A self-organizing lexical system in hypertext*

*Burghard B. Rieger***
Constantin Thiopoulos
Department of Computational Linguistics
FB II: LDV/CL – University of Trier

1 Introduction

1.1 Knowledge-based Semantics

Our understanding of the bunch of complex intellectual activities subsumed under the notion of *cognition* is still very limited, particularly in how knowledge is acquired from texts and what processes are responsible for it. Recent achievements in wordsemantics, conceptual structuring, and knowledge representation within the intersection of cognitive psychology, artificial intelligence and computational linguistics have shown some agreement that cognition is (among others) responsible for, if not identifiable with, the processes according to which for a cognitive system previously unstructured surroundings may be tranformed to its perceived environment whose identifiable portions and their relatedness does not only constitute structures but also allow for their permanent revision according to the system's capabilities.
The common ground and widely accepted frame for modelling the semantics of natural language is to be found in the dualism of the rationalistic tradition of thought as exemplified in its notions of some independent (objective) reality and the (subjective) conception of it. According to this *realistic* view, the meaning of a language term (i.e. text, sentence, phrase, word, syllable) is conceived as something being related somehow to (and partly derivable from) certain other entities, called signs, a term is composed of. As a sign and its meaning is to be related by some function, called interpretation, language *terms*, composed of *signs*, and related *meanings* are understood to form some structures of entities which appear to be at the same time part of the (objective) reality and its (subjective) interpretation of it. In order to let signs and their meanings be identified as part of language terms whose interpretations may then be derived, some knowledge of these structures has to be presupposed and accessible for any symbolic information processing. Accordingly, *understanding* of language expressions can basically be identified with a of matching some input strings with supposedly predefined configurations of word meaning and/or world structure whose representations have to be available to the (natural or artificial) understanding system's particular (though limited) *knowledge*. The so-called *cognitive paradigm* of advanced procedural linguistics can easily be traced back to stem from this fundamental duality, according to which natural language understanding can be modelled as the *knowledge-based* processing of information.

*Published in: Köhler, R./Rieger, B.B. (Eds.): Contributions to Quantitative Linguistics. Dordrecht/Boston (Kluwer), 1993, pp. 67 - 78.
**partly by support of *The German Marshall Fund of the United States*

Subscribing to this notion of understanding, however, tends to be tantamount to accepting certain unwarranted presuppositions of theoretical linguistics (and particularly some of its model-theoretical semantics) which have been exemplified elsewhere[1] by way of the formal and representational tools developed and used so far in cognitive psychology (*CP*), artificial intelligence (*AI*), and computational linguistics (*CL*). In accordance with these tools, *word meaning* and/or *world knowledge* is uniformly represented as a (more or less complex) labelled graph with the (tacid) understanding that associating its vertices and edges with symbols from some established system of sign-entity-relationship (like e.g. that of natural language) will render such graph-theoretical configurations a model of structures or properties which are believed to be those of either the sign-system that provided the graphs' labels or the system of entities that was to be depicted. Obviously, these representational formats are not meant to model the *emergence* of structures and the *processes* that constitute such structures as part of word meaning and/or world, but instead are merely making use of them[2].

1.2 Cognitive Semiotics

It has long been overlooked that relating arc-and-node structures with sign-and-term labels in symbolic knowledge representation formats is but another illustration of the traditional *mind-matter*-duality presupposing a realm of *meanings* very much like the structures of the *real world*. This duality does neither allow to explain where the structures nor where the labels come from. Their emergence, therefore, never occurred to be in need of some explanatory modelling because the existence of *objects, signs* and *meanings* seemed to be out of all scrutiny and hence was accepted unquestioned. Under this presupposition, fundamental *semiotic* questions of *semantics*—simply did not come up, they have hardly been asked yet[3], and are still far from being solved.

