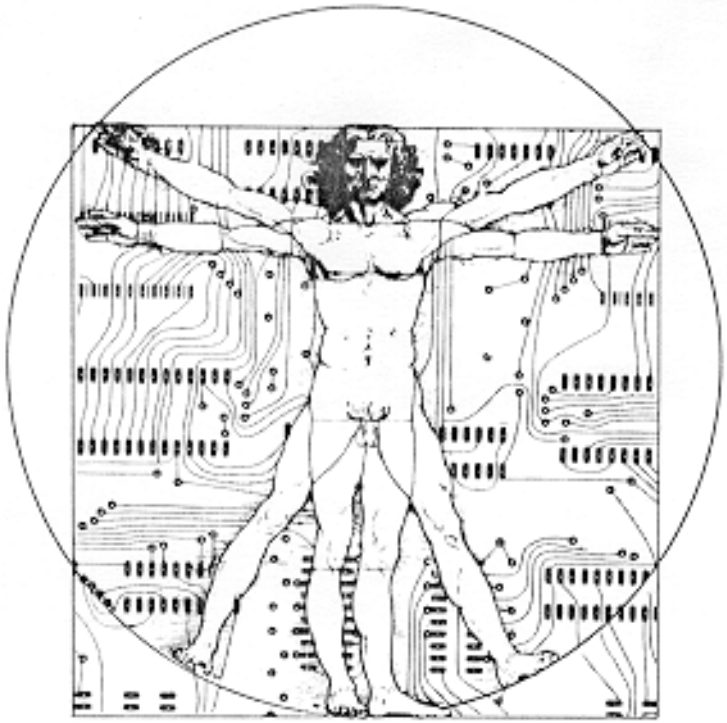

Actes du Congrès international informatique et sciences humaines

Liège, 18-19-20-21 novembre 1981

XX

Laboratoire d'Analyse Statistique
des Langues Anciennes

Clusters in Semantic Space

Analysing natural language texts to model word meaning as a procedural representation*

Burghard B. Rieger

This paper will report on one of the objectives of a project in Computational Semantics currently being worked on by the MESY-group at the Technical University of Aachen. Among others, the project is concerned with the development of automatic frame construction from natural language discourse. Both, in linguistic semantics and in artificial intelligence, most of the language data processed is still obtained introspectively or by way of questioning test-persons. Based mainly on the investigator's or the system-designer's own linguistic competence and / or world knowledge, the relevant data for the modelling of semantic and / or conceptual structures has a more or less ad hoc character and often lacks intersubjective control. Therefore, we have been and are trying to circumvent this deficiency by developing an algorithmic procedure that takes natural language texts on a certain subject domain as *input* and produces as *output* a distance-like datastructure of linguistically labeled space points whose positions represent (connotative) meanings.

1. As outlined and discussed elsewhere (RIEGER 1979, 1980, 1981a) statistical means can be applied for the empirical analysis of discourse and the formal representation of vague word meanings in natural language texts. These procedures allow for the systematic modelling of a fragment of the lexical structure constituted by the vocabulary employed in the texts as part of the concomitantly conveyed world knowledge concerned. The modified correlation coefficients used will map the lexical items onto fuzzy subsets of the vocabulary according to the numerically specified requlrities these items have been used with in the discourse analysed. The resulting system of sets of fuzzy subsets is a relational datastructure which may be interpreted topologically as a hyperspace with a natural metric. Its linguistically labeled elements represent *meaning points,* and their mutual distances represent *meaning differences.*

The position of a meaning point may be described by its semantic environment. This is determined by those other points in the semantic hyperspace which — within a given diameter — are most adjacent to the first one. Figure 1 shows the topological environments of two meaning points ALPEN (alps) and INDUSTRIE (industry) as computed from a corpus of German newspaper texts comprising some 8 ooo tokens of 360 types in 175 texts from the 1964 editions of the daily DIE WELT.

