Computational Semiotics and Fuzzy Linguistics

On Meaning Constitution and Soft Categories[*]

Burghard B. Rieger
FB II: Department of Computational Linguistics – University of Trier
rieger@ldv01.Uni-Trier.de

Abstract

Whereas most cognitive approaches in the study of language have been developing hypotheses concerning the principles of knowing and understanding natural languages (i.e. competence) without bothering too much about communicative language usages in realworld situations (i.e. performance), new semiotic *approaches in cognitive computational linguistics explore the* procedures *believed to underlie* processes *of language learning and understanding. They do so by simulating these capabilities as system behaviour under recourse to modeled structures, observable in very large samples of situated natural language discourse and represented in vector space formats via numerically specified by quantitative methods of dynamic (re-)construction. It will be argued that the ecological understanding of informational systems in* Computational Semiotics *corresponds well to the procedural modeling and numerical reconstruction of processes that simulate the constitution of meanings and the interpretation of signs (*semiosis*). The theories of* fuzzy sets *[24] and* possibility distributions *[23] together with their derivatives in* soft computing *[25] appear to be promising in providing suitable formats for computational approaches to natural language processing without the obligation neither to reject nor to accept traditional formal and modeltheoretic concepts or ontologies. Examples from* fuzzy linguistic *research [10] [18] will be given to illustrate these points.*

1 Cognitive Information Processing

For the majority of researchers in knowledge representation and natural language semantics the common ground and widely accepted frame for their modeling may be found in the dualism of the rationalistic tradition of thought as exemplified in its *matter-mind* notion of an independent (objective) reality and some (subjective) conception of it.

1.1 The traditional approach

According to the *realistic* view, the meaning of any portion of language material (like e.g. discourse, utterance, word(token), morph, phone, etc.) is interpreted as being an instantiation of (or partly derivable from) certain other entities, called linguistic categories (like e.g. text, sentence, word(type), morpheme, phoneme, etc.), with the understanding that these categories structure natural language material according to their compositional functions. It is by these functions that language material (strings of *terms*) appear to be composed of linguistic entities (aggregates of *categories*) to form structures and it is also by these functions that the quality of language structures (having *meaning*) is conceived as being part of both, the physical reality of language material and the semantic significance of linguistic signs. Illustrating this twofold membership are the graph-theoretical formats which have become standard representations for natural language meanings, both as *relational structure* and as *referential denotation*. Thus, relating arc-and-node configurations with sign-and-term labels in graphs like trees and nets appears to be but another aspect of the traditional *mind-matter*-duality according to which a realm of *meanings* is presupposed very much like the assumption of the pre-given structures of the *real world* related by *signs*. Accepting this duality has neither allowed to explain where the structures or where the labels come from nor how their mutual relatedness as *meanings* of *signs* can be derived. The emergence of the *meaning relation*, therefore, never occurred to be in need of some explanatory modeling because the existence of *signs*, *objects* and *meanings* were taken for granted and hence seemed to be out of all scrutiny. Under this presupposition, fundamental *semiotic* questions of *semantics* simply did not come up, they have hardly been asked yet, and are still far from being solved.

1.2 Modeling cognition

Extending an earlier attempt [22] to classify approaches in cognitive science, we may roughly discern four[1] types of approaches in modeling cognition:

▷ the *cognitive* approaches presuppose the existence of the external world, structured by given objects and properties and the existence of representations of (fragments of) this world internal to the system, so that the cognitive systems' (observable) behaviour of action and reaction may be modelled by processes

[*] Published in: Meystel, A. (Ed.): A Learning Perspective. *Proceedings of the 1997 International Conference on Intelligent Systems and Semiotics* (ISAS-97) [NIST Special Publication 918], Washington (US Government Printing Office) 1997, pp. 541–551; also in: Albus, J./Meystel, A. (Eds.): Intelligent Systems: A Semiotic Perspective (Vol. I: Formation of Symbols), Washington/New York/London (Wiley) 1997 [in print]

[1] There were only the first three of these four approaches distinguished by Varela/Thompson/Rosch (1991).

operating on these structures;

- ▷ the *associative* approach is described as a dynamic structuring based on the model concept of self-organization with cognitive systems constantly adapting to changing environmental conditions by modifying their internal representation of them.

Whereas both these approaches apparently draw on the traditional rationalistic paradigm of mind-matter-duality—*static* the former, *dynamic* the latter—in presupposing the *external* world structure and an *internal* representation of it, the third and fourth category do not:

- ▷ the *enactive* approaches may be characterized as being based upon the notion of *strcutural coupling*. Instead of assuming an external world and the systems' internal representations of it, some unity of structural relatedness is considered to be fundamental of—and the (only) condition for—any abstracted or acquired duality in notzions of the external and internal, object and subject, reality and its experience;
- ▷ the *semiotic* approaches focus on the notion of *semiosis* and may be characterized by the process of *enactment* too, supplemented, however, by the representational impact. It is considered fundamental to the distinction of e.g. *cognitive processes* from their *structural results* which—due to the traces these processes leave behind—may emerge in some form of *knowledge* whose different representational modes comply with the distinction of *internal* or *tacid* knowledge (i.e. *memory*) on the one hand and of *external* or *declarative* knowledge (i.e. *language*) on the other.

According to these types of cognitive modeling, *computational semiotics* can be characterized as aiming at the dynamics of of meaning constitution by simulating processes of multi-resolutional representation [5] within the frame of an ecological *information processing* paradigm [18].

As we take human beings to be *systems* whose knowledge based processing of represented *information* makes them *cognitive*, and whose sign and symbol generation, manipulation, and understanding capabilities render them *semiotic*, we may do so due to our own daily experience of these systems' outstanding ability for representing results of cognitive processes, organize these representations, and modify them according to changing conditions and states of system-environment adaptedness.

2 Computational Semiotics

For the *semiotic approach* to human cognition (constituting *computational semiotics*) such representations resulting from complex semiotic cognitive information processing may be found in any natural language discourse. In an aggregated form of *pragmatically homogeneous* text corpora [7] communicatively performative natural language discourse provides a cognitively highly interesting and empirically accessible system whose extreme structuredness may serve as a guideline for the cognitively motivated, empirically based, and computationally realized research in the semiotics of language, too.

