentence, and for the graphical representation of our example we use the conventions explained in the legend (see below). The names of the roles in the view and in the background knowledge have been chosen for mnemotechnical reasons only, they are not to be confused with the conceptual modeling of prepositions.



LEGEND

Figure 1: "The electronic dictionaries"

"The electronic dictionaries": In the *sentence structure* the reference expression "the electronic dictionaries" occurs. Since so far no corresponding *reference object* exists, it must be created and conceptually defined. No previous textual *context* has been established before this state, therefore immediate access to the global and unspecified background concepts is allowed. [COBUILD87]

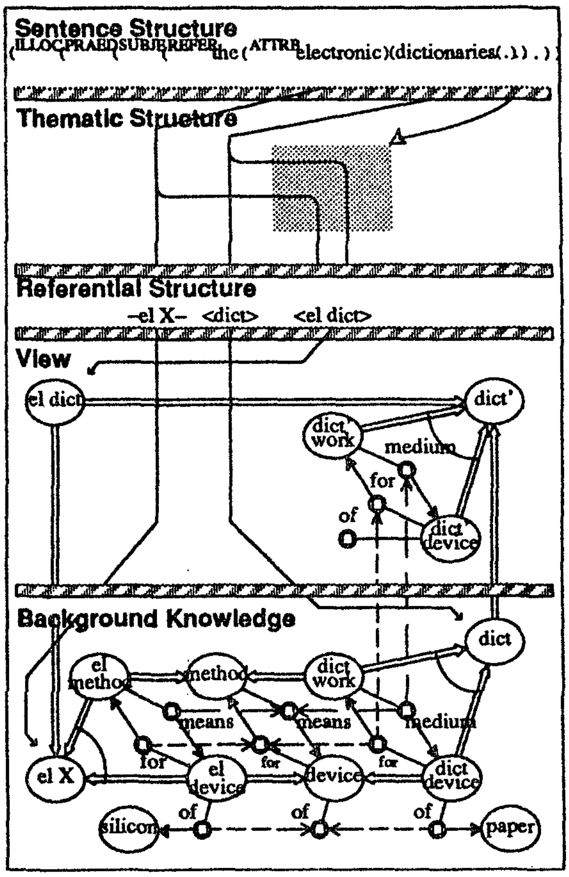


Fig. 1 : Discourse state after the interpretation of *"The electronic dictionaries . . . . "*

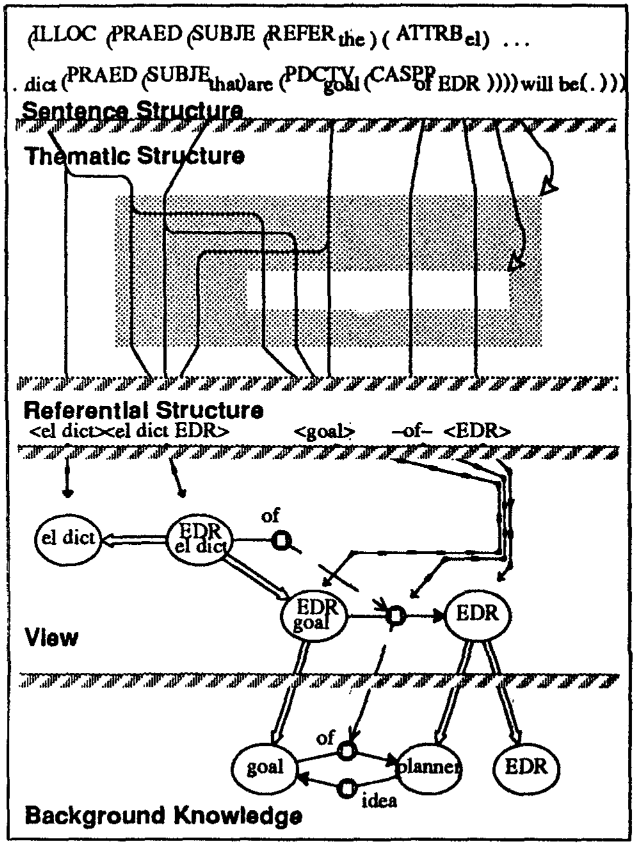
does not have an entry "electronic dictionary", which means that in the *background knowledge* no corresponding concept exists.

"electronic": As an adjective, "electronic" refers to the *reference item* –elX–, which does not select a concept, but a conceptual structure which is used to extend or to modify the dominating noun's concept. In [COBUILD87] there are two *conceptual* aspects of "electronic", which are related to each other. At first "electronic" can be 'a device, which has silicon chips and is used as a means for electronic methods'. Secondly 'a method' can be referred to as "electronic".

