Fuzzy Structural Semantics
On a generative model of vague natural language meaning*

Burghard B. Rieger

Illustrating the theoretical background of this chapter, I will *first* make some preliminary remarks, covering referential and structural semantic theory in linguistics, before *second* I will sketch the course of my approach in analyzing and describing natural language meaning within the frame of a pragmatically based generative model of structural semantics. Finally and *third* I will give some examples from computation of a corpus of 19th and 20th Century German students' poetry.

1. As a linguist, who thinks his discipline an empirical science, I will be not so much concerned with either language philosophy, formal logics or mathematics, but mainly with the study of meaning as it is constituted in spoken or written texts used in the process of *communication.* Rather than focussing on the fiction of an 'ideal speaker' or the formal rules of an abstract and mere theoretical language usage, my linguistic point of view implies that I am much more interested in the *analysis* and *description* of natural language regularities that real speakers/hearers follow and/or establish when they interact verbally by means of texts in order to communicate.

For any *description* of natural language meaning, however, we are in need of a formally adequate meta-language to depict semantic phenomena, and for any *analysis* of natural language meaning we need methods and procedures which are empirically adequate. Both, the postulates of *formal* and *empirical* adequacy will have to be met by a communicative theory of semantics that is comprehensive and satisfactory. Such a theory — that should be stressed here and kept in mind throughout the following — does not exist and no one has yet presented even the outlines of it — and I shall not either. But I think that the concept of fuzzy sets may prove to serve as an at least formally satisfactory and numerically flexible link or joint to connect the two main, seemingly divergent lines of research in modern semantics: namely, the more *theoretically oriented models* of what formal semanticists feel an 'ideal' speaker should, or would do when he produces meaningful sentences and the more *empirically oriented methods and procedures* of experimental semanticists that try to find out what real speakers actually do when they produce texts for communicative purposes.

In general, most linguists will probably agree that — whatever else has to be dealt with — natural language meaning presents two major problems:

[*] Published in: Trappl, R./Hanika, P./Pichler, F.R. (Eds.): Progress in Cybernetics and Systems Research (Vol. V), Washington/New York/London (Wiley & Sons) 1979, pp. 495–503.

firstly, what is known as the connotational or *structural* aspect of how the words and sentences of a language are related to one another;

and secondly, what is known as the denotational or *referential* aspect of how the words or sentences of a language are related to the objects and/or processes they refer to.

To start with the latter, *referential semantic theory* has developed along the line of Frege, Russel, early Wittgenstein and Carnap. Their relevance to linguistics and to linguistic semantics in particular has been recognized during recent years only. In the meantime, the increasing interest in formal semantics among linguists has produced quite a number of different models which share the fiction though, that natural language sentences ought to be either 'true' or 'false' or at worst have a third value like 'undetermined'. Like the truth-conditions for predicates, those for natural language sentences are analogously introduced in terms of classical set theory. Accordingly, the meaning of a word is basically identified with a set of points of reference in the universe of discourse, allowing a truth-value to be assigued to any (declarative) natural language sentence. These truth-value models now tend to exhibit all the formalisms and abstractions mathematical rigor calls for. They do so, however, at the price of a rather limited coverage of basic and very obvious characteristics of natural language meaning, one of which had to be excluded totally: that is *vagueness*.

Unlike referential theory, *structural semantics* has considered this very notion of vagueness to be fundamental to natural language meaning. Structuralists have therefore been concerned with the question of how the lexical meanings of words — rather than being related to extra-lingual sets of objects — are intra-lingually related to one another, constituting relational systems which people obviously make use of when communicating. According to structural theory, the meaning ('sense') of each term is to some extent depending on the position it occupies in that system. It is argued, that, although the terms may referentially be vague, the position of each term in the system relative to each other, will nevertheless be defined with precision.

