TOPIC Essentials*

Udo Hahn / Ulrich Reimer

Universitaet Konstanz
Informationswissenschaft
Postfach 5560
D-7750 Konstanz, F.R.G.

Abstract

An overview of TOPIC is provided, a knowledge-based text information system for the analysis of German-language texts. TOPIC supplies text condensates (summaries) on variable degrees of generality and makes available facts acquired from the texts. The presentation focuses on the major methodological principles underlying the design of TOPIC: a frame representation model that incorporates various integrity constraints, text parsing with focus on text cohesion and text coherence properties of expository texts, a lexically distributed semantic text grammar in the format of word experts, a model of partial text parsing, and text graphs as appropriate representation structures for text condensates.

1. Introduction

This paper provides an overview of TOPIC, a text understanding and text condensation system which analyzes German-language texts: complete magazine articles in the domain of information technology products. TOPIC performs the following functions:

* Text summarization (abstracting)
  TOPIC produces a graph representation of the most relevant topics dealt with in a text. This summary is derived from text representation structures and its level of generality varies from quite generic descriptions (similar to a system of index terms) to rather detailed information concerning facts, newly acquired concepts and their properties. Due to the flexibility inherent to this cascaded approach to text summarization (cf. KUHLEN 84) we refer to it as text condensation. This is opposed to invariant forms of text summarization based on summary schemata (DeJONG 79, TAIT 82) or structural features of the text representations (TAYLOR 74, LEHNERT 81), and dynamic abstracting procedures which depend on a priori specifications of appropriate parameters (FUM et al. 82) or rule sets for importance evaluation (FUM et al. 85) prior to text analysis.

* Extraction of facts / acquisition of new concepts
  Knowledge extraction resulting from text analysis not only leads TOPIC to the assignment of specific properties to concepts already known to the system, but also comprises the acquisition of new concepts and corresponding properties.

* Linking thematic descriptions with text passages
  TOPIC's analytic devices are by no means exhaustive to capture all the knowledge encoded in a text. Thus, the text representation structures provided might be incomplete. However, the thematic descriptions generated are linked to the corresponding text passages so that querying a text knowledge base may end up in the retrieval of relevant fragments of the original text (cf. similar approaches in LOEF 80, HOBBS et al. 82).

To perform these functions, the design of TOPIC is based on the following methodological principles:

* a method for making strategic decisions to control the depth of text understanding according to the functional level of system performance desired
* a knowledge representation model whose expressive power primarily comes from various integrity constraints which control the validity of the knowledge representation structures during text analysis
* a parsing model adapted to the specific constructive requirements of expository prose (local text cohesion and global text coherence phenomena)
* a text condensation model based on empirical well-formedness conditions on texts (text grammatical macro rules) and criteria derived from the knowledge representation model (complex operations)

2. Methodological Principles of Text Analysis Underlying the TOPIC Text Condensation System

Partial Text Parsing

The current version of TOPIC acts as a shallow understander of the original text (cf. the approach to 'integrated parsing' in SCHANK et al. 80). It concentrates on the thematic foci of texts and significant facts related to them and thus establishes an indicative level of text understanding. Partial parsing is realized by restricting the text analysis to taxonomic knowledge representation structures and by providing only those limited amounts of linguistic specifications which are needed for a text parser with respect to a taxonomic representation level. Primarily, the concepts which are available in the knowledge base correspond to nouns or nominal groups and their attributes (adjectives, numerical values).

A Frame Representation Model that Incorporates Various Integrity Constraints

The world knowledge underlying text analysis is represented by means of a frame representation model [REIMER/HAHN 85]. The large degree of schematization inherent to frame representations provides knowledge of the immediate semantic context of a concept (lexical cohesion). Additionally supplied integrity constraints formally restrict the execution of various operations (e.g. property assignment to support knowledge extraction from a text) in order to keep the knowledge base valid.

Text Parsing with Focus on Text Cohesion and Text Coherence Patterns

Text linguists seriously argue that texts constitute an object to be modeled differently from sentences in isolation. This is due to the occurrence of phenomena which establish textuality above the sentence level. A common distinction is made between local text

* The development of the TOPIC system is supported by BMFT/GID under contract 1020016 0. We want to thank D. Soergel for his contributions to this paper.

cohesion for immediate connectivity among adjacent text items (due to anaphora, lexical cohesion, co-ordination, etc.; see HALLIDAY/HASAN 76) and the thematic organization in terms of text coherence which primarily concerns the global structuring of texts according to pragmatic well-formedness constraints. Instances of global text structuring through text coherence phenomena are given by regular patterns of thematic progression in a text [DANES 74], or by various additional functional coherence relations, such as contrast, generalization, explanation, compatibility [REICHMAN 78, HOBBS 83]. Disregarding textual cohesion and coherence structures will inevitably result in invalid (text cohesion) and understructured (text coherence) text knowledge comparable to mere sentence level accumulations of knowledge structures which completely lack indicators of text structure. Therefore, there should be no question that specially tuned text grammars are needed. Unfortunately, the overwhelming majority of grammar/parser specifications currently available is unable to provide broad coverage of textual phenomena on the level of text cohesion and coherence, so that the format of text grammars and corresponding parsing devices is still far from being settled.

