DISCOURSE ANALYSIS AND COMPUTER ANALYSIS. A FEW NOTES FOR DISCUSSION

Teun A. van Dijk, Department of General Literary Studies, University of Amsterdam, The Netherlands,

1. Introduction

This paper sketches some ideas for a theoretical framework underlying the application of discourse analysis in the computerized description of texts. As the narre of this conference suggests, such descriptions have traditionally been given in linguistic terms. For all practical purposes this usually implies a grammatical analysis of phrases, clauses or sentences of texts, and involves the development of a more or less powerful parser or semi-automatic coding procedures (Aarts & Meijs, 1984). Since complete and fully explicit grammars, even of English, don't exist, such parsers have at most partial and practical applicability: they analyze only a subset of actually used sentence patterns. This holds both for syntactic and semantic analyses, which of course are mutually dependent.

Obviously, sets of texts (corpora) cannot be analyzed automatically without at least a simple syntactic parser or without some form of semantic interpretation of sentence structures. At the same time, if discourse analysis in general, and good old text grammars of more than 15 years ago in particular, have taught us anything at all, it is that texts are no more merely sequences of sentences than sentences are sequences of words. The arguments that support this statement need not be repeated here. They can be found in hundreds of books and thousands of articles (see e.g. van Dijk, 1972, 1977, 1985, and de Beaugrande & Dressler, 1981, for references). The conclusion for a serious attempt at a full-fledged computerized analysis of texts is that we need insight into the various levels *and* dimensions of discourse structure.

This paper, however, intends to spell out more than the obvious. It will not survey the advances of discourse analysis, which has repeatedly been done elsewhere, nor those in the computer analysis of texts, with which I am only indirectly familiar. Rather, it will select a few topics that in my opinion are important components in future developments in computerized text analysis.

2. The interdisciplinary approach

A first major perspective that deserves some attention pertains to the necessary interdisciplinary nature of both manual and automatic analyses of

discourse. To be sure, a discourse grammar will be able to account for structures beyond the sentence that are necessary components in any form of analysis and that cannot be provided by grammars for isolated sentences. However, this still leaves us with a narrow-minded form of (text) linguistics. Obviously, there is more to language use and communication than the isolated account of such discourse structures, whatever its importance for much practical analysis of text corpora.

We all know that parallel to the development of discourse analysis since the end of the 1960's, we have witnessed the emergence of pragmatics, accounting for speech acts, or illocutionary functions of utterances. Discourse analysis has extended pragmatics with the usual requirement that also structures of speech act sequences and their appropriateness conditions must be made explica (van Dijk, 1981). After all, dialogues usually consist of more than one speech act. Somewhat less pronounced than in spoken interaction, but nevertheless relevant for written texts as well, therefore, is the conclusion that also such written texts should be analyzed in terms of speech act sequences, both at the micro-level of local speech act coherence, and at the macro-level of global speech acts and their appropriateness conditions.

For the automatic analysis of text and talk, this conclusion is no less relevant. The parsing and interpretation of some aspects of word order, verb morphology and tense (such as the use of declarative, imperative or interrogative sentence types), of many pragmatic particles (especially in Dutch, German, Greek, and Russian), or other indicators of intended illocutionary function, vitally depends on an explication of discourse in terms of local and global speech act structures and their contextual conditions of appropriateness as well as on conditions of strategic interaction (Franck, 1980).

This also holds for an even broader analysis of the social acts and interaction accomplished by the use of discourse in the social situation. Verbal interaction sequences involve strategic moves that go beyond the classical analysis in terms of speech acts. In our own research on the expression and communication of ethnic prejudice in discourse, for instance, we found that (white) speakers --and writers-- have recourse to semantically, pragmatically and rhetorically based strategic moves (van Dijk, 1984, 1986). These try to combine two major goals, viz. the negative presentation of the 'others' (ethnic minority groups), and the positive presentation of 'self, in the form of the usual face-saving tactics that must avoid one's being considered a racist when saying negative things about 'foreigners'. The point for the present discussion is that such interaction strategies are coded at many levels of discourse structure. Appropriate interpretation of text and talk must thus take into account such social interaction functions, also in the automatic analysis of such discourse.

