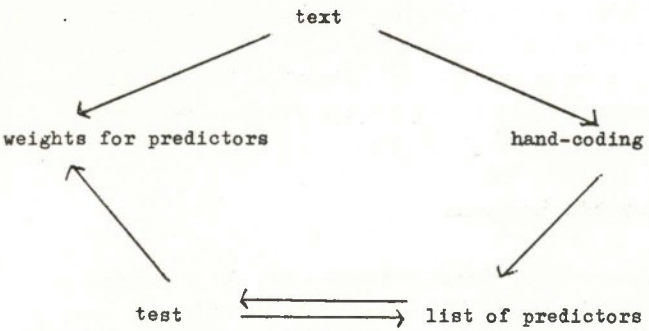
In search of semantic characteristics for machine coding[x]

J.Z. Namenwirth, W.E. Saris, I.N. Gallhofer, J. Kleinnijenhuis

Machine coding of content has many advantages as documented by Holsti (1969). In contrast to hand-coding, machine coding can speedily process large amounts of texts while leaving reliability unaffected with increasing size. Coding massive amounts of text would readily bore expensive expert content analysts making them less reliable in the process. Therefore, practical considerations alone favor machine coding. Accordingly, this paper reports on a search for an automatic procedure raising questions of a methodological nature: "How can one detect discriminating characteristics of texts which were used previously by hand-coders in their content analytic efforts. This paper presents a procedure for this purpose while illustrating its use on texts collected and alternatively analyzed by I.N. Gallhofer (Coders'reliability).
Figure 1 summarizes this procedure.

Figure 1: Procedure of analysis



The analysis proceeds in a number of steps. First, a relevant sample of text is selected. Second, the text is hand-coded, i.e.decomposed

x The research of this paper was in part supported by a fellowship to the senior author from the Nederlandse Organisatie voor Zuiver Wetenschappelijk Onderzoek, Z.W.O.

into part phrases or categories of kernel sentences. Third, if the latter coding is sufficiently reliable a first list of predictors is formulated. Fourth, one tests how well these predictors reproduce the original hand-coding efforts ex post facto. Fifth,for this predictive effort discriminant analysis is an obvious choice (1). It provides an indication of the correctness of the chosen predictors for the text at hand. Sixth, if the chosen list of predictors produces an unsatisfactory result, the list must be modified until satisfactory results are obtained.
Seventh, once such results are optimized, the discriminant analysis produces specific weights for each of the predictors indicating their relative importance for the prediction ex post facto.(Other procedures would produce weights of either one or zero thus indicating whether a predictor is or is not relevant and should or should not be used).
Eighth, having completed these efforts, a new text is selected and hand-coded and the results thereof are predicted upon the basis of the former model. Ninth, if these predictions prove satisfactory, then one has validated the predictors and hence the pertinent content characteristics. If these predictions prove unsatisfactory, the whole process commences anew, changing the list of predictors and retesting this list on other texts.
Even though seemingly straight forward this procedure encountered various problems in actual content analytic applications and these problems will be discussed in the following.

1. The categories of kernel sentences

Table 1 lists the categories of kernel sentences and their marginal frequencies (2). An explanation of the classification of kernel sentences the reader will find in I.N. Gallhofer (Coders'reliability).

Table 1 : Categories of kernel sentences and their frequencies (3)

Category of kernel sentence	frequency
Action Netherland	106
Action Opposition	20
Probability Statements	49
Value Statements	100
Outcomes	85
New Developments	62
Motivation Phrases	38
Undetermined	151
Total	611[x]

x In a later stage of this investigation these counts were modified as documented below.

2. The semantic predictors

Since the kernel sentences consist of concatenations of words, it seemed plausible that specific words might discriminate among the several phrases. Therefore we searched for words as our predictors (4). Unfortunately, the frequency distribution of these words was badly skewed and this skewness is a major source of problems for the subsequent analysis. First let us turn however to the question why our word frequencies were so badly skewed.

