An Associative-Categorical Model of Word Meaning

Robert M. Haralick

Department of Electrical Engineering, University of Kansas, Lawrence, Kan. 66044, U.S.A.

and

Knut Ripken

Mathematisches Institut, Technische Universität München, Munich, Federal Republic of Germany

Recommended by E. Sandewall

ABSTRACT

A new dual categorical-associative model for the representation of word meaning is proposed. In it, concepts are described by the values they have on a set of given variables (categories). A statistical relatedness measure (concomitant variation) is computed for these values on the basis of the specified word universe. An association measure between the words is defined, and the generalization of word clusters is introduced. A comparison with associative and categorical models is made and the application of the dual model to verbal analogy problems is described. Possible applications in Artificial Intelligence and Natural Language Processing are discussed.

1. Introduction

The representation of word meaning is still an obstacle which researchers in Natural Language Processing must remove in order to create information processing systems which understand and generate natural language. Word meaning has been represented by either associative or categorical models. In this paper we propose a dual associative-categorical model which combines the good qualities of both.

An important purpose of this work is to find some well-defined basis on which it would be possible to implement the mechanisms of assimilation, generalization and reasoning by analogy. For it is our contention that, already on the word level where there is some basic semantic knowledge

Artificial Intelligence 6 (1975), 75–99

involved, it will be necessary to apply these mechanisms to avoid any "combinatorial explosion" in dealing with large knowledge bases.

We shall first briefly explain both former models of word meaning in order to contrast them with our model. Then, we shall present the dual model, demonstrate its capability to structure the word universe and validate it by solving verbal analogy problems.

2. Former Models to Represent Word Meaning

Existing models to represent word meaning can be roughly characterized as either associative or categorization models. Since we assume that the reader is familiar with the well known models mentioned in this section, we are not going to explain the models in detail but rather discuss only the relevant aspects.

2.1. Associative models

As "associative models" we denote the following models which:

(1) Take the statement "The meaning of a word is its use in the language" (Wittgenstein [17, §43]) literally and define words by the actual context in which they occur, e.g., the models of Raphael [10], McCalla and Sampson [8], and Siklossy [14].

(2) Define a word—viewing it as a "stimulus word"—by a list of associated words, which, for example, could be gathered in psychological experiments. Reitman [11] developed such a model based on the ideas of Hebb [5].

(3) Use a method which is applied in dictionaries where tokens of other words constitute the definition of a word. This is a method which might resemble the way humans usually define concepts. It is incorporated in the "semantic memory" of Quillian [9]. It is our contention that the model of the first type will hardly be applicable to a large universe of discourse since its semantic structure is too simple so that an abounding mass of details will make it impossible to effectively generate relevant associations.

We want to consider more closely Quillian's model, as a type-3 model and the probably best-known example of an associative model. The meaning of a word ("type node") is defined by a respective set of other words ("token nodes") and their interrelations. The set of token nodes with all its links, which express six different relations, constitutes the "plane" of the type node. Since, however, each token node has an associative link to the type node of the same name, the plane of an original patriarchial type node considered is also linked to other planes. The full content of a word concept in such a memory then is—"as distinguished from its plane or 'immediate definition' " —"all the type and token nodes one can get to by starting at the initial type node, or patriarch, and moving first within its immediate definition plane to all the token nodes found there, then on 'through' to the type nodes named

by each of these nodes, then on to all the token nodes in each of their immediate definition planes, and so on until every token and type node that can be reached by this process has been traced through at least once." (See Quillian [9, p. 237].)

If we view Quillian's semantic memory as a graph U where the nodes are the planes of type nodes and the edges are the associative links from a token node within a plane to its type node which is represented by another plane. Then we can consider the following cases which are evident (we do not give the proofs):

(a) If the graph U is strongly connected, then each word's full concept, as defined above, is the entire U.

(b) If U is only connected, but not strongly, then there is at least one type node whose full concept is only a subgraph of U.

(c) If U is not connected, i.e., if there are isolated subgraphs, then there exists no word which has U as its full concept.

(d) If the set of token nodes is limited and fixed though U may expand, then the full concept of each word is never larger than a subgraph formed by the plane of the word, its edges and the subgraph of those planes which define the type nodes of all the token nodes existing.

Since in Quillian's model the set of token nodes bears no restriction, case (d) does not apply to it. Probably cases (a), (b) and (c) could apply to the graph U. Here the question arises concerning what concept the "memory builder" has of a word's full context. What is the recipe for constructing the planes of concepts? This question becomes difficult to answer if the universe of words gets larger and larger. Since the interesting aspect of Quillian's memory is the existence of associative paths between different planes and not only the retrieval of stored information within one plane, the links between the planes and thus the structure of U is essential.

How far then is the definition of a word's concept independent of the definition of other concepts? If the words were defined independently while the set of token nodes is arbitrary, then there would be no guaranty given for the right, i.e. meaningful, associations or any associations at all; then the paths from "cry" and from "comfort" needed not necessarily reach the same node "sad", which is what they do in Quillian [9, p. 250]. If the words were, however, defined dependently, then the associations would be preprogrammed and less interesting.

These questions set the problems of the semantic memory in perspective: a clear-cut recipe to build the definitions of concepts is missing; as a consequence, the derived associations do not have a reliable basis.

The same argument obviously applied to the type-2 associative model. Unfortunately, Reitman [11] has not explicitly given results of the experiments with his model.

2.2. Categorization models

The case (d) of the above cases (see Section 2.1) leads us to the categorization model where the token nodes degenerate to categories which form the attribute lists of word concepts while the different relations between the token nodes are abolished. Thus a word concept is described by a set of primitives.

Since Katz and Fodor [7] introduced the concept of semantic markers, such a model of categorization has been used by several researchers. Schank and Tesler [12] wrote a conceptual parser, and Winograd [16] used a simple hierarchical system of categories to describe the meaning of the objects in the robot's toy world. Tuggle et al. [15] most recently used categorization in their "test program" for verbal analogy problems: the description list of a word consists of categories as well as pointers to other words in case of the relations "part/whole", "contained/container", "opposite/similar meaning". The success of these programs points to categorization as a useful method. However, as it was the case with the associative model programs, the programs working with categorization only deal with extremely small universes so far.

In a categorization model, the process of describing a word concept is well-defined: words are defined independently and only with reference to the categories. (We separate certain binary relations between two concepts A and B, e.g., "has as a part/is a part of", "contains/is contained by", from the categorization model since they require a direct reference from A to B and vice versa.) Therefore, the categorization model seems to be more practical in a larger universe of discourse than an associative model; an expansion of the universe is easily possible without any effect on the part previously defined.

In such a categorization model associations between any two word concepts, however, are only found on the trivial basis of a comparison of the respective attribute lists. A matching of categories here corresponds to the existence of an intersection node in Quillian's associative model. If the category system is ordered in some fashion (e.g., hierarchy), then word concepts can at most be associated with the help of attribute comparison in a way which reflects this order (e.g., lattice structure).

2.3. The dual associative-categorical model

So far, we have seen that the associative model lacks a prescription for the construction of definitions which have a great influence on the possible associations. On the other hand, in a categorization model the independent construction of definitions is well-defined, but "interesting" associations between the single word concepts are not possible.

