GERT J. VAN DER STEEN

A TREATMENT OF INDEPENDENT SEMANTIC COMPONENTS

1. A TREATMENT OF INDEPENDENT SEMANTIC COMPONENTS

To distinguish things, we use terms which characterize them for us. For two balls it may be their color, for two people it may be their height, or their manner of speaking. In order to illustrate the differences in meaning for many words J. J. KATZ and J. A. FODOR (1963) proposed the use of " semantic characteristics ". They give an example for the meanings of *man* and *ball:*

man — ... — (physical object) — (human) — (adult) — (male)
ball$_1$ — ... — (social activity) — (large) — (assembly)
ball$_2$ — ... — (physical object)

D. BOLINGER (1965) proposed the systematizing of these characteristics, with hierarchic structures, so that the meaning of the word *bachelor* could be represented by a row of characteristics (fig. 1).

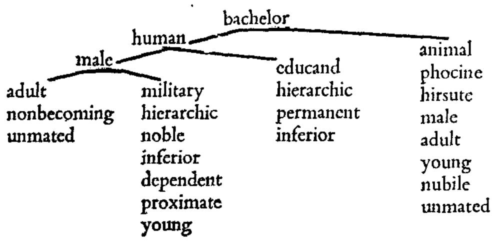

Fig. 1.

Here the meaning of a word is given by a refering it to other words. These words, in their turn, can be referenced by other words. There is the feeling that from here endless references will originate.

Let us suppose that there are a number of elementary characteristics which can not be expressed in other characteristics. We shall call them $e_1 \ldots e_I$. The question if they correspond with any existing word or expression let be leaved as it is, just as the limitation of I. We shall represent the meaning of a word by the intensity of the presence of specific characteristics e_i. If we construct a model in an I-dimensional vector-space with unit-vectors $\underset{-}{e}_1 \ldots \underset{-}{e}_I$ we may represent the meaning of a word or expression W by the vector $\underline{W} = w_1\, \underset{-}{e}_1 + w_2 \underset{-}{e}_2 + \ldots + w_I \underset{-}{e}_I$ with $w_i \geqslant 0$ for $i = 1 \ldots I$.

The common in the meaning of two words is the sum of the common in each of the basis characteristics.

In our model this is for the vectors

$$\underline{V} = \sum_{i=1}^{I} v_i \underset{-}{e}_i \quad \text{and} \quad \underline{W} = \sum_{i=1}^{I} w_i \underset{-}{e}_i \; : \; \underline{V} \cap \underline{W} = \sum_{i=1}^{I} min(v_i, w_i)\, \underset{-}{e}$$

For $I = 2$ refer to fig. 2.

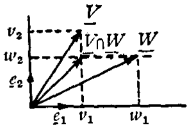

Fig. 2.

For the determination of the norm of the vectors we consider that the common of $\underline{V}$ and $\underline{W}$ is determined via their characteristics. Our consciousness can evaluate the factors v_i and w_i only one by one. Therefore we put as norm:

$$\|\underline{V}\| = \sum_{i=1}^{I} v_i.$$

Therewith is

$$\|\underline{V} \cap \underline{W}\| = \sum_{i=1}^{I} min(v_i, w_i),$$

by definition called the measure of association between $\underline{V}$ and $\underline{W}$. (*min* is the minimum-function, e.g. *min* (5, 7) = 5).

To test this model we designed two tests (G. J. VAN DER STEEN, 1971).

In the first test, individuals are asked to write down 12 words, starting with the word *bird*, and, relative to the associations between them, to indicate the measure of the association. This has to be a number between 0 and 10; ' 0 ' for: " no association ", ' 10 ' for: " synonym". For an example: see table 1. For the words W_1 to W_{12} in our model, we use the equations:

$$(1) \qquad \sum_{i=1}^{I} min\ (wn1_i,\ wn2_i) = v_{n1,n2}$$

$$\text{for } 1 \leqslant n1 \leqslant N\text{-}1$$

$$n1 < n2 \leqslant N \qquad (\text{here } N = 12)$$

wherein the numbers $v_{n1,n2}$ are given. From these equations the unknowns wn_i ($i = 1, ..., I$; $n = 1, ..., N$) have to be solved. At the same time, the number of characteristics I has to be determined. An upper limit for I is the number of equations: each association runs over a separate characteristic.