Skewness is not solely a property of the actual word frequency distributions of ministerial debate(s); it is a general characteristic of word frequency distributions (Zipf, G.K.). Indeed, these distributions are generally L-shaped with a few words having very high frequencies and most words having very low frequencies thus occurring most infrequently in the vocabulary of spoken or written language. In addition, high frequency words have generally low semantic content as for instance in the case of "of", "the", "a" and "that" while it is equally true that the greater the specificity of content the lower the frequency of word usage. This very fact further complicates our task because it is rather likely that words

with specificity of content will more readily discriminate among the different types of phrases than words with very low specificity. But , in a statistical analysis such as ours, words (or variables) , which rarely occur (having therefore predominantly extremely skewed distributions) are useless for statistical analysis. The solution for this problem is dictionary construction, i.e. the creation of new variables which consist of categories combining words and therefore collapsing distinctions in semantic meaning. Therefore, the question arises how and by what rules to combine infrequently occurring words. The following procedures were used: Of all words with a frequency of four or more a cross-sort produced a frequency distribution of each word by eight types of phrases. Thirty six words discriminated among types of phrases. Among the remaining words, some seemed randomly distributed across the types of phrases and were therefore dropped from further consideration, the remainder had frequencies which were too low to make such a determination in a reliable manner (5). In the latter case words which performed similarly (or had analogous distributions as well having some semantic meaning in common) were combined in categories of words. Table 2 illustrates the procedures. These remaining words were thus combined in originally 24 categories of words.

3. Discriminant analyses and recoding

According to the procedure described above each phrase was characterized by 60 variables constituted by 36 single words and 24 categories of words. These dichotomous variables were given one of two values: namely, 0 if the word or category of words did not occur in the kernel sentence and 1 if it did occur (one or more times). In this manner our procedure produced a data matrix each row representing each phrase, each column representing one of 60 variables namely the 36 single words and 24 categories of words with the quantities 0 and 1 in the intersections indicating whether these variables did or did not occur in each of the phrases. Finally, the matrix contained one more column, namely the nominal variable: "type of phrase" which varied from 1-8 and thus indicating which one of eight types of phrases each phrase belonged to.

Table 2 : Frequency distribution of three single words with high low frequencies across eight types of phrases.

Category of phrase	Words[+]						
	"the"		"chance"		"probable"		number of sentences
	absent %	present %	absent %	present %	absent %	present %	
Action Netherland	49.5	50.5	100.0	0.0	100.0	0.0	107
Action Opposition	42.1	57.9	100.0	0.0	100.0	0.0	19
Probability Statements	68.0	32.0	82.0	18.0	88.0	12.0	50
Value Statements	91.4	8.6	100.0	0.0	100.0	0.0	105
New Developments	36.5	63.5	100.0	0.0	100.0	0.0	63
Outcomes	57.1	42.9	100.0	0.0	100.0	0.0	84
Motivation Phrases	66.7	33.3	97.2	2.8	100.0	0.0	36
Undetermined	32.7	67.3	99.3	0.7	100.0	0.0	153
total frequency	336	281	606	11	611	6	617[x]

+ The word "the" is a single word variable, the entries "chance" and "probable" are both included in the category: Probability-Words.

x Some phrases were added, others reclassified as documented below.

The described data matrix was subjected to a discriminant analysis whereby each phrase was considered as a case and the eight types of phrases as eight different groups. As stands to reason, this procedure produced seven discriminant functions containing redundant variables which did not contribute to their predictive power. In the first test about 70 % of all phrases were correctly postdicted. This result suggested a number of alterations which will be discussed to explicate our procedures.