In following a *semiotic paradigm* this inadequacy can be overcome, hopefully allowing to avoid (if not to solve) a number of spin-off problems, which originate in the traditional distinction and/or the methodological separation of the meaning of a language's term from the way it is employed in discourse. It appears that failing to mediate between these two sides of natural language semantics, phenomena like *creativity, dynamism, efficiency, vagueness,* and *variability* of meaning—to name only the most salient—have fallen in between, stayed (or be kept) out of the focus of interest, or were being overlooked altogether, sofar. Moreover, there is some chance to bridge the gap between the formal theories of language description (*competence*) and the empirical analysis of language usage (*performance*) that is increasingly felt to be responsible for some unwarranted abstractions of fundamental properties of natural languages.

Approaching the problem from a *cognitive* point-of-view, identification and interpretation of external structures has to be conceived as some form of *information processing* which (natural/artificial) systems—due to their own structuredness—are (or ought to be)

[1] Rieger 1991

[2] For illustrative examples and a detailed discussion see Rieger 1989, pp.103–132.

[3] see however Rieger (1977)

able to perform. These processes or the structures underlying them, however, ought to be derivable from—rather than presupposed to—procedural models of meaning[4]. Based upon a phenomenological reinterpretation of the analytical concept of *situation* as expressed by BARWISE/PERRY (1983) and the synthetical notion of *language game* as advanced by the late WITTGENSTEIN (1958), the combination of both lends itself easily to operational extensions in empirical analysis and procedural simulation of associative meaning constitution which may grasp essential parts of what PEIRCE named *semiosis*[5].
Modelling the meaning of an expression along reference-theoretical lines had to presuppose the structured sets of entities to serve as range of the denotational function which provided the expression's interpretation. However, it appears feasible to have this very range be constituted as a result of exactly those cognitive functions by way of which understanding is produced. It will have to be modelled as a dynamic generation which reconstructs the possible structural connections of an expression towards cognitive systems (that may both intend/produce and realize/understand it) and in respect to their *situational* settings, being specified by the expressions' pragmatics.
In phenomenological terms, the set of structural constraints defines any cognitive (natural or artificial) system's possible range in constituting its schemata whose instantiations will determine the system's actual interpretations of what it perceives. As such, these cannot be characterized as a domain of objective entities, external to and standing in contrast with a system's internal, subjective domain; instead, the links between these two domains are to be thought of as *ontologically fundamental*[6] or pre-theoretical. They constitute—from a *semiotic* point-of-view—a system's primary means of access to and interpretation of what may be called its "world" as the system's particular apprehension of its environment. Being fundamental to any cognitive activity, this basal identification appears to provide the grounding framework which underlies the duality of categorial-type rationalistic mind-world or subject-object separation.
From a systems-theoretical point-of-view, this is tantamount to a shift from linear to non-linear systems in modelling cognitive and semiotic behaviour. The simplest way to distinguish these approaches is by identifying the behaviour of *linear systems* as being equal to the sum of the behaviour of its parts, whereas the behaviour of *non-linear systems* is more than than the sum of its parts. FREGEs principle of *compositionality* as well as CHOMSKEYs hypotheses of independance of syntax are concepts in point of the *linear*-systems'-view: by studying the parts of a system in isolation first, will then allow for a full understanding of the complete system by composition. This collides with the *non-linear*-systems'-view according to which the primary interest is not in the behaviour of parts as properties of a system but rather in the behaviour of the *interaction* between parts of a system. Such interaction-based properties necessarily disappear when the parts

[4]It has been argued elsewhere (Rieger 1990, 1991) that *meaning* need not be introduced as a presupposition of *semantics* but may instead be derived as a result of semiotic modelling.

[5]"By *semiosis* I mean [...] an action, or influence, which is, or involves, a cooperation of *three* subjects, such as sign, its object, and its interpretant, this tri-relative influence not being in any way resolvable into actions between pairs." (Peirce 1906, p.282)

[6]Heidegger (1927)

are studied in isolation, as can be witnessed in referencial and model-theoretic semantics where phenomena like *vagueness, contextual variability* and *creative dynamism* cannot be dealt with, or in competence theoretical syntax where grades of *grammaticality, adaptive change* and *discourse adequacy* cannot be addressed.
The *self-organizing* property of the non-linear system introduced here has formally been derived elsewhere[7] from mathematical *topos theory*[8] and *category theory*[9]. This implementation of the system and its organisation as a dynamic *hypertext* structure is to simulate the emergence of lexical meanings by way of word co-occurrence constrains of—as yet—rather coarse syntagmatic/paradigmatic regularities in natural language texts.

