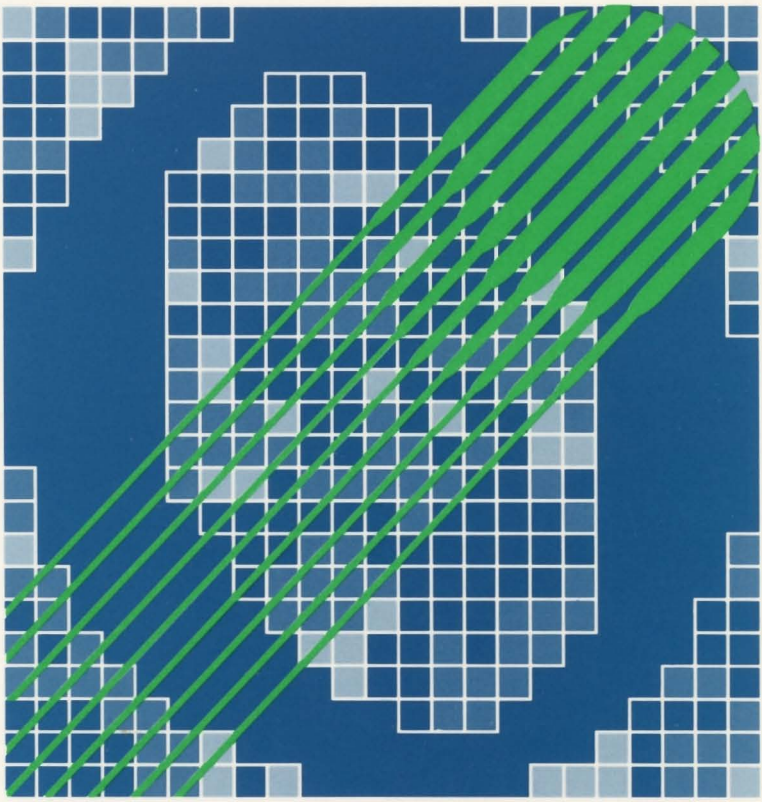
First publ. in:

ONLINE 89 INFORMATION



13th International Online Information Meeting

Proceedings

12–14 December 1989
London • England

Learned Information • Oxford and New Jersey

HYPER-TOPIC – A SYSTEM FOR THE AUTOMATIC CONSTRUCTION OF A HYPERTEXT-BASE WITH INTERTEXTUAL RELATIONS

R. Kuhlen, F. Yetim, Universität Konstanz, FRG

ABSTRACT
This paper reports on the extension of an automatic text analysis system (TOPIC) to a hypertext-system (HYPER-TOPIC) which automatically builds up intertextual relations between semantic units from different texts. These units, which we call information units, are provided by TOPIC on the basis of coherent knowledge structures within texts and can be represented in flexible ways, i.e. as graphs, tables, pictures, automatically generated textual abstracts, or as passages of text from the original full text. By realizing intertextual relations between the different presentations of text units, HYPER-TOPIC allows comfortable and flexible navigation in the hypertextbase.

1. REPRESENTATION OF TEXT KNOWLEDGE

Methods of automatic text analysis together with techniques of knowledge representation make it possible to identify knowledge structures within texts and to present them in a flexible way. The traditional linear presentation of knowledge, as is common in printed publications or in full-text banks, can be completed and partially replaced by non-linear and non-verbal means. For instance, knowledge can be presented as networks, graphics, tables, etc. in addition to natural language. And relations between chunks of knowledge or semantically coherent information units can be realized in ways different from those which rely on textual coherence and cohesion. Moreover, these relations are not necessarily restricted to a single text.