A Lexically Distributed Semantic Text Grammar

Since major linguistic processes provide textual cohesion by immediate reference to conceptual structures of the world knowledge, and since many of the text coherence relations can be attributed to these semantic sources, a semantic approach to text parsing has been adopted which primarily incorporates the conceptual constraints inherent to the domain of discourse as well as structural properties of the text class considered (for an opposite view of text parsing, primarily based on syntactic considerations, cf. POLANYI/SCHA 84). Thus, the result of a text parse are knowledge structures in terms of the frame representation model, i.e. valid extensions of the semantic representation of the applicational domain in terms of text-specific knowledge.

Text parsing, although crucially depending on semantic knowledge, demands that additional knowledge sources (focus indications, parsing memory, etc.) be accessible without delay. This can best be achieved by highly modularized grammatical processes (actors) which take over/give up control and communicate with each other and with the knowledge sources mentioned above. Since the semantic foundation of text understanding is most evidently reflected by the interaction of the senses of the various lexical items that make up a text, these modular elements themselves provide the most natural point of departure to propose a lexical distribution of grammatical knowledge [HAHN 86] when deciding on the linguistic organization of a semantic text grammar (ALTERMAN 85 argues in a similar vein).

Text Graphs as Representation Structures for Text Condensates

Knowledge representation structures built up during text parsing are submitted to a condensation process which transforms them into a condensate representation on different levels of thematic specialization or explicitness. The structure resulting from corresponding complex operations is a text graph (its visualized form resembles an idea first introduced by STRONG 74). It is a hyper graph which is composed of

* leaf nodes each of which contains a semantic net that indicates the topic description of a thematically coherent text passage
* the text passages that correspond to these topic descriptions
* the higher-order nodes which comprise generalized topic descriptions

From this condensate representation of a text access can also be provided to the factual knowledge acquired during text analysis. TOPIC does not include natural language text generation devices since the retrieval interface to TOPIC, TOPOGRAPHIC [HAMMWOEHNER/THIEL 84], is exclusively based on an interactive-graphical access mode.

3. An Outline of the Text Model

Despite the apparent diversity of linguistic phenomena occurring in expository texts, a large degree of the corresponding variety can be attributed to two basic processes (cf. HALLIDAY/HASAN 76): various forms of anaphora (and cataphora), and processes incorporating lexical cohesion. Both serve as basic text cohesion preserving mechanisms on the local level of textuality. Their repeated application yields global text coherence patterns which either follow the principle of constant theme, linear thematization of rhemes, or derived themes (see DANES 74). In Fig_1 we give a fairly abstracted presentation of these coherence patterns which should be considered together with the linguistic examples provided in Fig_2 and their graphical reinterpretation in Fig_3. The notions of frames, slots, and slot entries occurring in Fig_1 correspond to concepts of the world knowledge, their property domains, and associated properties, which may be frames again.

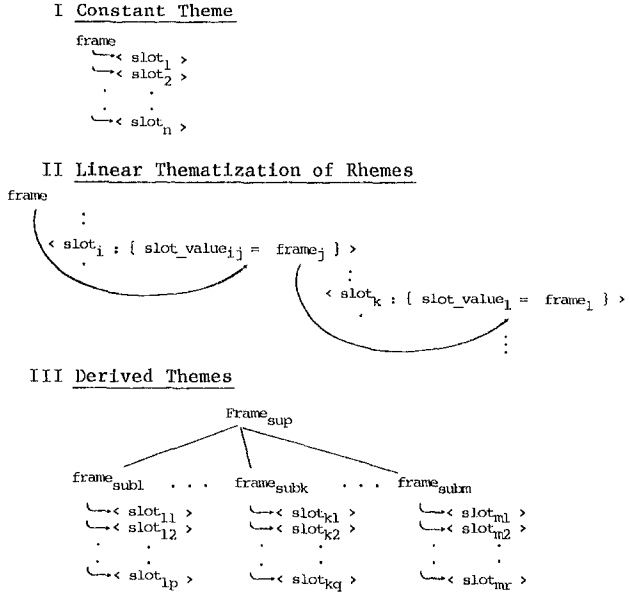


Fig 1: Basic Patterns of Thematic Progression

The interpretation of coherence patterns as given in Fig_1 refers to two kinds of knowledge structures:

* concept specialization corresponds to the phenomena of anaphora
* aggregation of slots to frames corresponds to the phenomena of lexical cohesion

This tight coupling of text linking processes and representation structures of the underlying world knowledge strongly supports the hypothesis that text

understanding is basically a semantic process which, as a consequence, requires a semantic text parser.