In order to overcome these disadvantages of the former two models, it is

therefore desirable to find a model which combines the advantages of a categorization model, i.e., the extendability and the clear definition procedure, with a method to associate words in a meaningful way. In this paper, we suggest such a model. We describe the words of a given universe with their values on a fixed category set, e.g., a "knife" could be described by the categories "man-made", "out of metal", "sharp", etc. Then we calculate a statistical measure of relatedness (concomitant variation) for all the possible pairs of values. With the help of this measure, an association measure for each two words of the universe can finally be calculated.

It is interesting that Deese [1] went the reverse way to discover—from the psychologists' point of view—"the categories of association, if they exist". He carried out association experiments with stimulus words, which were chosen so that there was some degree of related associative meaning. The relative common frequencies of responses to each word of this set of words were then factored by the centroid method. Deese found out that the resultant factors could be described as, e.g., "having to do with animate creation", "having to do with inanimate creation", etc.

Since, in our model, we actually compute an associative net, starting out with words, which are separately defined by categories, our model bridges the gap between categorization and associative model.

3. From Categorization to Association by Statistical Computation

In this section, we present the mathematical basis of our approach, discuss a simple instructive application and report about a validity test of the model with verbal analogy problems.

3.1. The mathematical basis

Let $U = \{u_1, \ldots, u_N\}$ be the specified universe of words. Let $X = \{x_1, \ldots, x_K\}$ be the set of categories of variables which can describe the words in U. Each variable in X is, therefore, a function $x_k: U \rightarrow L_k$, where the range L_k of x_k is the set of possible values a word can be described by when measured by variable or category x_k.

Our problem is to structure the universe U by defining a binary relation on it. We divide this task in two parts: first, we try to find a measure of statistical relatedness, concomitant variation, between any value in one range set with any value in another range set; then, we define a measure of statistical association between any pair of words in U on the basis of the concomitant variation between the values by which each word is described.[1]

[1] Hunt et al. [6] accidentally use similar definitions in their work on concept learning.

There are numerous measures of statistical association which could be used for measuring concomitant variation and the articles by Goodman and Kruskal [2, 3] give an excellent discussion and summary of the often used ones. Here, we introduce a measure of concomitant variation which is deducible from a few important properties.

3.1.1. Concomitant variation

We distinguish between concomitant variation as a symmetric and asymmetric concept as follows: Asymmetric concomitant variation we call conditional concomitance and it measures the control or dominance one set of events has on another. If R and S denote two arbitrary sets of events, then $C(R \mid S)$ denotes the conditional concomitance of R given S. Symmetric concomitant variation we call just concomitant variation and it measures the interaction between the two sets of events. We denote the concomitant variation between R and S by $C(R, S)$. We naturally expect that the concomitant variation between R and S should be the average conditional concomitance of R given S plus the conditional concomitance of S given R:

$$C(R, S) = \tfrac{1}{2}[C(R \mid S) + C(S \mid R)].$$

We determine a measure of conditional concomitance having the following four properties: (1) it is a linear combination of the four probabilities $\mathbf{P}(R \cap S)$, $\mathbf{P}(R^c \cap S)$, $\mathbf{P}(R \cap S^c)$ and $\mathbf{P}(R^c \cap S^c)$; (2) the conditional concomitance of R given S equals the conditional concomitance of S^c given R^c; (3) if the event $R^c \cap S$ has zero probability, then the conditional concomitance of R given S equals 1; (4) if the conditional probability of R given S plus the conditional probability of S^c given R^c is 1, then event S does not control event R and the conditional concomitance of R given S equals 0. These four properties imply that

$$C(R \mid S) = \tfrac{1}{2}[\mathbf{P}(R \mid S) - \mathbf{P}(R^c \mid S) + \mathbf{P}(S^c \mid R) - \mathbf{P}(S \mid R^c)].$$

We briefly sketch the proof of this. Without loss of generality, let

$$C(R \mid S) = \alpha(R, S)\,\mathbf{P}(R \cap S) + \beta(R, S)\,\mathbf{P}(R^c \cap S^c) + \gamma(R, S)\,\mathbf{P}(R \cap S^c).$$

Then, $C(R \mid S) = C(S^c \mid R^c)$ implies

$$\mathbf{P}(R \cap S)[\alpha(R, S) - \beta(R^c, S^c)] + \mathbf{P}(R^c \cap S^c)[\beta(R, S) - \alpha(S^c, R^c)] + \mathbf{P}(R^c \cap S)[\gamma(R, S) - (S^c, R^c)] = 0. \quad (1)$$

This combined with $C(R \mid S) = 1$ when $\mathbf{P}(R^c \cap S) = 0$ implies

$$\alpha(R, S)\,\mathbf{P}(S) + \beta(R, S)\,\mathbf{P}(R^c) = 1. \quad (2)$$

$C(R \mid S) = 0$ when $\mathbf{P}(R \mid S) + \mathbf{P}(S^c \mid R^c) = 1$ implies

$$\gamma(R, S) = -\left[\alpha(R, S)\frac{\mathbf{P}(S)}{\mathbf{P}(R^c)} + \beta(R, S)\frac{\mathbf{P}(R^c)}{\mathbf{P}(S)}\right]. \quad (3)$$

Equation (3) can be used to determine $\gamma(R, S) - \gamma(S^c, R^c)$:

$$\gamma(R, S) - \gamma(S^c, R^c)$$

$$= \frac{\mathbf{P}(S)}{\mathbf{P}(R^c)}[\beta(S^c, R^c) - \alpha(R, S)] + \frac{\mathbf{P}(R^c)}{\mathbf{P}(S)}[\alpha(S^c, R^c) - \beta(R, S)]. \quad (4)$$

Using equation (1), equation (4) can be reduced to

$$\gamma(R, S) - \gamma(S^c, R^c) = \left[\frac{1}{\mathbf{P}(R^c)} - \frac{1}{\mathbf{P}(S)}\right][\mathbf{P}(S)\beta(S^c, R^c) + \mathbf{P}(R^c)\beta(R, S) - 1]. \quad (5)$$

Using equation (1),

$$\beta(S^c, R^c) - \alpha(R, S) = \frac{1}{\mathbf{P}(S)}[\beta(S^c, R^c)\,\mathbf{P}(S) + \beta(R, S)\,\mathbf{P}(R^c) - 1] \quad (6)$$

and

$$\alpha(S^c, R^c) - \beta(R, S) = \frac{1}{\mathbf{P}(R^c)}[1 - \beta(R, S)\,\mathbf{P}(R^c) - \beta(S^c, R^c)\,\mathbf{P}(S)]. \quad (7)$$

Substituting (5), (6) and (7) into (1) there results

$$\beta(R, S)\,\mathbf{P}(R^c) + \beta(S^c, R^c)\,\mathbf{P}(S) = 1. \quad (8)$$

Since $1 = \alpha(R, S)\,\mathbf{P}(S) + \beta(R, S)\,\mathbf{P}(R^c)$, we obtain

$$\alpha(R, S) = \beta(S^c, R^c). \quad (9)$$

But $\beta(R, S)\,\mathbf{P}(R^c) + \beta(S^c, R^c)\,\mathbf{P}(S) = 1$ is an identity in R and S. This implies that $\beta(R, S) = \frac{1}{2}\mathbf{P}(R^c)$ and by (9) we must have $\alpha(R, S) = \frac{1}{2}\mathbf{P}(S)$.