Having seen that the environments do in fact assemble meaning points of a certain semantic affinity, three questions arise which will be discussed in the following:

- first, are there regions of point density in the semantic space, forming clouds and

*Published in: Delatte, L. (Ed.): Actes du Congrès International Informatique et Science Humaines, Liège (Laboratoire d'Analyse Statistique des Langues Anciennes) 1983, pp. 805-814

Topological environment $E(z_s, r)$; $s =$ ALPEN DIE WELT
Range $r = 9.8900$

URLAUB	3.9977	FAHR EN ER T	4.5023
AUTO	5.6149	GAST	7.0474
BAHN	7.2863	RETT EN UNG	7.3960
SPORT LER	8.0673	BERG	8.1442
SKI	8.2592	TOUR	8.8162
LUFT	8.8245	PISTE	9.2650
TOD	9.2724	GEFAHR LICH	9.3234
LIFT	9.4442	SICHER HEIT N UNG	9.3234
LAUT EN	9.4442	ALLE	9.6397
SCHNEE EIEN	9.7497	ABFAHR EN T	9.8102
GLÜCK LICH	9.8102		

Topological environment $E(z_s, r)$; $s =$ INDUSTRIE
Range $r = 18.000$

ELEKTRO NISCH	2.106	LEIT EN R UNG	2.369
BERUF LICH	2.507	SCHUL E R	3.229
SCHREIB EN	3.328	COMPUTER	3.667
FÄHIG KEIT	3.959	SYSTEM ATIK	4.040
ERFAHR EN UNG	4.294	KENN EN TNIS	5.285
DIPLOM	5.504	TECHN IK ISCH	5.882
UNTERRICHT EN	7.041	ORGANISATION	8.355
WUNSCH EN	8.280	ZONE	8.546
BITTE N	9.429	STELLE	11.708
UNTERNEHME R N	14.430	STADT	16.330
GEBIET	17.389	VERBAND	17.569

Figure 1

clusters which might indicate a semantic (paradigmatic and / or syntagmatic) structuredness;

- second, can these be detected and described automatically by methods of cluster analysis, and, if so;
- third, how do these clusters look like and of what meaning points are they composed of?

2. According to BOCK (1974) cluster analysis is a collection of methods for automatic classification. Automatic classification refers to a number of mathematical-statistical procedures which aim to detect inherent similarities among sets or elements of sets in order to decide on the formation of greater partitions among them. These are considered classifications of elements or objects of a certain kind which can be achieved by objective, or rather intersubjectively controllable operations which do not necessitate an analysing subject's knowledge or ability.

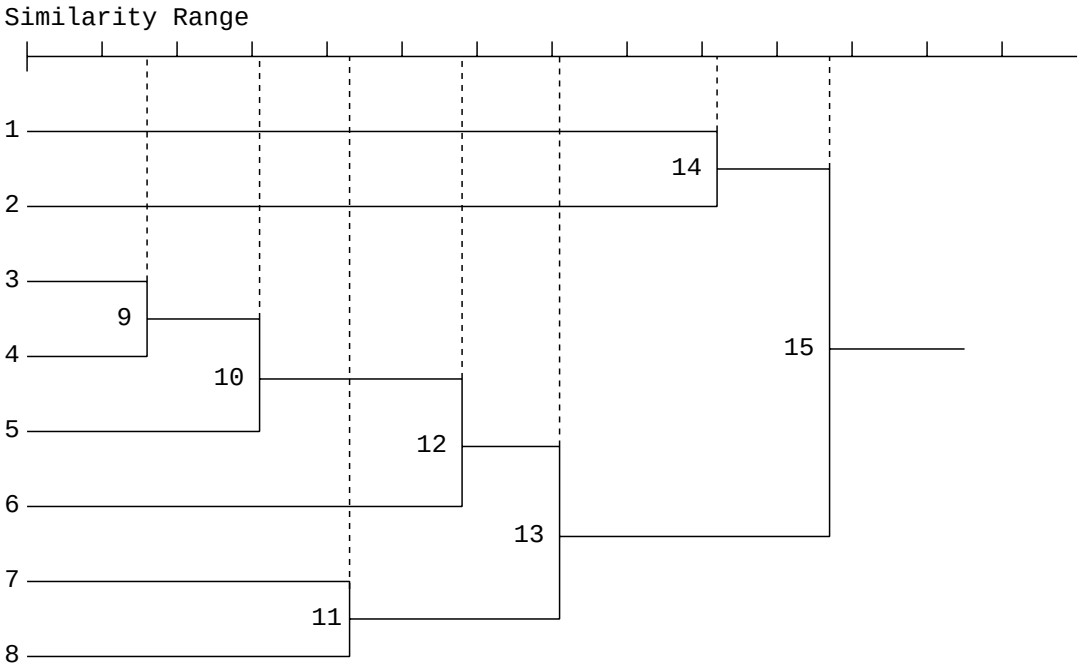

Figure 2

Methods of automatic classification require distance-like data to be processed, i.e. objects or entities whose similarity may be measured by a numerical expression the values of which satisfy the condition of a metric. This is the case with what we have analysed and defined to be the fuzzy subsets of the vocabulary or, equivalently, meaning points in the semantic hyperspace. Submitted to the cluster analysing algorithms provided by, for instance, the IMSL-program-library, the meaning points of one topological environment would be grouped together according to the least differences between them in classes of successive agglomeration.