Following this line, however, will necessitate to pass on from traditional approaches in *competence oriented* linguistics analysing introspectively the propositional contents of singular sentences as conceived by *ideal speakers/writers* towards a new understanding of meaning constitution as a dynamic process based upon the semiotic cognitive information processing the traces of which are to be identified and systematically reconstructed on the basis of empirically well founded observation and rigorous mathematical description of universal regularities that structure and constitute different levels of representations in masses of pragmatically homogeneous texts produced by *real speakers/writers* in actual situations of either performed or intended communicative interaction. Only such a *performance oriented* semiotic approach will give a chance to formally reconstruct and model procedurally both, the *significance* of entities and the meanings of *signs* as a function of a first and second order semiotic embedding relation of *situations* (or contexts) and of *language games* (or cotexts) which corresponds to the two-level actualisation of cognitive processes in language understanding [18].

2.1 Ecological information systems

Life may be understood as the ability to survive by adapting to changing requirements in the real world. In terms of the theory of information systems, faculties like perception, identification, and interpretation of structures (external or internal to a system) may be conceived as a form of *information processing* which (natural or artificial) systems—due to their own structeredness—are able to perform. Thus, living systems receive or derive information from relevant portions only of their surrounding environments, they learn from experience, and change their behaviour accordingly. In contrast to other living systems which transmit experiencial results of environmental adaptation only biogenetically[2] to their descendants, human information processing systems have additional means to convey their knowledge to others. In addition to the *vertical* transmission of system specific (*intraneous*) experience through (biogenetically successive) generations, mankind has complementally developed *horizontal* means of mediating specific and foreign (*extraneous*) experience and knowledge to (biogenetically unrelated) fellow systems within their own or any later generation. This is made possible by a *semiotic* move that allows not only to distinguish *processes* from *results* of experience but also to convert the latter to *knowledge* facilitating it to be re-used, modified and improved in

[2] According to standard theory there is no direct genetic coding of experiencial results but rather indirect transmission of them by selectional advantages which organisms with certain genetic mutations gain over others without them to survive under changing environmental conditions.

learning. Vehicle and medium of this move are *representations*, i.e. complex sign systems which constitute *languages* and form structures, called *texts* which may be realized in communicative processes, called *actualisation*.

In terms of the theory of information systems, *texts*—whether internal or external to the systems—function like virtual environments[3].

2.2 Modes of Representation

Considering the system-environment relation, *virtuality* may be characterized by the fact that it dispenses with the identity of space-time coordinates for system-environment pairs which normally prevails for this relation when qualified to be indexed *real*. It appears, that this dispensation of identity—for short: space-time-dispensation—is not only conditional for the possible distinction of (mutually and relatively independent) *systems* from their *environments*, but establishes also the notion of *representation*. Accordingly, *immediate* or space-time-identical system-environments existing in their space-time-identity may well be distinguished from *mediate* or space-time-dispensed system-environments whose particular representational form (*texts*) corresponds to their particular status both, as language material (being *signs*), and as language structure (having *meaning*). This double identity calls for a particular modus of actualisation (*understanding*) that may be characterized as follows:
For systems appropriately adapted and tuned to such environments *actualisation* consists essentially in a twofold embedding to realize

- ▷ the spaciotemporal identity of pairs of *immediate* system-environment coordinates which will let the system experience the material properties of texts as *signs* (i.e. by functions of *physical access* and *mutually homomorphic* appearance). These properties apply to the percepts of language structures presented to a system in a particular *discourse situation*, and
- ▷ the representational identity of pairs of *mediate* system-environment parameters which will let the system experience the semantic properties of texts as *meanings* (i.e. by functions of *emergence, identification, organisation, representation* of structures). These apply to the comprehension of language structures recognized by a system to form the *described situation*

Hence, according to the theory of information systems, functions like *interpreting* signs and *understanding* meanings translate to processes which extend the fragments of reality accesssible to a living (natural and possibly artificial) information processing system. This extension applies to both, the *immediate* and *mediate* relations a system may establish according to its own evolved adaptedness or dispositions (i.e. innate and acquired *structuredness*, processing *capabilities*, represented *knowledge*).

2.3 Semiotic enactment

Semiotic systems' ability to actualize environmental *representations* does not merely add to the amount of experiencial results available, but constitutes also a significant change in experiencial modus. This change is characterized by the fact that only now the *processes* of experience may be realized as being different and hence be separable from the *results* of experience which in *immediate* system-environments appear to be indistinguishable. Splitting up experience in experiencial processes and experiencial results—the latter being representational and in need for procedural actualisation by the former—is tantamount to the emergence of *virtual* experiences which have not to be *made* but can instead just be *tried*, very much like hypotheses in an experimental setting of a testbed. These *results*—like in *immediate* system-environments—may become part of a system's adaptive knowledge but may also—other than in *immediate* system-environments—be neglected or tested, accepted or dismissed, repeatedly actualized and re-used without any risk for the system's own survival, stability or adaptedness.

This in a way experimental quality of textual representations which increases the potentials of adaptive information processing beyond the system's lifespan, is constrained simultaneously by *dynamic* structures corresponding to *knowledge*. The built-up, employment, and modification of these structural constraints[4] is controlled by procedures whose processes determine *cognition* and whose results constitute *adaptation*. Systems properly attuned to textual system-environments have acquired these structural constraints (language learning) and can perform certain operations efficiently on them (language understanding). These are prerequisites to recognizing *mediate* (textual) environments and to identify their need for and the systems' own ability to *actualize* the mutual (and trifold) relatedness constituting what PEIRCE called *semiosis*[5]. Systems capable of and tuned to such knowledge-based processes of actualisation will in the sequel be referred to as *semiotic cognitive information processing systems (SCIPS)* [17, 19].

[3] Simon's (1982) remark "There is a certain arbitrariness in drawing the boundary between inner and outer environments of artificial systems. . . . Long-term memory operates like a second environment, parallel to the environment sensed through eyes and ears" (pp. 104) is not a case in point here. Primarily concerned with where to place the boundary, he does not seem to see its placing in need to be justified or derived as a consequence of some possibly representational processes we call *semiotic*. As will become clear in what follows, Simon's distinction of *inner* (memory structure) and *outer* (world structure) environments is not concerned with the special quality of language signs whose twofold environmental embedding (textual structure) cuts accross that distinction, resolving both in becoming representational for each other.

[4] What Simon (1982) calls *memory* in accordance with his questioning of the inner-outer-distiction of cognitive systems and their environments.