The interrelation between texts represented in a hypertext basis is the major concern of this paper. In the last six years in the department for Information Science at the University of Constance we have developed the protytype of an automatic text analysis system (TOPIC), which produces text condensation structures from texts in the domain of information technology (Ref. 1). The system is based on a frame model (FRM) (Ref. 2) and a semantic partial parser (Ref. 3). Its performance consists mainly in partitioning the current text into semantically coherent parts, so-called text constituents. These units are presented as graphs or frame-networks (cf. Fig. 1).

According to the underlying frame model the nodes contain frames, slots or slot entries and are interrelated via mainly hierarchical relations such as instance, is-a, part-of. From Fig. 1 one can see that TOPIC has originally been designed as an abstracting/text condensation system. On the lowest level the basic text constituents represent the semantic structure of elementary text units, normally identical with text paragraphs. In the following, when we use the term "information unit", we refer to the different ways of representing these basic text constituents, be it as graphs, as tables, as text passages, as abstracts, or as pictures (cf. Fig. 2). Hierarchically higher nodes (in Fig. 1) represent more abstracts concepts which can be derived from the lower ones if these lower concepts have some features (frame, slot, slot entries) in common.

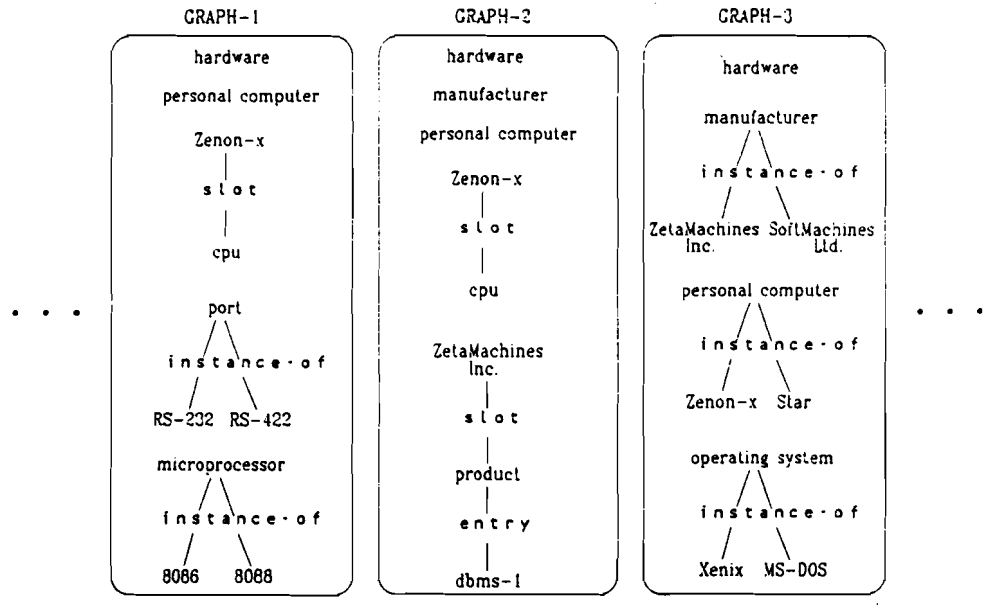


Fig. 1.: Presentation of semantic information units as conceptual graphs

The graphs are the basis for the other system, TWRM-TOPOGRAPHIC (Ref. 4), which has been designed as a graphic retrieval machine and which presents the inherent knowledge of the information units in a flexible and multi-medial way. For this flexible presentation we have introduced the concept of *cascaded condensation and presentation*. At the moment, the system can present knowledge, according to different users' needs, in the following ways: a) graphically as frame-based concepts of the pertinent domain of discourse; b) verbally as automatically generated textual abstracts; c) graphically as the thematic structure of a single part or different parts of the text; d) in tabular form as frames, slot, and slot entries; e) as pictures stemming from the original text; f) textual passages as excerpts from the underlying full text (cf. Fig. 2).