Wrongly classified sentences were examined in order to detect what errors in word classifications and/or omissions might have produced the wrong classifications (or predictions). This process of exami-

nations made us consider what words should be added to the analysis, which ones dropped, which ones might require reclassifications and what new categories of words should be tried. For instance, it was thus discovered that words such as destruction, annexation and usurpation were almost always found in Value Phrases but not in other types of phrases. This prompted the creation of the category: Identity Building or Destruction. These latter words, however, have very low frequencies indeed, sometimes occurring only once in the complete text maximizing at times chance error in the predictive effort and surely lowering replicability of the predictive model for other samples of texts. To minimize this possibility the previous classifications of all words were reconsidered. Having arrived at a substantive interpretation of each of the categories, words which did not seem to fit semantically in the category were eliminated irrespective of their predictive power while on the same grounds other words were included. The modified system of predictive words and word categories will be discussed briefly below and more extensively by Namenwirth (Contrasting themes).

At the same time, we asked the original coders what kinds of words they believed did discriminate among the eight types of phrases. Many suggestions were offered. For instance some coders were sure that Value Phrases were characterized by evaluative adjectives such as good, bad, and so on. Hence, we created a category of such words appearing in the text.

Finally there is the problem of homographs. Dutch as much as any other language has frequent homographs in its vocabulary, i.e. words with the same spelling (even prononciation) which have more than one distinct meaning. As an example, consider the word "kind". What kind of word is that ? Some misclassifications of the preliminary discriminant analyses were caused by misclassified homographs. As yet there is no Dutch version of English disambiguation routines (Kelley and Stone) and disambiguation errors were corrected by "manual" changes in the data matrix (6). This work is tiresome and one is never certain that the changes are an improvement. Especially there is a danger of ad hoc solutions which do not hold for other texts.

On the basis of the revised 41 content analytic variables, some single words, other categories of semantically related words,a

discriminant analysis (7) postdicted correctly 82,7 % of the kernel sentences into eight categories of phrases. In order to correct for agreement due to chance Scotts π (8) was computed on the results of table 3. This measure produced an agreement coefficient of .80 between the hand codings and their postdictions which is quite satisfactory.

Table 3 : Distribution of actual and postdicted phrases by eight types according to discriminant analysis of 41 content analytic variables.

Actual phrases	Postdicted phrases								N
	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	
1. Action Netherland	79.4	1.9	1.9	3.7	4.7	2.8	2.8	2.8	107
2. Action Opposition	-	73.7	-	-	15.8	5.3	5.3	-	19
3. Probability Statements	4.0	-	88.0	2.0	-	-	2.0	4.0	50
4. Value Statements	1.9	1.9	3.8	89.5	1.0	1.0	1.0	-	105
5. Outcomes	1.2	3.6	-	1.2	86.9	3.6	2.4	1.2	84
6. New Developments	3.2	1.6	3.2	11.2	1.6	79.4	-	-	63
7. Motivation Phrases	8.3	5.6	5.6	2.8	-	-	72.2	5.6	36
8. Undetermined	3.9	1.3	3.3	5.2	0.7	-	4.6	81.0	153

This overall result, however, hides important differences: some types of phrases were far better predicted than others as illustrated in table 3. For instance, of the 105 Value Phrases about 90 % were correctly postdicted. In general, the larger the number of phrases of a particular kind, the more likely was their correct postdiction. Probability Phrases are a conspicuous exception to this rule and this, more likely than not, results from their very homogeneous semantics.

In this case, discriminant analysis creates a multi-dimensional semantic space where each type of phrase is located in a unique corner of this space while the nature of the dimensions is indicated by the predictive variables. Table 4 provides the relevant information.

Table 4: Standardized discriminant function analysis coefficients of 41 predictors (content analytic variables) and centroids of eight types of phrases in reduced space. Coefficients < |.15| omitted.