2 The formalism

2.1 The self-organizing mechanism

A numerical measure expressing the dependency between two lexems can be calculated by taking the number of common contexts to be a representation of their mutual use. Thus for $\mathcal{O}(a)$ set of contexts of a, i.e. texts, where an instantiation of a appears, we define:

Definition 2.1 $conf(a,b) = \frac{|\mathcal{O}(a)\cap\mathcal{O}(b)|}{|\mathcal{O}(b)|}$

For L set of the considered lexems, the actual state of the structure is given by a matrix $CONF = (conf(a,b))_{a,b\in L}$. The self-organizing modification can be obtained by recomputing $conf$ after each new context. There are three cases:

1. a and b are both in the new text. Then: $conf(a,b)_{new} = \frac{|\mathcal{O}(a)\cap\mathcal{O}(b)|+1}{|\mathcal{O}(b)|+1} = \frac{\frac{|\mathcal{O}(a)\cap\mathcal{O}(b)|+1}{|\mathcal{O}(b)|}}{\frac{|\mathcal{O}(b)|+1}{|\mathcal{O}(b)|}} = \frac{conf_{old}+\frac{1}{|\mathcal{O}(b)|}}{1+\frac{1}{|\mathcal{O}(b)|}}$. In this case $conf(a,b)_{new} \geq conf(a,b)_{old}$, i.e the intensity of the connection between a and b increases.

2. Only b is in the new text. Then: $conf(a,b)_{new} = \frac{conf_{old}}{1+\frac{1}{|\mathcal{O}(b)|}}$. In this case $conf(a,b)_{new} \leq conf(a,b)_{old}$, i.e the intensity of the connection between a and b decreases.

3. Only a is in the new text. Then: $conf(a,b)_{new} = conf(a,b)_{old}$.

2.2 Categories

In order to capture structural features of the actual state of the system the $CONF$ matrix is transformed to a category[10]. A category is a directed graph with some additional features.

[7] Thiopoulos 1992 forthcoming
[8] Goldblatt 1979
[9] Bell 1981; Lambek/Scott 1986
[10] For a complete description of the theoretical framework see (Thiopoulos,1991).

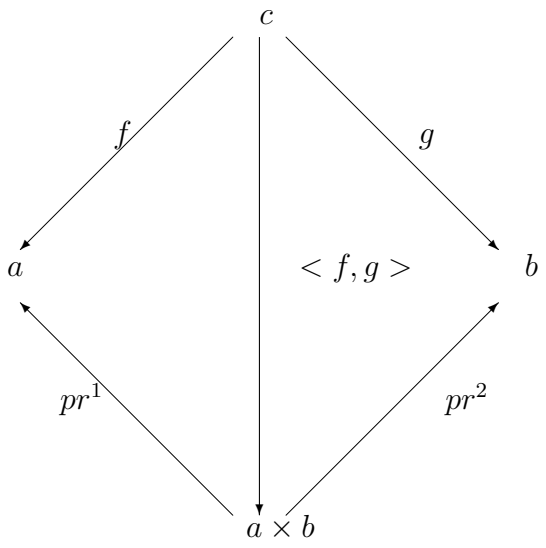

Figure 1: Product

A full subcategory is thus a subcategory that contains all the morphisms of the original category between its objects, i.e. it is a function closed under functional application. A special class of categories is the class of cartesian closed categories. They are characterized by the fact that some structural operation are defined of them. Here we consider two of them:

Definition 2.2 *The product from* $a, b \in OBJ(A)$ *is* $a \times b \in OBJ(A)$ *together with* $pr^1 : a \times b \to a, pr^2 : a \times b \to b \in MORPH(A)$*, so that* $\forall c \in OBJ(A) \exists! < f, g >: c \to a \times b \in MORPH(A)$ *with* $pr^1 \circ < f, g >= f \wedge pr^2 \circ < f, g >= g$.