This fiction of 'structural preciseness' as opposed to 'referential impreciseness' has inspired linguists since Saussure, Trier and Weisgerber up to Coseriu, Greimas and Lyons and even scholars from nonlinguistic disciplines like Osgood, Goodenough or Wallace — to mention only these few. Their models and methods have undoubtedly been fertile and influencial for some time and/or discipline. But as they were based mainly upon intuitive introspection and probands questioning, they do not seem to have achieved either the theoretical consistency or the methodological objectivity that empirical theory calls for. Thus, apart from the ethno-sciences or experimental psychology, structuralistic ideas seem to be of decreasing influence in modern linguistics and its recent semantic theories.

If, however, it is agreed that on the one hand natural language semantics should be an empirical science and as such be an integral part of modern linguistics, one obviously cannot be content to rely on traditional structural methods and related procedures of people looking into their minds, each into his own and some into others'. If on the other hand one is just as malcontent with the highly theoretical and most abstract concepts formal semantics can offer to cope with very real and concrete problems of natural language meaning, one is apt to think of the aforementioned comprehensive semantic theory which is both, empirically and formally adequate.

These issues became involved, when the concept of fuzzy sets was introduced into linguistic semantics. Basic [1] to the notion of fuzzy sets is — other than in classical set theory — that the elements of fuzzy sets show gradual rather than abrupt transition from non- to full-membership. Fuzzy sets are defined by characteristic- or membership-functions which associate with each element a real, nonnegative number between 0 and 1 with 0 equaling 'non-membership' and 1 equaling 'full-membership' in the classical sets-theoretical sense. Let A be a subset of X, then A can be defined by a membership-function

$$\mu_A \colon X \rightarrow [0, 1]$$

that will map X onto the interval $[0, 1]$. Hence, the fuzzy set A is defined to be the set of ordered pairs

$$A := \big\{ \big(x, \mu_A(x)\big) \big\} \text{ for all } x \in X.$$

Now, let X for instance be the continuous range of possible human ages from 0 to 100, then the meaning of a term like 'middle-aged' may referentially be represented as a fuzzy set, defined by a membership-function μ_m that associates with each possible age $x \in X$ a numeric value $\mu_m(x)$, giving the membership-grade of x in the fuzzy subset m ('middle-aged') of X, illustrated in Fig. 1.

In his 1971 paper on 'Quantitative Fuzzy Semantics' Zadeh adopted a strictly reference-theoretical model, into which he successfully incorporated the notion of fuzziness. He was able to show that the meaning of a word or term may well be vague in the sense that it refers to a set of reference-points whose boundary is not sharply defined, thus constituting a fuzzy set in the universe of discourse.

> "In fact it may be argued that in the case of natural languages, most of the words occurring in a sentence are names of fuzzy rather than non-fuzzy sets, with the sentence as a whole constituting a composite name for a fuzzy subset of the universe of discourse" [2]

The second aspect raised by Zadeh in the same paper, whether

> "fuzziness of meaning can be treated quantitatively, at least in principle" [3]

has however been dealt with only formally. The empirical side of it, concerning questions of how the meaning of a term described as a fuzzy set may be detected, or how the membership-grades may be ascertained and associated with the elements of a descriptor set in a particular case, these questions have not even been touched upon. We are informed instead that membership-functions

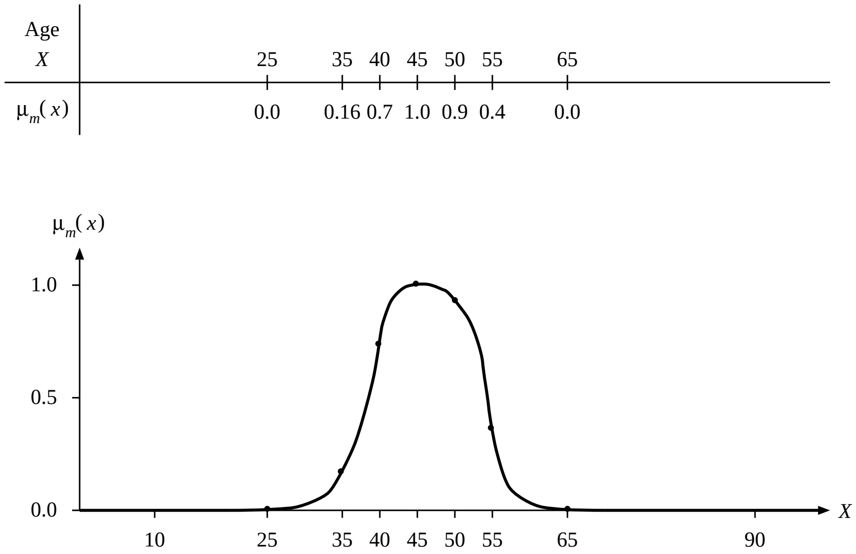

Figure 1: The meaning of ‘middle-aged’, referentially represented by the fuzzy set m, defined on the descriptor set X of possible ages x (subjectively).