A linguistic illustration of the coherence patterns introduced above is given by the following text passages. For convenience, the examples in this paper are in English, although the TOPIC system deals with German textual input only.

I Constant Theme
The PC2000 is equipped with a 8086 cpu as opposed to the 8088 of the previous model. The standard amount of dynamic RAM is 256K bytes. One of the two RS-232C ports also serves as a higher-speed RS-422 port.

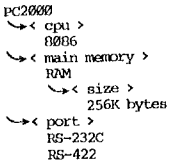
II Linear Thematization of Rhemes
A floppy disk drive by StorComp is available which holds around 710K bytes. Also available by StorComp is a hard disk drive which provides 20M bytes of mass storage.

III Derived Themes
Compared to the FS-190 by DP-Products which comes with Concurrent CP/M the PC2000 runs UNIX just like the new UNPC by PCP Inc.

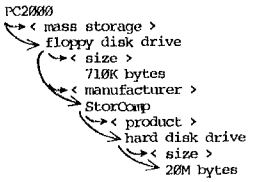
Fig 2: Linguistic Examples of the Basic Patterns of Thematic Progression

Fig 3 shows an interpretation of the text passages of Fig 2 in terms of thematic progression patterns.

I Constant Theme (PC2000)



II Linear Thematization of Rhemes (disk drives from StorComp)



III Derived Themes (personal computers)

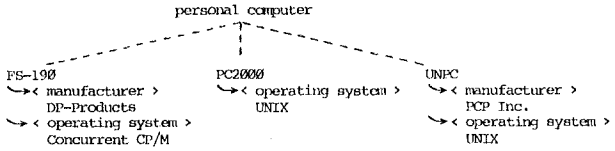


Fig 3: Interpretation of the Text Passages of Fig 2 in Terms of Thematic Progression Patterns

4. The Process of Text Parsing

TOPIC is a knowledge-based system with focus on semantic parsing. Accordingly, incoming text is directly mapped onto the frame representation structures of the system's predefined world knowledge without considering in-depth intermediate linguistic descriptions. Basically, these mappings perform continuous activations of frames and slots in order to provide operational indicators for text summarization. Together with slot filling procedures they build up the thematic structure of the text under analysis in the system's world knowledge base. To account for linguistic phenomena these concept activation and property assignment processes are controlled by a set of decision procedures which test for certain structural patterns in the world knowledge and the text to occur. Consequently, TOPIC's text parser consists of two main components: the world knowledge which provides the means of correctly associating concepts with each other (see sec.4.1) and the decision procedures (word experts) which utilize this foreknowledge to relate the concepts that are actually referred to by lexical items in a text, thus determining the patterns of thematic progression (see sec.4.2).

4.1 Representation of World Knowledge by a Frame Representation Model

Knowledge of the underlying domain of discourse is provided through a frame representation model [REIMER/HAHN 85] which supports relationally connected frames. A frame can be considered as providing highly stereotyped and pre-structured pieces of knowledge about the corresponding concept of the world. It describes the semantic context of a concept by associating slots to it which either refer to semantically closely related frames or which simply describe basic properties. A slot may roughly be considered as a property domain while actual properties of a frame are represented by entries in these slots (Fig 4). An entry may only be assigned to a slot if it is declared as being a permitted entry (see below).

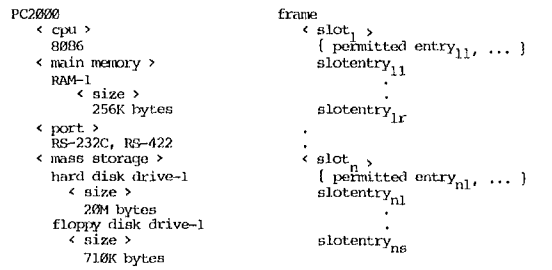
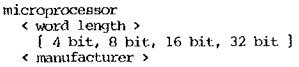


Fig 4: Examples of Frames, Slots and Slot Entries

Two kinds of frames are distinguished. A prototype frame acts as a representative of a concept class consisting of instance frames which all have the same slots but differ from the prototype in that they are further characterized by slot entries. Thus, instance frames stand for individual concepts of a domain of discourse. This point may be illustrated by a microprocessor frame which represents as a prototype the set of all microprocessor instances (Fig 5).