Substitution in (3) then yields

$$\gamma(R, S) = -\frac{1}{2}\frac{\mathbf{P}(S) + \mathbf{P}(R^c)}{\mathbf{P}(S)\,\mathbf{P}(R^c)}.$$

Substitution for α, β and γ in (1), then gives the formula for $C(R \mid S)$:

$$C(R \mid S) = \tfrac{1}{2}[\mathbf{P}(R \mid S) - \mathbf{P}(R^c \mid S) + \mathbf{P}(S^c \mid R^c) - \mathbf{P}(S \mid R^c)].$$

It is easily verified that concomitant variation has the following properties:

(1) $C(R, S) = C(S, R)$,
(2) $C(R, S^c) = -C(S, R)$,
(3) $C(R, S) = C(R^c, S^c)$,
(4) $C(R, R) = 1$,
(5) $C(R, S) = 0$ if and only if R and S are independent events.

3.1.2. A measure of association

In our model, the measure of concomitant variation, as derived in Section 3.1.1, is applied to the values of the variables which describe the words. Inserting $x_l(u_1)$ and $x_m(u_j)$ for R and S in the equation for concomitant variation yields a measure of relatedness between the values the ith word

takes on for the lth category and the value the jth word takes on for the mth category.

Having related the values to each other in this way, we can define a measure of association between words. Such a measure between words u_i and u_j should take into account the relatedness of each of the values, which describe u_i, to each of the values which describe u_j. Thus, we define as a measure of association between any two units $u_i, u_j \in U$ the association $\mathrm{A}(u_i, u_j)$:

$$\mathrm{A}(u_i, u_j) = \sum_{l=1}^{K} \sum_{m=1}^{K} C(x_l(u_i), x_m(u_j)),$$

where C is the concomitant variation and K the number of variables. Since $-1 \leqslant \mathrm{C}(R, S) \leqslant +1$, A can take on positive and negative values.

3.1.3. Structuring the universe

We structure the universe U of words by defining a binary relation on it. In working a clustering problem, Haralick and Haralick [4] defined such a binary relation by

$$R_1 = \{(u_i, u_j) \in U \times U \mid \mathrm{A}(u_i, u_j) \geqslant \theta\}.$$

This definition has the property that it relates together those words having only highest associations. Unfortunately, some words may have relatively small associations with all words and would therefore never appear related to anything through R. This property can be modified by using relative association ranks instead of raw associations. Let $r(u_i, u_j)$ be the number of words whose associations with word u_i is greater than $\mathrm{A}(u_i, u_j)$. Formally,

$$r(u_i, u_j) = \#\{u \in U \mid \mathrm{A}(u_i, u_j) < \mathrm{A}(u_i, u)\}.$$

We can then define the binary relation R_2 by

$$R_2 = \{(u_i, u_j) \in U \times U \mid r(u_i, u_j) \leqslant p\}.$$

Note that the rank measure is an asymmetric measure $r(u_i, u_j) \neq r(u_i, u_i)$. The asymmetry arises because the associations u_i has with the rest of the words is not necessarily the same as the associations u_j has with the rest of the words.

Clusters of associated units can be determined from R_2 as those units having a relatively high number of interconnections through R_2.

In Sections 3.2 and 3.3 we examine these structures which the defined relations yield in instructive examples involving practical problems, in our case the association of meaningful words.

3.2. A concrete application

The task is to choose a set of words as the universe and to describe these words with terms which are subsumed under certain categories. One of our

TABLE I. Categories 1–17

1 WHOLE	2 PART1	3 PART2	4 NUMBER	5 AGE
1	1	1	1	1
2 human	2 human	2 vegetable	2 single	2 baby
3 animal	3 animal	3 manmade	3 pair	3 child
4 vegetable			4 group	4 adult
5 manmade				5 old
6 natural				

6 SEX	7 COMPLEXITY	8 SIZE	9 MATERIAL
1	1 element/compound	1	1
2 male	2 simple object	2 $\leqslant$ 0.01 m	2 metal
3 female	3 simple machine	3 $\leqslant$ 0.1	3 glass
	4 sophisticated mach.	4 $\leqslant$ 0.3	4 plastic
	5 organism	5 $\leqslant$ 0.6	5 wood
		6 $\leqslant$ 1	6 paper
		7 $\leqslant$ 2	7 textile
		8 $\leqslant$ 3	8 leather
		9 $>$ 3	9 organic product
			10 mineral

10 COLOR	11 SHAPE	12 WEIGHT	13 HARDNESS
1	1	1	1
2 red	2 point	2 $\leqslant$ 10 g	2 powder
3 orange	3 line	3 $\leqslant$ 100	3 marshmallow
4 yellow	4 triangle	4 $\leqslant$ 500	4 sponge
5 green	5 rectangle	5 $\leqslant$ 2.5 kg	5 flesh
6 blue	6 circle	6 $\leqslant$ 10	6 basketball
7 brown	7 polygon	7 $\leqslant$ 50	7 wood/glass
8 white	8 parallelepiped	8 $>$ 50	8 metal
9 grey	9 ellipsoid/sphere		
10 black	10 cylinder		
11 silvery			
12 golden			
13 opaque			

14 AGGREGATE STATE	15 BRIGHTNESS	16 TEMPERATURE	17 SOUND
1	1	1	1
2 gaseous	2 sunlight	2 freezing	2 noise
3 fluid	3 electr. light	3 cold	3 harmonic sound
4 solid	4 candle	4 room temperature	
		5 warm	
		6 hot	

TABLE II. Categories 18–39

18	19
TASTE	TOUCH (primarily)
1	1
2 sweet	2 rough
3 bitter	3 smooth
4 salty	4 sharp
5 sour	
6 spicy	

ACTIVE FUNCTIONS PERFORMED BY HUMAN AND INSTRUMENT

20	21	22	23	24
1	1	1	1	1
2 cut	2 store	2 protect	2 cover	2 smoke
3 hold/support	3 screw	3 carve	3 mash/grind	3 see
4 paint	4 cook	4 heat	4 cool	4 wash
5 communicate	5 dry	5 do sport	5 sew	5 pick up
6 drink	6 write	6 entertain	6 hear	6 show movies/
7 dwell	7 eat (context)	7 sit	7 clean	pictures
8 lie	8 make music		8 open	7 hit
9 make light				

ACTIVE FUNCTION OF OBJECT ONLY

25
1
2 contain solid thing
3 contain fluid thing

PASSIVE FUNCTION (n: normally being . . .; m: must be . . .; p: for the purpose of being . . .)

26	27	28	29	30
1	1	1	1	1
2 cut nm	2 drunk p	2 held/supported m	2 contained m	2 colled n
3 heard p	3 written on n			3 sewn n
				4 smoked p

31	32	33	34
1	1	1	1
2 heated n	2 cooked n	2 eaten p	2 protected m
	3 read p	3 put on (dress) p	3 dried n
	4 washed/cleaned n	4 electr.-operated	

ENVIRONMENT

35	36	37	38	39
1	1	1	1	1

categories could, for instance, be "color", which as a variable could take on as values all the different colors like "red", "orange", "yellow", etc., but only one value for each word. If we cannot decide the color for a word or if it does not make sense to assign a color to a word—abstract words are an example—the variable "color" has to have a value which means "not applicable or not decidable".