Discourse structures, speech acts and social interaction in contexts are literally meaningless without planning, intentions, interpretation and representation, and hence need a cognitive foundation. Again, such a cognitive account is crucial for both the theory and the automatic analysis of texts. Rapid developments in psychology and Artificial Intelligence in the past decade 'have shown that the understanding of texts (and hence of sentences) cannot just be accounted for in terms of linguistic-semantic interpretation rules and some kind of lexicon. Rather, we need a cognitive theory featuring flexible strategies of analysis, interpretation, representation and retrieval, at all levels of linguistic and social structure of the communicative event. Such a strategic acc-unt is fundamentally different from a grammar or an algorithmic parser (van Dijk & Kintsch, 1983).

The second major component of such a cognitive foundation is the role of knowledge, which has received i: major attention in AI, and which has resulted in the proposal of knowledge structures in memory in terms of 'scripts', 'frames' and 'models' (Schank & Abelson, 1977). We return to these cognitive structures and strategic7 below. That these are relevant for automatic analysis has been amply demonstrated by a decade of Al work on text and dialogue, which hardly needs tr; be spelled out here (see, e.g., Schank, 1982, for references).

The crucial conclusion from tL e developments is that the natural or automatic (simulated) analysis of sente-tial and textual structures as well as their interpretation is impossible without a full-fledged cognitive theory of strategic processing and representation in memory. To put it bluntly: automatic text analysis might be feasible without grammar, but is unthinkable without a cognitive component. Trivially, we all know that interpretation without knowledge and beliefs is impossible.

From the assumptions and statements in this section we conclude that automatic text analysis that is limited to a linguistic account of sentence or even of text structures will never lead us to a successful simulation of what we are all after, viz. the understanding of discourse by (if not communication with) computers. And without such an understanding of individual texts, automatic analysis of text corpora must remain severely limited and ad hoc. At all levels text and talk signal their links with pragmatic, interactional, social and cognitive structures and processes. To understand and represent such structural signals means that our computer programs must simulate these contextual structures and strategies.

3. Textual structures

Of course, such a programmatic statement carnes us across into the third millenium, given the present limitations of computer science (and computers), linguistics, discourse analysis, psychology, AI and the social

sciences. To remain with our feet on the ground of realistic research programs, we need to focus on a few central fines of investigation. In the remainder of this talk, therefore, I attended to in these posible research programs.

We have stress' !he relevance of a complex set of contextual theories for the development of a theory of natural or automatic analysis and understanding of text and -l[11]c. This does not mean that we should abandon the investigation of relevant text structures, no more iban we should forget about sentence grammars parsers. Only, we now have a more adequate framework to develop an account of such structures, because we are able to analyze and explain t, relation to various contextual structures and processes. Let us examine a few examples that are directly relevant for the computer analysis of text corpora.

Topics

Few analyses of text and talk are interesting without an account of the overall meanings, themes or topics of discourse. Both for a pragmatic ínterpretation of intended global speech acts, and for an account of local meanings of sentences, it ís crucial that we know what a text 'is about', globally speaking. In our own work since the end of the 1960s we have repeatedly stressed this dímension of díscourse, and we proposed an interdisciplínary theory of such themes or topics in terms of (semantic) macrostructures (van Dijk, 1980). In formal terms, such macrostructures are defined in tercos of an (hierarchically) ordered set of (macro-)propositions that are derived from sequences of textually expressed proposítions, so-called 'episodes', by means of recursive macrorules (deletion, generalization and construction). Such tules are a formal equivalent of principies that determine what is most 'important' or 'relevant' in a text. They define the overall coherence of a text, which in turn allows us to define texts as structured units beyond a simple concatenation of sentences or propositions. At various levels, thus, macropropositions make explicit what we understand by the topics of a text.

Interestingly, these macrostructures are not merely abstract, underlying structures of discourse. They are also routinely signalled or expressed in text and talk, viz., by way of titles, headlines, thematic sentences or paragraphs (such as leads in news reports in the press), and ín general by summaries. For automatic text processing, the possibilíty to derive 'abstracts' from texts is of course of crucial importance. Macrorules are essentially semantic reduction rules, which map long and complex proposition sequences onto a much smaller and manageable *sequence* of macroproposítions that underly what we call the summary, abstract, gist or upshot of text ór talk. Many analyses have shown that both the notion of macrostructure, as well as the rules that generate them, are crucial in the description of texts. Similarly, the cognitive counterpart of these structures

and rules, viz. more flexible strategies of macroprocessing, have proved to forro the core of an account of discourse understanding. And computer simulations have demonstrated that such rules or strategies can be successfully implemented in working programs (see e.g., Miller, 1985).