Definition 2.3 *A category A consists of:*

1. *a class of objects OBJ(A)*
2. *a class of morphisms MORPH(A)*
3. *two operations dom,cod:MORPH(A)→OBJ(A), with f:a → b iff dom(f)=a and cod(f)=b*
4. *an operation comp* : $MORPH(A) \times MORPH(A) \to MORPH(A)$ *with comp(f,g)=f∘ g such that* $f \circ (g \circ h) = (f \circ g) \circ h$
5. *an operation id:OBJ(A)* $\to MORPH(A)$ *with* $id(a) = id_a : a \to a$*, where* id_a *is the identity function on a .*

The nodes of the graph are the objects and the links the morphisms. For $f : a \to b$, a is the domain of f and b the codomain. *comp* is the (associative) composition of morphisms and *id* maps each object to the corresponding identity morphism.

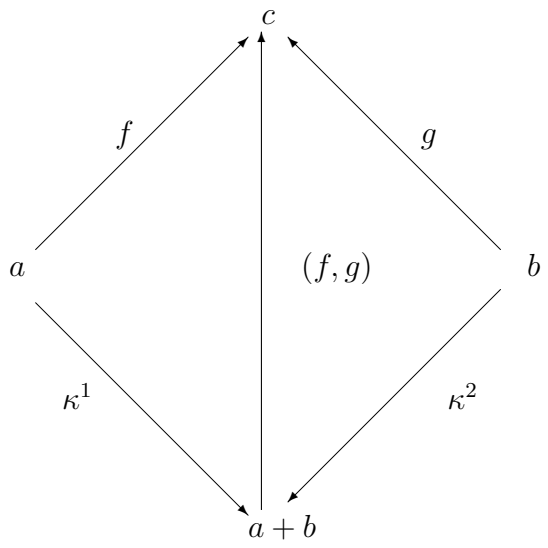

Figure 2: Coproduct

Definition 2.4 $A(a,b) = \{f \mid f \in MORPH(A) \wedge f : a \rightarrow b\}$

Definition 2.5 *B is a subcategory of $A (B \subseteq A)$ iff*

1. $OBJ(B) \subseteq OBJ(A)$
2. $\forall a, b \in OBJ(B) B(a,b) \subseteq A(a,b)$.

Definition 2.6 *B is a full subcategory of A iff*

1. $B \subseteq A$
2. $\forall a, b \in OBJ(B) B(a,b) = A(a,b)$.

Definition 2.7 *The coproduct from $a, b \in OBJ(A)$ is $a + b \in OBJ(A)$ together with $\kappa^1 : a \rightarrow a + b, \kappa^2 : b \rightarrow a + b \in MORPH(A)$, so that $\forall c \in OBJ(A) \exists! (f,g) : a + b \rightarrow c \in MORPH(A)$ with $(f,g) \circ \kappa^1 = f \wedge (f,g) \circ \kappa^2 = g$.*

The transformation of the matrix $CONF$ to a category $C(CONF)$ is defined by:

- $OBJ(C(CONF)) = L$
- $conf(a,b) \geq conf(b,a) \Rightarrow f : a \rightarrow b \in MORPH(C(CONF))$
- $conf(a,b) < conf(b,a) \Rightarrow f : b \rightarrow a \in MORPH(C(CONF))$

The weighting of the morphisms is thus given as a partial function:

$$conf(f) = conf(a,b) \text{ iff } dom(f) = a \wedge cod(f) = b \wedge conf(a,b) \geq conf(b,a).$$

that can be extended to morphism combination as follows:

for the composition: $conf(f \circ g) = conf(f)conf(g)$

for the product: $conf(< f, g >) = minimum(conf(f), conf(g))$.

for the coproduct: $conf((f, g)) = minimum(conf(f), conf(g))$.