In order to have a nice world for the first investigation, we chose a relatively coherent subset of concrete objects of the real world, a set of 283 objects which occur in an ordinary household, e.g., "alarm-clock", "bedstead", "beer", "bracelet", "boy", etc. The category system which we developed is just detailed enough to provide a different description for each of the units. It is not claimed that this system is complete or optimal in any other respect. Table I and Table II show the category system with a total of 39 variables and 194 values, an average of 5 values per variable.

Variables 20 through 34 show collections of properties which are mutually exclusive in regard to their application to words in our universe. The activities under variables 20 through 24 are interpreted in the following way: if the word u is a human or an animal, it means that the human or the animal executes the activity, e.g., $x_{21}(u) = 6$ (write) means "a human can write" or "a human is properly described by the activity of writing"; if, however, u is an inanimate object, it can either mean that the object executes the activity—if this is possible—or that the object aids a human (or animal) in performing the activity; thus, if u is an inanimate object, then $x_{21}(u) = 6$ (write) means: "the object is instrumental to a human (or animal) performing the activity", e.g., x_{21}(pencil) = 6.

The passive functions 25 through 34 can be interpreted in different ways so that the proper interpretation is suitably given by a code letter. E.g., value 2 for variable 30 means "normally being cooled"; this would apply to things which we keep in the refrigerator; and value 2 for variable 27 means "for the purpose of being drunk"; value 2 for variable 34 means "must be protected"; e.g., meat must be protected in some way or it rots.

These few explanations show already that we describe some standard world in which unusual events like "drinking HCl" are not contained. Of course, we could equally well describe a crazy world; our results, however, would not give so much insight in the usefulness of our methods. Since the category system is fairly detailed a lot of decisions have to be made when one describes the units. In order to get some standardization, we used a pictorial dictionary [18] which displayed standard exemplars of the things we described. Furthermore, our description assumes that things are where they belong in the household: the proper things in the kitchen, in the living area, etc. Table III shows the descriptions of some words. The values of the variables are listed from the left to the right, variables 1 through 39.

It remains to stress that certain relationships of the real world are not covered by the category system. Since each unit is described independently, links from one unit to another—either indirectly by certain relations between categories or directly by pointers—are not contained. For example, the fact that a "snout" is a part of an animal is contained, but not that it is a part of a dog. And the data base does not reflect the "opposite", "final", "causal" or "idiomatic" relationships as in the examples "oven"/"freezer", "bottle-opener"/"bottle", "painter"/"picture", and "needle"/"thread". These relations could be either taken care of with the help of categories which specify a context or environment explicitly like "fine arts" in the case of "painter"/"picture", or they could be contained in a special dictionary: e.g., the entry "hot"–opposite–"cold". In the first case they would be automatically reflected in the associative structure which we want to compute; in the second case, however, the description of the considered units would have to be checked for such special relations which are contained in the special dictionary, and then an appropriate weight would have to be applied to the corresponding values in the computation process of the association measure. Up to now the only relationships contained are those given by some activity like "sewing"—e.g., "needle"/"thread"—or by the environment categories.

TABLE III. Examples of word descriptions

1[a]	baby-boy	21122	255	9	1	8	6541	42131	11111	11111	11112	2222
5	binocular	51121	124	2	10	10	5841	41111	11131	11111	11111	2112
11	butter	51111	111	9	4	1	1541	41431	11111	11122	11222	2111
14	cat	31121	154	9	1	8	5541	42111	11111	11111	11112	2112
18	cup	51121	123	10	1	9	3741	41116	12113	11111	14132	2111
21	ear (human)	12131	153	9	3	6	1541	41135	11611	11211	11112	2222
26	flower	41121	125	9	1	7	3441	41131	16111	21221	11122	2111
34	juice	61221	111	9	1	1	1131	11211	11111	12222	11122	2111

[a] The number in the first column denotes the unit according to Table IV.

The 283 descriptions of units, arranged as in Table III, were the input to a FORTRAN program which calculated the concomitant variations of all the values of the variables. These calculated measures were stored on magnetic tape for their use in the computation of the association measures.

3.3. The results of the application

First we give the reader a short impression of the concomitant variation between the values of the variables, before we deal with the relations between the words as the units.

3.3.1. The concomitant variations

Fig. 1 displays some of the graphs of the concomitant variations which are greater than 0.6. That is, an edge between two nodes, which are labeled with the code of variable and value, say "i, j" and "k, 1", means that $C(x_i(u) = j, x_k(u) = 1) < 0.6$. For the meaning of the values see Tables I and II. A few larger clusters are: that of values which describe fluids (no size, no shape, no

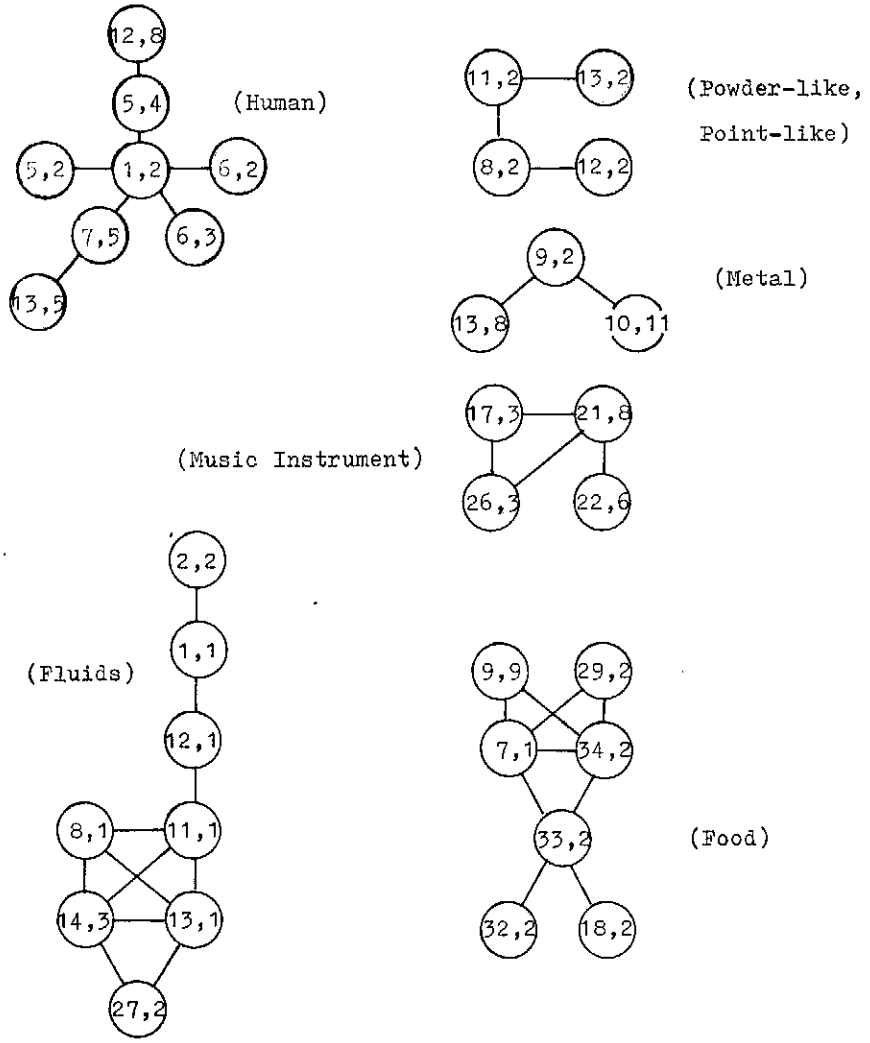


FIG. 1. The concomitant variations which are greater than 0.6. The node labels denote variable and value, e.g., "i, j" and "k, 1"; an edge between these nodes means that $C(t_{ij}, t_{k1}) > 0.6$. In parentheses are given names of unit classes to which the clustered values apply.

hardness, fluid, drunk p) with a connection to parts whose weight is undecided; that of values which describe food, then that of values describing objects that make music, and finally the cluster of values which are proper for humans.