As usual there are problems, and these need our attention in the future. Like all formal interpretation rules and cognitive strategies of understanding, macroprocesses presuppose vast amounts of knowledge. Texts have the well-known and in practice most annoying property of being highly implicit: for both pragmatic and cognitive reasons, speakers or writers in principie only assert information the recipient doesn't know. To infer overall topics from text, language users apply flexible strategies that work fine with incomplete information from text and context: the rest of the many 'missing links' is supplied by their extant knowledge, e.g. by scripts or situation modeis (see below). We only have to read that Laura wanted to visit her aunt in Paris, that she went to the station, bought a ticket and boarded a train, in order to infer that she is making a traintrip to Paris to visit her aunt, which would be an acceptable summary of a possibly long story. Such macro-inferences are derived from both textual and scriptal information represented in memory.

For automatic derivation of summaries and abstracta, therefore, we need an explicit simulation program of such macro-strategies. Such a program should of course specify the relevant knowledge structures --which is the most daunting task for any work on automatic text processing now and in the near future. Secondly, the program must specify the flexible strategies that access, activate, retrieve, apply and deactivate such knowledge. After all, to understand a newspaper story about a car accident, we do not, and need not activate all our knowledge about cars, driving, traffic rules, roads, cities, actors or movement. Strategic operations will activate and apply only the relevant information, a process monitored by the level and extension of actually occurring topics in tests. Thus, thirdly, the program must indicate the textual conditions that allow macro-interpretation, such as the level, specificity or number of propositions and their ordering. The vital role of macro-interpretation in understanding has taught language users to be experts in making fast (but tentative) inferences about textual topics, as soon as they have heard or read the first few words or sentences.

Textual organization will often help. Hence the relevance o' initial titles, headlines, summaries or announcements in many forms of text or talk, e.g. in news reports, everyday stories and scholarly publications (Flammer & Kintsch, 1982; Mandl, Stein & Trabasso, 1984; Britton & Black, 1985.) Once readers have a plausible topic, it will play a central role in the top-down interpretation of more local information of words and sentences and their connections. Given the topic 'John went to the movies', a reader knows that the sequence 'He bought a ticket and went to his sea', is

fine, whereas the sequence 'He bought a ticket and boarded a train' is somehow strange or wrong, which however would be OK if the topic and its concomitant script involved travel, as in the example we gave aboye.

We conclude that one very important module in a computer program for text analysis and text understanding is such a complex macro theory. This theory will have both textual and cognitive dimensions, and features complicated strategies that make use of different types of information, both from text (and context) and from knowledge and belief structures in memory. Satisfactory automated abstraction will be feasible only when we have spelled out all these representations and strategies in explicit detall. This won't be easy for another reason: we have stressed that cognitive strategies are unlike rules and algorithms. Yet, these are the kind of formal procedures our present day computers want in their programs. The more tentative, flexible, goal oriented, context dependent nature of strategies is much harder to program. This, then, will be one important fine of research for the future.

Schemata

A similar story may be told for the understanding of the schematic structures of texts, so-called 'superstructures' (van Dijk, 1980). In a way that resembles the organization of sentences, a text also requires an overall form in order to organize and categorize its overall 'content' (i.e. its topics or macro-structure). This global form or schema, which can be formally characterized in terms of conventional categories and (formation) rules, can be found in many types of discourse, e.g. in conversation, stories, newspaper reports, argumentations, and, again, scholarly articles. We have mentioned headlines and leads, and also initial summaries, and these are one (important) type of schematic category. Similarly, we have conclusions, endings, closings, final summaries, evaluations, morals and other forms of 'coda' that close discourse or communicative events. Depending on text type, we find many different categories in between such beginnings' and 'ends'. These categories have a functional nature:

a story fragment summarized by the macroproposition (topic) 'The train had an accident', could fun-tion as the Complication category in a story about the trip to Paris.

Cognitively, schematic global structures operate in collaboration with macrostructurcs. But since the categories are conventional, and hence their possible patterns known, they have a potent top-down role in understanding text We recognize news reports or stories from the stars, retrieve the schema, and use this for the interpretation and representation of their topics, and hence of the propositional sequences that realize these topics.