"dictionary": Initially "dictionary" refers to the *reference object* <dict>. *Conceptually* "dictionary" can refer to two aspects: It can refer to 'a physical device, which is made of paper and serves as a medium for recording symbols; it has been compiled by an author and is used for reference purposes.' It can also refer to 'the recorded symbols as a work'.

"electronic dictionary": In order to find a *conceptual definition* of the introduced reference object <eldict> we create a less specific abstract concept of dictionary. On the one hand it must be as specific as possible, and on the other hand it must be compatible with what is known conceptually about the referential item –elX–. 'Electronic dictionary' then is a combination of 'electronic' and 'dictionary' leaving open e.g. the incompatible device 'paper'. A more specific concept of "dictionary" is introduced. This means that from now on the text will not deal with "dictionaries" in general, but with "dictionaries" in the restricted context of "electronic dictionaries". Therefore a new *context* is opened, and in this new context "dictionaries" refers to a new *reference object* <eldict> which can be the theme of the further ongoing discourse.

Figure 2: "that are the goal of EDR"



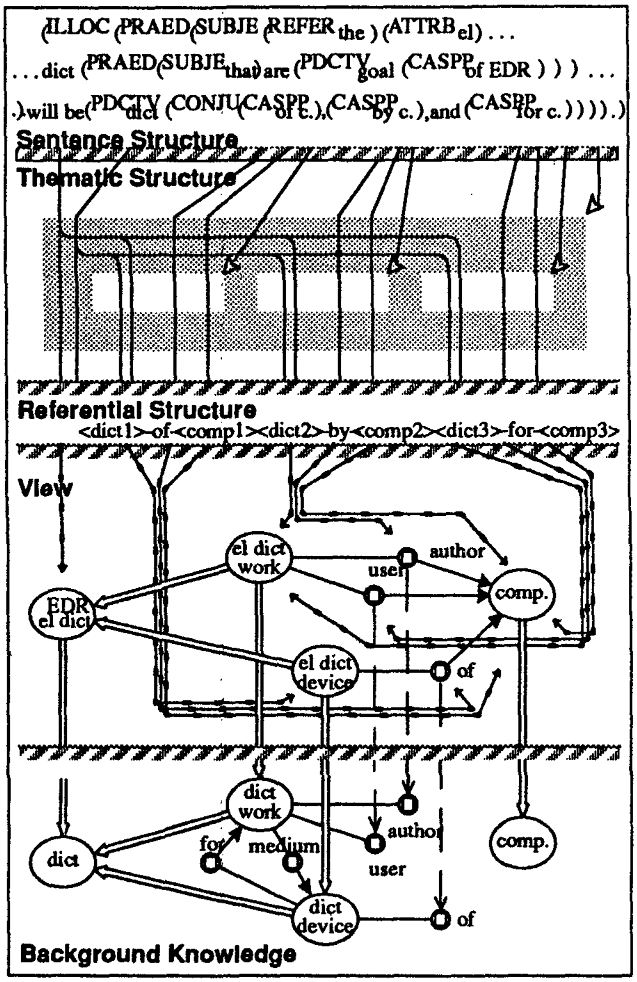
(ILLOC (PRAED (SUBJE (REFER the) (ATTRB el) ...
. dict (PRAED (SUBJE that) are (PDCTY goal (CASPP of EDR))))will be(.))
Sentence Structure
Thematic Structure
Referential Structure
<el dict><el dict EDR> <goal> –of– <EDR>
el dict
EDR el dict
of
EDR goal
EDR
View
of
goal
planner
EDR
idea
Background Knowledge

Fig. 2 : Discourse state after the interpretation of "*The electronic dictionaries that are the goal of EDR . . .* ."

"that": This relative pronoun, again, forces the creation of a new context. A new *context* is opened which is restricted to those "electronic dictionaries" only, which "are the goal of EDR". The pronoun also has the function of a connexion instruction [Kallmeyer/etal77] and effects a referential equation of "electronic dictionaries" and what is predicated about "that". Both expressions and also "that" then refer to <eldictEDR> in this new context.

"are": It is the textual function of the copula to form a unified context of the contexts of its subject ("that") and its predicative complement ("the goal of EDR"). The unified context defines the reference object <eldictEDR>.