[5] By *semiosis* I mean [. . .] an action, or influence, which is, or involves, a coöperation 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)

Representation, therefore, has to be considered fundamental to the distinction of the *processes* of cognition from their *results* which may emerge—due to the traces these processes leave behind—in some structure (*knowledge*). Different representational modes of this structure not only comply with the distinction of *internal* or *tacid* knowledge (i.e. *memory*) on the one hand and of *external* or *declarative* knowledge (i.e. *texts*) on the other[6], but these modes also relate to different types of formats (*distributional* vs. *symbolic*), modeling (*connectionist* vs. *rule-based*) and processing (*stochastic* vs. *deterministic*). It is this range of correspondences that *Fuzzy Linguistics* is based upon and tries to exploit to come up with a unifying framework for most of the different approaches followed sofar.
Soft categorising appears to be a prerequisite for fuzzy linguistic modeling examples of which will illustrate the notion of dynamic structures emerging from corpora of natural language discourse.

3 Fuzzy Linguistic Modelling

It is far from certain yet whether—and if so, how—*semiotic* models will help to understand how structures may emerge from orders of some kind and how these orders evolve from regularities which multitudes of repeatedly observable entities show. Recent research findings, however, give rise to expect that processes which determine regularities and assemble them to form (intermediate) representations whose properties resemble (or can even be identified with) those of observable entities may indeed be responsible for (if not identical with) the emergence and usage of sign-functional structures in language understanding systems, both natural and artificial. As more abstract (theoretical) levels of representation for these processes —other than their procedural modeling— are not (yet) to be assumed, and as any (formal) means of deriving their possible results—other than by their (operational) enactment—are (still) lacking, it has to be postulated that these processes—independent of all other explanatory paradigms—will not only relate but produce different representational levels in a way that is formally controlled or *computable*, that can be modeled procedurally or *algorithmized*,and that may empirically be tested or *implemented* [4]. *Procedural models* are understood to denote a class of (re-)presentational or modeled (re-)constructions of entities whose interpretation is not (yet) tied to an underlying theory which would privide the semantics for the entities (or expressions) that these type of models present. Instantiating their defining *procedures* as implemented algorithms will result in *processes* which produce some observable *structures* that can only then be compared to those of the modeled original.

As as some of these procedural characteristics have also be claimed by cognitive linguistic approaches and computational models of language understanding, their main traits may help to illustrate the different positions of semiotic modeling in fuzzy linguistics.

3.1 Cognitive linguistic strata

Cognitive theory has long identified the complex of language understanding to be a modular system of subsystems of information processing. The idea of symbolic representation and the computer metaphor offered a frame for modeling cognitive processes, formally grounded by logical calculi and procedurally on algorithms operating on representational structures. Following and partly supplementing strata of semiotic description and analysis of signs, different levels of modular aggregation of external and internal information have been distinguished in cognitive models of language understanding. These partly correspond to and partly cut accross the *syntactics-semantics-pragmatics* distinction in the semiotic relatedness of signs, the *utterance-discourse-corpus* levels of performative language processing, and the hierarchy of *morpho-phonological*, *syntax-sentencial* and *lexico-semantic* descriptions in traditional models of structural linguistics.

In one of the rare ventures on discussing of how cognitive, i.e. knowledge based information processing mechanisms may be provided with the knowledge bases they are meant to operate on, and how these knowledge structures may be related to observable language data, Bierwisch (1981) sketches a hierarchy of information processing mechanisms whose representational format (sets of rewrite rules operating on structured data) allows algorithms be formulated and implementations be found to guaratee their computability. According to this schema (*Fig.* 1) and starting with the morpho-phonological level, an information processing mechanism M_1 is postulated which receives *utterances* as input and produces some associated *structures* as output. In doing so, however, the mechanism's performance will be determined not only by the external input strings but also by some internal knowledge of elements and rules which allow to agglomerate the structures identified. The acquisition and representation of this internal knowledge is hypothesized as resulting from a process M_2 which also includes a multitude of rudimentary, incomplete, and tentative M_1-kind processes. M_2 is assumed to be a complex information processing mechanism whose inputs are *corpora* of utterances together with some environmental information, and whose outputs will be the *grammars* underying these utterances. Again, this mechanism's results will not only and completely be determined by the external inputs but also by some internal structures which are believed to control the human language faculty in a fundamental way

[6] Whereas *tacid knowledge* cannot be represented other than by the *immediate* system-environments' corresponding states, *explicit knowledge* is bound to acquire some formal properties in order to become externally presented and thereby part of *mediate* system-environments. Natural languages obviously provide these formal properties—as partly identified by research in linguistic competence (principles knowledge and acquisition of language)—whose enactment—as investigated in studies on natural language performance (production and understanding of texts)—draws cognitively on both bases of (explicit and tacid) knowledge.

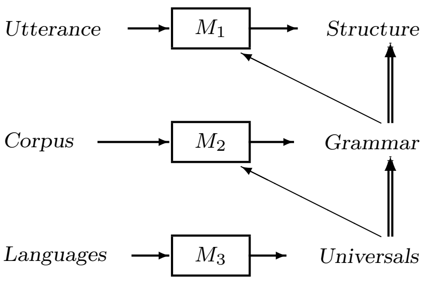

Figure 1

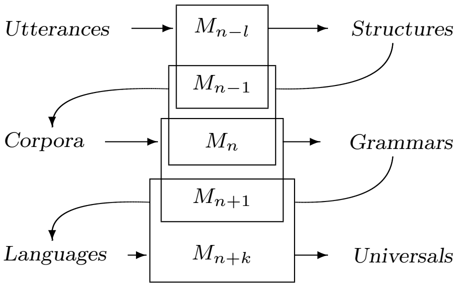

Figure 2

Schemata of model hierarchy of *cognitiv linguistic* strata of mechanisms (Bierwisch) as compared to model tiling of *computational semiotic* coverage of procedures (Rieger) for the analysis and representation of (abstracted and observable) language phenomena.

as so-called *linguistic universals*. These may (or may not) be assumed to be derivable as results of an information processing mechanism M_3 whose input is as comprehensive (or unspecified) as the term *languages* might allow.

Taking the relation of inclusion for $M_1 \subset M_2$ to hold also for $M_2 \subset M_3$, and considering M_1, M_2, M_3 computationally specifyable procedures of language analysing processes instead of mere metaphors for some (more or less plausible) mechanisms of the human mind, then it appears reasonable to consider M_3 a collection of all the processes of methodical analysis, representation, and comparison of structured sets of utterances from different languages, including the processes in M_1 as a device that explicidly specifies an utterance's structure relative to a given grammar, and the processes in M_2 as a system that generates a grammar from a corpus of utterances relative to the given set of universals. This modeling view allows for the notions of *Universals* ⇒ *Grammar* ⇒ *Structure* to be understood as variables of theoretical constructs hinged on empirical regularities observed in *Languages, Corpus, Utterance* respectively. Whereas the latter are external representations,the former are internal to any *SCIP*-system and considered external representations only under the competence linguistic approach to cognitive modeling. As such they are hypothesized to form a hierarchy of *linguistic*–not *language*—entities which formally specify a class of other linguistic entities (following the double arrows in *Fig.* 1).