> ”can be defined in a variety of ways: in particular (a) by a formula, (b) by a table, (c) by an algorithm (recursively), and (d) in terms of other membership functions (as in a dictionary)” [4]

From the empirical linguist’s point of view this is rather unsatisfactory. As clearly as he does recognize the relevance of fuzzy sets theory for the *description* of natural language meaning, he also will find that — what its *analysis* is concerned — fuzzy sets theory does not offer any new method. It seems that it merely allows a somehow quantified notation of more or less subjective, more or less acceptable results, which traditional methods of linguistic introspection may yield anyway.

I therefore would like to propose fuzzy sets *theory* to be combined with *methods* of statistical text-analysis in order to arrive at a generative model of structural semantics for which the notion of vagueness is constitutive.

2. It is assumed that the structural meaning of any lexical item (word, lexeme, stem, etc.) depends on its pragmatics and hence may be detected from sets of natural language texts according to the use the speakers/writers make of an item when they produce utterances in order to communicate. Such utterances are called ‘pragmatically homogeneous’ if they were written or spoken by real communicants in sufficiently similar situations of actually performed or at least intended verbal interaction.

It has been shown elsewhere [5] that in a sufficiently large sample of pragmatically homogeneous texts, called *corpus*, only a restricted vocabulary, i.e. a limited number of lexical items will be used by the communicants however comprehensive their personal vocabularies in general might be. Consequently, the lexical items employed in these texts will be distributed according to their communicative properties, constituting *semantic regularities* which may be detected empirically [6]. For this purpose a modified correlation coefficient has experimentally been used. It allows to compute the relational interdependence of any two lexical items from their textual frequencies. Those items which co-occur frequently in 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'. Different degrees of word-repugnancy and word-affinity — indicated by numeric values ranging from -1 to $+1$ — may thus be ascertained without recurring to an investigator's or his probands' knowledge of the language (competence), but solely from the regularities observed in a corpus of texts spoken or written by real speakers/writers in actual communication (performance).

Let T be such a corpus that consists of a number of texts t satisfying the conditions of pragmatic homogeneity. For illustrative purposes, we will consider a simplified case where the vocabulary V employed in these texts shall be restricted to only three word-types, say, i, j, and k which have a certain overall token-frequency. The correlation-coefficient α will measure the regularities of usage by the 'affinities' and 'repugnancies' that may hold between any one lexical item and all the others used in the texts. That will yield for any item an n-tuple of correlation-values, in this case for the lexical item i with $n = 3$ the tripel of values ii, ji, and ik. These correlation-values are now interpreted as being coordinates, that will define for each lexical item i, j or k one point α_i, α_j or α_k in a three-dimensional space.

This is illustrated in Fig. 2. There we have three axes representing the three word-types i, j and k which cross in front of the three planes cutting the axes at their $+1$ values. The point α_i is defined by the correlation-values $ii = +1$, $ij = -.25$ and $ik = -.75$; it is therefore situated in the i-plane with the interrupted lines (parallel to the j- and k-axis) representing the ii- and ik-values. The other points α_j and α_k are defined analogously. The position of α_i, in this space now obviously depends on the regularities the lexical item i has been used with in the texts of the corpus. α_i therefore is called *corpus-point* of i in the α- or *corpus-space*.

Two α-points in this space will consequently be the more adjacent to each other, the less their regularities of usage differ. This difference may be calculated now by a distance measure δ between any two α-points, illustrated in this figure by the dotted lines.