Figure 3: "will be dictionaries of computers, by computers, and for computers"



(ILLOC (PRAED (SUBJE (REFER the) (ATTRB el) . . .
. . . dict (PRAED (SUBJE that) are (PDCTY goal (CASPP of EDR))) . . .
.)will be(PDCTY dict (CONJU(CASPP c.),(CASPP by c.),and (CASPP for c.))))).)
Sentence Structure
Thematic Structure
Referential Structure
<dict1>-of-<comp1><dict2>-by-<comp2><dict3>-for-<comp3>
View
el dict work
author
user
comp.
EDR el dict
el dict device
of
dict work
author
for
medium
user
comp.
dict
dict device
of
Background Knowledge

Fig. 3 : Discourse state after the interpretation of " *.will be dictionaries of computers, by c., and for c.*"

"dictionaries": The expression "dictionaries of computers, by computers, and for computers" refers to three *reference objects* <dict1>, <dict2> and <dict3> (namely "dictionary" in the context of "of", "by", and "for"). The three contexts established for these reference objects are textually focused on and thus provide the basis for further textual progression.

"will be": The copula, again, forms a unified context of the contexts of its subject and its predicative complement. This also effects a referential equivalence of "electronic dictionary" and "dictionary". Therefore "dictionary" must at this state of the discourse no longer access the concept of "dictionary" of the background knowledge as freely as at the beginning of the text, when there was no restriction in interpretation. Now it rather must access the concept which meanwhile has been established by the text (namely 'dictionary' in the sense in which it has been modified and defined by 'electronic').

"of, by, for": make further conceptual contributions to the concept of "electronic dictionaries" by refining the concept by the aspects denoted by "of", "by" and "for".

4 CONCLUSION

The model described in this contribution serves as a theoretical foundation of a computer implementation of a text analysis system. It enables us to model a discourse which can simulate the communication of new concepts. In this simulation concepts are constituted sequentially by means of state transitions which are the effect of the interpretation of the actual textual usage of a limited set of linguistic means. This technique offers the possibility to create actual concepts on the basis of globally and unspecifically defined concepts. Thus texts are regarded as construction instructions which guide the incremental construction of *views* on conceptual knowledge bases.

5 REFERENCES

[COBUILD87] Sinclair, John (ed. in chief): Collins COBUILD English Language Dictionary. London, Stuttgart: 1987.

[Danes70] Danes, Frantisek: Zur linguistischen Analyse der Textstruktur. In: Folia Linguistica 4, 1970, pp. 72–78

[Hajicová/Sgall88] Hajicová, Eva; Sgall, Petr: Topic and Focus of a Sentence and the Patterning of a Text. In: Petöfi, János S. (ed.): Text and Discourse Constitution. Berlin: 1988. pp. 70–96

[Hajicová/Vrbová82] Hajicová, Eva; Vrbová, Jarka: On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding. In: Horecky, J. (ed.): Proc. of COLING 1982, pp. 107–113

[Hellwig84] Hellwig, Peter: Grundzüge einer Theorie des Textzusammenhangs. In: Rothkegel, A.; Sandig, B. (eds.): Text–Textsorten–Semantik: linguistische Modelle und maschinelle Anwendung. Hamburg, 1984. pp.51–59

[Hellwig80] Hellwig, Peter:
Bausteine des Deutschen. Germanistisches Seminar, Universität Heidelberg 1980

[Kallmayer/etal77] Kallmeyer, Werner; Klein, Wolfgang; Meyer–Hermann, Reinhard; Netzer, Klaus; Siebert, Hans–Jürgen: Lektürekolleg zur Textlinguistik. Band 1: Einführung. Kronberg/Ts.: 2. Aufl. 1977 (1.Aufl. 1974)

[Kunze90] Kunze, Jürgen: Kasusrelationen und Semantische Emphase. to appear in: Studia Grammatica 1990

[Mann/Thompson87] Mann, William C.; Thompson, Sandra A.: Rhetorical Structure Theory: A Theory of Text Organization. In: Livia Polanyi (ed.): The Structure of Discourse. Norwood, N.J.: 1987

[Polanyi88] Polanyi, Livia: A Formal Model of the Structure of Discourse. In: Journal of Pragmatics, Vol.12, 1988, pp. 601–638

[Melcuk87] Melcuk, Igor A.; Polguère, Alain: A Formal Lexicon in the Meaning–Text Theory (or How to Do Lexica with Words). In: CL, Volume 13, Numbers 3–4, July–December 1987

[Reichman85] Reichman, Rachel: Getting Computers to Talk like You and Me. Cambridge, Mass. 1985

[Scha/Bruce/Polanyi87] Scha, Remko J.H.; Bruce, B.C.; Polanyi, Livia: Discourse Understanding. in: Shapiro, S. C. (Ed. in chief); Eckroth, D. (manag. editor): Encyclopedia of Artificial Intelligence. New York/Chicester/Brisbane/Toronto/Singapore: 1987, pp. 233–245