The model theoretical and operational problems inherent in this setup concern the (non universal and highly restrictive) representational format which is assumed to enable the denotation of *universals, grammar* and *structure*, and the essentially top-down, nonrecursive propagation of externally presented but internally processed results of these mechanisms. Thus, M_3 whose performance in identifying universals and representing them externally depends crucially on the efficient performance of M_2 which is said to employ these universals as internal procedural constraints in order to identify syntactic regularities and represent them externally in a rule based format as grammars. Grammars, in turn, have to be employed as internal procedural constraints by M_1 if this mechanism's identification processes and the external representation of their findings is meant to be successful.

Distinguishing between these two kinds of structures either external or internal to the mechanisms M introduced so far, is indicative of the systems theoretical view proposed in semiotic modeling. It easily allows to translate these mechanisms as sets of procedures which allow to describe and simulate a living systems' abilility to process environmental input (external structures) according to procedural constraints known to the system (internal structures) in order to produce some results of this processing. However,it appears not at all conclusively compelling to assume that these procedural constraints and the processing results need to be represented in a rule-based format. According to an ecologically motivated systems theoretical view, systems enacting these processes under boundary conditions as determined by their surrounding environments, or their internal structuredness, or both, will have to process certain inputs to produce specified output structures. But identifying their status of being at the same time internal and external to the processing system is tantamount to the methodological dilemma which can solely be solved on the grounds of revising the representational mode and the formatting constraints which the model construction has to be decided on to allow.

Following Chomsky these modes have been restricted to abstract principles of language competence by processes [2] whose assumed rule-based determinacy consequently led to formal representations of these rules giving rise to the above model hierarchy of discrete strata [1]. In trying to relate these strata to observable performative language data structures in order to mediate observable language regularities with theoretical constructs supposedly representing principles underlying these constructs, the methodological shortcomings of the cognitive linguistic approach are revealed. It suffers from competence theoretically inspired idealisations of regularities and theoretical abstractions (like *universals, grammars, sentences*) whose symbolic notations and formal expressions may be scrutinized for their syntactic correctness but lack empirically observable and experimentally testable procedures of language representation which are independant from competent speakers' understanding of that language.

3.2 Fuzzy linguistic tiling

Other than cognitive linguistic and competence theoretical *mechanisms*, we propose cognitive semiotic and language performative *procedures* which redefine the modularity of language understanding as an overlapping covering of computational processes. The classical and coarse three-stage description and modeling of linguistic regularities will be replaced or rather complemented [11] by a multi-stage covering of *semiotic* procedures M_{n-l} to M_{n+k} (*Fig.* 2) which allows for the definition of more adequate soft categories or intermediate representations to fit regularities of entities on any level.Their essentially cognitive character will not be borrowed from any predefined strata (and their puportedly related abstract categories) but is to be derived as a result of their performance, i.e. the ability to transform linearly structured entities (strings) of one level to multi-dimensional structures of entities (vectors) on another.

This is achieved by analysing the linear or *syntagmatic* and selective or *paradigmatic* constraints which natural language structure imposes on the formation of (strings of) linguistic entities on whatever level of entity formation. It has been shown and illustrated elsewhere [15], [9] in some detail, that *fuzzy linguistic modeling* allows to derive the representational means (e.g. *soft categories, continuous gradation, variable granularity, flexible plasticity, dynamic approximation*, etc.) which crisp categories and competence theoretically inspired idealisations of performative regularities lack. The (numerical) specificity and (procedural) definiteness of sub-symbolic, distributed formats in entity formation appear to provide for higher phenomenological compatibility and more cognitive adequacy than traditional levels of categorial representation whose symbolic mediation and syntactic correctness could only formally be scrutinized but not empirically or experimentally be tested [11].

4 Procedural (Re-)Construction

The success of computational language analysis and generation is based upon adequate structural descriptions of input strings and their semantic interpretations. This was assumed to be made possible by the correctness of rule based representations of (syntactic and lexical) knowledge of language and of (referential and situative) world knowledge on which grammar formalisms and deductive inferencial mechanisms can operate on. Notwithstanding the essentially static representations of structures in Although this kind of cognitive modeling of language processing (based on monotonic logics, symbolic representations, rule-based operations, serial processing, etc.), has produced considerable advances in formal theory and the consistent development of increasingly more complex systems, their idea of representing language entities as essentially static categorial type linguistic entities proves to become increasingly problematic. As the processing of very large language corpora (VLLC)[7], has made clear, traditional linguistic categories do not reduce but increase model complexity when applied to regularities and structures which quantitative-numerical means may easily identify and represent. Trying to map such sub-rule regularities and sub-symbolic structures to inadequate categories will generally result in a large number of borderline cases, variations, and ambiguities which then have to be dealt with, but possibly could be avoided from the very start.

4.1 Exploiting constraints

Structural linguistics has given substantial hints[8] on how language items come about to be employed in communicative discourse the way they are. They have identified the fundamental constraints that control the multi-level combinability and formation of language entities by distinguishing the restrictions on linear aggregation of elements (*syntagmatics*) from restrictions on their selective replacement (*paradigmatics*). This distinction allows within any sufficiently large set of strings of natural language discourse to ascertain syntagmatic regularities of element aggregations on level n whose characteristic patterns form paradigmatic regularities tantamount to their aggregational status on level $n + 1$. As has been illustrated above, the distinction of representational levels is tantamount to the categorial constraints applied when identifying regularities. Fully deterministic if-then rules will result in a rather coarse three-level hierarchy of categorial description whereas as probabilistic or possibilistic dependency produces a continuous, multi-level covering of distributional representations. Thus, it can be distinguished sharply between cognitive *linguistic* and *semiotic* procedures whose computations transform structured input data according to its immanent regularities to yield new, structural representations emerging from that computation (as hypothesized by *performative* linguistics and realized in procedural models of

[7] The Trier *dpa*-Corpus for instance comprises the complete textual materials from the so-called *basic news real service* of 1990–1993 (720.000 documents) which the Deutschen Presseagentur (*dpa*), Hamburg, deserves thanks to have the author provided with for research purposes. After deletion of editing commands the Trier-*dpa*-Corpus consists of approx. 180 Mio. $(18 \cdot 10^7)$ running words (*tokens*) for which an automatic tagging and lemmatising tool is under development. It is this corpus which provides the performative data of written language use for the current (and planned) *fuzzy*-linguistic projects at the our department.