Prototype frame (concept class):



Associated instance frames (individual concepts):

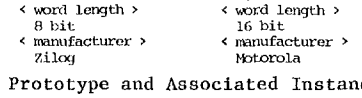


Fig 5: A Prototype and Associated Instance Frames

Frames are connected with each other by semantic relations (cf. Fig 6). Concept specialization between prototypes (is-a relation) is of fundamental importance to anaphora resolution. Concept specialization between a prototype and its instances (instance-of)

requires the instances to have the same slots as the prototype with the same set of permitted entries, resp. This property supports learning of new concepts from the text (i.e. incorporating new data in the knowledge base). When a new concept occurs in the text and it is possible to determine its concept class the structural description of the new concept is taken from the prototype that stands for the concept class. Indicators of what concept class a new concept belongs to are e.g. given by composite nouns, which are particularly characteristic of German language (8-Bit-Cpu, Sirius-Computer), attributions (serial interface, monochromatic display), or specific noun phrases (laser printer LA-9).

The semantic relation part-of is a special kind of aggregation which expresses a particularly tight semantic closeness.

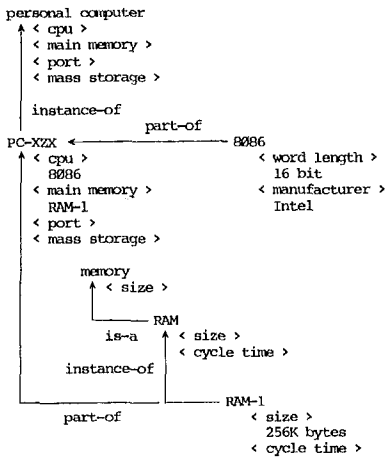


Fig 6: Semantic Relations among Frames

While the learning of new concepts is supported by the distinction of prototypes and instances, the acquisition of new facts from the text is possible by utilizing knowledge about the permitted entries of a slot. Two cases can be distinguished which correspond to two slot types. Non-terminal slots are slots whose name is identical to the name of a frame in the knowledge base. Permitted entries for them are defined implicitly and are given by all those frames which are specializations of the frame whose name equals the slot name (cf. the slot ´operating system´ in Fig_7). On the other hand, entries of the complementary class of terminal slots must be specified explicitly (cf. the slot ´word length´ in Fig_7).

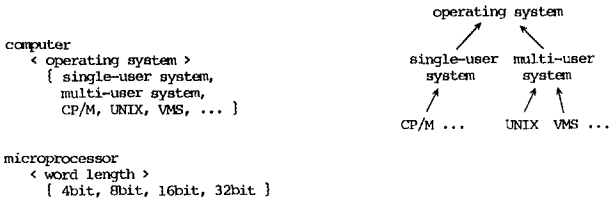


Fig 7: Permitted Entries for (Non-)Terminal Slots

Further devices for controlling slot filling are given by the construct of singleton slots which may hold at most one entry (e.g. the slots ´cpu´ and ´size´ in Fig_4). Singleton slots are of use when several permitted entries for a slot occur at adjacent text positions. Only if that slot is a singleton slot, the filling is constrained to one of those candidates; linguistic knowledge has to account for the selection of the appropriate one. Moreover, such a situation is interpreted as an indication of comparison (see Fig_2/I and the parsing effects occurring with respect to ´cpu´ and the candidate entries ´8086´ and ´8088´ in Fig_10).

Control of slot filling is also supported by an inferential construct called cross reference filling. When two frames, frame-1 and frame-2 (Fig_8), refer to each other in such a way that each has a non-terminal slot for which the other frame is a permitted entry, then assigning frame-1 to the appropriate slot of frame-2 automatically results in assigning frame-2 to the appropriate slot of frame-1. Now, if the second slot assignment is not permitted and therefore blocked, the primary assignment is blocked, too. The following sentence gives an example (Fig_8): "Compared to the FS-190 by DP-Products the PC2000 runs UNIX". The concept ´PC2000´ is a permitted entry of the product slot of the manufacturer ´DP-Products´. Its assignment would trigger the assignment of ´DP-Products´ in the manufacturer slot of ´PC2000´ which is a singleton slot and already occupied. Therefore no slot filling at all is performed.

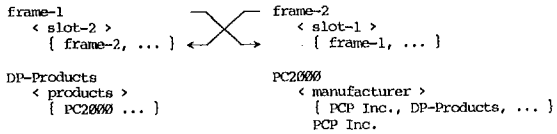


Fig 8: Cross Reference Filling

The structural features of the frame representation model are extended by activation weights attached to frames and slots. They serve the purpose of indicating the frequency of reference to the corresponding concepts in a text and are of significant importance for the summarization procedures.

Currently, TOPIC´s frame knowledge base comprises about 120 frames, an average of 6 slots per frame.

