Computing Fuzzy Semantic Granules from Natural Language Texts.
A computational semiotics approach to understanding word meanings.

Burghard B. Rieger*
FB II: Department of Computational Linguistics,
University of Trier, Germany

Abstract
The notion of Computing with Words hinges crucially on the employment of natural language expressions. As meaning representations, these are considered observable and accessible evidence of processes of human cognition, represented by textual structures and actualized in processes of understanding. Cognitive processes and language structures are characterized by information granulation, organization, and causation which can be modeled both, in their crisp as well as fuzzy modes of structural and functional processing. Allowing this are intrinsic constraints which may be exploited, analyzed, and represented in a procedural way.

1 Introduction
In his keynote lecture on Information Granulation and its Centrality in Human and Machine Intelligence, ZADEH related significant properties of natural languages to human perceptions which lie at the base of the meanings of words. Underlying natural language understanding are the same remarkable human capabilities as shared by processes of perceiving the world, of constituting meanings and/or (parts of) reality respectively. These allow a wide variety of physical and mental tasks to be performed by humans without detailed measurement and/or numeric computation that artificial information processing systems obviously are unable to solve. Vehicle of such outstanding performance is the particular way representations of results of these processes are formed, employed, and processed by forming representations of results of such processes. For the recursive way this is achieved, the centrality of information granulation and organization has been identified. These are core concepts in the theory of fuzzy information granulation (TFIG) which are believed to play a fundamental role also in the computational theory of perception (CTP) under development for the successful design and utilization of advanced intelligent information systems.

As the notion of computing with words (CW) hinges crucially on the communicative employment of natural language expressions, it has been found that these may provide not only the representational structures but also some valuable hints for the operational processing allowing for decomposing wholes into their constituent parts (granulation), and for composing or integrating parts into wholes (organization). From a linguistic point-of-view, natural languages themselves may be taken as the salient paradigm for information granulation both, in its crisp as well as its fuzzy modes of structural representations and the way their processing can be modeled in machine simulation.

2 Computational Semiotics
According to information systems theory, human beings may be taken as living systems whose knowledge based processing of represented information makes them cognitive, and whose sign and symbol generation, manipulation, and understanding capabilities render them semiotic. Due to our own daily experience of these systems’ performance and ability in representing results of cognitive processes, in organizing these representations, and in modifying them according to changing conditions and states of system-environment adaptedness, it is argued that the semiotic approach to modeling human cognition—constituting computational semiotics—will have to be grounded in such complex semiotic cognitive information processing. Consequently, it has to be based upon the representational structures resulting from and initiating such processing, i.e. natural language discourse. In the aggregated form of pragmatically homogeneous text (PHT) corpora, natural language discourse, as performed for communicative purposes, provides a cognitively revealing and empirically accessible system whose multi faceted structuredness may serve as guideline for the cognitively motivated, empirically based, and computationally realized research in the semiotics of language.

In a rather sharp departure from Computational Linguistics (CL) and Artificial Intelligence (AI) approaches, Computational Semiotics (CS) modeling neither presupposes rule-based or symbolic formats for linguistic knowledge representations, nor does it subscribe to the notion of world knowledge as some static structures that may
be abstracted from and represented independently of the way they are processed. Consequently, knowledge structures and the processes operating on them are modeled procedurally and implemented as algorithms for computational simulation. They determine Semiotic Cognitive Information Processing Systems (SCIPS) [5] as a collection of cognitive information processing devices whose semiotic character consists in a multi-level representational system of (working) structures emerging from and being modified by such processing. Corresponding to these levels of emerging structures are different degrees of resolution [1] that account for varying levels of representational granularity [15].

3 Processing PHT Corpora
In earlier attempts, semantic meaning functions have been modeled and computed as results of the same (semiotic) procedures by way of which (representational) structures emerge [12]. Their actualization (interpretation) can be simulated by analyzing the probabilistic constraints found to be imposed upon the linear ordering (syntagmatics) and the selective combination (paradigmatics) of natural language entities (word-types) in discourse [4]. In a fuzzy linguistics approach to lexical semantics this is tantamount to (re-)construct an entity’s semiotic potential (meaning function) by a weighted graph (fuzzy distributional pattern) representing a particular state of the modeled system’s lexical state space rather than by a single symbol whose interpretation would have to be extrinsic to that system [8]. In this view, the emergence of semantic structure can be represented and studied as a self-organizing process based upon word usage regularities in natural language discourse [5]. In its course, the linearly agglomerative (or syntagmatic) as well as the distributionally selective (or paradigmatic) constraints are exploited by text analyzing algorithms [6]. These accept natural language text corpora as input and produce—via levels of intermediate processing and representation—a vector space structure as output. As semantic hyperspace (SHS) it may be interpreted as an internal (end) representation of the SCIP system’s states of adaptation to the external (exo) structures of its environment as mediated by the discourse processed [7]. The degree of correspondence between these two is determined by the granularity that the texts provide in depicting an exo-view, and the resolution that the SCIP system is able to acquire as its endo-view in the course of that discourse’ processing [9].