The meaning of a lexem a, as a structural description of how a is interlinked in the network of lexems, according to a numerical boundary GLB that determines the depth of the activation, is given by:

Definition 2.8 $a^{\star_{GLB}} = \{(b, conf(f)) \mid \exists f \in MORPH(ST) f : a \rightarrow b \wedge conf(f) \geq GLB\}$

The meaning of two or more lexems can be represented as a full subcategory generated by the product and coproduct constructions.

Definition 2.9 $prod_{GLB}(a, b) = \{(C, conf(f)) \mid \exists f : C \rightarrow a \times b \wedge conf(f) \geq GLB\}$
$coprod_{GLB}(a, b = \{(C, conf(f)) \mid \exists f : a + b \rightarrow C \wedge conf(f) \geq GLB\}$

Definition 2.10 *The situation generated out of a, b in a cartesian closed category C is the full subcategory $SIT(a, b)$ with*

$$OBJ(SIT(a, b)) = \{c \in OBJ(C) \mid c \text{ collected by } prod(a, b) \text{ and } coprod(a, b) \text{ for a specific } GLB\}.$$

A situation is thus a substructure of the original category that is closed under functional application and since it is a category again it is possible to apply the same mechanisms as in the original category. The meaning of a lexem *a relative* to a situation, $SIT(b, c)$ is $a^\star$ in $SIT(b, c)$.

3 The Implementation

Hypertext seems to be the most suitable tool for capturing the dynamic nature of the formalism. The category $C(CONF)$ can, as a directed graph, be mapped into a hypertext structure[11], in the following way:

-Each object of $C(CONF)$ is implemented as a card named after this object that contains a text field with the name (see figure 3).

- Each morphism $f : a \rightarrow b$ is implemented as a button on the card named a that leads, when activated, to the card named b. The name of the button is formed by concatenating b with $conf(f)$ (see figure 3). The user can navigate through the network by clicking on the buttons.

industrie

dienst, 0.79	wissenschaft, 0.50	kontrollieren, 0.66
entwickeln, 0.79	wirtschaft, 1.00	konstruktion, 1.00
handeln, 0.79	technik, 0.66	plan, 0.66
zielen, 0.80	qualität, 0.50	
wunsch, 0.66	produktion, 0.66	

Figure 3: A card containing a text field with the name of the corresponding lexem and buttons with the names of the lexems that are codomains of morphisms starting from this lexem.