The program produces a so-called *dendrogram* of this agglomerative process which has the form of a tree (Fig. 2), generated from its leaves up to the root. Here we have eight elements at the bottom, numbered from 1 to 8. On the first level these are evaluated according to their mutual similarities, resulting in elements 3 and 4 being found least different, to form the new class number 9. Then the whole procedure is repeated on the second level, resulting in the merge of class number 9 with element number 5, forming the new class number 10, and so forth. Now, at each level, the cluster analysing algorithm produces a new partition of the original set of elements whose similarities will numerically be specified on the similarity range.

Within each cycle of the cluster algorithm a decision is made on the highest similarity or least difference between the elements or classes concerned. This decision depends on the *cluster-criterion* employed which hence will partly determine the results.

We have tested only the three most commonly applied criteria, namely *single, complete* and *average linkage* to compare their differing performance on our data. To give an idea in what respect these three differ, Figure 3 (taken with kind permission from WICKMANN 1980) shows an example of three element-classes A_1, A_2 and A_3 that merge differently

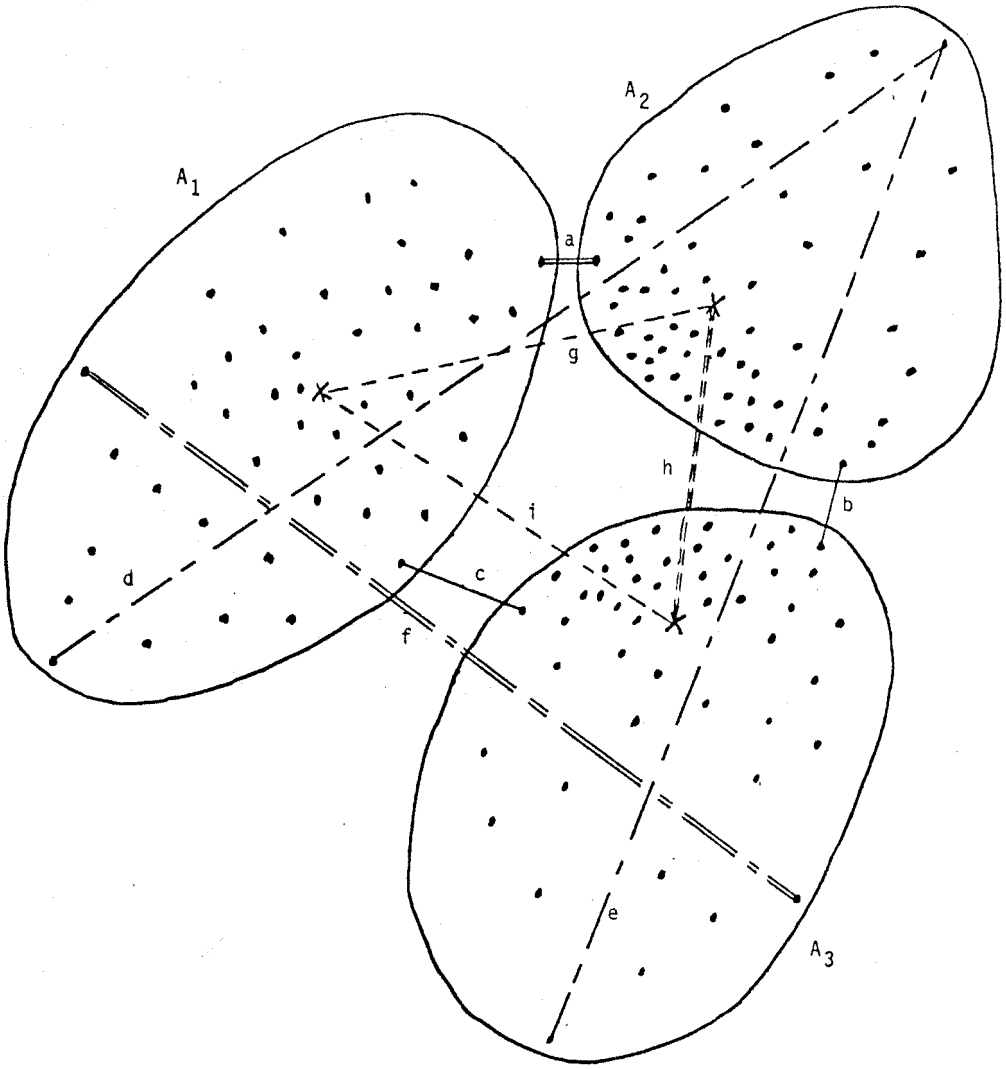

Figure 3

when *single, complete* or *average linkage* is applied.

For *single linkage* the smallest distances between all pairs of elements of mutually exclusive class-memberships are ascertained, in our case the distances *a, b* and *c*. The minimum distance value found to be *a*, this indicates that on the next level class A_1 will merge with A_2.

For *complete linkage* not the smallest but the greatest distances of all pairs of elements of mutually exclusive class-memberships are ascertained, namely *d, e* and *f*, the minimum of which, *f*, again determines the merge of the class A_1 with A_3 on the next level.

For *average linkage* not only singular pairs of elements are considered but all elements of a class contribute to a mean value which can be interpreted as the centre-of-gravity of the data cloud constituting a class. From the distances between these centres, namely *g, h,* and *i*, the minimum distance *h* is selected to determine the merge of classes A_2 with A_3 in this example.