These clusters show that the concomitant variation, indeed, relates highly associated values to each other. We have not investigated the effectiveness of our category system in the light of these value clusters. Such an investigation might be useful in order to compact the category system, especially when one deals with a larger universe which probably requires more values to describe the units.

3.3.2. The association of words

As well as the concomitant variations, the associations are most suitably presented as partial graphs, namely of the relations R_1 and R_2 which have been defined in Section 3.1.3. These graphs show that the defined measure of association is meaningful in that it associates units which are highly related to each other in the real "universe" of our household either in regard to their physical appearance, their functions, or the environment in which they occur; for these are the factors which are reflected in the category system of Section 3.2. For demonstration purposes, the results in this section refer only to a subset S of the universe U, consisting of those 60 words, which are given in Table IV.

TABLE IV

1	baby-boy	21	ear (human)	41	milk
2	baby-girl	22	eye (animal)	42	mother
3	bed	23	eye (human)	43	mouth
4	beer	24	father	44	needle
5	binocular	25	feather bed	45	newspaper
6	bird	26	flower	46	pastry cutter
7	book	27	football	47	pickle
8	boy	28	fork	48	picture (wall)
9	bread	29	frisbee	49	pitcher
10	bread knife	30	girl	50	poster
11	butter	31	glasses	51	punch-bowl
12	candle	32	green salad	52	radio
13	carrots	33	hand	53	record player
14	cat	34	juice	54	socks
15	cigar	35	knife	55	suit
16	cigarette	36	lamp (ceiling)	56	tape recorder
17	creamjug	37	leg (animal)	57	tennis racket
18	cup	38	leg (human)	58	thread
19	dog	39	linen sheet	59	tomato
20	ear (animal)	40	mattress	60	tv set

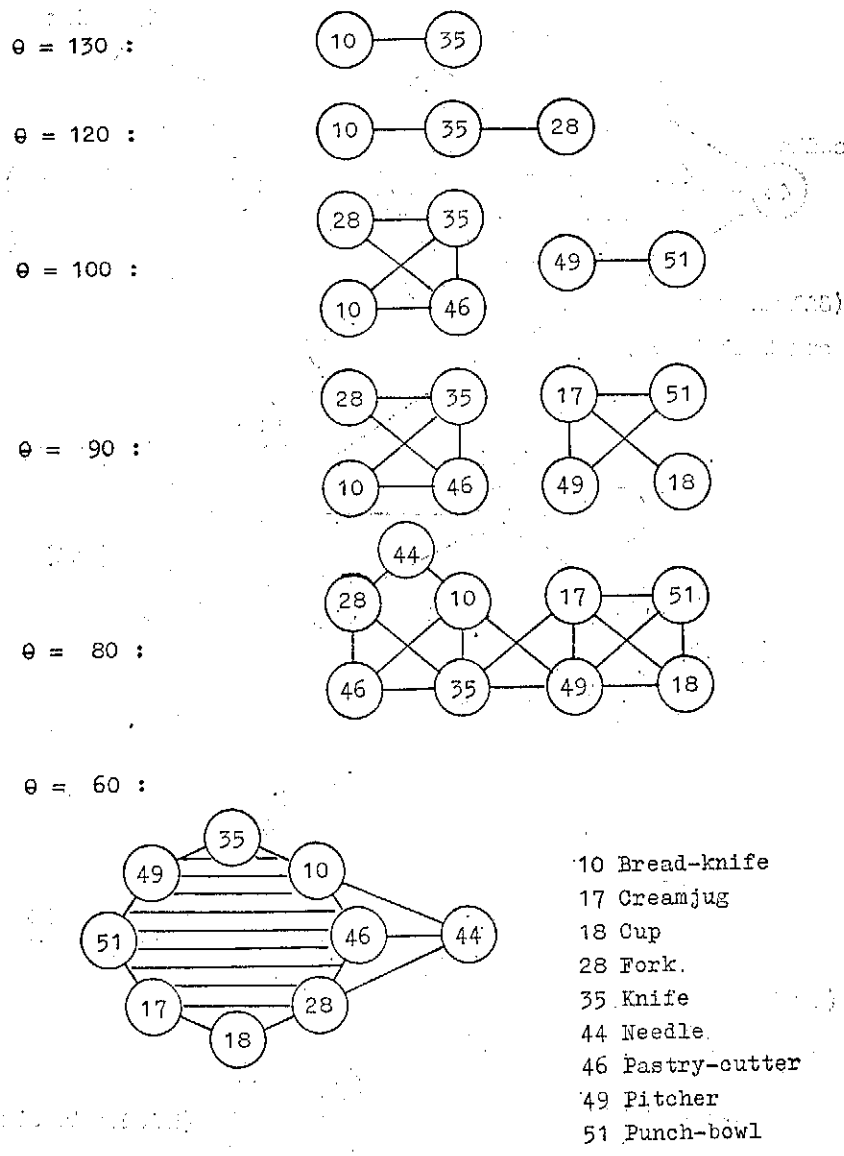


FIG. 2. Growth of the graph, showing the relations between kitchen instruments, with decreasing θ. The shaded partial graph is complete.