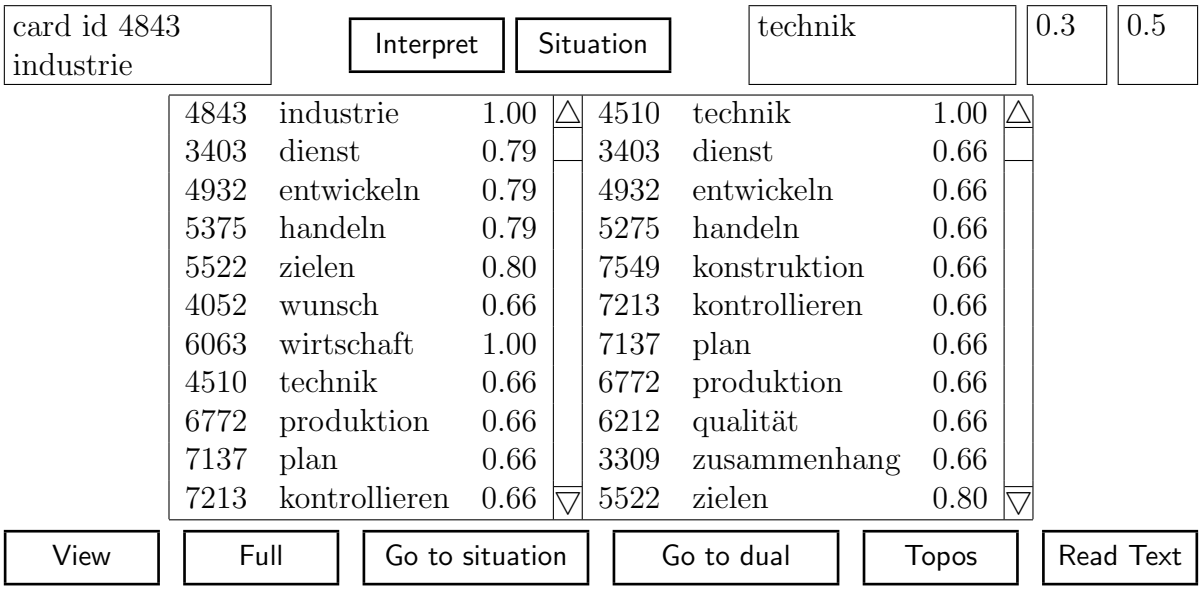

Figure 4: The control card.

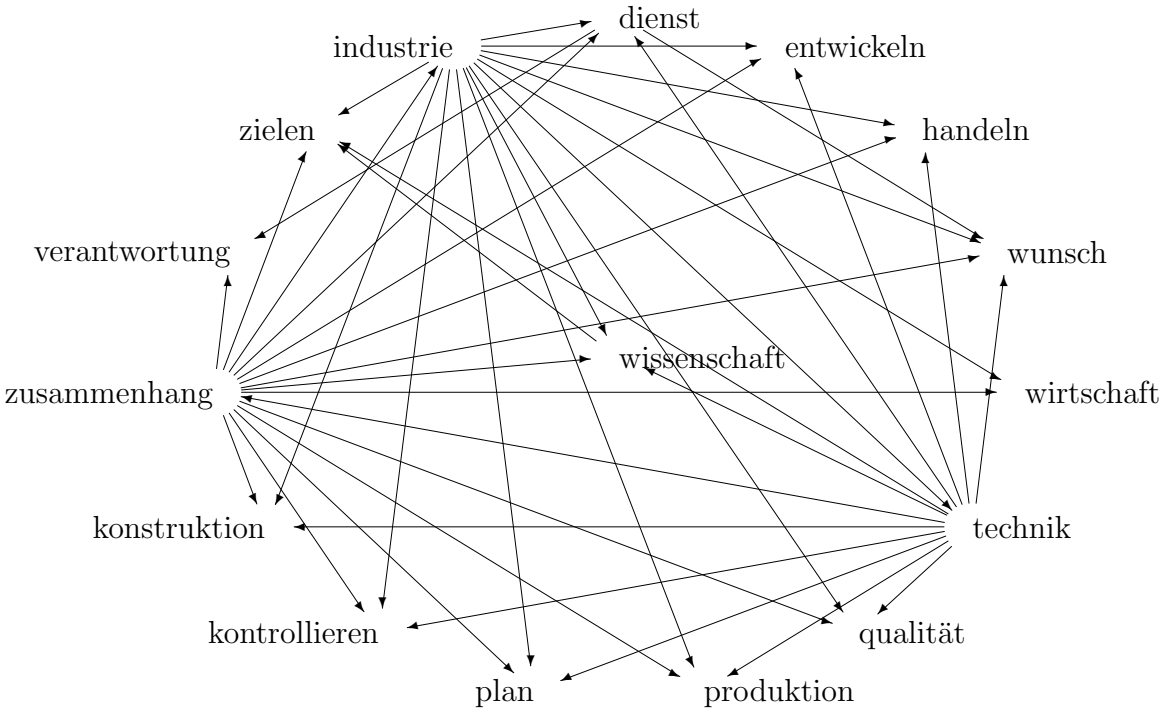

Figure 5: The view of the hypertext file representing $C(CONF)$.