Content categories	Discriminant functions						
	1	2	3	4	5	6	7
I,me mine	-.2	-	-	-	-	-	-
Third Person	-.2	.2	-	-	-	-	-
Authority(ies)	-.3		-.2	-	-	-	-
Understate Words	-.4	.3	-	-	-	.2	-
Time/Space	-.2	-	-	-	-	.2	-
Deliberation Words	-.7	.5	.2	.2	.2	.2	-
Embedded Indicators	.6	.7	-.3	.2	-	-	-
Development Words	.3	.3	-.2	-	-	-	-.2
Effect Words	-	.3	-.2	-	-	-	-.2
First Person Pl.	-	.3	-.2	-.2	-	-	-
One, him,her	-	.2	-	-	-	-	-
Propositions	-	.2	-	-	-	-	-
Termination Words	.2	-.2	.6	.4	-	-	-.2
War Words	.3	-	.5	.3	-	-	-.3
Status Quo Words	.2	-	.2	-	-	-	-
Possibility Words	-	-.2	.2	-	-	--	-
That	-	-	.2	-	-	-	.2
Is	-	-.2	-.2	.2	.2	-	-
Have/has	-.2	-	-	.2	-	-	-
Neutrality	-	-	-	-.3	.2	-	-
At/in	-	-	-	-.2	-	-	-
Interaction Words	-	-	-	-.6	.3	-	.2
Moral Policy Words	-	-	-	-.6	.2	-	.3
Probability Words	-.2	-.4	-.4	.5	.8	.3	-
The	-	.2	-	.2	.3	-	-
Of	-	.2	.2	-	-.3	-	-.2
Identity Builders	-	.3	-.3	-	-.8	-	-
On	-	-	-	-	-	.2	-
Not	-	-	-	-	-	.2	-
Overstate Words	-	-	-	-	-	-.2	-
Would	-	.2	-	.2	-	-.3	-
Moment	-	-	-	-	-	-.3	-
Motivation Words	-	-	-.2	-	-.2	-1.0	-
Let	-	-	-	-	-	-.2	.6
Foreign Actors	.3	-	.4	.2	-	-	.5
Dutch	-	-	-	-	-	-	.3
Will	-	-	-	-	-	-	.3
On	-	-	-	-	-	-	-.2
Before/for	-	-	-	-	-	.2	-.2
Become	-	-	-	-	-	-	-.3
National Interest	-	-	-	-	-	-	-
Action Netherland	.3	-.2	.5	-2.0	.6	.4	-.2
Action Opposition	1.0	.4	.8	-	.9	-.5	3.0
Probability	-.4	-1.9	-1.4	1.4	2.0	.8	-
Value Statements	.4	-1.5	-.8	-	-1.5	-	-
Outcomes	2.2	2.3	-1.1	.4	-	-	.2
New Developments	1.7	-.7	2.5	1.3	-	-	-.4
Motivation Phrases	-.8	-1.6	-.5	-.3	.6	-3.3	-
Undetermined	-2.2	1.0	.3	.4	-.3	.2	-

The first discriminant function indicates that in general Undetermined Phrases are located on its negative end with a centroid value of -2.2. Such sentences are characterized by frequent usage of Deliberation (-.7), Understate (-.4) and to a far lesser extent Authority Words (-.3). Hence these must be sentences which refer to the minister (Authority) who may (Understate) consider (Deliberation) what he (Third Person) would suggest (Deliberation) today (Time/Space). In this respect Outcome Phrases (and to a lesser extent New Development and Action Opposition Phrases) differ the most having a centroid value of 2.2. This is to say that Deliberation and Understand Words are likely to be absent from Outcome Phrases as are references to Authorities and so on. In contrast, this type of phrase contains references to words of the categories War, Foreign Actors, Termination and Development Words. Most characteristic for this phrase, however, is the fact that Value and Probability Phrases were originally embedded in Outcome Phrases. Therefore, it is Outcome Phrases which are most likely to contain evaluations as well as chance estimates. In contrast, Indeterminate Phrases do not contain either evaluations or chance attributions for one particular reason:Coders were instructed to disregard evaluations and chance attributions in Indeterminate Phrases because they are of no consequence for the decision making process. In this case, the content analytic finding merely reproduces the coding instructions! As table 4 further illustrates the discrimination among most types of phrases is multi-dimensional, i.e. two types of phrases may be very similar with regard to one dimension (or discriminant function) but different with respect to another. Undetermined Phrases and Action Opposition Phrases are exceptions to this rule, the former singularly characterized by their centroid position on the first discriminant function and the latter by their centroid on the last. A fuller interpretation of the semantic space which discriminates among the eight types of phrases (including definitions of the predictive content variables) will not be attempted here since the validity of this result has to be established first by replicating the procedures on a new text.