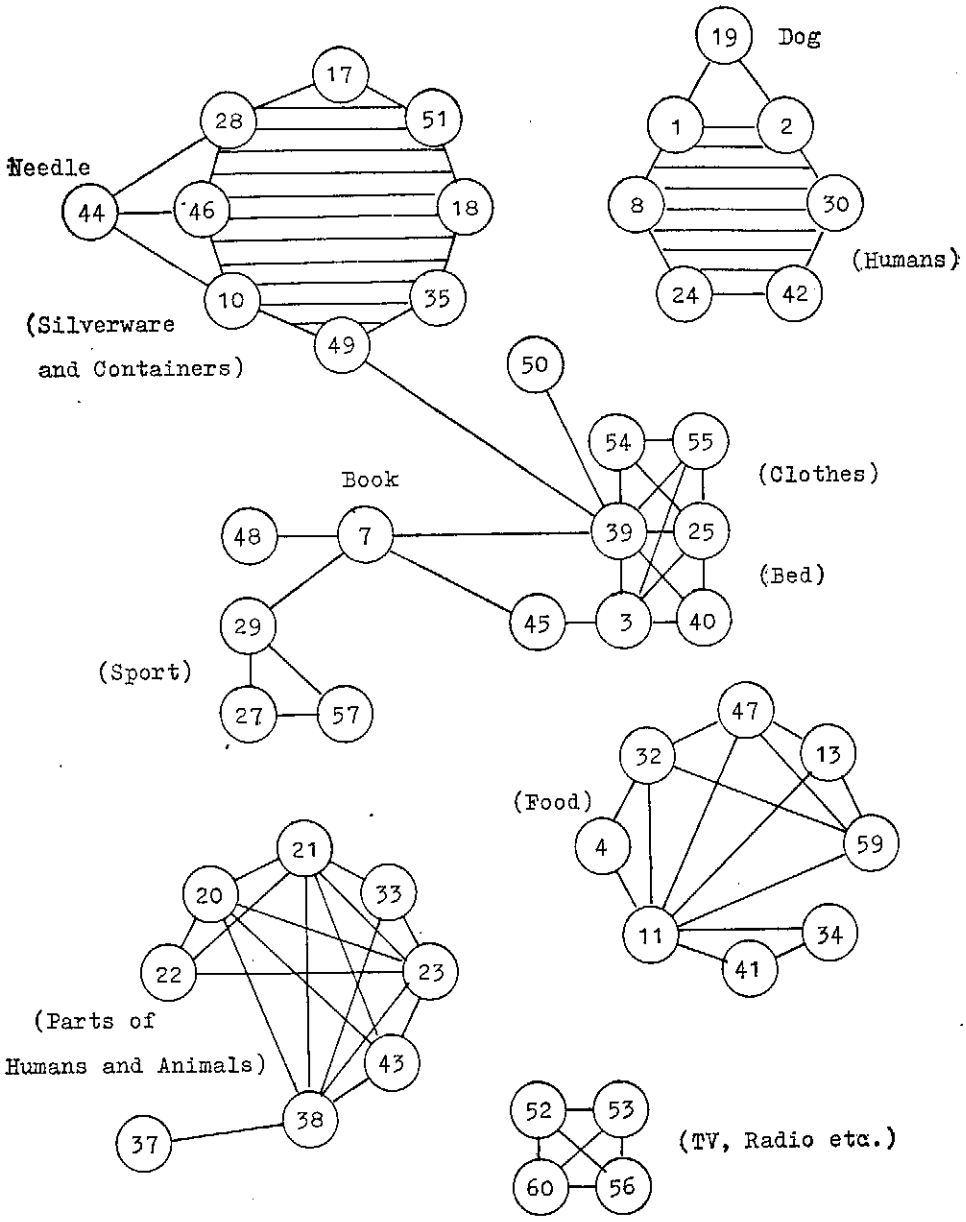


FIG. 3. The graphs of R_1 for $\theta = 60$. The numbers refer to the units of Table IV. Characterization of the clusters are given in parentheses. Complete graphs are shaded.

In regard to these words, for example, the highest association measures for the word "baby-boy" are with the words:

Baby-girl	133.02
Boy	120.01
Father	119.81
Girl	117.47
Mother	117.26
Dog	62.12
Cat	54.03

This rank order is quite plausible. For a threshold value $\theta = 115$, relation R_1 contains a complete subgraph whose nodes are the humans of the sample set S. In regard to R_2, one finds that the complete subgraph of the humans is a complete cluster for a threshold p of 5.

Fig. 2 gives an illustration of the dynamic growth of graphs when more and more association links are added; it shows the associations between kitchen instruments like containers and silverware. For $\theta = 60$ the result is a large complete graph with three edges to "needle".

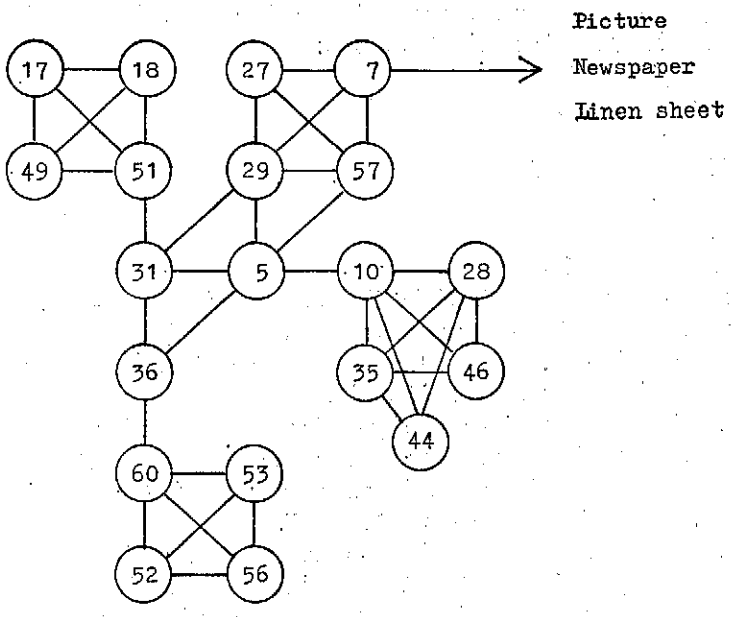
Fig. 3 shows R_1 on $S \times S$ with $\theta = 60$. The numbers in the nodes refer to the units as given in the list of S. The main clusters are given a label for the units in the cluster. We notice several complete and nearly complete subgraphs of R_1 which contain highly related units of our universe.

However, we do not only want to find clusters, but also to relate them to each other in a lively way. An interesting connection in this respect is that one in Fig. 3 which goes from the "sport"-cluster over "book" to the "clothes"-cluster. If we drew the graphs for R_1 with $\theta < 60$, of course, more such edges would become visible.

In Fig. 3 we notice that several words, e.g., "lamp", "glasses", "candle", etc., have not yet appeared. This means that their association measures are all less than 60. According to the discussion in Section 3.1.3, we expect them to show up in graphs of relation R_2. Fig. 4 shows such a graph with the ranks less or equal $p = 3$. This graph was started with the unit "lamp". Again we see clusters familiar from previous graphs, but we also see their connection with "lamp".

3.3.3. Generalizations

The occurrence of complete subgraphs suggests that a concept of generalization be introduced. Words of a complete subgraph are highly interrelated so that it might be possible to represent them by a single concept which is substituted for the whole cluster and represents this cluster in its relations with the rest of the universe. For the cluster of containers "creamjug", "cup", "pitcher" and "punchbowl", for example, a generalization would be something like "container for nutritive fluids".



5 Binocular
7 Book
10 Bread-knife
17 Creamjug
18 Cup
27 Football
28 Fork
29 Frisbee
31 Glasses
35 Knife
36 Lamp
44 Needle
46 Pastry-cutter
49 Pitcher
51 Punch-bowl
52 Radio
53 Recordplayer
56 Tape Recorder
60 TV Set

FIG. 4. Subgraph of R_2 for $p = 3$, containing the unit "lamp."

A good generalization of words of a complete graph should keep this graph complete when it itself is added. Furthermore, it should represent the cluster in its relations with the "outside world" in the same way the words of the cluster do it. Formally, we thus define a generalization in the following way:

A word u_c, not necessarily already in U, is a *generalization* of a complete cluster of rank p of words $u_1, \ldots, u_{p+1} \in U$ if and only if the following conditions hold:

(1) $r(u_i, u_j) \leqslant p + 1, \quad i, j = 1, 2, \ldots, p + 1, c.$
(2) $r(u_c, u_k) \approx r(u_i, u_k), \quad u_k \in U, \quad i = 1, \ldots, p + 1.$

We leave the second condition in the vague form since the universe will not be that neatly structured as to make an equal sign feasible. Besides, one might also want to apply the concept of a generalization when a graph is only nearly complete, but not quite. Therefore, we take the definition as a guide rather than an obligation. For we feel that the real world and the relations of meaning in it cannot totally be forced into a rigid scheme.