4.2 A Generalized Word Expert Model of Lexically Distributed Text Parsing

Characterizations of what texts are about are carried predominantly in domain-specific keywords as designators of contents (cf. SMETACEK/KOENIGOVA 77 for the task domain of abstracting) - in linguistic terminology: nominals or nominal groups. Accordingly, TOPIC´s parsing system is essentially based on a noun phrase grammar adapted to the requirements of text phenomena. Its shallow text parsing performance can be attributed to the exhaustive recognition of all relevant keywords and the semantic and thematic relationships holding among them. This is sufficient for the provision of indicative text condensates.

Accordingly, word experts [SMALL/RIEGER 82] have been designed which reflect the specific role of nominals in the process of making up connected text. The current section illustrates this idea through a discussion of a word expert for lexical cohesion (for a more detailed account cf. HAHN 86). Together with various forms of anaphora (not considered here, although we refer to the effects of a corresponding expert in Fig_10 by NA) it provides for a continuous cohesion stream and a corresponding thematic development in (expository) texts. Exceptions to this basic rule are due to special linguistic markers in terms of quantifiers, connectors, etc. As a consequence,

supplementary word experts have to be provided which reflect the influence these markers have on the basic text cohesion and text coherence processes: experts applying to quantifiers and comparative expressions typically block simple text cohesion processes (for an example cf. Fig_10), experts for conjunctions trigger them, and experts referring to negation particles provide appropriately modified assignments of properties to frames.

This kind of selective parsing is based on strategic considerations which, however, do not affect the linguistic generality of the approach at all. On the contrary, due to the high degree of modularization inherent to word expert specifications a word expert grammar can easily be extended to incrementally cover more and more linguistic phenomena. Moreover, the partial specifications of grammatical knowledge in the format of word experts lead to a highly robust parsing system, while full-fledged text grammars accounting for the whole range of propositional and pragmatic implications of a comprehensive understanding of texts are simply not available (not even in sublanguage domains). In other words, current text analysis systems must cope with linguistic descriptions that will reveal specification lags in the course of a text analysis if 'realistic texts' [RIESBECK 82] are being processed. Therefore, the text parser carries the burden of recovering even in cases of severe under-specification of lexical, grammatical, and pragmatic knowledge. Unlike question-answering systems, this problem cannot be side-stepped by asking a user to rephrase unparsable input, since the input to text understanding systems is entirely fixed. Distributing knowledge over various interacting knowledge sources allows easy recovery mechanisms since the agents which are executable take over the initiative while those lacking of appropriate information simply shut down.

Summing up, each of the word expert specifications supplied (those for nominals, quantifiers, conjunctions, etc.) is not bound to a particular lexical item and its idiosyncrasies, but reflects functionally regular linguistic processes (anaphora, lexical cohesion, coordination, etc.). Accordingly, a relatively small number of general grammatical descriptions encapsulated in highly modularized communities of agents form the declarative base of lexically distributed text parsing.

By word experts (consider the word expert prototype provided below) we refer to a declarative organization of linguistic knowledge in terms of a decision net whose root is assigned the name of a lexical class or a specific word. Appropriate occurrences of lexical items in the text prompt the execution of corresponding word experts. Non-terminal nodes of a word expert's decision net are constructed of boolean expressions of query predicates or messages while its terminal nodes are composed of readings. With respect to non-terminal nodes word experts

- query the frame knowledge base,e.g. testing for semantic relations (e.g. is-a, instance-of) to hold, for the existence and activation weight of concepts in the knowledge base, or for integrity criteria that restrict the assignment of slot entries
- investigate the current state of text analysis, e.g. the types of operations already performed in the knowledge base (activation, slot entry assignment, creation of new concepts, etc.)
- consider the immediate textual environment, e.g. testing co-occurrences of lexical items under qualified conditions, e.g. within sentence or noun phrase boundaries
- have message sending facilities to force direct communication among the running experts for blocking, canceling, or re-starting companion experts

According to the path actually taken in the decision net of a word expert, readings are worked out which either demand various actions to be performed on the knowledge base in order to keep it valid in terms of text cohesion (incrementing/decrementing activation weights of concepts, assignment of slot entries, creation of new frames as specializations of already existing ones, etc.), or which indicate functional coherence relations (e.g. contrast, classificatory relations) and demand overlaying the knowledge base by the corresponding textual macro structure. Apparently, the basic constructs of the word expert model (query predicates, messages, and readings) do not refer to any particular domain of discourse. This guarantees a high degree of transportability of a corresponding word expert grammar.

The word expert collection currently comprises about 15 word expert prototypes, i.e. word experts for lexical classes, like frames, quantifiers, negation particles, etc. Word expert modules encapsulating knowledge common to different word experts amount to 20 items. The word expert system is implemented in C and running under UNIX. Grammatical knowledge is represented using a high-level word expert specification language, and it is inserted and modified using an interactive graphical word expert editor.