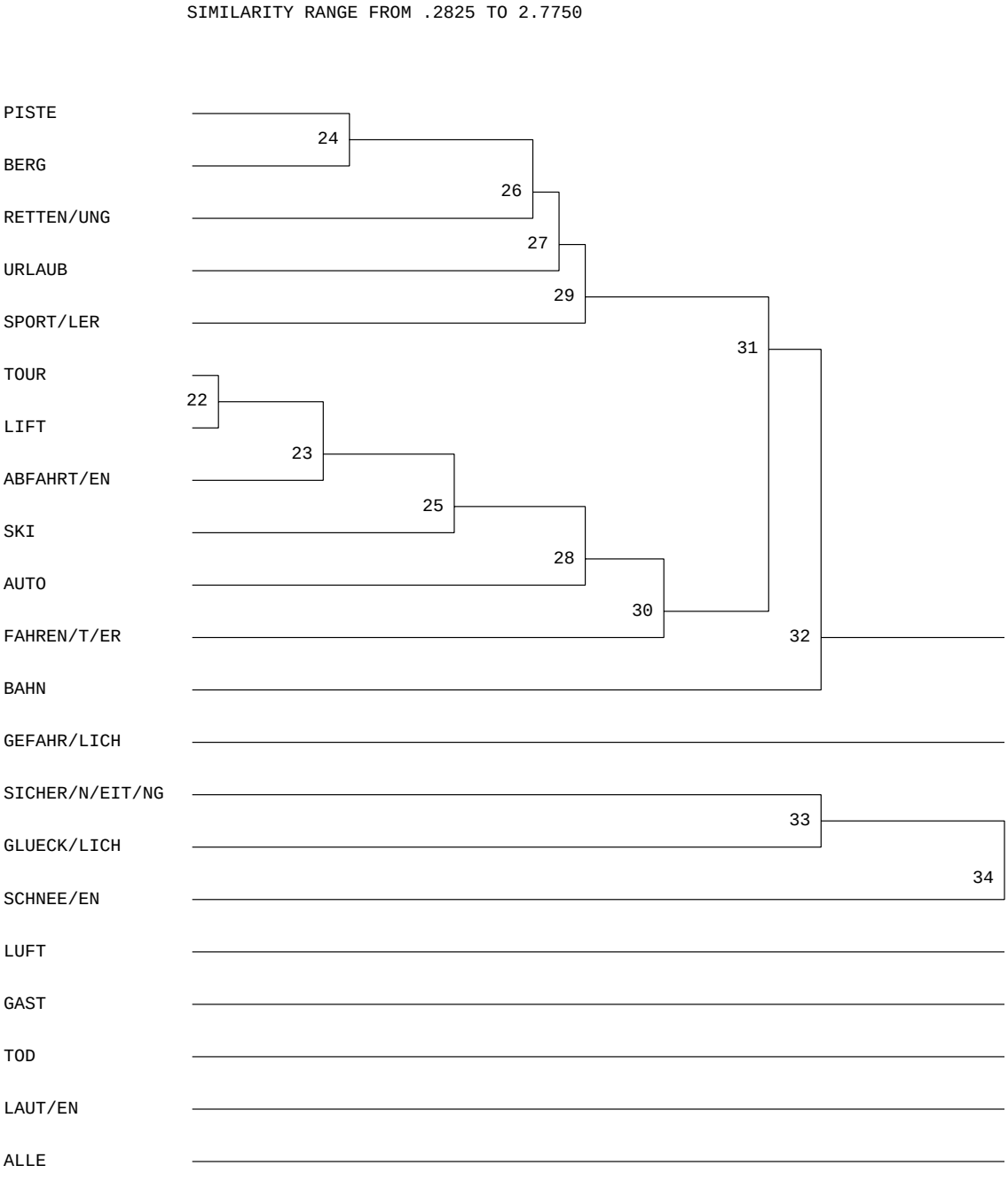

Figure 4. ALPEN Cluster Single Linkage

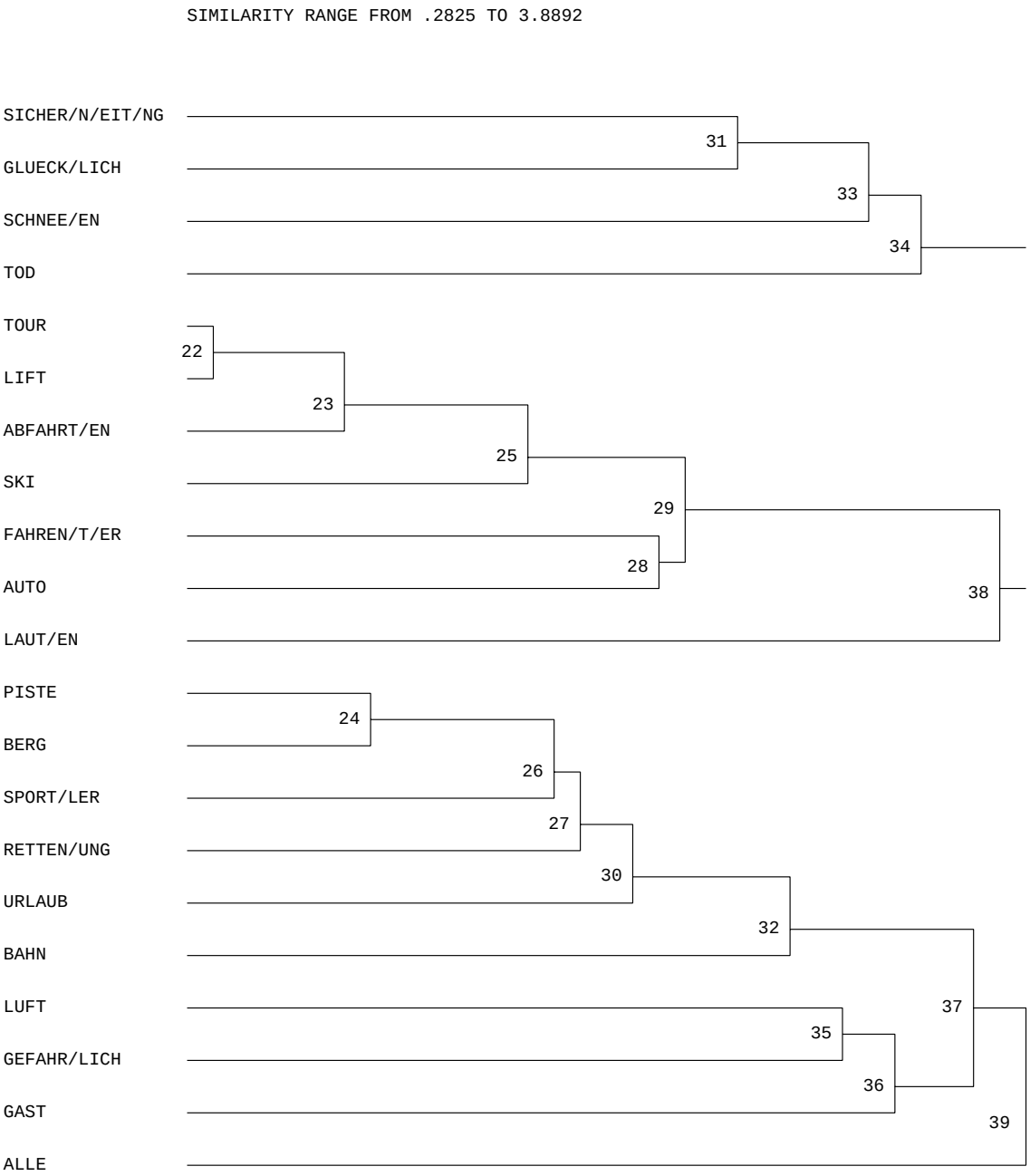

Figure 5. ALPEN Cluster Complete Linkage

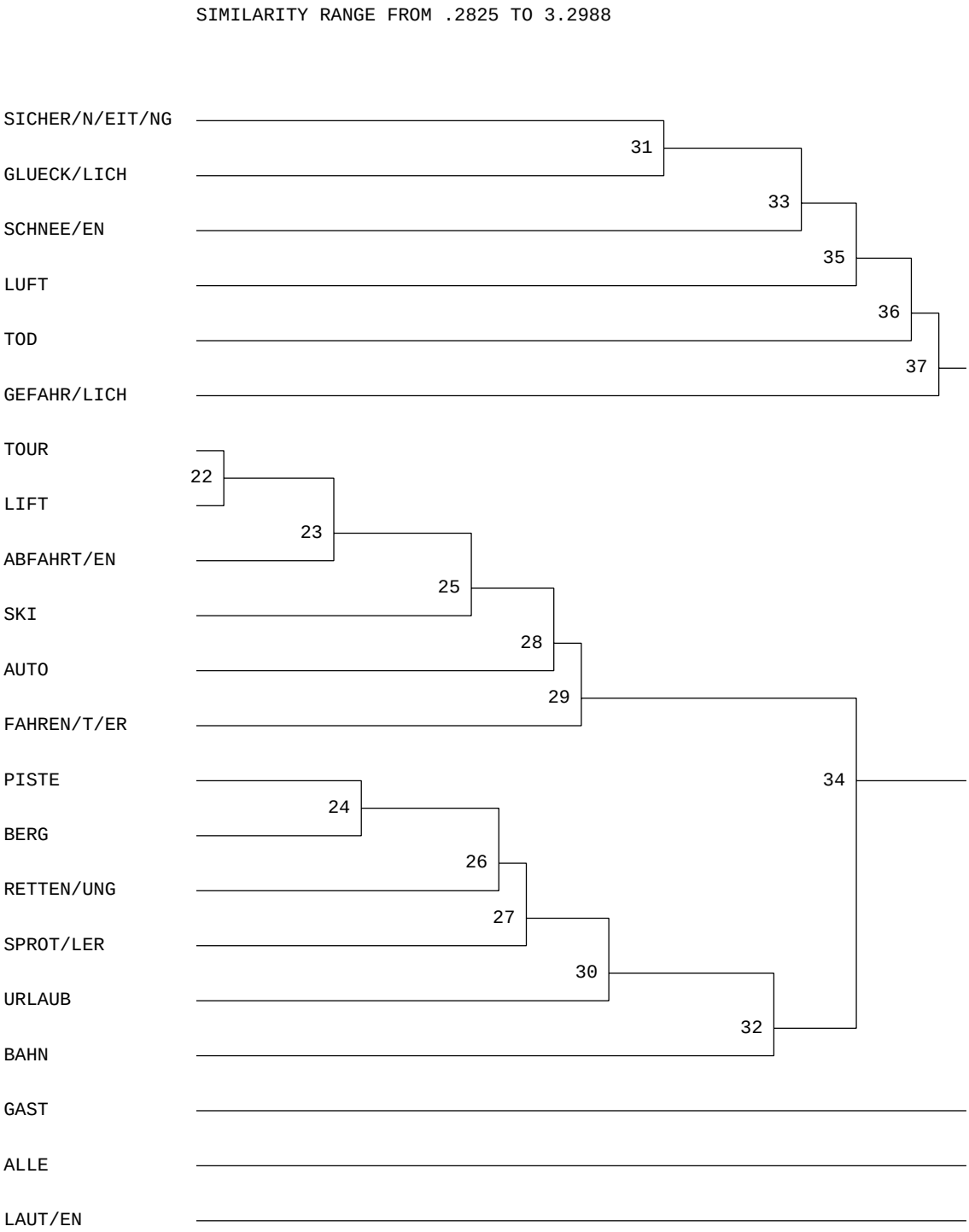