Generalizations can either be obtained by defining them explicitly, i.e., by constructing their representation by hand, or by intersecting the description lists of the words in the cluster. This intersection can be defined variable wise for the generalization u_c of units in a subset U_s.

$$x_i(u_c) = 1 \text{ if } \left.\begin{array}{l} x_i(u_l) \neq x_i(u_m) \text{ for some } u_l, u_m \in U_s \\ x_i(u) \text{ for some } u \in U_s, \text{ otherwise} \end{array}\right\}.$$

We tested units which we expected to be generalizations and which were obtained by both methods, with the good result that they satisfied the above definition, i.e., that they were indeed acceptable generalizations.

3.4. A validity test: Verbal analogy problems

As a validity test for our measure of association, we have written a program which solves verbal analogy problems of the type "u_1 is to u_2 as u_3 is to which of the following: u_4, u_5, u_6 and u_7?". We chose analogy problems as a possible test since they are well-defined tasks and since there exist earlier programs which deal with these problems but apply different solution mechanisms according to their different representation of meaning. A discussion and comparison of the programs will follow the presentation of our solution mechanism and its results.

3.4.1. The solution mechanism

The problem is to choose a u_4 out of a set of several choices so that the equation

$$u_1 : u_2 = u_3 : u_4 \tag{10}$$

is satisfied, where the operator ":" is some measure of association between two units, which expresses their relationships. Since, in our case, the u_i are words with a more or less complex meaning, we cannot expect an exact solution of the equation. We might think of using

$$|\mathrm{A}(u_1, u_2) - \mathrm{A}(u_3, u_x)| = \text{minimum} \qquad (x \in \{4, 5, 6, 7\})$$

as a possible way to find the solution u_x, but we would realize that highly related units need not have the same absolute association score—as we remarked in Section 3.1.3—and that this approach therefore fails.

In terms of our normalized asymmetric measure of association, the ranks, we can interpret $u_1 : u_2$ as $r(u_1, u_2)$, i.e., we measure the directed "distance" from u_1 to u_2 on the association scale of u_1. But since eq. (10) has four other possible forms given by the four possible ways in which we can write eq. (10), keeping the operator ":", we suggest that the solution of an analogy problem may be that u_x ($x \in \{4, 5, 6, 7\}$) for which the following score $S(x)$ is minimal:

$$\begin{aligned} S(x) = \quad & |r(u_1, u_2) - r(u_3, u_x)| \cdot (r(u_1, u_2) + r(u_3, u_x)) \\ & + |r(u_2, u_1) - r(u_x, u_3)| \cdot (r(u_2, u_1) + r(u_x, u_3)) \\ & + |r(u_1, u_3) - r(u_2, u_x)| \cdot (r(u_1, u_3) + r(u_2, u_x)) \\ & + |r(u_3, u_1) - r(u_x, u_2)| \cdot (r(u_3, u_1) + r(u_x, u_2)). \end{aligned}$$

Here the sum of the ranks as a factor of the absolute difference takes care of our preference for a u_x, for which the rank difference in the low ranks is small, compared with the same rank difference between high ranks. We have the feeling that strong associations have more weight in the analogy than weak associations. Large ranks of a word often vary considerably in regard to highly related words. Since, in an analogy problem, u_1 is either more associated with u_2 or with u_3, and since the less association is likely to be represented by a rank in that range of the association scale where the ranks are not so meaningful, we reduce the expression for the score $S(x)$ to a score $S'(x)$, by summing only over those products which are determined by the two lowest of the ranks $r(u_1, u_2)$, $r(u_2, u_1)$, $r(u_1, u_3)$ and $r(u_3, u_1)$. The solution of the problem then is that u_x with the least score $S'(x)$.

3.4.2. The results

A FORTRAN program was written which uses the solution mechanism of 3.4.1 to solve analogy problems. The possible words were those 60 words which are listed in Section 3.3.2. The program can generate its own analogy problems with the help of a random number generator.

The solutions the program gives are always the most plausible solutions in the case of meaningful problems. The reader is invited to try the problems in Table V himself. We add the scores on the basis of which the program made its choice. A comparison of the programs results with those of human subjects was not performed.

3.4.3. Discussion

Unlike other analogy programs with specifically designed word stores, our program acts on a set of words which were not primarily described in regard to their use in analogy tests. Therefore, analogy in our program can only relate to the appearance, function, environment or association of an object which is subject of the test. As a consequence, in some examples the analogy

seems to be rather weak. With a more specifically designed universe we can expect more interesting analogy problems. Their solution might, however, need a slightly refined mechanism where several classes of association measures and ranks are computed and evaluated which relate to different classes of categories. Nevertheless, such an analogy is a valid test in an

TABLE V. Analogy problems and their solution (the solution is in italics)

u_1 :	u_2 =	u_3 :	?	Score $S'(u_x)$
baby-boy	baby-girl	father	*mother*	0
			newspaper	2355
			dog	98
			girl	16
frisbee	flower	tennis racket	lamp	5971
			carrots	75
			pastry-cutter	1613
			radio	2357
knife	pastry-cutter	pitcher	record-player	819
			radio	1200
			punch-bowl	11
			socks	397
tennis-racket	frisbee	lamp	*candle*	15
			binocular	146
			football	1680
			pickle	6392
tape recorder	binocular	tv set	baby-boy	2987
			glasses	117
			bird	1447
			football	131
tape recorder	radio	newspaper	bird	3219
			book	10
			bed	16
			mattress	187
tomato	carrots	beer	needle	3672
			mouth	4995
			mother	5535
			milk	13
cigar	cigarette	ear (animal)	*ear (human)*	95
			father	144
			girl	150
			green salad	1889
flower	green salad	needle	thread	348
			bread	2276
			bread knife	51
			butter	4029
hand	human leg	needle	*thread*	315
			mouth	2588
			juice	4220
			bed	621

attempt to examine the qualities of our association measure and its validity, i.e., whether it is able to structure the meaning space of a set of words. On the basis of the results, one can say that the used association measure passes the test. It is able to relate the words to each other in a natural way.

The results of the analogy test are quite surprising if one remembers the simple solution mechanism which is solely based on the $\frac{1}{2}m(m-1)$ association measures between the m words used in the test. All the information needed is compactly contained in these measures so that it is not necessary to look at the basic descriptions of the words in order to compare them. Reitman [11] with "ARGUS" and Tuggle et al. [15] work with lists attached to the used words in order to solve analogy problems of the same type. "ARGUS" uses an associative network (see Section 2.1) with parallel processing; Tuggle et al. use a category system to describe the words; the description lists of the words in the problem are then compared by the program in order to find the right choice. Our approach appears to be appealing in its simplicity and elegance compared with these earlier approaches. We concede, however, that important dimensions of meaning are not yet contained in our association measure, and thus neglected in the analogy program. This lead us to a general discussion of our approach, its possible applications and its current shortcomings, in the next section.

4. Discussion

The discussion concerns two questions, possible applications of our proposed model and its current shortcomings. In Section 2, we have already reported about former research of other people. So far we could only tell about the very first experiments with the method itself, not yet about applications—except the verbal analogy test. Therefore, the following section on possible applications is only a sketch and, at the same time, an enumeration of problems whose solution is a future task.

4.1. Possible applications

It should be clear that a method to represent the meaning of words and their relations is only relevant in regard to its application in a larger program which is to process natural language.