System performance indicates that, on the basis of current knowledge and in addition to other experimental intelligent retrieval systems, this is a realistic alternative to existing information retrieval systems, which tend to be rather rigid and linear. System performance also makes it possible to regard this as a hypertext system (Ref. 5) but, different from most existing or experimental hypertext systems, the hypertext base of TOPIC is built up automatically by an analysis of the underlying texts.

2. DESIGN OF A HYPER-TOPIC SYSTEM

The work we would like to report on here concentrates on overcoming the present restrictions in the system, in particular building up intertextual relations between information units and their respective presentation forms (Ref. 6). This is necessary because users of hypertext systems are typically not interested in the knowledge structure of a single text but regard the whole domain of knowledge as a base in which they can navigate. The new HYPER-TOPIC system allows intertextual navigation not only via pre-established system relations but also via relations which are built up *in question time* according to actually articulated user queries.

We differentiate between syntagmatic, paradigmatic, and pragmatic relations, which can be constructed automatically by taking advantage of the formal properties of frame-based partial text graphs (the representation of the information units). Syntagmatic relations, such as "next-passage-within-the-same-text", are comparable to

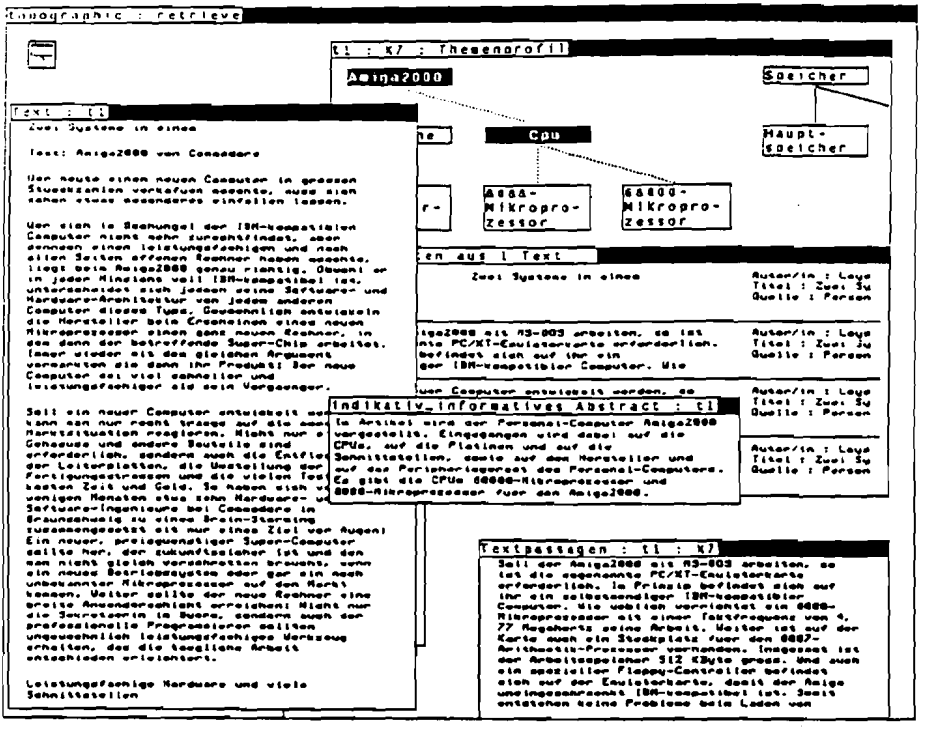


Fig.2.: Presentation of an entire screen with the objects: Text passage, abstract and full text (in the background the list of the relevant constituents and a theme profile)

[Source: Ref. 4]

context operators, well known in the domain of full-text retrieval. Paradigmatic relations, such as "have-same-feature", represent semantic relations between information units both within a single text and between units of different texts. And finally, pragmatic relations reflect the dependencies between information units and user intentions on the basis of the actual dialogue context.