These principles will be illustrated by considering an informal specification of a word expert (a more formal treatment gives HAHN/REIMER 85) which accounts for lexical cohesion that is due to relations between a concept and its corresponding properties.

Fig_10 shows a sample parse of text I (Fig_2) which gives an impression of the way text parsing is realized by word experts that incorporate the linguistic phenomena just mentioned.

With respect to text summarization (cf. HAHN/REIMER 84) it is an important point to determine the proper extension of the world knowledge actually considered in a text as well as its conceptual foci. This is achieved by incrementing activation weights associated to frames and slots whenever they are referred to in the text (this default activation process is denoted DA in Fig_10). In order to guarantee valid activation values their assignment must be independent from linguistic interferences. As an example for a process that causes illegal activation values consider the case of nominal anaphora which holds for [17] in Fig_10 (the associated word expert NA is not considered here, cf. HAHN 86).

Recognizing lexical cohesion phenomena contributes to associating concepts with each other in terms of aggregation. The word expert for lexical cohesion, an extremely simplified version of which is given in Fig_9, tests if a frame refers to a slot or to an actual or permitted entry of a frame preceding in the text. In the case of a slot or of an actual entry the activation weight of the slot (entry) is incremented; in the case of a permitted entry the appropriate slot filling is performed, thus acquiring new knowledge from the text. Examples of lexical cohesion processes

are given by positions [02/07], [07.1/24], [24/26], [26.2/32], and [32.1/38] in Fig 10.

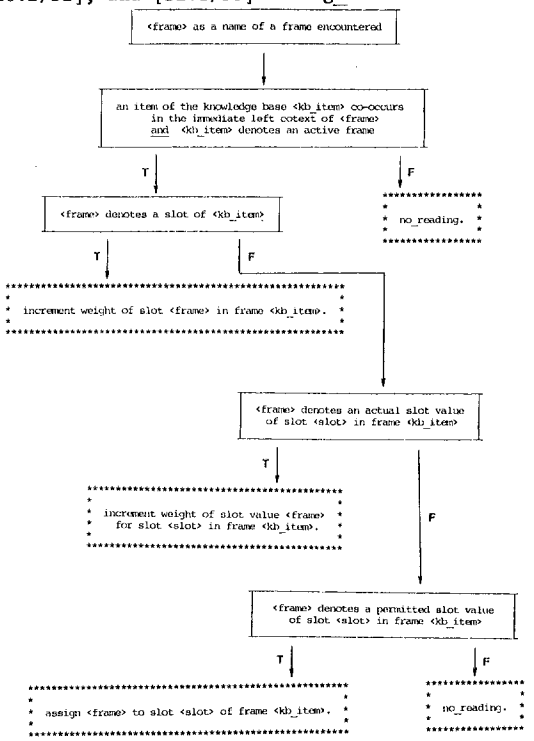


Fig 9: Word Expert for Lexical Cohesion (= LC)

Fig 10 shows a sample parse with respect to the text I given in Fig 2. It includes all actions taken with respect to nominal anaphora resolution (NA) and lexical cohesion recognition (LC).

```
[Ø2]    PC2ØØØ              DA: 'PC2ØØØ': Ø ---> 1
[Ø7]    8Ø86                DA: '8Ø86':   Ø ---> 1
[Ø7.1]                      LC: PC2ØØØ < cpu: 8Ø86 >
[Ø9/1Ø] opposed to          < start blocking of LC >
[13]    8Ø88                DA: '8Ø88': Ø ---> 1
[13.1]                      LC: PC-Ø1 < cpu : 8Ø88 >
[17]    model               DA: 'model': Ø ---> 1
[17.1]                      NA: 'model': 1 ---> Ø, 'PC-Ø1': 1 ---> 2
                            < stop blocking of LC >
[24]    RAM                 DA: 'RAM': Ø ---> 1
[24.1]                      LC: PC2ØØØ < main memory : RAM >
[26]    256K bytes
[26.1]                      LC: RAM-Ø1 < size : 256K bytes >
[26.2]                      LC: PC2ØØØ < main memory : RAM-Ø1 >
[32]    RS-232C             DA: 'RS-232C': Ø ---> 1
[32.1]                      LC: PC2ØØØ  < port : RS-232C >
[38]    RS-422              DA: 'RS-422': Ø ---> 1
[38.1]                      LC: PC2ØØØ  < port : RS-232C, RS-422 >
```

Fig 10: Sample Parse of Text Fragment I in Fig 2 Applying the Experts LC and NA

Some comments seem worthy:

1) [13.1]: The frame ´8088´ is not considered as an entry of the slot <cpu> of ´PC2000´ since it already has been assigned an entry and it is a singleton slot (cf. sec.4.1). Instead, a new instance of a personal computer is created (´PC-01´) to which ´8088´ is assigned as a slot entry
2) [24.1]: ´RAM´ does not refer to ´PC-01´ as might be expected from the specification of LC because a comparative expression ([09/10]) occurs in the text. This blocks the execution of the LC expert with respect to the noun phrase occurring immediately after that expression.
3) [26.1/26.2]: The instance created (´RAM-01´) describes the main memory of the ´PC2000´. Therefore it is assigned as an entry to ´PC2000´ and readjusts the previous assignment of ´RAM´.