We determine I and the unknown wn_i by an iteration-process. Suppose that the factors wn_i ($1 \leqslant n \leqslant N$; $1 \leqslant i < I_1$) are determined ($I_1 = 1, 2..I$-1). Then we may write:

$$min\ (wn1_{I_1},\ wn2_{I_1}) + v^{(I_1+1)}_{n1,n2} = v^{(I_1)}_{n1,n2} \qquad (1 \leqslant n1 \leqslant N\text{-}1,\ n1 < n2 \leqslant N)$$

$$\text{with } v^{(1)}_{n1,n2} = v_{n1,n2} \text{ and } v^{(I_1+1)}_{n1,n2} = \sum_{i=I_1+1}^{I} min\ \ (wn1,\ wn2).$$

Let us denote the sum of all v's in step $I_1 + 1$ with S, so

$$S = \sum_{n1=1}^{N-1} \sum_{n2=n1+1}^{N} v^{(I_1+1)}_{n1,n2}$$

To minimize I we try to solve the system with S as small as possible. By successively assuming that a specific wn_{I_1} is the smallest of all wn_{I_1}'s we can determine for each of the suppositions the sum S. We choose now the wn_{I_1} which belongs to the smallest sum S. If there are more sums S with this value then there are more refined criteria available. Suppose this is $wn1_{I_1}$. In the equation

$$min\ (wn1_{I_1},\ wn2_{I_1}) + v^{(I_1+1)}_{n1,n2} = v^{(I_1)}_{n1,n2}$$

we choose then $wn1_{I_1} = v^{(I_1)}_{n1,n2}$. Therewith $v^{(I_1+1)}_{n1,n2} = 0$. $wn2_{I_1}$ will be determined later.

In all equations wherein $wn1_{I_1}$ appears $v^{(I_1+1)}_{n1,n2}$ can now be determined. In the remaining equations we now apply the same process till there stays at last one equation, for instance

$$min\ (wn3_{I_1},\ wn4_{I_1}) + v^{(I_1+1)}_{n3,n4} = v^{(I_1)}_{n3,n4}$$

Here we choose $wn3_{I_1} = wn4_{I_1} = v^{(I_1)}_{n3,n4}$. Therewith our iteration step for I_1 has been ended. When all $v^{(I_1+1)}_{ni,nj} = 0$ then $I = I_1$ and the whole iteration-process has come to an end.

The process is illustrated in table 2 for 4 words with associations $v_{1,2} = 3$, $v_{1,3} = 6$, $v_{1,4} = 8$, $v_{2,3} = 2$, $v_{2,4} = 4$ and $v_{3,4} = 5$ (randomly chosen). With s_j we denote the sum which belongs to a w_j which appears in a line with the lowest v.

With some small modifications this solution-scheme can be used also if some equations have been deleted in the beginning; in others words: when some associations are not given. This is illustrated in table 3 with the same v's as in table 2, except for $v_{1,3}$ which is omitted. The small modification concerns the calculation of s_j: we divide s_j by the number of lines minus 1 in which w_j appears.

We now try our model by omitting some of the given associations from one individual. According to the foregoing method we determine the number of characteristics I and the vector-representations of the words. From them we calculate the omitted associations with the aid of formula (1). For the discrepancies between the thus predicted and the omitted associations we can determine statistically an estimation.

If the associations are randomly given then the mean and the standard-deviation do agree indeed with their calculated values. If the associations are given by test-persons these numbers are significantly lower.

There are interesting discrepancies if associations are left out which express an extra aspect of meaning. In a specific case the words *bird*, *leg*, *table* and *chair* were given among others. If the association between *leg* and *bird* was left out an association of 0 was predicted, as it should be. The number of characteristics was decreased by one.

The evaluation of associations between given words by test-persons is subjective. This, however, plays no role here: the relations between a number of consciousness-contents are concerned. If the test-individual is not consistent in his evaluations then the discrepancies between the given and the predicted associations become greater. In practice there seems to be a good correlation of consistency in evaluation and the intelligence-level of the test-person.

On a more reliable level are the extended observations of the language expressions of a test-individual. For this purpose a second test was designed. As source material 20 pages of the novel *De verliezers* of Anna Blaman were taken. With the aid of the programming language

SNOBOL a frequency-table was made for all words in that piece of the text. From them 20 words with a high frequency were chosen which are relevant with regard to each other. Then it was determined how many times each of the 20 words was found together with each of the other words in the same sentence. This number was taken as a measure of the associations between the words. If we leave out the 0-associations and predict their associations by the method of the first test our model will not be unreliable if we predict the value 0.

Indeed, it appears that there are 0's predicted, except in associations between nouns and the words *mine* and *your* which give values 2 and 3, and some, randomly distributed, exceptions. The method of prediction of 0-associations was chosen to avoid the rather crude measure of association. This measure was used because of the absence of a well-defined method to detect coherent subphrases. If two words never occur together in a phrase they will certainly never appear in the same sub-phrase.