[8] In subscribing to a structuralistic view of natural languages, the distinction of *langue–parole* and *competence–performance* in modern linguistics allowes for different levels of language description and linguistic analysis. Being able to *segment* strings of language discourse and to *categorize* types of linguistic entities, however, is but making analytical use of *structural couplings* presented by natural language discourse to semiotic systems properly attuned.

n- grams	$\|F_n\|)$ (fact.occurr.)	$\|T_n\| = \|Z^n\|$ (theor.possib.)	$100 \cdot \frac{F_n}{T_n}$ percent	$A_n = m \cdot \|F_{n-1}\|)$ (act.possib.)	$100 \cdot \frac{F_n}{A_n}$ percent
1	31	31	100,000	31	100,000
2	817	961	85,015	961	85,015
3	10.175	29.791	34,154	25.327	40,174
4	54.470	923.521	5,898	315.425	17,268
5	164.045	28.629.151	0,572	1.688.570	9,715
6	357.632	887.503.681	0,040	5.085.395	7,032
7	634.767	27.512.614.111	0,002	11.086.592	5,725
Size of test-corpus :		3.648.326	(signs)	502.587	(words)

Table 1: Graph-(letter-)combinatorics with (theoretically and faktually) possible and actually occurring *types* of n-grams in a subset of the Trier *dpa*-Korpus

computational semiotics). The elements of these new structures are value distributions or vectors of input entities that depict properties of their structural relatedness, constituting multi-dimensional (metric) space structures (*semiotic spaces*). The elements may also be interpreted as *fuzzy sets* allowing set theoretical operations be exercised on these representations that do not require categorial type (*crisp*) definitions of concept formations. Computation of letter (*morphic*) vectors in *word space*, derived from n-grams of letters *graphemes* [10] [11] as well as of word (*semic*) vectors in *semantic space* [12] [13], derived from wordtype correlations of their tokens in discourse will serve to illustrate the operational flexibility and fine granularity of vector notations [9], [16] to identify regularities of semiotic meaning constitution in language performance which traditional linguistic categories fail to represent.

4.2 The word space

The following notations will be used to outline the computational semiotic approach on the morphic level:

n-grams are n-elementary strings of entities. For $n \geq 2$ they may be analysed as ordered pais of adjacent items (letters, graphs, sign-strings, word-strings, etc.) which are the basis of

abstractions over such items may procedurally be determined as *soft* categorial types (corresponding to characters, graphems, morphems, syllables, words, etc.). These have been introduced as *dispositional dependency structures (DDS)* [14] [8] and formally declared as

fuzzy (sub-)sets of multi-dimensional sign inventories Z^n with $n \geq 1$

$$\tilde{X}_n := \{(x, \mu_n(x)) \colon x \in Z^n\} \subseteq Z^n \times [0,1] \qquad (1)$$

whose elements' grades of membership are defined by the membership-function

$$\mu_n \colon Z^n \to [0,1] \qquad (2)$$

membership-values $\mu_n(x)$ may be computed inductively as the overall tendency of linear chaining of items in language corpora. For an n-elementary string $x \in Z^n$ be $H_n(x)$ the frequency of x occurring in a corpus. Then, for any

bi-gram $x = (y,z) \in Z^n$, $y \in Z^{n-1}$, $z \in Z$, the coefficient

$$\mu_n(x) = \frac{H_n(x)}{H_{n-1}(y)}. \qquad (3)$$

with $Z = \{z_1, \ldots, z_m\}$ will yield for each $y \in Z^{n-1}$ a vector

$$(\mu_n(y, z_1), \ldots, \mu_n(y, z_m))^T \in I\!R^m. \qquad (4)$$

The set of all vectors reflect the morphological structure of the corpus analysed which is the numerically specified basis for the procedural definition of

soft categories which are defined as a system of *fuzzy* sub-sets of observed chaining regularities. They may be interpreted to represent *elastic constraints* operating on the language items' chaining tendencies which structure the corresponding corpus.

The presentation of the development of *soft* categories as *elastic constraints* (operating on different levels) can be simplified by their formal introduction as (n-ary) *fuzzy* relations and their corresponding numerical formats of transition matrices (of higher orders).

For written German discourse analysed on typesetting level with m discernable types of signs (letters) and maximum lengths n of strings there are quite a number of theoretically possible (*Tab.* 1, col. T_n) crisp n-ary relations $T_n = Z^n$, i.e.

$$\begin{array}{rcll}
Z = T_1 & = & \{x_1 & : x_1 \in Z\} \\
T_2 & = & \{(x_1, x_2) & : x_1, x_2 \in Z\} \\
T_3 & = & \{(x_1, x_2, x_3) & : x_1, x_2, x_3 \in Z\} \\
\vdots & & \quad\vdots & \quad\vdots \\
T_{n-1} & = & \{(x_1, \ldots, x_{n-1}) & : x_1, \ldots, x_n \in Z\} \\
T_n & = & \{(x_1, \ldots, x_n) & : x_1, \ldots, x_n \in Z\}.
\end{array}$$

Out of these, however, only those have to be computed which are not only actually possible (col. A_n) but which have indeed been observed to factually occur, i.e. $F_n \subseteq F_{n-1} \times Z$ (*Tab* 1, col. F_n), i.e.

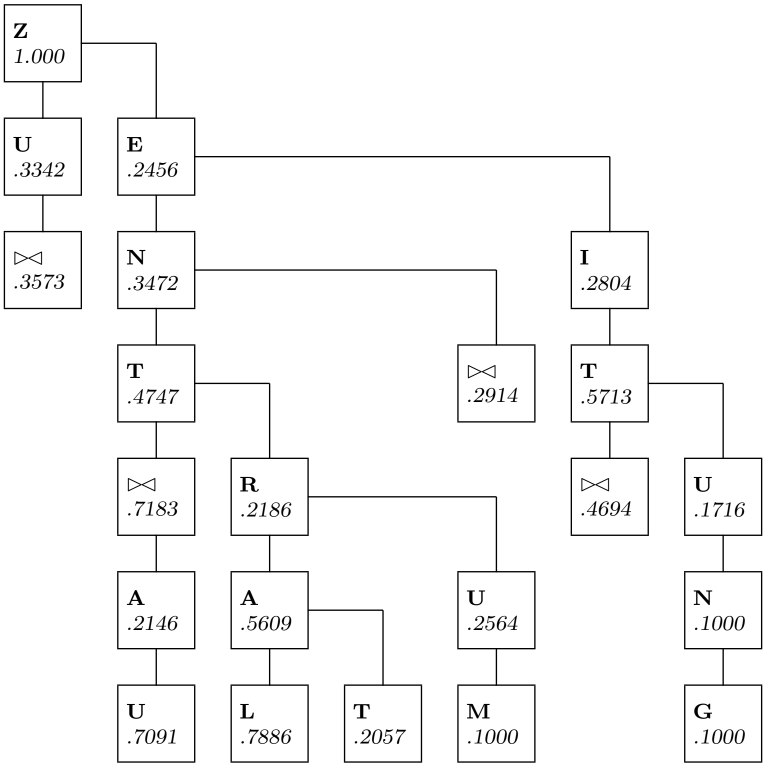

Figure 4: Tree representation of procedural *soft* category $\tilde{\mathbf{Z}}$ depicting the hierarchy of graded letter agglomeration according to decreasing transition tendencies (in 7-grams) of German newspaper texts.