These distance-values, which are real, non-negative numbers, do represent a new characteristic which may be interpreted in two ways:

firstly: die dotted distances between any one α-point and all the others are interpreted as new coordinates: then these coordinates will again define a point in a new n-dimensional space, called *semantic-space*. The position of such a *meaning-point* in the semantic space will depend on all the differences (δ- or distance-values) in all the regularities of usage (α- or correlation-values) any lexical item shows in the texts analyzed;

secondly: the dotted distances between any one α-point and all the others are interpreted as membership-grades: then — after these δ-values have been trans-

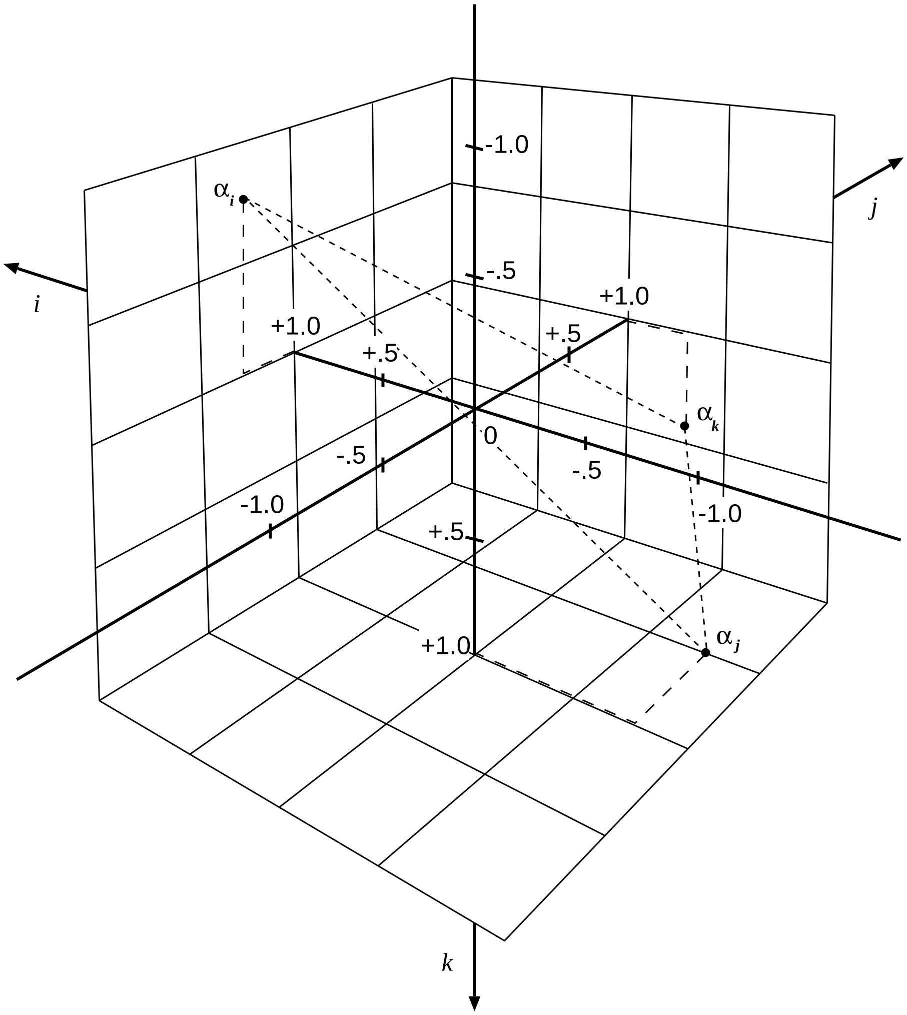

Figure 2: *Corpus* or *α-space*, representing usage of terms i, j, and k by *corpus points* α_i, α_j, and α_k, the δ distances (dotted lines) between which indicate usage differences of terms according to the texts analysed.

formed appropriately into μ-values, ranging from 0 to +1 — the differences of a lexical item's usage-regularities may well be represented by a fuzzy set with the vocabulary serving as its descriptor set.