4. Test on the new text

To test the validity of the discriminant analysis as a predictive model, the analysis was replicated on a different text, namely a debate by the Dutch Council of Ministers concerning the temporal reduction of Dutch troops in the spring of 1916 (9).
The results of this replication were most disappointing since the discriminant analysis model classified only 34,8 % of all sentences correctly. Table 5 shows the distribution of actual and predicted phrases of this test.

Table 5: Distribution of actual and predicted phrases of the test on a new text

Actual Phrases	Predicted Phrases								
	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	N
1. Action Netherland	18.8	-	-	12.5	37.4	18.8	12.5	-	16
2. Action Opposition	15.5	7.6	15.5	7.6	23.1	23.1	-	7.6	13
3. Probability Statements	10.5	10.5	36.9	31.6	-	-	-	10.5	19
4. Value Statements	33.3	-	-	61.1	-	-	-	5.6	18
5. Outcomes	14.3	14.3	-	-	71.4	-	-	-	7
6. New Developments	10.5	5.3	-	10.5	36.8	15.8	5.3	15.8	19
7. Motivation Phrases	25.0	-	-	-	-	-	50.0	25.0	4
8. Undetermined	17.8	4.4	-	4.4	4.4	11.1	20.0	37.8	45
N	25	7	9	24	25	14	14	25	141

When correcting for chance agreement by computing Scott's Π the agreement between the hand-codings and their predictors was even lower, namely .23. Clearly the predictor model is falsified. This result shows that successful postdiction does not automatically garanty good predictions for new texts. A partial explanation of this result might be that the attempted corrections, while maximizing the fit for the first text, introduced error in the analysis of the general case which resulted from the use of ad hoc categories. Although this might be the case, it would not explain why Outcome, Value and Motivation Statements are predicted better than Action Opposition, Action Netherland and New Development Phrases. An explanation thereof might be, that the seman-

tic nature of the different types of phrases would vary from topic to topic but to a greater extend for some types of phrases than for others. Particularly, it is possible that the semantic content, i.e. the kind of words used in Outcome, Value and Motivation Phrases, does not vary greatly from one topic to another, e.g. from a discussion of the impending occupation of Antwerp to the risks involved in the reduction of troop strength. With respect to Action Opposition, Action Netherland and New Development Phrases the situation might be quite different because the semantics of these types of phrases could be far more topical and situation specific. If this were the case, the semantic predictors found to be successful in one text would not work satisfactorily for another one. Since the results of the study indicated that the classification of certain types of phrases can not be based on semantic content alone a classification based on more formal characteristics such as syntax might produce more satisfactory results. Therefore,a further exploration of the successful postdiction by syntactical characteristics of the two sets of original documents and tests on new texts are still in order.

Conclusion

Although the procedures used in this study seem useful for this kind of inquiries, we encountered a number of problems:

1) skewness of the frequency distribution of words;
2) problems in dictionary construction;
3) the danger of maximizing success by ad hoc classification introducing error for the general case;
4) the problem of content dependency of semantic predictors.

Our approach substituted the skewness problem of the frequency distribution (problem 1) with the problems of dictionary construction. In addition, this latter problem was approached not systematically but in a trial and error fashion leading to problems 3) and 4), namely the danger of maximizing success by ad hoc classification and the resulting topic dependency of semantic predictors.
The major methodological conclusions of this study are therefore that tests on new material with different content are always essential and can prevent the type of error mentioned under 3) above whi-

le the utility of syntactical characteristics has to be tested; such characteristics are most likely to be more content independent than semantic ones. In this respect the elsewhere reported work of M. Boot (Ambiguity and automated content analysis) could be most helpful.