Similar conclusions hold for programming superstructures for automatic analysis and interpretation. A story about a car accident may be

formally different from a police report about the 'same' accident, and this may also come out in summaries, and in the way the program understands such texts. In other words, for each sequence of sentences and their underlying propositions, we must know both its topic and its overall schematic function. Representation, summarization, question answering, translation, and many other processes vitally depend on these forms of global organization of texts, as well as on their cognitively represented presuppositions (knowledge of discourse schemata, knowledge about prototypical events and situations).

Local coherence

What has been stressed for the global levels of discourse naturally also holds for local interpretations which crucially depend on such global structures and strategies. The interpretation of words and especially of clauses and sentences and their connections is feasible only when monitored by higher level information, if we want to avoid the well-known combinatorial explosion (among other explosions).

One issue that is relevant here is the interpretation of topic/comment structures of sentences (where the notion of 'sentence topic' should not be confused with that of 'discourse topic' discussed aboye). The functional semantics of such structures must formulate the many conditions that organize the analysis and understanding of word, clause and sentence ordering (Givón, 1979). Without these conditions, no automatic question answering system could even come close to simulating natural communication. An explicit account of discourse dependent sentential topics is one of the preconditions for solving the well-known problems surrounding the automatic interpretation of corefential pronouns. Simple algorithms that look for last mentioned NP's or similar rules won't work, of course. We found in experimental research that language users interpret coreferential pronouns strategically along a topic continuity dimension: topical pronouns are predominantly interpreted as co-referring with the last previous topic, independent of position or grammatical function (van Dijk & Kintsch, 1983, also for a survey of the literature on pronoun and sentential topic interpretation). Topic continuity itself again depends on higher level macrostructure arguments, so that interpretation can fall back on co-reference with overall text arguments (e.g. a story about 'Peter') if local co-topical interpretation fails. Note that this is a strategic procedure: it is tentative but effective. The decisive interpretation of co-reference is of course based on local semantic coherence.

Like all semantic interpretations, pronoun interpretation and topic assignment has a cognitive basis (Reichman, 1981). Coreference and topicality are monitored and constrained by processing textual information in short term memory. Due to the limited storage capacity of STM, only a few new propositions are construed and connected in STM in each

interpretative cycle, after which storage in Episodic Memory becomes necessary. In order to establish local coherence the language user must thus keep at least one concept in the STM buffer, and this is what we call the sentence topic. The active macroproposition controls this process and may supply information to the local interpretation strategies, e.g. a dominant topic for a sequence of sentences.

Coherence and reference

Coreferential interpretation and topic assignment are elements of a more general process of local discourse interpretation. Any computer analysis of texts presupposes that we make explicit how sentences or propositions are connected in coherent sequences (Rollinger, 1984). The usual surface expressions of this underlying (semantic) coherence, viz., cohesion markers such as definite articles, pronouns, demonstratives or verb tenses, are merely strategically used signals of this coherence. What is involved in semantic coherence, however, are not meaning relations between propositions, but 'referential' relations between 'facts' denoted by propositions or the clauses that express them (van Dijk, 1977). To put it simply: two propositions are coherent if they refer to related facts. However, grammars do not feature notions like 'facts' or other referential elements. Originally, psycho-linguistics and the psychology of text processing didn't either: only conceptual links, for instance in a mental representation of the meaning of a text in STM or Episodic Memory, were assumed to exist.

Apparently, we need an additional type of knowledge representation in memory, viz. what we have called a (situation) model. These models are in a sense the mental correlate of models or model structures in a formal semantics: they represent the way language users subjectively store information about real or fictitious (world) events in memory (see below). Reference, and hence co-reference, is relative to such a model. If propositions of a text refer to, e.g. conditionally, related facts in such a model, they are interpreted as coherent. Since models may be subjective, coherence may similarly be subjective, and hence fail to be understood by the reader if no corresponding model or model fragment can be retrieved or constructed on the basis of textual information (Johnson-Laird, 1983; van Dijk & Kintsch, 1983; van Dijk, 1986c).