As already mentioned, HYPER-TOPIC relies on the performance of the systems TOPIC and TWRM-TOPOGRAPHIC. Accordingly, the hypertext knowledge base (HTKB) of HYPER-TOPIC can be defined as a 6-tupel:

$$HTKB = <FT, PTG, TAB, ABS, GRA, f>$$

HTKB consists of the sets of full texts (FT) analyzed in the system, of partial text graphs (PTG), of derived tables (TAB), of generated abstracts (ABS), and of graphics available in the full-text version. f is a function which assigns to each information unit one element of the other sets, whereby the elements of TAB, ABS, and GRA are sets by themselves; they may contain multiple entries, such as different figures in one information unit. (We have to admit that using ABS as a *set* of automatically generated natural language abstracts is somewhat of an over-simplification because in reality it is one of the genuine performance characteristics of TWRM-TOPOGRAPHIC (Ref. 7) that the abstracts are not prefabricated in analysis time but are derived, in question time, by mapping actual user queries onto text graphs or parts of them. These abstracts are thus, in documentation terminology, slanted abstracts which occur on the text surface differently according to different users' needs.)

As already mentioned, an information unit is an abstract concept which refers to a semantically coherent piece of text knowledge and which, in combining the different presentation forms, is the main node type in the hypertext knowledge base. It is defined as a 5-tupel:

IU = <ft, ptg, tab, abs, gra>

with ft: full text (ft FT); ptg: partial text graph (ptg PTG); tab: tabular presentation of a frame structure (tab TAB); abs: natural language abstract (abs ABS); gra: graphic taken from the full text (gra GRA).

3. INTERTEXTUAL RELATIONS

The main idea of a HYPER-TOPIC system consists in interrelating these abstract concepts (information units) by using syntagmatic, paradigmatic and pragmatic relations. The concrete shallow realization of the information units, be it as graphs, as abstracts, as tables, or as text passages, depends on the concrete retrieval dialogue situation or the actual user's query, respectively.

Fig. 3 gives examples of intertextual relations between information units. Whereas syntagmatic relations ('SYN-RELi' or 'SYN-RELj' in Fig. 3) refer to surface phenomena in full texts, paradigmatic relations ('PAR-RELi', 'PAR-RELj' etc. in Fig. 3) refer to partial text graphs (PTG).

3.1 SYNTAGMATIC RELATIONS

The following formal syntagmatic relations are realized. They correspond roughly to the context operators well known in the domain of full text retrieval:

next-passage-within-the-same-text
previous-passage-within-the-same-text
first-passage-within-the-same-text
last-passage-within-the-same-text.

3.2 PARADIGMATIC RELATIONS

Paradigmatic relations provide more information because they are based on semantic similarities and differences between information units. They can be built up automatically by taking advantage of the formal semantic structures of frame-based partial text graphs (for example, having the same or different frames or slots or slot entries). The usage of these relations depends on the retrieval results and the intentions articulated by the user of the system. The system firstly presents the thematic description of an information unit, and the user, starting from this information, selects the appropriate paradigmatic relations (appropriate to his/her needs) and may thus navigate in the set of all information units which have been selected as relevant to the original query.

Paradigmatic relations can be divided into two main groups: relations which are based on slots (properties of frame concepts) and relations which are additionnally based on concrete slot-entries. The following relations have been defined so far:

share-concept
have-same-features
have-same-info
have-additional-info
have-complement-info
have-additional-feature
have-alternative-feature
feature-coincidence
property-coincidence
additional-info-to-feature
alternative-info-to-feature
same-property.