Notes

(1) Another possible method would have been J.N. Morgan's and J.A. Sonquist's "Automatic interaction detector" (AID) which has the advantage of lower measurement level but is more timeconsuming and does not lead to such clear predictions.

(2) Because two categories (Actual State, Evaluation of the Actual State, in I.N. Gallhofer, Coders'reliability) had very low frequencies (9,2) they were omitted from the research as their inclusion would certainly lead to spurious results.The data mentioned in table 1 are forthcoming from a Dutch ministerial debate concerning the impending occupation of Antwerp during the beginning of World War 1 (Algemeen Rijksarchief, 's-Gravenhage, RA 2^{e}, 3 october 1914).

(3) For computer content analysis purposes a kernel sentence is defined as a string of words numbered from 1 to n, each word defined as a string of machine readable characters preceded and followed by either one or more empty spaces or punctuation symbols, such as period, comma, semi-colon, and so on. In this manner each kernel sentence is a separate record or unit and each different word a variable and its presence or absence characterizing the unit.

(4) Using the program RIQS, described in I.N. Gallhofer (Programs). we determined which words occured in the ministerial debate and their frequencies.

(5) This experience has further confirmed our belief that Iker's (Iker and Harway) celebration of the virtues of single word analysis is misguided; only by combining words into categories and thus using dictionaries of one kind or another as advocated by Stone, et al. (1966) we can include important low frequency words in content analysis.

(6) These tasks were facilitated by the program RIQS, namely the indices INVERT and KWIC, see I.N. Gallhofer (Programs).

(7) Each type of phrase was assigned an equal chance of occurrence, the extreme unequal distribution of the marginals were not considered since they seemed a poor estimate of some true distribution of a population of decision debate sentences. See on this score also the marginal frequencies of table 5.

(8) See K. Krippendorff (1970), p.144.

(9) Algemeen Rijksarchief, 's-Gravenhage, dossier RA 2^e, Archief Ministerraad, april 1916, Ontwerp vermindering troepenmacht.

Bibliography

Algemeen Rijksarchief, 's-Gravenhage, dossier RA 2^e, Archief Ministerraad, 3 october 1914, april 1916, ontwerp vermindering troepenmacht.

Boot, M., Ambiguity and automated content analysis, in MDN, february 1978.

Gallhofer, I.N., Coders' reliability in the study of decision making concepts, replications in time and across topics, MDN, february, 1978.
Computer programs for content analysis available in the Netherlands, MDN, february, 1978.

Holsti, O.R.,Content analysis for the Social Sciences and the Humanities, Reading, Mass., 1969, Addison-Wesley.

Iker, H.P. and Harway, N.I., A computer systems approach toward the recognition and analysis of content, in Gerbner G. et al., The analysis of content; development in scientific theories and computer techniques, New York, 1969, John Wiley.

Kelly, E.F. and Stone, P.J., Computer recognition of English word senses, Amsterdam 1975, North Holland.

Morgan, J.N.and Sonquist, J.A., "Automatic interaction detector" (AID), in The detection of interaction effects:A report on a computer program for the selection of optimal combinations of explanatory variables, monograph no. 35, Survey Research Centre, Institue for Social Research, University of Michigan, 1964.

Krippendorff, K., Bivariate agreement coefficients for reliability of data in Sociological Methodology, Borgatta, E.F. and Bohrnstedt, G.W. eds., San Francisco, 1970, Jossey Bass.

Namenwirth, J.Z., An analysis of British newspaper editorials, AJS, LXXIV; 4, january 1969: 343-360.

Stone, P.J. et al., The General Inquirer: a computer approach to content analysis, Cambridge, Mass., 1966, M.I.T.

Zipf G.K., The psycho-biology of language, Boston, 1935, Houghton-Mifflin.