3.1 Empirical quantitative analysis
Following the procedural approach in computational semiotics, the reconstruction of linguistic functions or meanings of words is based upon a fundamental analytical as well as representational formalism. It can be characterized as a two-level process of abstraction (called α- and δ-abstraction) on the set of fuzzy subsets of the vocabulary—providing the word-types’ usage regularities or corpus points—and on the set of fuzzy subsets of these—providing the corresponding meaning points. These may be understood to interpret semantically (by way of the meaning function) those word-types which are being instantiated by word-tokens as employed in pragmatically homogeneous corpora of natural language texts.

The basically descriptive statistics used to grasp these relations on the level of words in discourse is centered around a correlational measure (1) to specify intensities of co-occurring lexical items in texts, and a measure of similarity (or rather, dissimilarity) (4) to specify these correlation value distributions’ differences. Simultaneously, these two measures may also be interpreted semiotically as providing for the set theoretical constraints or formal mappings α (2) and δ (5) which model the meanings of words as a function of these words’ differences of usage regularities as produced in discourse and analysed in the PHT corpus.

αi,j allows to express pairwise relatedness of word-types \((x_i, x_j) \in V \times V\) in a number of texts while being instantiated by word-tokens as employed in the PHT corpus.

\[
\alpha_{i,j} = \frac{\sum_{t=1}^{T}(h_{it} - e_{it})(h_{jt} - e_{jt})}{\left(\sum_{t=1}^{T}(h_{it} - e_{it})^2 \sum_{t=1}^{T}(h_{jt} - e_{jt})^2\right)^{\frac{1}{2}}};
\]

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

Evidently, pairs of items which frequently either co-occur in, or are both absent from, a number of texts will positively be correlated and hence called affined, those of which only one (and not the other) frequently occurs in a number of texts will negatively be correlated and hence called repugnant.

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

\[
\tilde{\alpha} | x_n : V \rightarrow C; \{y_n\} =: C
\]

where the tuples \(((x_{n_1}, \tilde{\alpha}(n_1)), \ldots, (x_{n_N}, \tilde{\alpha}(n_N)))\) form a matrix representing the numerically specified, generalized syntagmatic usage regularities that have been observed for each word-type \(x_i\) against all other \(x_n \in V\). The α-abstraction over one of the components in each ordered pair defines

\[
x_i(\tilde{\alpha}(i, 1), \ldots, \tilde{\alpha}(i, N)) =: y_i \subset C
\]

Hence, the regularities of usage of any lexical item will be determined by the tuple of its affinity/repugnancy-values towards each other item of the vocabulary which—interpreted as coordinates—can be represented by points
again yields a crisp mapping, \( \tilde{\alpha} \) against those of all other items. As a fuzzy binary relation any item’s differences of usage regularities measured by a distance measure \( \delta \) of \( \alpha \)- and \( \delta \)-regularities. These may be calculated as an empirically founded and functionally derived representation of a lexically labeled knowledge structure (Tab. 1).

### 4 Processing SHS Structures

Thus, the SCIP system’s architecture is a two-level consecutive mapping of distributed representations of systems of (fuzzy) linguistic entities. Being derived from usage regularities as observed in texts, these representations provide for the aspect driven generation of formal dependencies and their interrelations in a format of structured stereotypes. Corresponding algorithms select and represent fuzzy subsets (word meanings) as dispositional hierarchies that render only those relations accessible to perspective processing which can—under differing aspects differently—be considered relevant. Such dynamic dispositional dependency structures (DDS) have proved to be an operational prerequisite to and a promising candidate for the simulation of content-driven (analogically-associative) reasoning instead of formal (logically-deductive) inferences in semantic processing [10]. Considered as states which the SCIP system can enter, certain properties of these structures can be identified as results of symbolic functions which were shown to correspond to basal referential predicates [5] [7].