To give a bit more information on how these relations are defined we look at two of them more in detail (for further information see Ref. 6):

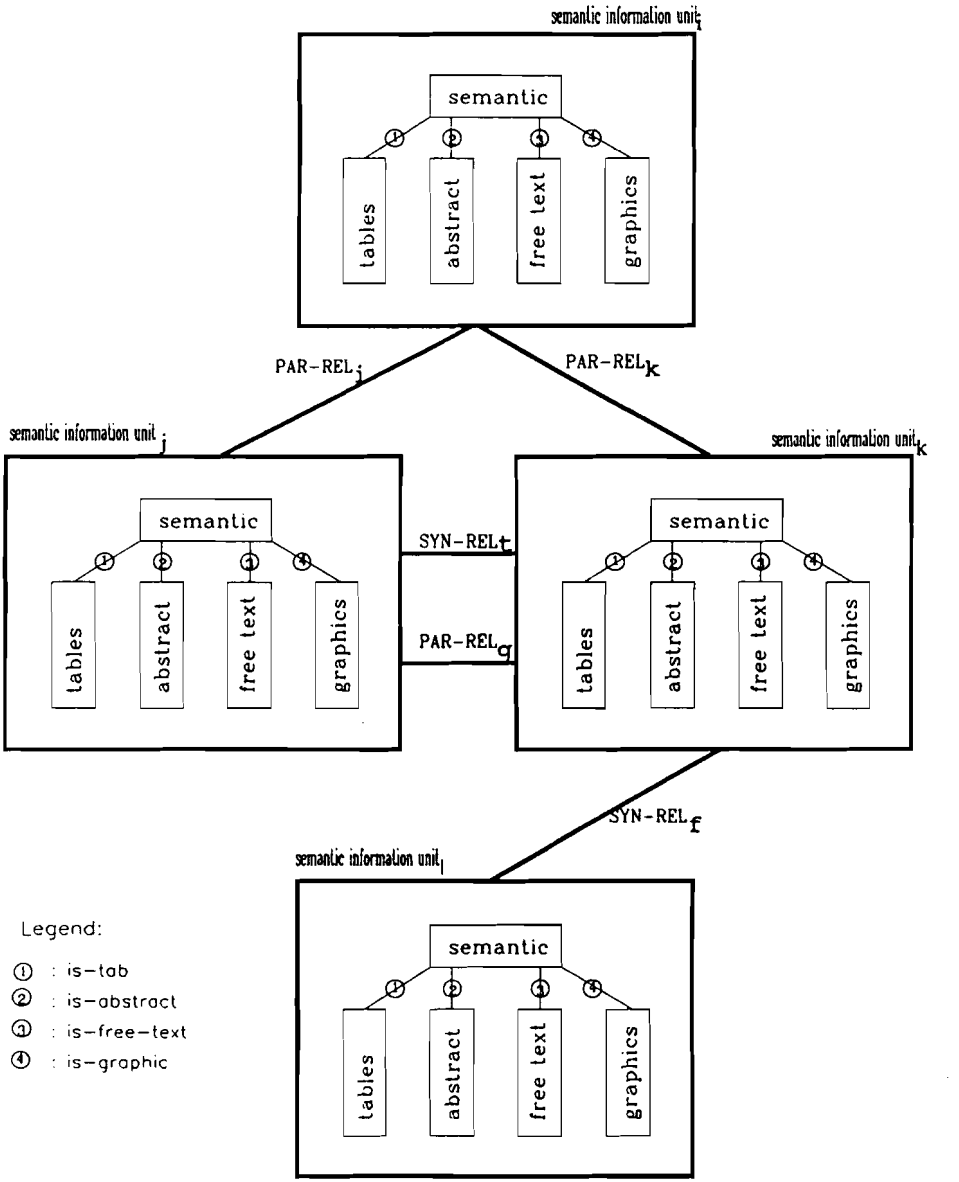
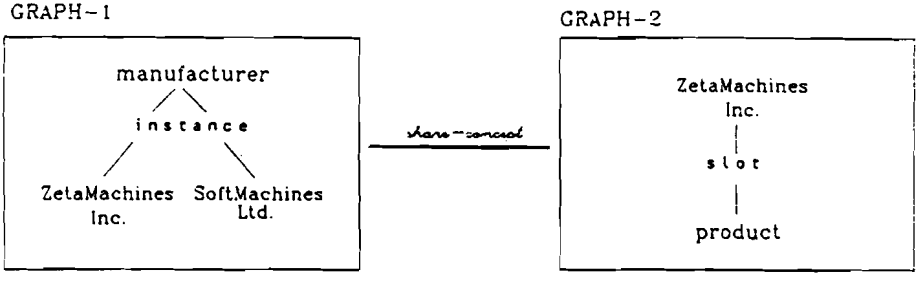