$$
\begin{array}{lll}
Z = & & \\
F_1 \subseteq \{x_1 & : x_1 \in Z\} \\
F_2 \subseteq \{(x_1, x_2) & : x_1 \in F_1, x_2 \in Z\} \\
F_3 \subseteq \{(x_1, x_2, x_3) & : (x_1, x_2) \in F_2, x_3 \in Z\} \\
\vdots \quad \vdots & \quad \vdots \\
F_{n-1} \subseteq \{(x_1, \ldots, x_{n-1}) & : (x_1, \ldots, x_{n-2}) \in F_{n-2}, x_{n-1} \in Z\} \\
F_n \subseteq \{(x_1, \ldots, x_n) & : (x_1, \ldots, x_{n-1}) \in F_{n-1}, x_n \in Z\}.
\end{array}
$$

The *fuzzy* relational modeling (*Eqns.* 3 and 2) shows that even for higher n only *bi-grams* have to be traced and computed due the $(n-1)$-ary relations computed on the previous level of representation. It is this principle of procedural *self-similarity* of n-ary agglomerative steps which allows for the trie-like representation [3] of entities that are labeled (by soft categorial n-relative letter transitions) and are an outcome of procedural constraints (over n levels of processing) which produce a dynamically structured system of fuzzy relations that depicts the overall transition tendencies of signs. For the letter Z this structure is given in Fig. 4 illustrating sub-regularities of morphic word formation. bzw. des Verst”andnisses dies erzwingt.

4.3 The semantic space

Based upon the language entity word, its different types, and their frequencies of occurrence in natural language discourse, the fundamental distinction of agglomerative or *syntagmatic* and selective or *paradigmatic* can also be employed to reconstruct a relational system structure which will serve as base for a procedural model generating tree-like representations of dynamic semantic constraints. As these techniques have been introduced and elaborated elsewhere [9] [15] [20] their concise description may suffice here.

The core of the representational formalism can be characterized as a two-level process of abstractions. The first (called α-abstraction) on the set of *fuzzy* subsets of the vocabulary provides the word-types' usage regularities or *corpus points*, the second (called δ-abstraction) on this set of *fuzzy* subsets of corpus points provides the corresponding *meaning points* as a function of all differences of all usage regularities which a set of word-types may produce by its word-tokens' frequencies as observed in *pragmatically homogeneous* corpora of natural language texts.

The basically descriptive statistics to specify intensities of co-occurring lexical items in texts is centred around the correlational measure

$$
\alpha(x_i, x_j) = \frac{\sum_{t=1}^{T}(h_{it} - e_{it})(h_{jt} - e_{jt})}{\left(\sum_{t=1}^{T}(h_{it} - e_{it})^2 \sum_{t=1}^{T}(h_{jt} - e_{jt})^2\right)^{\frac{1}{2}}}; \quad (5)
$$
$$
-1 \leq \alpha(x_i, x_j) \leq +1
$$

where $e_{it} = \frac{H_i}{L} l_t$ and $e_{jt} = \frac{H_j}{L} l_t$, computed over a textcorpus $K = \{k_t\}; t = 1, \ldots, T$ having an overall length $L = \sum_{t=1}^{T} l_t; 1 \leq l_t \leq L$ measured by the number of word-tokens per text, and a vocabulary $V = \{x_n\}; n = 1, \ldots, i, j, \ldots, N$ of word-types whose frequencies are denoted by $H_i = \sum_{t=1}^{T} h_{it}; 0 \leq h_{it} \leq H_i$.

To specify these correlational value distributions' differences, a measure of similarity (or rather, dissimilarity) is used

$$\delta(y_i, y_j) = \left(\sum_{n=1}^{N} (\alpha(x_i, x_n) - \alpha(x_j, x_n))^2 \right)^{\frac{1}{2}}; \qquad (6)$$

$$0 \leq \delta(y_i, y_j) \leq 2\sqrt{n}$$

The consecutive application of (*Eqns.* 7) on input texts and (*Eqns.* 9) on the output data of (*Eqns.* 7) allows to model the meanings of words as a function of differences of usage regularities (*Fig.* 6).

Thus, $\alpha_{i,j}$ allows to express pairwise relatedness of word-types $(x_i, x_j) \in V \times V$ in numerical values ranging from -1 to $+1$ by calculating co-occurring word-token frequencies (*Eqn.* 5) for pairs of items.
As a fuzzy binary relation, $\tilde{\alpha} : V \times V \rightarrow I$ can be conditioned on $x_n \in V$ which yields a crisp mapping

$$\tilde{\alpha} \mid x_n : V \rightarrow C; \{y_n\} =: C \qquad (7)$$

where the tupels $\langle (x_{n,1}, \tilde{\alpha}(n,1)), \ldots, (x_{n,N}, \tilde{\alpha}(n,N)) \rangle$ represent the numerically specified, *syntagmatic* usage regularities that have been observed for each word-type x_i against all other $x_n \in V$. *α-abstraction* over one of the components in each ordered pair defines

$$x_i(\tilde{\alpha}(i,1), \ldots, \tilde{\alpha}(i,N)) =: y_i \in C \qquad (8)$$

Hence, the regularities of usage of any lexical item will be determined by the tupel of its α-values which for all word types can be represented as vector space C.

$$\begin{tikzcd}
 & C \arrow[loop above, "\tilde{\delta}"] \arrow[dr, "\tilde{\delta}|y_j"] & \\
V \arrow[loop below, "\tilde{\alpha}"] \arrow[ur, "\tilde{\alpha}|x_i"] \arrow[rr, "\tilde{\delta}|y_n \circ \tilde{\alpha}|x_n"'] & & S \arrow[loop right, "\zeta"]
\end{tikzcd}$$

Figure 6: Fuzzy mapping relations $\tilde{\alpha}$ and $\tilde{\delta}$ between the structured sets of vocabulary items $x_n \in V$, of corpus points $y_n \in C$, and of meaning points $z_n \in S$.