Both these interpretations of δ-values, as coordinates of points in the semantic space *or* as membership-grades of fuzzy subsets in the vocabulary, are equivalent: they will equally map the 'meaning' of a word as a function of *all* its differences in *all* its regularities onto the vocabulary, according to the usage a lexical item is made of by the speakers/writers in a corpus of pragmatically homogeneous texts.

Apart from that, the fuzzy-sets-theoretical interpretation allows an considerable extension of this analytical model of structural meaning. Some basic definitions and formal operations may now be introduced which will allow an empirically based and formally satisfactory explication of linguistic *sense-relations* and — even more important than that — the formal *generation* of (at least in principle) infinitely many new meanings from the finite number of those lexical meanings, which prior to that have been analyzed empirically from the text-corpus.

Assuming that the definition — as proposed by Zadeh [7] — are well known, I will confine myself to show their semantic correspondences in this linguistic model of structural lexical meanings.

Synonymy (equality):

$$i = j \quad \text{iff} \quad \mu_i(k) = \mu_j(k) \quad \text{for all } k = 1, \ldots, n$$

Partial Synonymy (similarity):

$$i \approx j \quad \text{iff} \quad |\mu_i(k) - \mu_j(k)| \leq s \quad \text{for all } k = 1, \ldots, n$$

Hypnoymy (containment):

$$i \subset j \quad \text{iff} \quad \mu_i(k) \leq \mu_j(k) \quad \text{for all } k = 1, \ldots, n$$

Negation (complement):

$$\sim i := \mu_{i'}(k) = 1 - \mu_i(k) \quad \text{for all } k = 1, \ldots, n$$

Conjunction (intersection):

$$i \wedge j := \mu_{i \cap j}(k) = \min\bigl[\mu_i(k); \mu_j(k)\bigr] \quad \text{for all } k = 1, \ldots, n$$

Adjunction (union):

$$i \vee j := \mu_{i \cup j}(k) = \max\bigl[\mu_i(k); \mu_j(k)\bigr] \quad \text{for all } k = 1, \ldots, n$$

Synonymy of meanings may be explicated as equality of two fuzzy sets;

Partial synonymy of meanings may be defined in terms of a similarity-formula, introducing a threshold-value s;

Hyponymy of a meaning relative to another may be explicated as containment of fuzzy sets.

What the operations of *negation, adjunction* and *conjunction* are concerned, there has been quite a bit of critical discussion lately, particularly from the empiricists' point of view. For the generation of new meaning-points in the semantic space, I have so far gone back on those definitions proposed by Zadeh. Modified definitions of *adjunction* and *conjunction* proposed are, however, experimented with at the moment.

3. Coming to the end, I would like to give you some examples from the computer-analysis of a corpus of 19th and 20th Century German students poetry, the first part of which covering the early 19th Century comprises some 500 texts and a vocabulary of 315 lemmatized word-types/21000 tokens.

As there are serious difficulties in visualizing a 315-dimensional *semantic space* on the one hand, and, as there is, on the other, but little illustrative use in reproducing an n-tupel of 315 δ-values, defining a meaning-point or fuzzy set respectively of, say, the lexical item BAUM/tree (Fig. 3), I have thought of some other means to give an impression of the lexical structure.

To illustrate the position of a meaning-point, I have tabulated those points which are nearest to it in the semantic space, constituting something like a meaning-point's topological environment As I have shown elsewhere [8], these environments prove to be very similar to what linguists have called *paradigmatic* or *semantic fields*.

When you let your eyes pass along the meanings-points listed in Fig. 4 and Fig. 5, showing the environments of BAUM/tree and FRIEDHOF/graveyard, you will get an idea of the semantic fields of these words as used in the German poems of the early 19th Century. What the paradigmatic relations are concerned, I think they are rather self-evident to a native speaker of German, or, to say the least, they are not contra-intuitive [9].

As I have only started with the testing [10] of the operations defined to generate *new* meanings, I have chosen two lexical items, namely BAUM/tree and BLÜTE/blossom which paradigmatically are closely related. The idea was, that the new meaning-points 'BAUM/tree $\wedge$ BLÜTE/blossom' (Fig. 6) and 'BAUM/tree $\vee$ BLÜTE/blossom' (Fig. 7) resulting from conjunction and adjunction of these two items, should be positioned somewhere in the same region of the semantic space, which in fact they are.