Fig. 3.: Presentation of semantic information units and their relationship

▪ share concept

Between two (partial) text graphs (ptg_1 and ptg_2) the relation *share concept* exists if they have at least one frame node in common (a frame node is a complex structure consisting of a frame name and its set of slots).

Example:

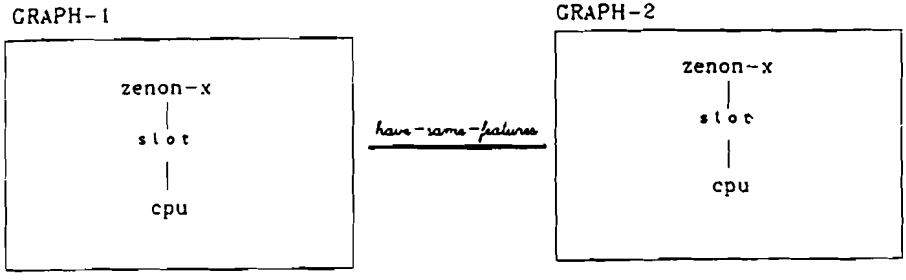


The *share concept* relation is very weakly constrained. By repeated selections of this relation the user may inspect text passages which are somehow thematically related. This relation causes - in terms of information retrieval theory - high recall but eventually low precision.

▪ have-same-features

Between the two partial graphs ptg_1 and ptg_2 in the following example the system can identify the relation *have-same-features* if the ptg_i have common frame nodes with the same slot nodes. By selecting this relation several times it is possible to look at many information units in different texts which deal with the same topic seen from different authors' point of view.

Example:



3.3 PRAGMATIC RELATIONS

Pragmatic relations determine the relevance of information units with respect to the actual retrieval or dialogue context and the user's intention. The evaluation of the relevance also allows a hierarchical ordering (*ranking*) of the different information units selected. The relevance relation determines to what extent a query matches the semantic representation of a partial text graph (ptg).

The set of relevant information units (RIU) is thus determined by the relation "share-concept" which exists between the concepts (nodes) of a query Q and the concepts (nodes) of the pertinent graph.

$$RIU := \{G \mid \exists Q \ (share - concept\,(Q, G))\}$$

This relation is very weakly determined. By applying it, every information unit is considered as relevant whose graphs have at least one concept in common with the query. In case of a large set of relevant information units it is useful to evaluate the

relevance of each single unit in order to produce a hierarchical order (a ranking) in the hypertext graph. The relevance is defined as follows[1]:

$$\forall Q\, \forall G\, \forall r:\ \text{relevance}(Q, G, r) \Leftrightarrow r = \sum_{n_1 \in Q} \sum_{n_2 \in G} \text{weight}(n_1) \times \text{equal}(n_1, n_2) + + \log\left(\sum_{G' \in RIU} \text{rel}(G, G') \right)$$

The relevance relation *relevance(Q, G, r)* determines to what extent the query (Q) overlaps with the semantic representation (G) of an information unit. r represents the measure of the relevance. The relevance of an information unit is calculated by adding the weight of those concepts of Q which are identical with the concepts from G. The function *weight* gives the weight of a concept. The weight (between 1 and 10) is given by the user in retrieval time. The function *equal* examines the identity of two nodes. Its value is 1 if they are identical, 0 if not. In addition, the relevance relation takes into account the subset of information units (G') of the set RIU which are related with the information unit (G) over the *share concept* relation. Because this relation may produce large sets, the value is given in its logarithmic form.