Figure 6. ALPEN Cluster Average Linkage

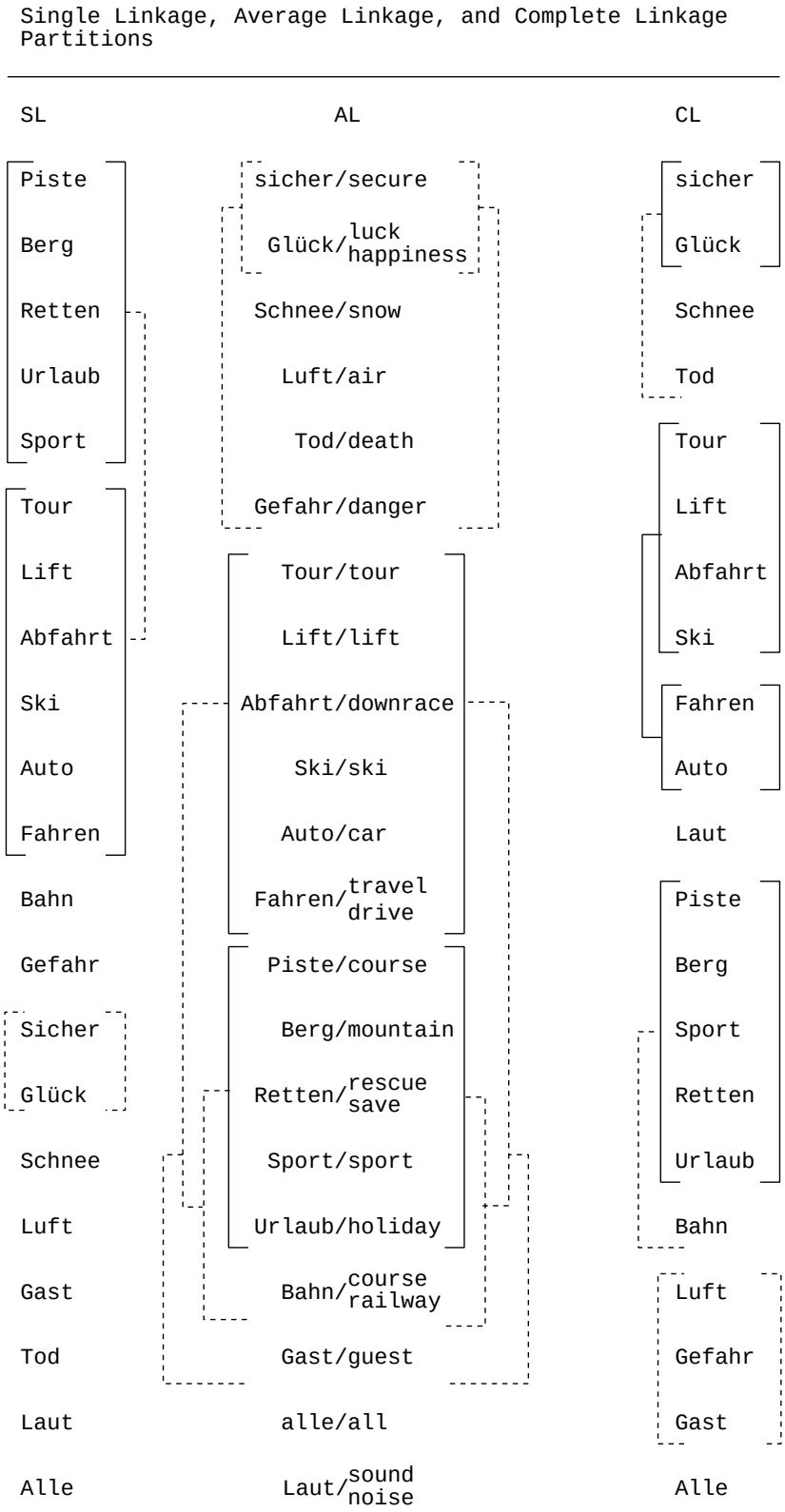
Single Linkage, Average Linkage, and Complete Linkage Partitions
SL
AL
CL
Piste
Berg
Retten
Urlaub
Sport
Tour
Lift
Abfahrt
Ski
Auto
Fahren
Bahn
Gefahr
Sicher
Glück
Schnee
Luft
Gast
Tod
Laut
Alle
sicher/secure
Glück/luck happiness
Schnee/snow
Luft/air
Tod/death
Gefahr/danger
Tour/tour
Lift/lift
Abfahrt/downrace
Ski/ski
Auto/car
Fahren/travel drive
Piste/course
Berg/mountain
Retten/rescue save
Sport/sport
Urlaub/holiday
Bahn/course railway
Gast/guest
alle/all
Laut/sound noise
sicher
Glück
Schnee
Tod
Tour
Lift
Abfahrt
Ski
Fahren
Auto
Laut
Piste
Berg
Sport
Retten
Urlaub
Bahn
Luft
Gefahr
Gast
Alle