Considering C as representational structure of abstract entities constituted by *syntagmatic* regularities of word-token occurrences in *pragmatically homogeneous* discourse, then the similarities and/or dissimilarities of these entities will capture their corresponding word-types' *paradigmatic* regularities calculated by δ *Eqn.* 6 serving as *second* mapping function, As a fuzzy binary relation, $\tilde{\delta} : C \times C \rightarrow I$ can be conditioned on $y_n \in C$ which again yields a crisp mapping

$$\tilde{\delta} \mid y_n : C \rightarrow S; \{z_n\} =: S \qquad (9)$$

where the tupels $\langle (y_{n,1}, \tilde{\delta}(n,1)), \ldots, (y_{n,N}\tilde{\delta}(n,N)) \rangle$ represents the numerically specified *paradigmatic* structure that has been derived for each abstract *syntagmatic* usage regularity y_j against all other $y_n \in C$. The distance values can therefore be abstracted analogous to *Eqn.* 8, this time, however, over the other of the components in each ordered pair, thus defining an element $z_j \in S$ called *meaning point* by

$$y_j(\tilde{\delta}(j,1), \ldots, \tilde{\delta}(j,N)) =: z_j \in S \qquad (10)$$

Identifying $z_n \in S$ with the numerically specified elements of potential paradigms, the set of possible combinations $S \times S$ may structurally be constrained and evaluated without (direct or indirect) recourse to any pre-existent external world. Introducing a EUCLIDian metric

$$\zeta : S \times S \rightarrow I \qquad (11)$$

the hyperstructure $\langle S, \zeta \rangle$ or *semantic space* (*SS*) is declared constituting the system of *meaning points* as an empirically founded and functionally derived representation of a lexically labelled knowledge structure.

Weighted numerically as a function of an element's distance values and its associated node's level and position in the tree, $Cr(z_i)$ either is an expression of the head-node's z_i meaning-dependencies on the daughter-nodes z_n or, inversely, expresses their meaning-criterialities adding up to an aspect's interpretation determined by that head [15]. To illustrate the feasibility of the Δ-operation's generative procedure, the substructure of relevant constraints (related meaning points) $DDS(z_i) \subseteq \langle S, \zeta \rangle$ anchored with the lexical item x_i, $i =$ COMPUTER is shown in *Fig.* 4.3.

5 Conclusion

It has been outlined here that the morphic sign or the semantic meaning functions' ranges may be computed and simulated as a result of exactly those (semiotic) procedures by way of which (representational) structures emerge and their (interpreting) actualisation is produced from observing and analyzing the domain's possibilistically determined constraints as imposed on the linear ordering (*syntagmatics*) and the selective combination (*paradigmatics*) of natural language entities (morph-types, word-types) in communicative language performance. For *fuzzy linguistic* morhology and lexical semantics this is tantamount to (re-)present an entity's *semiotic* potential (function, meaning] by a fuzzy *distributional pattern* of the modelled system's state rather than a *single symbol*. The representational system's dynamic structure modeled by the procedures outlined is to represent a *semiotic cognitive information processing system*'s interpretation of its environment.