### 4.1 Structuring information granules

Dispositional dependency structures (DDS) (Fig. 2) can be viewed as alternative procedural format of fuzzy information granulation which extends the rule-based frame
Figure 2: The semantic inference procedure is a parallel process activated from start nodes (premises) generating DDS graphs and stopped by first node common to all (conclusion). Subtrees constitute perspective determined information granules of differing connotative, resolutional, and dependency structure.

as introduced by the concept of generalized constraint [13] and exemplified in [15] as unconditional constraints. According to Zadeh (1997), a generalized constraint on values of $X$ is expressed as $X \text{ isr } R$, where $X$ is a variable which takes values in a universe of discourse $U$, $\text{isr}$ is a variable copula with $r$ being a discrete variable whose values define the way in which $R$ constrains $X$, and $R$ is the constraining relation. For $r$ different values may be defined as equality, possibility, verity, probability, random set, and fuzzy graph and their related (definitional, operational, procedural, computational) interpretations can be given. From our perspective it is important to observe that $r$ is a means to enrich the copula’s interpretations in a controlled and operationally defined way which relates to $R$ in a predicative sense, i.e. specifying the interpretation of $R$ (generally a distribution of grades of membership) as being possibilities, truth values, probabilities or composites thereof. As these functional types of $r$ need to be distinguished in order to determine their interpretation for $R$ in rule-based mechanisms of inferential processing, this necessity may be relaxed or even become obsolete when the rule-based inference mechanism is replaced by an algorithmic procedure, operating on a well-defined structure like SHS as specified numerically by the value distributions which constitute the meaning points’ interpretations.

In addition to the types of constraints defined above there are many others that are more specialized and less common. A question that arises is: What purpose is served by having a large variety of constraints to choose from? A basic reason is that, in general setting, information may be viewed as a constraint on a variable. (Zadeh 1997, p. 117)

4.2 Generating granular structures

Such constraints are induced not only by predicative expressions of truth-functional propositions but also by word meanings in natural language situated contexts. To model these constraints, word meanings are represented as procedurally determined numerically weighted graphs or dispositional dependency structures (DDS) as computed from natural language discourse in fuzzy linguistics [8]. Taking the concept of a generalized constraint to hold likewise for sentence meanings (propositional structure) as well as for word meanings (DDS), then the TFIG notational format translates to $X \simeq \{x_i\}$ where $X$ is a variable which takes values—via $\alpha$- and $\delta$-abstraction—of $z_i \in S$ with $S \subseteq U$. A semiotically generalized constraint on values of $X$ is expressed by $X \text{ dds}_S$, $S$ where $\text{dds}$ relates $x_i$ via $z_i$ to $S$ by restricting SHS procedurally in generating the tree structure from meaning point $z_i$.
as its root, and \( z_n \) as its discrete variables whose values determine different structures (dependency paths) which constrain the topology of \( S \) in a semantically perspective way.

It should be noted here that the notion of dependency path is a structural representation of a dynamic concept of granular word meaning which induces a reflexive, symmetric, and weakly transitive relation between relevant meaning points as its components. It allows for the procedural definition and computational enactment of semantic inferencing on the word level \([10]\), very much like the rule-based models of inferencing in granular fuzzy information processing based on fuzzy rules, or the syntactically defined propositional formats of symbolic processing in (cognitive linguistic) sentence semantics based on crisp logic calculi.

In Fig. 2 the semantic hyperspace \( \langle S \rangle \) was computed from a corpus of Reuters 1987 newswire articles\(^3\). Two vocabulary items \( x_i = \text{administration}, x_j = \text{deposit} \), corresponding to meaning points \( z_i, z_j \) were chosen as premises for the semantic inference process. It restricts \( \langle S \rangle \) simultaneously by generating the graphs \( DDS_i, DDS_j \) in parallel. The inferred conclusion is the first common node \( z_k = \text{estate} \) whose different dependency paths \( \text{dep}_i(z_k), \text{dep}_j(z_k) \) are given (center column). Depending on the semantic perspectives, however, as determined by the root node \( z_i, z_j \) respectively, the subtrees or information granules \( ig_i(k), ig_j(k) \), headed by \( z_k = \text{estate} \) (left and right column) demonstrate the \( i \) and \( j \) induced differences both, in connotative meaning and in semantic resolution of these fuzzy information granules.

## 5 Conclusion

The dynamics of semiotic knowledge structures and the processes operating on them essentially consist in their recursively applied mappings of multilevel representations resulting in a multiresolutional granularity of fuzzy word meanings which emerge from and are modified by such text processing. Numerous computational results from experimental test settings (in semantically different discourse environments) will be produced to illustrate the SCIP system’s granular meaning acquisition and language understanding capacity without any explicit initial morphological, lexical, syntactic, or semantic knowledge.

### References


\(^3\)Reuters-21578 (1.0) Text Categorization Test Collection, prepared by D.D.Lewis (AT&T Labs) and thankfully acknowledged here (www.research.att.com/~lewis/reuters21578.html).