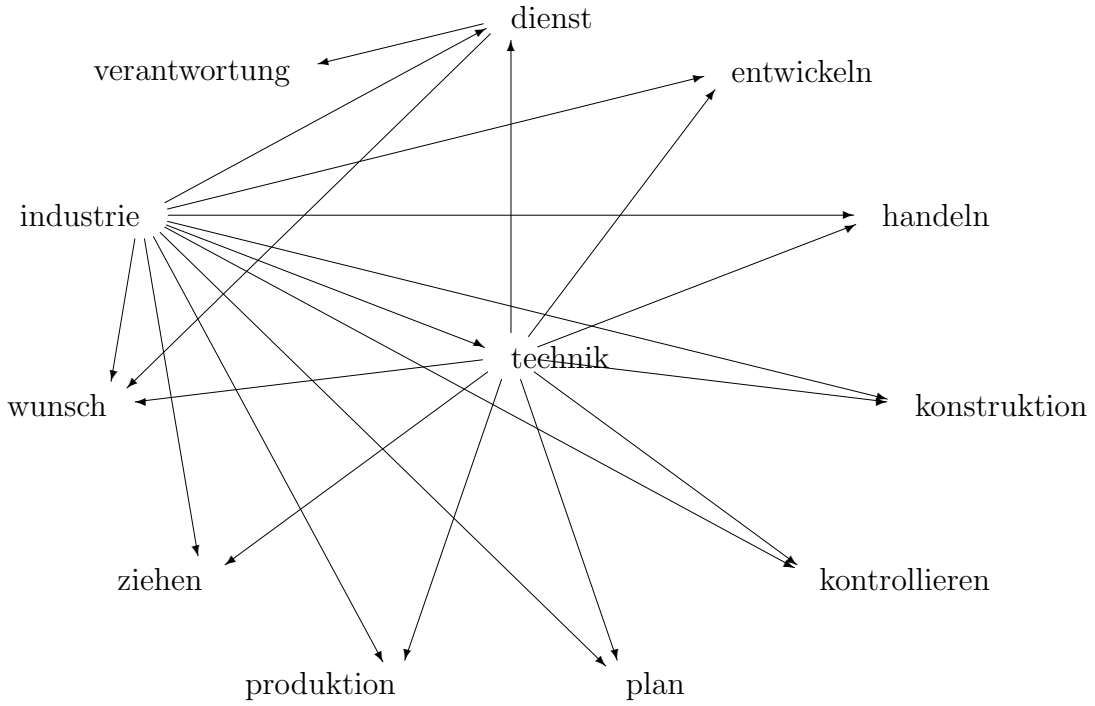

Figure 6: The view of the hypertext file representing $sit(industrie, technik)$.

- The structural operations defined on $C(CONF)$ can be implemented as *browsers* and the determined portions of the network can be accessed via *hyperviews*.

These mechanisms can be activated by clicking on the buttons of the control card:

1. *Read Text* calls a C program that reads the actual text and recomputes the $CONF$ matrix.

2. *Topos* generates from the $CONF$ matrix the corresponding hypertext file.

3. *View* produces a global view of the file.

4. *Interpret* generates $*_{GLB}$ for the lexem that is given in the left text field at the top of the card (industrie), where the depth of the activation (GLB) can be specified by the user as entry in a text field (here is 0.5). *Interpret* activates a browser that avoids cycles by keeping a list of visited cards. The collected lexems are listed, together with the corresponding weight and card number, in the left scrolling field in the center of the card.

5. *Situation* activates the *prod* and *coprod* mechanisms for the lexem contained in the left text field at the top of the card (industrie) and the lexem (or lexems) contained in the scrolling field at the top of the card (technik). The right srolling field in the center of the card is tereby used to keep track of the lexems acessed by the different interpretations.

6. *Full* generates the corresponding full subcategory (here sit(industrie,technik), i.e. determines all the morphisms between the collected lexems.

7. *Go to situation* leads to the control card of a hypertext file that corresponds to the actual situation, where by using the *Topos* button the full subcategory is mapped - using the same mapping operation as for $C(CONF)$ - into this file.