Figure 7

3. The figures 4, 5 and 6 show fragments of dendrograms, analysing an environmental set of meaning points centred around the one with the label ALPEN (the alps) under all three cluster-criteria. As these examples are hard to follow up, a synopsis of the three partitioning hierarchies are given in Figure 7. Apparently, all three linkage procedures produce certain core groupings of meaning-points, i.e. partitions merging on a relatively low level of similarity; which seem to be uneffected by whatever cluster criterion is employed for their descripitve representation. This applies to the two clusters of what we may call the alpine winter holiday field (TOUR, LIFT, ABFAHRT, SKI, AUTO, FAHREN) and (PISTE, BERG, RETTEN, SPORTLER, URLAUB, BAHN) as well as to the two others concerning different aspects of it (SICHER, GLÜCK) and (SCHNEE, TOD, LUFT, GEFAHR) As may be seen from a comparison of all three dendrograms, however, there are a few meaning-points (ALLE, BAHN), GAST, (GEFAHR), LAUT, whose merging behaviour exhibits that their distances relative to the others must be rather high. Depending on the cluster criteria, this causes changing memberships of these meaning points in partitions of relatively high merging level as shown in the dendrograms. The different performances of the three cluster criteria employed are obvious, and the advantage of AL over SL and CL is apparent, so we accepted average linkage (AL) to be preferred for our purpose.

Applying cluster analysis on the semantic space data on a large scale proved that the distribution of meaning points in the hyperspace are in fact structured in semantically relevant clouds and clusters of higher density. This presupposition being ascertained objectively, the semantic hyperspace may serve as data base for a dynamic meaning representation by generating connotative dependency structures (CDS) as introduced in RIEGER (1981 b).

Acknowledgement

The research project reported on here is supported by the Northrhine-Westfalia Ministry of Science and Research under grant FA 8600. I would like to thank my colleagues Dr. D. Wickmann for his problem solving assistance and H.U. Block, M.A. who did part of the programming of the CDC Cyper 175 at the Aachen Technical University Computing Centre.

References

BOCK, H.H. (1974): Automatische Klassifikation. Göttingen

RIEGER, B. (1979): "Linguistic Semantics and the Problem of Vagueness." in: Advances in Computer-aided Literary and Linguistic Research, ed. by D.E. Ager, F.E. Knowles, J. Smith, Birmingham, pp. 271–288

RIEGER, B. (1980): Fuzzy Word Meaning Analysis and Representation in Linguistic Semantics. An empirical approach to the reconstruction of lexical meanings". COLING 80 — Proceedings of the 8th Intern. Conference on Computational Linguistics, Tokyo, pp. 76–84

RIEGER, B. (1981a): "Feasible Fuzzy Semantics. On some Problems of How to Handle Word Meaning Empirically" in: Words, Worlds, and Contexts. New Approaches in Word Semantics, ed. by H.J. Eikmeyer, H. Rieser, Berlin / New York, pp. 193–209

RIEGER, B. (1981b): ”Connotative Dependency Structures in Semantic Space” in: Empirical Semantics. A Collection of New Approaches, vol. II, ed. by B. Rieger, Bochum (in print)

WICKMANN, D. (1980): ”An automatic analysis of semantic relationship between words in texts: factor or cluster analysis — what fits best?”, ALLC-Bulletin 2 (1980) pp. 152–165