As you might have noticed, my approach to the analysis and description of natural language meaning is still very tentative and far away from a consistent theory of semantics; but it is hoped, that this approach will arrive at a model which in its abstract (algebraic) parts may linguistically be interpreted as a corpus-independent theory of semantic competence ('langue'), whereas its empirical (quantitative) parts will represent the performative data ('parole') which are corpus-dependent and hence will vary according to the texts analyzed.

1.905	2.575	1.995	2.397	2.240	2.249	2.142	2.392
2.221	2.159	2.038	2.263	1.994	1.919	2.353	0.000
0.000	1.850	2.301	2.008	2.101	2.062	2.394	1.709
2.404	0.000	1.710	2.292	2.164	2.113	1.896	1.884
2.070	2.227	2.108	2.524	2.152	1.617	2.473	2.312
2.227	1.895	2.262	2.602	2.329	2.200	2.125	2.154
1.742	2.040	2.115	2.123	2.404	2.760	1.917	2.098
2.208	2.330	2.076	2.337	2.003	2.594	2.182	1.652
2.053	1.459	2.038	2.191	1.982	1.955	2.135	2.062
2.154	2.066	2.516	2.762	2.307	1.753	2.336	2.009
2.177	1.925	2.510	2.341	1.906	2.114	2.274	1.933
2.449	2.328	2.342	1.801	2.263	1.947	2.162	2.048
2.022	1.880	2.064	1.907	2.048	2.078	2.092	2.338
1.983	2.159	2.368	2.280	2.278	2.249	2.188	2.376
2.143	2.040	2.054	2.154	2.082	2.014	2.016	2.046
2.187	1.981	2.203	2.481	1.927	1.771	2.017	2.367
2.112	1.993	1.721	1.851	2.220	2.046	1.949	2.142
1.838	2.426	2.288	2.464	2.131	0.000	2.255	2.123
2.133	2.384	1.796	2.194	2.354	2.008	2.094	2.122
2.381	2.201	2.289	2.098	1.870	1.994	2.066	2.149
2.081	1.648	2.020	2.126	2.088	2.124	2.131	2.352
2.080	2.047	1.923	1.874	2.144	2.191	2.292	2.248
2.284	2.054	1.981	2.191	2.001	2.354	2.185	1.769
2.186	1.893	2.572	2.417	2.021	2.166	2.000	2.133
2.144	2.093	2.046	2.255	2.205	2.174	1.978	2.140
2.081	2.135	1.956	2.186	1.997	2.057	2.837	2.242
2.217	1.728	2.059	2.057	2.248	2.323	2.302	2.265
2.114	2.252	2.115	2.124	2.162	0.000	2.135	2.328
2.119	2.357	2.080	2.526	2.127	1.908	2.016	1.958
1.802	2.064	1.792	2.013	2.253	1.527	2.144	2.102
1.829	2.052	2.134	1.892	2.095	2.355	2.268	2.347
1.986	2.273	2.097	2.340	2.082			

Figure 3: δ value n tuple ($n = 315$) of α or corpus point BAUM/tree

Zweig/Ast bough/branch	2.970	Garten garden	3.166	Blüte blossom	3.339
Blatt leaf	3.508	Grün green	3.664	Frühling spring	3.736
Duft fragrance	3.833	Leise low/faint	3.891	Vogel bird	3.910
Laub leaves	3.962	Blume flower	3.981	Gras grass	4.025
Herbst autumn	4.053	Frühe early	4.065	Traum dream	4.068
Wiese/Aue lea/meadow	4.102	Wunder miracle	4.214	Lenz spring (poet.)	4.226

Figure 4: BAUM/tree

Grab/Gruft grave/tomb	2.945	kalt cold	3.192	Stunde hour	3.494
Tod death	3.636	fahl/welk dim/faded	3.669	finster dark/sad	3.980
bleich pale/dim	4.141	hohl hollow	4.436	Schein shine	4.533
schwarz black	4.595	Abgrund gulf/depth	4.642	grau grey	4.718
Angst terror/fright	5.078	heilig holy	5.115	blaß colourless	5.487
Schweben hover	5.543	weiß white	5.977	gelb yellow	5.992