Our constructive approach to text cohesion and coherence provides a great amount of flexibility, since the identification of variable patterns of thematic organization of topics is solely based on generalized, dynamically re-combinable cohesion devices yielding fairly regular coherence patterns. This is in contrast to the analytic approach of story grammars [RUMELHART 75] which depend completely on pre-defined global text structures and thus can only account for fairly idealized texts in static domains.

5. Text Condensation

During the process of text parsing, activation patterns and patterns of property assignment (slot filling) are continuously evolving in the knowledge base, which consequently exhibits an increasing degree of connectivity between the frames involved (text cohesion). If the analysis of a whole text would proceed this way, we would finally get an amorphous mass of activation and slot filling data in the knowledge base without any structural organization, although the original text does not lack appropriate organizational indicators. In order to avoid this deficiency, it is essential in text parsing to recognize topic shifts and breaks in texts to delimit the extension of topics exactly and to relate different topics properly. For this purpose every paragraph boundary triggers a condensation process which determines the topic of the latest paragraph (in the sublanguage domain we are working in topic shifts occur predominantly at paragraph boundaries). If its topic description matches with the topic description of the preceding paragraph(s), both descriptions are merged; thus they form a text passage of a coherent thematic characterization, called a text constituent. If the topic descriptions do not match a new text constituent is created. After the topic of a paragraph has been determined, the activation weights in the world knowledge are reset, except of a residual activation of the frame(s) in focus. This way the thematic characterization of a paragraph can be exactly determined without any interference with knowledge structures that result from parsing preceding paragraphs.

The next section presents the main ideas underlying the process of determining the thematic characterization of a text passage. Sec.5.2 concludes by giving a very concise discussion of the concept of a text graph which is the representational device for text condensates in the TOPIC system.

5.1 Determination of Text Constituents

The condensation process (for details cf. HAHN/REIMER 84) completely depends on the knowledge structures generated in the course of text analysis. As outlined above, this text knowledge consists of frame structures which have been extended by an activation counter associated to each concept (frame, slot, or slot entry) to indicate the frequency of reference to a concept in the text under analysis. These activation weights as well as distribution patterns of slot filling among frames together with connectivity patterns of frames via semantic relations provide the major estimation parameters for computing text constituents (connectivity approaches to text summarization are also described in TAYLOR 74, LEHNERT 81).

These indicators are evaluated in a first condensation step where the significantly salient concepts are determined. We distinguish between dominant

frames, dominant slots and dominant clusters of frames, the latter being represented by the common superordinate frame (for a detailed discussion see HAHN/REIMER 84). The determination of dominant concepts can be viewed as a complex query operation on a frame knowledge base. In a subsequent step the dominant concepts are related to each other with respect to concept specialization as well as the frame/slot relationship. The topic of a text passage is thus represented by a semantic net all of whose elements are given by the dominant concepts (cf. the nodes of the text graph in Fig 11).

5.2 The Text Graph

The text graph (Fig 11) is a hierarchical hyper graph whose leaf nodes are the text constituents (as given above) and whose higher-order nodes represent generalizations of their topics. Similar to the distinction of micro and macro propositions [CORREIRA 80] its nodes are associated by different kinds of relationships which are based on the frame representation model (is-a, instance-of, is-slot, identity) or which are constituted by the coherence relations (e.g. contrast).

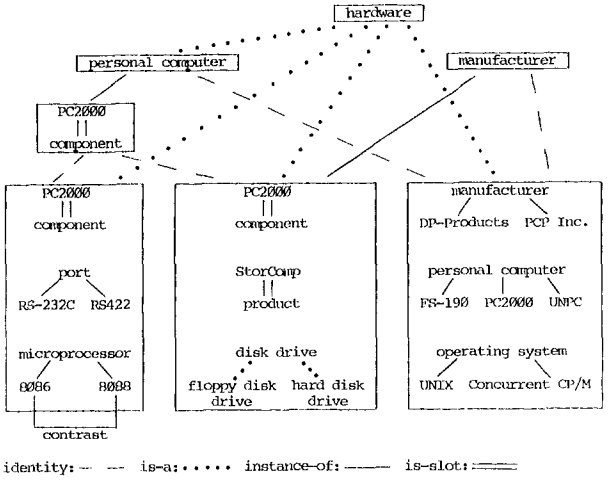


Fig 11: Text Graph for Text Fragments I-III (Fig 2)

6. Conclusions

A comprehensive description of the text condensation system TOPIC has been provided which serves for the conceptual analysis of textual input of a knowledge-based full-text information system. The following issues are most characteristic of it:

- a frame representation model which incorporates various integrity constraints
- a text grammar with focus on text cohesion and text coherence properties of expository texts
- a lexically distributed semantic text grammar in the format of word experts
- partial text parsing based on a noun phrase word expert parser and a taxonomic knowledge representation
- text graphs as representation structures of text condensates which provide different layers of informational specifity

References

Alterman, R.: A Dictionary Based on Concept Coherence. In: Art. Intell. 25. 1985, pp.153-186.