4. The cognitive dimension

From the remarks made aboye about the interpretation of discourse structures, it has become obvious that at each point we have to fall back on a cognitive account. This is not surprising, because interpretation processes are of course cognitive
rules are merely p iaibstractions of such cognitive processes. Therefore,

if we want the computer to analyze and understand text, we have no other alternative than to build in a full-fledged cognitive program that simulates how language users interpret texts. such a cognitive program, we have no way of accounting for knowledge , both general (semantic) and particular (episodio), and without knowledge , no interpretation is possible.

This is not the place to detail the complexities of a theory of cognitive processing. We only highlight a few features that have hitherto received little attention and that are of crucial relevance for the future development of computer-based textprocessing.

Strategies vs. rules or algorithms

A first typical feature of the cognitive approach has been briefly indicated aboye: understanding is not (only) rule based or algorithmic, but strategic. This means, among many other things, that language users (i) process information at several levels at the same time, which may mutually 'help' each other (hence no syntactic analysis without semantic analysis, and vise versa), (ii) use information from both text and perceptual, pragmatic or social contexts, (iii) use both 'external' information from text and context, and 'internar information from memory, (iv) make interpretative steps that are provisional or hypothetical, in need of later (dis-)confirmation, and (v) may set specific interpretation 'goals', which monitor depth, type or extent of interpretation. In other words, strategies are context dependent, goal directed, tentative but effective procedures to establish likely interpretations as quickly as possible, given the available resources and the operative time constraints. Further information may necessitate revisions of such hypothetical interpretations. The derivation of macrostructures, the establishment of local coherence, the interpretation of co-reference, or the assignment of topics and topical continuity, are all examples of such strategic processes (see van Dijk & Kintsch, 1983, for details).

We have repeatedly stressed aboye that discourse interpretation is knowledge based. Psychology and Al in the past decade have therefore focused on the development of theories and computer programs that specify what knowledge must be represented in memory, and especially how this knowledge is organized. The notion of 'script', among others, has been proposed to account for knowledge structures in 'semantic' (or rather 'social') memory about prototypical events, such as 'eating in a restaurant' or 'going to a party' (Schank Abelson, 1977; Schank, 1982). Apart from many other problems that must be solved in such a theory of knowledge representation, there is the problem of effective management: when, how, and how much of this knowledge is actually accessed and used in interpretation? We have suggested aboye that this process too is strategic. Depending on goals, current macropropositions (text topics) and local coherence, only relevant information need be retrieved and applied in interpretation.

Until a few years ago, it was assumed that cognitive text processing basically results in a (mostly semantic) 'text representation' (TR) in episodic memory, gradually built up through interpretation cycles taking place in STM. During understanding, parts of this TR could be 're-instated' to supply missing links or presuppositions for the interpretation of later sentences in the text. TR has a hierarchical structure, with macropropositions on top, and coherently connected (micro-)propositions at the bottom. This theoretical set-up nicely explained many properties of discourse understanding and recall, for instance why people remember macropropositions much better than. local details (Kintsch & van Dijk, 1978).

However, this approach also met with difficulties. What about the role of knowledge? It was assumed that such knowledge was simply derleved from scripts, and if needed inserted into TR to establish macrostructures or to provide the usual missing links in local coherence construction. *Yet,* for many reasons, this will not do. The textual representation would be inflated with a lot of knowledge that was not actually expressed in the text, and up to a point language users do know what was actually said, and what wasn't. More importantly, however, crucial notions such as coherence couldn't be explained in this way. Unlike many text grammars and text theories still maintain, propositions do not refer to other propositions. Hence, an independent type of memory representation was necessary, viz., the (situation) models we mentioned aboye.

Models

According to our present analysis, such models play a pivotal role in discourse understanding. They are the ultimate goal of understanding. We have 'understood' a text, if and only if we have been able to retrieve or build a model of the events or situation this text 'is about'. Reference, coreference and hence coherence are relative to such a model (Johnson-Laird, 1983; van Dijk & Kintsch, 1983). Cognitively speaking, reference to 'the world' or to indlviduals, events or properties in such a world, is of course meaningless: in a quasi-platonic way, we should say that we refer to our (subjective) interpretation and hence to our representation of world fragments, viz. to cognitive models of situations. The knowledge we presuppose during understanding is the ▾ nadan stored in such models. Models are unique, ad hoc, and about a single, particular situation construed for a particular text interpretation, or for particular communicative acts, speech acts or social interactions in general. Models are the episodic representations of our personal experiences. Text representations are merely a semantic interface between text and mor` 1, Especially after longer delays, when we say we recall a text, usually ;ve recall is (part of) the model bulelt from textual information. Hence, 1" 1, well-known incidence of 'false recall' after longer delays.