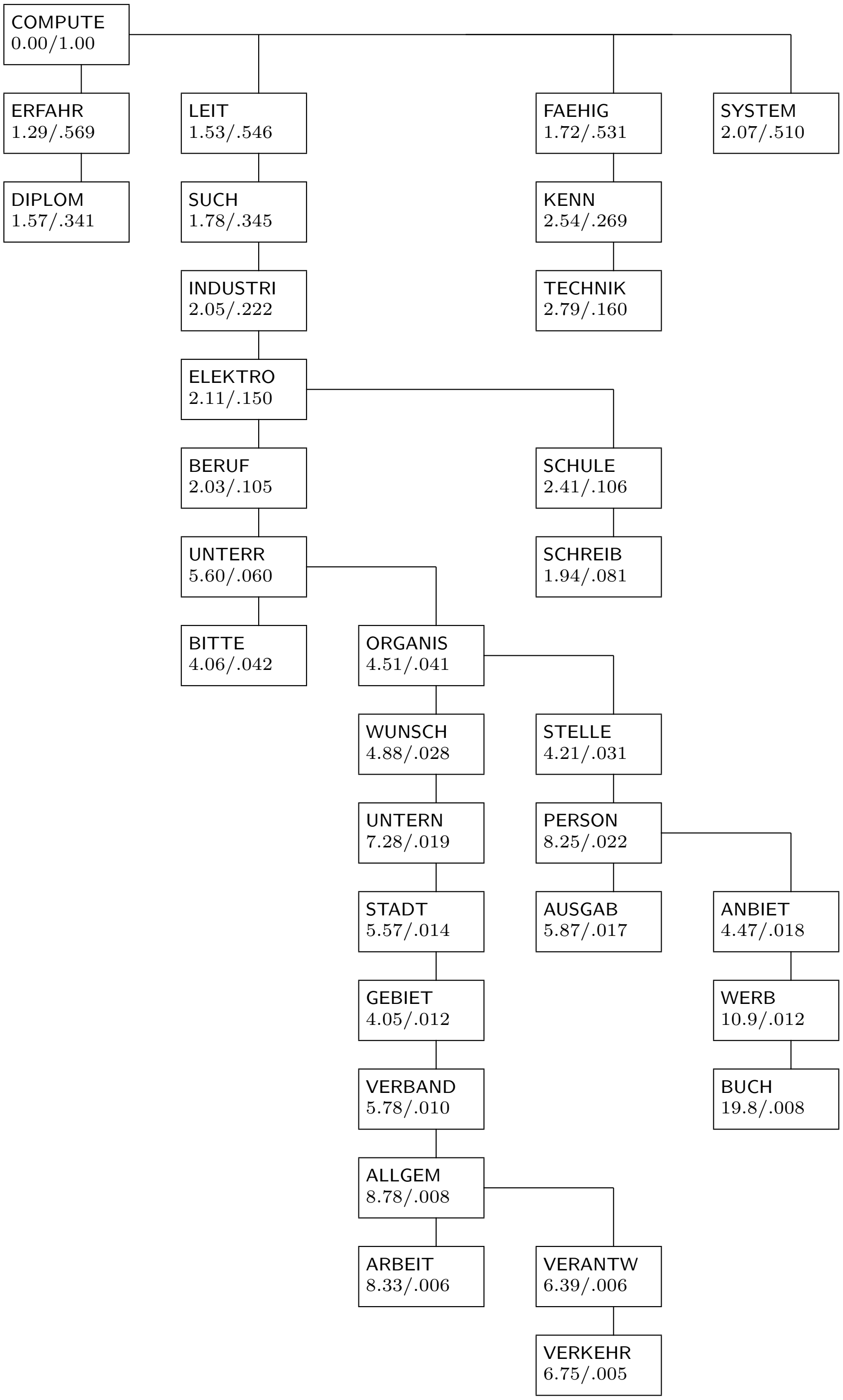

Figure 5: *DDS*-tree representation of meaning of COMPUTER as assembled for relevant meaning points' distances (first value) and criterialities (second value) on the basis of the semantic space $\langle S, \zeta \rangle$ intermediate as computed from a subcorpus of German newspaper texts (Die Welt, 1964 Berlin Edition).

References

[1] M. Bierwisch: Logische und psychologische Determinanten der Struktur natürlicher Sprachen. In: J. Scharf (Ed): *Naturwissenschaftliche Linguistik*, Nova Acta Leopoldina, NF 54, Nr.245, pp. 176–187; 841–851 (J.A. Barth) Halle, 1981.

[2] N. Chomsky: *The Logical Structure of Linguistic Theory.* (Plenum) NewYork, 1975.

[3] D. E. Knuth: *The Art of Computer Programming. Vol.3*, (Addison-Wesley) Reading, MA, 1973.

[4] D. Marr: *Vision.* (Freeman) SanFrancisco, 1982.

[5] A. Meystel: *Semiotic Modeling and Situation Analysis: an Introduction.* (AdRem Inc) Bala Cynwyd, PA, 1995.

[6] C. Peirce: Pragmatism in Retrospect: A Last Formulation. In: J. Buchler (Ed): *The Philosophical Writings of Peirce*, pp. 269–289. (Dover) New York, 1906.

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

[8] B. Rieger: *Unscharfe Semantik. Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten.* (Lang) Frankfurt a.Main/ Bern/ Paris, 1989.

[9] B. Rieger: Unscharfe Semantik: zur numerischen Modellierung vager Bedeutungen von Wörtern als fuzzy Mengen. In: H. Friemel, et.al. (Eds): *Forum-90: Wissenschaft und Technik. (Informatik-Fachberichte 259)*, pp. 80–104. (Springer) Berlin/ Heildeberg/ NewYork, 1990.

[10] B. Rieger: Fuzzy Modellierung linguistischer Kategorien. In: W. Feldmann, H. Hinrichs (Eds): *Text und Lexikon*, [in print] (Niemeyer) Tübingen, 1996.

[11] B. Rieger: Warum Fuzzy Linguistik? Überlegungen und Ansätze einer computerlinguistischen Neuorientierung. In: D. Krallmann, W. Schmitz (Eds): *Intern. Gerold Ungeheuer Symposium*, [in print] (Nodus) Münster, 1996.

[12] B. B. Rieger: Feasible Fuzzy Semantics. On some problems of how to handle word meaning empirically. In: H. Eikmeyer, H. Rieser (Eds): *Words, Worlds, and Contexts. New Approaches in Word Semantics*, pp. 193–209. (de Gruyter) Berlin/ NewYork, 1981.

[13] B. B. Rieger: Fuzzy Representation Systems in Linguistic Semantics. In: R. Trappl, et.al. (Eds): *Progress in Cybernetics and Systems Research*, Vol. XI, pp 249–256. (McGraw-Hill), Washington/ NewYork/ London, 1982.

[14] B. B. Rieger: Lexical Relevance and Semantic Disposition. On stereotype word meaning representation in procedural semantics. In: G. Hoppenbrouwes, et.al. (Eds): *Meaning and the Lexicon*, pp 387–400. (Foris), Dordrecht, 1985.

[15] B. B. Rieger: Distributed Semantic Representation of Word Meanings. In: J. D. Becker, et.al. (Eds): *Parallelism, Learning, Evolution. Evolutionary Models and Strategies WOPPLOT-89*, [Lecture Notes in Artificial Intelligence 565], pp. 243–273. (Springer) Berlin/ Heidelberg/ New York, 1991.

[16] B. B. Rieger: On Distributed Representation in Word Semantics. ICSI-TR-91-012, (ICSI), Berkeley, CA, 1991.

[17] B. B. Rieger: Meaning Acquisition by SCIPS. In: B. M. Ayyub (Ed): *ISUMA-NAFIPS-95*, [IEEE-Transactions: Joint Intern. Conf. on Uncertainty Modeling and Analysis, North American Fuzzy Information Processing Society], pp. 390–395, (IEEE Computer Society Press), Los Alamitos, CA, 1995.

[18] B. B. Rieger: Situation Semantics and Computational Linguistics: towards Informational Ecology. A semiotic perspective for cognitive information processing systems. In: K. Kornwachs, K. Jacoby (Eds): *Information. New Questions to a Multidisciplinary Concept*, pp 285–315, (Akademie), Berlin, 1995.

[19] B. B. Rieger: Situations, Language Games, and SCIPS. Modeling semiotic cognitive information processing systems. In: A. Meystel, N. Nerode (Eds): editors, *Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems*, pp. 130–138, (AdRem), Bala Cynwyd, PA, 1995.

[20] B. B. Rieger / C. Thiopoulos: Semiotic Dynamics: a self-organizing lexical system in hypertext. In: R. Köhler, B. Rieger (Eds): *Contributions to Quantitative Linguistics. Proceedings of the 1st Quantitative Linguistics Conference – QUALICO-91*, pp. 67–78. (Kluwer Academic) Publishers, Dordrecht, 1993.

[21] H. A. Simon: *The Sciences of the Artificial.* (MIT Press), Cambridge, MA, 1982.

[22] F. Varela / E. Thompson / E. Rosch: *The Embodies Mind. Cognitive Science and Human Experience.* (MIT Press), Cambridge, MA, 1991.

[23] L. Zadeh (Ed): *Fuzzy Sets and their Application to Cognitive and Decision Processes.* (Academic Press), New York/ San Francisco, 1975.

[24] L. A. Zadeh: Fuzzy Sets. *Information and Control*, 17(8):338–353, 1965.

[25] L. A. Zadeh: Fuzzy Logic, Neural Networks, and Soft Computing. *Comm. of the ACM*, 37(3):77–84, 1994.