Figure 5: FRIEDHOF (graveyard)

Zweig/Ast bough/branch	2.268	Blume flower	2.535	Blatt leaf	2.618
grün green	2.622	Garten garden	2.663	Frühling spring	2.673
Vogel bird	2.761	Duft fragrance	2.780	Wiese/Aue lea/meadow	2.865
Rose rose	2.885	leise low/faint	3.001	singen sing	3.121
Lenz spring (poet.)	3.127	Gold gold	3.158	Welle wave	3.221
Wunder miracle	3.224	Jubel joy	3.247	Luft air	3.302

Figure 6: BAUM ∧ BLÜTE (tree ∧ blossom)

Baum tree	2.489	Blüte blossom	2.489	Frühling spring	3.060
Garten garden	3.437	Duft fragrance	3.550	Zweig/Ast bough/branch	3.678
Gras grass	3.754	Lenz spring (poet.)	3.767	Traum dream	3.897
Laub leaves	3.937	Rose rose	3.941	Eiche oak tree	3.964
Blatt leaf	3.971	Vogel bird	3.972	Feld field	3.991
Pracht splendour	3.994	zart tender	4.017	Nachtigall nightingale	4.049

Figure 7: BAUM ∨ BLÜTE (tree ∨ blossom)

References

[1] Zadeh, L.A.:

1965 'Fuzzy Sets', *Information and Control* **8**, 338–353.

1971 'Quantitative Fuzzy Semantics', *Information Science* **3**, 159–176.

1972 'A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges', *Journal of Cybernetics* **2**, 4–34.

[2] Zadeh, L.A. (1971), 160.

[3] ibid.

[4] Zadeh, L.A. (1971), 161.

[5] Rieger, B.:

1971 'Wort- und Motivkreise als Konstituenten lyrischer Umgebungsfelder. Eine quantitative Analyse semantisch bestimmter Textelemente', *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik* **4**, 23–41.

1972 'Warum mengenorientierte Textwissenschaft? Zur Begründung der Statistik als Methode', *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik* **8**, 11–28.

[6] Salton, G.:

1970 'Automatic Text Analysis', *Science* **168**, 335–343.

1974 'A Theory of Term Importance in Automatic Text Analysis' (together with Yang, C.S./Yu, C.T.) Technical Report TR 74-208, Dep. of Computer Science, Cornell University Ithaca, N.Y. 14850.

1975 'On the Role of Words and Phrases in the Automatic Content Analysis of Texts', Paper presented on the Intern. Conference on Computers and the Humanities 1975 (ICCH/2), Los Angeles, Univ. of Southern California (mimeogr.).

[7] Zadeh, L.A. (1965), 340–42.

[8] Rieger, B.:

1974 'Eine 'tolerante' Lexikonstruktur. Zur Abbildung natürlich-sprachlicher Bedeutung auf 'unscharfe' Mengen in Toleranzraumen', *LiLi, Zeitschrift für Literaturwissenschaft und Linguistik* **16**, 31–47.

1975 'On a Tolerance Topology Model of Natural Language Meaning', paper presented on the International Conference on Computers and the Humanities (ICCH/2), Los Angeles: University of Southern California (mimeogr.).

1976 ‘Theorie der unscharfen Mengen und empirische Textanalyse’, paper presented on the ‘Deutsche Germanistentag 1976’, Düsseldorf (BRD) in: Klein, W. (ed.): Methoden der Textanalyse, Heidelberg 1977, 84–99.

[9] It should be noted that paradigmatic relations vary considerably from one language to another; the word–word translation of meaning-points from German into English in Fig. 4 to Fig. 7 might be rather inadequate and cannot be meant to depict comparable English paradigmatic relations, but has been given for illustration reasons only. For comparable results, one would have to analyze a similar corpus of English natural language texts.

[10] I would like to thank Dr. H.M. Dannhauer who is doing the programming for the CDC-Cyber 175 at the Technical University of Aachen Computing Center.