Importantly, the use of more general, script-like, knowledge in text understanding is never direct. It needs selection, specification and instantiation, that is, translation from general propositions to particular ones, and hence application within or through a concrete model. Depending on the other information in a model, this means that strategies are now able to activate, retrieve, and apply more or less the general information needed, which avoids the tricky explosion problems in automatic knowledge activation.

Models, thus, are the theoretical explanation of what we mean by saying that language users 'imagine' what a text is about during its understanding. They feature the 'facts' a text refers to, and their possible structures underly what we mean by the conditions of textual coherence.

For the effective processing of the vast amounts of textual and perceptual or interactional information that make up our daily experiences, models of situations must of course have a canonical schema that can be built, applied and filled with varleable information. It is therefore assumed that models are structured by a hierarchical schema, consisting of basic cognitive categories for the analysis of events or situations. There is psychological, linguistic and sociological evidence whlech suggests that these schemata are similar to the categorial structure we also flend in sentential semantics and narratives. After all, sentences and stories routinely describe events and situations, and it is plausible that some of the model structures are mapped into the semantic structures of the TR, and *hence* in those abstractly reconstructed in linguistic semantics for sentences and texts themselves. And vice versa: given specific semantic structures, it becomes easier to translate or insert their information into our episodic knowledge structures if their format is similar (van Dijk, 1986c).

Therefore, it is plausible that models feature categories such as Setting (Time and Location), Circumstances, Participants, and Action/ Event or Process, probably each with a possible Modifier, that is categories we also find in Case Grammars or Functional Grammars (Dik, 1978). This means, conversely, that the semantic structures we have become familiar with are in fact a result of more fundamental cognitive categories for the organization of experience in the form of episodic models.

We have mentioned these models in somewhat more detail because in my opinion they will become the core of future text analysis and comprehension programs for computers. That is, they serve as the essential knowledge base underlying all interpretation. They provide referents for (co-)referential expressions and feature the facts textual propositions refer to. They channel and control the strategic use of general (e.g. scriptal) knowledge. They are subjective and embody personal experiences and opinions besides other, socially shared, beliefs, and hence explain subjective understanding and communication problems. And since they also feature evaluative beliefs (opinions), they also are the interface between the

evaluative dimensions of understanding or communication and more fundamental social attitudes and ideologies. This type of intermediary function is necessary to explain biases in understanding. Conversely, particular models may be generalized and hence forro the basis of knowledge acquisition about the 'same' or 'similar' events and situations. Finally, models explain what knowledge 'updating' actually means: viz. the construction of a new model on the basis of previous ('old') models --and scripts-- , which will lead to the possible transformation of such extant models. If we read about Lebanon in the papen each day, what we do is not merely store hundreds of separate TR's, but each time contribute new elements to our extant model of the situation in that country (see van Dijk, 1986b, for details about the role of models in news comprehension).

These properties and functions of models are of course vital in any serious computer program of text interpretation and representation. In other words, in order to make explicit what a text means, we must spell out what a text is about. Question answering, inferences, co-reference interpretation, coherence assigment, topic continuity, summarization, or the interpolation of missing links, among many other tasks a computer program should be able to fulfill, presuppose an explicit model of the situation a text is about.

Strategically, the communicative context (which is also represented as a --specific-- model in episodic memory), may supply rich information about intentions, goals, setting, participants, speech acts, or social interaction that may be relevant in the provisional retrieval of old models and the partial con-otat ion of new models, even before a word has been heard or interpretad. Next, initial textual information at the macro-level (titles, summaries, annoannoncements introductions) provide the tentative macropropositions actívate and partly retrieve relevant models and scripts. At the same these topics 'define the situation', that is, they supply the higher :.macropropositions that dominate the model.