In Winograd's remarkable program [16], nouns, adjectives and verbs are defined by two functions "NMEANS" and "CMEANS" which contain the semantic markers for the noun or adjective, and for the possible subject and object of the verb, resp. The semantic definition of a "ball" is, for instance,

(NMEANS ((#MANIP #ROUND) ((#IS***BALL)))),

where the first argument of NMEANS contains the two most specific

semantic markers—the next higher markers in the hierarchy of semantic markers—the next higher markers in the hierarchy of semantic markers are implied—and the second argument explicitly says that the defined object "is" a "ball". An example for the definition of a verb is:

(CMEANS (((#PHYSOB) (#PHYSOB)) (#SUPPORT #1 #2) NIL)).

Again, "#PHYSOB" is a semantic marker, here for both the object (#2) and the subject (#1) of the verb "to support".

We propose to use the coded description lists of words instead of the semantic markers of Winograd or other similar definitions. This is still no change; when we, however, want to see whether a definition applies in a certain case, we propose to generally compute the measure of association in order to compare candidates. That means we would like not to literally compare the features of, e.g., a "ball" with the markers of the subject of "to support", but to compute the association measure of the full description lists in order to decide a match.

One problem emerges here: How can we determine a threshold for the association measure which it has to exceed for a successful match. This threshold clearly is a function of the composition of the universe and the category system. A threshold could be indirectly defined by a maximal allowed rank, but since this rank is a function of all the units, the adding of words to the universe would require the revision of the thresholds. This defect can be fixed by using a threshold between 0 and 1 and comparing it to the rank divided by the number of words in the universe.

We feel that the following advantages render worthwhile the efforts which the solution of this problem requires:

(1) The association measure provides us with a simple procedure to resolve semantic ambiguities; the highest association measure decides.

(2) We can get an idea of the context of a text by simply clustering the occurring words with the help of the association measure.

(3) We can use generalizations in the sense of Section 3.3.3 in the definition of verbs.

(4) We can—and this is the most important advantage—probably implement mechanisms of generalization, assimilation and reasoning by analogy, and thus build programs which are able to learn and to structure knowledge about the meaning of words in a reasonable way.

Instrumental to these processes is a useful interaction between the well-defined categorical descriptions and the association measure. Note that the categorical information would not only be used to generate the associative one but that it would also be involved in the processes of generalization and concept learning. A few comments should be made on the last suggestions.

The overall context of a discourse could be defined by clusters of occurring

words with the help of one of the relations R_1 or R_2 of 3.1.3. Eventual ambiguities which depend on the overall discourse could then be resolved by calculating the association measures with a cluster representative or with a cluster generalization. This idea, to define a context by the clusters of highly related words, resembles the approach of Sedelow and Sedelow [13] to stylistic analysis with the help of thesauri. The association mechanism could, of course, also be useful in Information Retrieval Systems, for example, to interactively find keywords.

Subject and object or other characteristics of a verb could be defined by generalizations as in 3.3.3. The subject of human activities could, for instance, be a generalization of "human being".

We can imagine a program which learns such generalizations from particular instances. Let a program encounter the commands "pick up a block" and "pick up a pyramid", it could use these particular occurrences of the verb "pick up" to find a generalization of "block" and "pyramid" to use this as a definition of the object of "pick up". On the basis of high association measures the program could try new candidates, and find new generalizations if it is given feedback signalizing correctness or falseness of a guess. Thus, the program could learn by guessing and experience, though to a modest extent. After new concepts have been added to the universe the concomitant variations of the values could be recalculated. Of course, this would change the associations; but humans also encounter a change in their view of things when they gain new knowledge about them. It might be interesting to see whether the dullness of the association mechanism could make the program come up with unexpected processes and answers. These ideas are still vaguely stated, however quite appealing.

4.2. Current shortcomings

Like nearly all Artificial Intelligence papers which deal with natural language processing, our paper is based on experiments with a subset—though a reasonably sized one for a first start—of concrete English words. A large universe will create some problems because of the amount of computation involved in the association measure. A larger universe will also require more specific categories in order to distinguish between highly similar objects. As already mentioned in 3.3.1, the efficiency of a category system should be investigated with a closer look at the concomitant variations of the values.

It is probably advisable to distinguish between the two cases that a variable cannot be decided or that it is not applicable, in the description of a word. Moreover, the category system does not contain important aspects of meaning like special "whole/part" relations, the "opposite/similar", "final" and "causal" relations. How to implement these in the category system or

how to combine the category system with a linked-list dictionary in a larger program, still has to be solved. One would also like to have a critical look at the influence of the composition of the universe on the statistical measures.

Finally, one could define a similarity measure between the words in order to find still more structure in U. This definition should be based on an asymmetric measure like the conditional concomitance.

Despite all these problems, we regard this approach as valuable since it bridges the gap between categorization and associations, and allows for many useful applications, though it only concerns a subset of all the problems on the way towards an efficient Natural Language Processing System.

REFERENCES

1. Deese, J. On the structure of associative meaning. *Psychological Review* **69** (1962), 161–175.
2. Goodman, L. A. and Kruskal, W. H. Measures of association for cross classifications. *J. Am. Statist. Assoc.* **49** (1954), 732–764.
3. Goodman, L. A. and Kruskal, W. H. Measures of association for cross classifications II: Further discussion and references. *J. Am. Statist. Assoc.* **54** (1959), 123–163.
4. Haralick, R. M. and Haralick, J. G. Behavioral problems of deaf children: Clustering of variables using measures of association and similarity. *Pattern Recognition* **3** (1971), 269–280.
5. Hebb, D. O. *The Organization of Behavior*. Wiley, New York (1949).
6. Hunt, E. B., Martin, J. and Stone, P. J. *Experiments in Induction*. Academic Press, New York (1966).
7. Katz, J. J. and Fodor, J. A. The structure of a semantic theory. *The Structure of Language*, J. J. Katz and J. A. Fodor (eds.), Prentice Hall, Englewood Cliffs, N.J. (1964).
8. McCalla, G. I. and Sampson, J. R. MUSE: A model to understand simple English. *Comm. ACM* **15** (January 1972), 29–40.
9. Quillian, M. R. Semantic memory. *Semantic Information Processing*, M. Minsky (ed.), MIT Press, Cambridge, Mass. (1968), 216–270.
10. Raphael, B. "SIR": A computer program for semantic information retrieval. *Semantic Information Processing*, M. Minsky (ed.), MIT Press, Cambridge, Mass. (1968), 33–134.
11. Reitman, W. R. *Cognition and Thought*. Wiley, New York (1965).
12. Schank, R. C. and Tesler, L. G. A conceptual parser for natural language. *Proceedings of the IJCAI* (1969), 569–578.
13. Sedelow, S. Y. and Sedelow, W. A. Categories and procedures for content analysis in the humanities. *The Analysis of Communication Content*, Gerbner et al. (eds.), Wiley, New York (1969), 487–499.
14. Siklossy, L. Natural language learning. *Representation and Meaning*, H. A. Simon and L. Siklossy (eds.), Prentice Hall, Englewood Cliffs, N.J. (1972), 288–328.
15. Tuggle, F. D., Moore, D., Vestal, S. C. and Isaacs, R. Computer solution of verbal analogy problems. To appear.
16. Winograd, T. Understanding natural language. *Cogn. Psych.* **3** (1) (1972).
17. Wittgenstein, L. *Philosophical Investigations*. Basil Blackwell, Oxford (1953).
18. *The English Duden*—Pictorial Dictionary. 2nd revised ed., Adler, Mannheim (1960).

Received December 1973; revised September 1974