It will be interesting to try this model on the common usage of languages. The obtained vector-representations can be transferred to other characteristic systems by means of matrix-manipulations. A further extension lies in the determination of the representation of words in different natural languages with a mutual comparison and eventual transformation of the representations. The measure of association is critical here. This should be refined by using more knowledge about the syntactic structure of the sentences.

A further restriction lies in the number of developed characteristics. For the first test this was approximately 13, for the second approximately 76. From these 76, the first 20 were the most relevant. The remaining characteristics served more to compensate for several small discrepancies. By chancing to a larger amount of language information the number of relevant characteristics will naturally increase.[1]

[1] Programming languages used: PL/I, FORTRAN and SNOBOL.
Machines used: IBM 360/65 and PDP-9.
A FORTRAN-II program for determining the semantic components from given associations is available upon request.

Table 1.

Words and associations given by a test-person

		2	3	4	5	6	7	8	9	10	11	12
		HUIS HOUSE	*NEST* NEST	*BOOM* TREE	*TAK* BRANCH	*HOUT* WOOD	*LEVEND* LIVING	*ETEN* TO EAT	*KIJKEN* TO LOOK	*DIER* ANIMAL	*POOT* LEG	*DAK* ROOF
1	*VOGEL* BIRD	2	7	5	3	2	8	6	6	8	8	3
2	*HUIS* HOUSE		3	3	0	2	0	0	0	0	0	6
3	*NEST* NEST			7	5	4	7	2	0	8	0	2
4	*BOOM* TREE				8	9	8	0	0	1	0	1
5	*TAK* BRANCH					8	5	0	0	0	1	0
6	*HOUT* WOOD						6	0	0	0	0	3
7	*LEVEND* LIVING							4	6	7	0	0
8	*ETEN* TO EAT								1	5	0	0
9	*KIJKEN* TO LOOK									5	0	0
10	*DIER* ANIMAL										4	0
11	*POOT* LEG											1
12	*DAK* ROOF											

REFERENCES

D. Bolinger, *The atomization of meaning*, in « Language », XLI (1965) 4, pp. 555-573.

J. J. Katz, J. A. Fodor, *The structure of a semantic theory*, in « Language », XXXIX (1963) 2, pp. 170-210.

G. J. van der Steen, *Semantic processes in artificial intelligence systems* (in Dutch), Delft, 1971.

TABLE 2.

$i =$		1			2			3			4		
		$S_2=7, S_3=11$	$S_3=6, S_4=8$		$S_1=2, S_2=3$ $S_3=1, S_4=2$	$S_1=1, S_4=2$		$S_1=2, S_2=1$ $S_3=1, S_4=0$	$S_1=2, S_2=1$ $S_3=1$		$S_1=1, S_2=1$ $S_3=0, S_4=0$	$S_1=1, S_2=1$ $S_4=0$	
INDICES OF V		$W2_1:=2$	$W3_1:=5$	$W1_1:=W4_1:=8$	$W3_2:=0$	$W1_2:=0$	$W2_2:=W4_2:=2$	$W4_3:=0$	$W2_3:=0$	$W1_3:=W3_3:=1$	$W3_4:=0$	$W4_4:=0$	$W1_4:W2_4:=1$
1,2	3			1	1		1	1		1	1	1	0
1,3	6	6		1			1	1	1	0			0
1,4	8	8	8	0	0		0			0	0		0
2,3	2			0			0	0		0			0
2,4	4			2	2	2	0			0	0		0
3,4	5	5		0			0			0			0

The representations are $W1 = (8, 0, 1, 1)$; $W2 = (2, 2, 0, 1)$; $W3 = (5, 0 1, 0)$; $W4$ $(8, 2, 0, 0)$.

TABLE 3.

$i =$		1			2			3
		$S_2=7/2$ $S_3=5/1$	$S_3=0/0$ $S_4=8/1$		$S_1=1, S_2=3/2$ $S_3=0, S_4=2/2$	$S_1=1$ $S_4=2$		
INDICES OF V		$W2_1:=2$	$W3_1:=5$	$W1_1:=W4_1:=8$	$W3_2:=0$	$W1_2:=0$	$W2_2:=W4_2:=2$	$W1_3:=W2_3:=W3_3:=W4_3:=0$
1,2	3			1	1		1	0
1,3		------	------	------	------	------	------	------
1,4	8	8	8	0	0		0	0
2,3	2			0			0	0
2,4	4			2	2	2	0	0
3,4	5	5		0			0	0

The representations are: $W1 = (8, 0, 1)$; $W2 = (2, 2, 1)$; $W3 = (5, 0, 0)$; $W4 = (8,2,0)$
The expectation for the association between $W1$ and $W3$ is then: 5.