Relevance in hypertext systems refers to single pieces of knowledge. This is the main difference to full text retrieval systems which, even in their passage retrieval form, primarily intend to inform users about whole documents. Hypertext is based on the de-linearization of text. Consequently, the evaluation of the relevance of an information unit in HYPER-TOPIC does not consider the relevance of the whole text. The relevance relation is a measure for discrete information units.

4. CONCLUSION AND OUTLOOK:

The results achieved so far show that a hypertext knowledge base can be built up automatically by text analysis procedures such as those provided by TOPIC. Furthermore, the knowledge structures of the pertinent information units can be presented in a flexible and multi-modal way. HYPER-TOPIC allows intertextual navigation in the related knowledge structures of a homogeneous, domain-specific text base. The main advantage of the system, although still in a very limited experimental stage, lies in the automatic construction and controlled maintenance of the complicated relational structure. Further work will concentrate a) on the development of text relations such as causal, conditional and temporal relations, b) on the development of methods by which the quantity barrier can be overcome; and c) on evaluating the acceptance of ambitious hypertext systems, in particular to what extent users of such systems are willing and able to handle the complex relational structure and the flexible knowledge presentation provided by systems such as HYPER-TOPIC.

TOPIC, TWRM-TOPOGRAPHIC are realized in C and PROLOG on UNIX machines; HYPER-TOPIC is in the process of being developed in the same programming environment.

Prof. Dr.Rainer Kuhlen / Dipl.-Inf.Wiss. Fahri Yetim
Department of Information Science at the University of Constance
Box 5560, D-7750 Konstanz

1 As a modified version of the formula proposed in Hammwöhner/Thiel 1987

REFERENCES

1. Hahn, U./ Reimer, U. 1986: Semantic Parsing and Summarizing of Technical Texts in the TOPIC System. In: Kuhlen, R. (ed.): Informationslinguistik. Theoretische, experimentelle, curriculare und prognostische Aspekte einer informationswissenschaftlichen Teildisziplinen. Tübingen, Niemeyer, 153-193.

2. Reimer, U. 1989: FRM: Ein Frame-Repräsentationsmodell und seine formale Semantik. Zur Integration von Datenbank- und Wissensrepräsentationsansätzen. In: Informatik-Fachberichte, Vol.198. Berlin et al.:Springer.

3. Hahn, U. 1987: Lexikalisch verteiltes Text-Parsing: Eine objektorientierte Spezifikation eines Wortexpertensystems auf der Grundlage des Aktorenmodells. Konstanz: Universität Konstanz, Sozialwissenschaftliche Fakultät (Dissertation).

4. Kuhlen, R. / Hammwöhner, R./ Sonnenberger, G / Thiel, U. 1989: TWRM-TOPOGRAPHIC. Ein wissensbasiertes System zur situationsgerechten Aufbereitung und Präsentation von Textinformation in graphischen Retrievaldialogen. In: Informatik Forschung und Entwicklung (1989)4: 89-107. Berlin et al.: Springer.

5. Hammwöhner, R./ Thiel, U. 1987: Content Oriented Relations between Text Units - A Structural Model for Hypertexts. In: Hypertext '87 Papers. Chapel Hill, N.C.: Univ. of North Carolina, 155-174.

6. Yetim, F. 1989: Ein Hypertextmodell für intertextuelle Relationen. Konstanz: Universität Konstanz, Informationswissenschaft, Februar 1989 (Diplomarbeit).

7. Sonnenberger, G. 1988: Flexible Generierung von natürlichsprachigen Abstracts aus Textrepräsentationsstrukturen. In: Trost, H. (ed.): 4. Österreichische Artificial-Intelligence-Tagung. Wiener Workshop - Wissensbasierte Sprachverarbeitung. Berlin et al., Springer, 259-268.