Besides the cards that correspond to the objects of $C(CONF)$ there is also a control card, from where the user controls the implemented navigation mechanisms (see figure 4).
The process of restricting the category $C(CONF)$ can be used recursively and reflects the focus of interest of the user of the system. In this example (that, since it corresponds to the first stages of the system has a rough structure) the view of the category $C(CONF)$ is given in figure 5 and the view of the full subcategory $sit(industrie, technik)$ is given in figure 6. The text field in figure 6 contains the lexems that led to this view and for successive determinations of situations, i.e. situations of situations of ..., it contains the history of the restrictions. By using the *Interpret* button of the hypertext file that represents the actual situation the user can determine the meaning of a lexem *relative* to this situation.

[11]The implementation is made in Hypercard

References

Barrett, E. (Ed.)(1988): *Text, ConText, and HyperText.* Cambridge, MA (MIT)

Barwise, J./ Perry, J.(1983): Situations and Attitudes. Cambridge, MA (MIT)

Bell, J. L. (1981): "Category theory and the foundation of mathematics" *British Journal of the Philosophy of Science,* 32,1981:349–358

Conklin,J. (1987): "Hypertext: an introduction and survey" *Computer*, Vol.20, No.9.

Frisse, M. (1987): "From text to hypertext" *Byte* October, 1987.

Goldblatt, R. (1984): Topoi. The Categorial Analysis of Logic. (Studies in Logic and the Foundations of Mathematics 98), Amsterdam (North Holland)

Heidegger, M. (1927): Sein und Zeit. Tübingen (M.Niemeyer)

Lambek, J./ Scott P. J. (1986): Introduction to higher order categorical logic. Cambridge (Cambridge University Press)

Maturana, H./ Varela, F. (1980): Autopoiesis and Cognition. The Realization of the Living. Dordrecht (Reidel)

Peirce, C.S. (1906): "Pragmatics in Retrospect: a last formulation" (CP 5.11 – 5.13), in: The Philosophical Writings of Peirce. Ed. by J. Buchler, NewYork (Dover), pp.269–289

Rieger, B. (1977): "Bedeutungskonstitution. Einige Bemerkungen zur semiotischen Problematik eines linguistischen Problems" *Zeitschrift für Literaturwissenschaft und Linguistik* 27/28, pp.55–68

Rieger, B. (1985b): "On Generating Semantic Dispositions in a Given Subject Domain" in: Agrawal, J.C./ Zunde, P. (Eds.): Empirical Foundation of Information and Software Science. NewYork/ London (Plenum Press), pp.273–291

Rieger, B. (1989): Unscharfe Semantik. Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten. Frankfurt/ Bern/ NewYork (P. Lang)

Rieger, B. (1990): "Situations and Dispositions. Some formal and empirical tools for semantic analysis" in: Bahner, W. / Schildt, J./ Viehweger, D. (Ed.): Proceedings of the XIV.International Congress of Linguists (CIPL), Vol. II, Berlin (Akademie Verlag), pp.1233–1235

Rieger, B. (1991): "On Distributed Representation in Word Semantics" (ICSI-Report TR-91-012) International Computer Science Institute, Berkeley, CA

Rieger, B.B./ Thiopoulos, C. (1989): Situations, Topoi, and Dispositions. On the phenomenological modelling of meaning. in: Retti, J./ Leidlmair, K. (Eds.): 5th Austrian Artificial Intelligence Conference. (ÖGAI 89) Innsbruck; (KI–Informatik-""Fachberichte Bd.208) Berlin/ Heidelberg/ NewYork (Springer), pp.365–375

Thiopoulos, C. (1991): Meaning metamorphosis in the semiotic topos. *Theoretical Linguistics* [in print]

Winograd,T./ Flores, F. (1986): Understanding Computers and Cognition: A New Foundation for Design. Norwood, NJ (Ablex)

Wittgenstein, L. (1958): The Blue and Brown Books. Ed. by R. Rhees, Oxford (Blackwell)