5. *Conclusion*

For computer programs of text understandleng all this implies that (i) local analysis of words and sentences must be complemented with global, macro-analysis of dominant, monitoring topics that define oyeran coherence and continuity, (ii) meaning semantics, whether linguistic or cognitive, must be based on a referential semantics, that is, on a model theory, and (iii) rule-based analysis and interpretation must be supplemented with flexible strategic processes. With these principles in mind, the computer interpretation of a text involves a double representation structure, viz. a textual representation of actually expressed information, and an underlying model embodying all presupposed as well as all resulting knowledge and beliefs. Each interpretation step takes place, tentatively, relative to this

model. Once this kind of 'double' interpretation of a text has been constructed, a number of specific tasks may be carried out on that basis, such as answering questions, providing summaries, or making (other) inferences. And many formal properties of texts which were difficult to deal with in current (linguistic) semantics, such as the use of pragmatic particles, pronouns, or certain features of word and clause ordering, can now be straightforwardly analyzed and interpreted relative to models.

These and other tasks will form the future goals of the automatic analysis of large text corpora. Therefore, besides *current* work on syntactic sentence parsers, semantic interpretation systems, or practical coding procedures, more fundamental research is necessary on the structures of discourse and communicative events. In order to automatize what language users do intuitively but effectively, this theoretical framework should have an important cognitive componen[, which specifies all processes of interpretation that are involved in corpus analysis. Until we have spelled out such a framework, our practical analyses will at most be useful, but necessarily limited *and ad* hoc.

References

Aarts, J., & Meijs, W. (Eds.). 1984. *Corpus linguistica. Recent developments in Me use of computer corpora in English language research.* Amsterdam: Rodopi.

Britton, B.K., & Black, J.B. (Eds.). 1985. *Understanding expository text.* Hillsdale, NJ: Erlbaum.

de Beaugrande, R., & Dressler, W.U. 1981. *Introduction to textlinguistics.* London: Longman.

Dik, S.C. 1981. *Functional grammar.* Amsterdam: North Holland.

Flammer, A., & Kintsch, W. (Eds.). 1982. *Discourse processing.* Amsterdam: North Holland.

Franck, D.M.L. 1980. *Grammatik und Konversation.* Kónigstein: Scriptor.

Givón, T. (Ed.). 1979. *Discourse and syntax.* New York: Academic Press.

Johnson-Laird, P.N. 1983. *Mental models.* Cambridge: Cambridge University Press.

Kintsch, W., & van Dijk, T.A. 1978. 'Toward a model of text comprehension and production'. *Psychological Review* 85, 363-394.

Mandl, H., Stein, N.L., & Trabasso, T. (Eds.). 1984. *Learning and comprehension of text.* Hillsdale, NJ: Erlbaum.

Miller, J. R. 1985. 'A knowledge-based model of prose comprehension: Applications to expository texts'. In B.K. Britton & J.B. Black (Eds.), *Understanding expository text,* pp. 199-226. Hillsdale, NJ: Erlbaum.

Reichman, R. 1981. *Plain speaking: A theory and grammar of spontaneous discourse.* Boston: Bolt, Beranek and Newman, Technical Report.

Rollinger, C-R. (Ed.). 1984. *Probleme des (Text-)verstehens. Ansdtze zur Künsdichen Intelligenz.* Tübingen: Niemeyer.

Schank, R.C. 1982. *Dynamic memory.* Cambridge: Cambridge University Press.

Schank, R.C., & Abelson, R.P. 1977. *Scripts, plans, goals and understanding.* Hillsdale, NJ: Erlbaum.

van Dijk, T.A. 1972. *Some aspects of text grammars.* The Hague: Mouton.

van Dijk, T.A. 1977. *Text and context.* London: Longman.

van Dijk, T.A. 1980. *Macrostructures.* Hillsdale, NJ: Erlbaum.

van Dijk, T.A. 1981. *Studies in the pragmatics of discourse.* Berlin: Mouton.

van Dijk, T.A. 1984. *Prejudice in discourse.* Amsterdam: Benjamins.

van Dijk, T.A. (Ed.). 1985. *Handbook of discourse analysis.* 4 vols London: Academic Press.

van Dijk, T.A. 1986a. *Communicating racism.* Beverly Hills, CA: Sage.

van Dijk, T.A. 1986b. News as discourse'. University of Amsterdam, Depi of General Literary Studies, Section of Discourse Studies. Unpublishe book ms.

van Dijk, T.A. 1986c. 'Episodic models in discourse processing'. In R. Horowitz & S.J. Samuels (Eds.), *Comprehending oral and written language.* New York: Academic Press.

van Dijk, T.A., & Kintsch, \V. 1983. *Strategies of discourse comprehension.* New York: Academic Press.