Correira, A.: Computing Story Trees. In: Amer. J. Comput. Ling. 6. 1980, pp.135-149.

Danes, F.: Functional Sentence Perspective and the Organization of the Text. In: Danes (ed): Papers on Functional Sentence Perspective. Academia, 1974, pp.106-128.

DeJong, G.: Skimming Stories in Real Time: an Experiment in Integrated Understanding. Yale Univ, 1979.

Fum, D. et al.: Forward and Backward Reasoning in Automatic Abstracting. In: Proc. COLING 82, pp.83-88.

Fum, D. et al.: Evaluating Importance: a Step towards Text Summarization. In: Proc. IJCAI-85, pp.840-844.

Hahn, U.: On Lexically Distributed Text Parsing: A Computational Model for the Analysis of Textuality on the Level of Text Cohesion and Text Coherence. In: Kiefer (ed): Linking in Text. Reidel, 1986.

Hahn, U.; U. Reimer: Computing Text Constituency: An Algorithmic Approach to the Generation of Text Graphs. In: Rijsbergen (ed): Research and Development in Information Retrieval. Cambridge U.P., 1984, pp.343-368.

Hahn, U.; U. Reimer: The TOPIC Project: Text-Oriented Procedures for Information Management and Condensation of Expository Texts. Final Report Univ. Konstanz, 1985 (TOPIC-17/85)

Halliday, M.; R. Hasan: Cohesion in English. Longman, 1976.

Hammwoehner, R.; U. Thiel: TOPOGRAPHIC: eine graphisch-interaktive Retrievalschnittstelle. In: Proc. MICROGRAPHICS. GI, 1984, pp.155-169.

Hobbs, J.: Why is Discourse Coherent? In: Neubauer (ed): Coherence in Natural-Language Texts. Buske, 1983, pp.29-70.

Hobbs, J. et al.: Natural Language Access to Structured Text. In: Proc. COLING 82, pp.127-132.

Kuhlen, R.: A Knowledge-Based Text Analysis System for the Graphically Supported Production of Cascaded Text Condensates. Univ. Konstanz, 1984 (TOPIC-9/84)

Lehnert, W.: Plot Units and Narrative Summarization. In: Cognitive Science 5. 1981, pp.293-331.

Loef, S.: The POLYTEXT/ARBIT Demonstration System. Umea/Sweden: Foersvarets Forskningsanstalt, FOA 4 rapport, C 40121-M7, 1980.

Polanyi, L.; R. Scha: A Syntactic Approach to Discourse Semantics. In: Proc. COLING 84, pp.413-419.

Reichman, R.: Conversational Coherency. In: Cognitive Science 2. 1978, pp.283-327.

Reimer, U.; U. Hahn: On Formal Semantic Properties of a Frame Data Model. In: Computers and Artificial Intelligence 4. 1985, pp.335-351.

Riesbeck, C.: Realistic Language Comprehension. In: Lehnert / Ringle (eds): Strategies for Natural Language Processing. Erlbaum, 1982, pp.37-54.

Rumelhart, D.: Notes on a Schema for Stories. In: Bobrow / Collins (eds): Representation and Understanding. Academic P., 1975, pp.211-236.

Schank, R. et al.: An Integrated Understander. In: Amer. J. Comput. Ling. 6. 1980, pp.13-30.

Small, S.; C. Rieger: Parsing and Comprehending with Word Experts (a Theory and its Realization). In: Lehnert / Ringle (eds): Strategies for Natural Language Processing. Erlbaum, 1982, pp.89-147.

Smetacek, V.; M. Koenigova: Vnimani odborneho textu: experiment. In: Ceskoslovenska Informatika 19. 1977, pp.40-46.

Strong,S: An Algorithm for Generating Structural Surrogates of English Text. In:JASIS 25.1974, pp.10-24

Tait, J.: Automatic Summarising of English Texts. Univ. of Cambridge, 1982 (= Technical Report 47)

Taylor, S.: Automatic Abstracting by Applying Graphical Techniques to Semantic Networks. Evanston/Ill.: Northwestern Univ., 1974.