MULTIPLE DISCRIMINANT ANALYSIS IN LINGUISTIC PROBLEMS

Seppo Mustonen

The purpose of this paper is to indicate the possibilities of applying statistical multivariate analysis to some linguistic problems. We shall discuss the problem of identifying a language using the statistical multiple discriminant technique. This technique will be illustrated with an example of teaching the computer to decide to which language a given word most probably belongs. The example has been originally planned only for fun and it is not to be taken too seriously, but nevertheless this example can be extended to have practical linguistic applications also.

The method considered, Multiple Discriminant Analysis, belongs to statistical multivariate methods and its principal tasks can be described as follows (see e.g. Anderson (1) p. 126, Cooley-Lohnes (2) pp. 116–145).

It is assumed that we have a sample of statistical data containing certain quantitative information for several individuals. We further assume that our sample is not homogeneous, but subdivided into several groups in each of which homogeneity will be achieved. If we now know the correct group of every individual in our sample, the first problem to be solved by Discriminant Analysis is to work out (on the ground of our quantitative data) how the groups can be distinguished from each other most efficiently. The second task of Discriminant Analysis is to answer the question: "What is the correct group of an individual whose quantitative properties we know?" This is a problem of classification.

In our example we tried to teach the computer to distinguish three languages, English (E), Swedish (S) and Finnish (F), from each other. In this tentative research our sample was 900 words, 300 for each language. These words were chosen at random from dictionaries.

Since a word itself is not a numeric quantity, our first task was to invent suitable devices for measuring properties of the words numerically. We chose altogether 43 quantitative variables to describe these properties. (See Table 1.) These data were computed from the original words by a special program.

TABLE 1. List of variables

		Always
1	different letters	5
2	different vowels	2
3	syllables (according to the rules of Finnish)	2
4	syllables with 1 letter.	0
5	syllables with 2 letters	1
6	syllables with 3 letters	0
7	letters in first syllable	2
8	letters in last syllable	4
9	syllables of type AB (A, I stand for vowels)	1
10	syllables of type BA (B, L stand for consonants)	0
11	syllables of type ALB	0
12	syllables of type BAI	0
13	syllables of type BAL	0
14	letter twins of type AA or BB	0
15	diphthongs	1
16	first letter (vowel=0, consonant=1)	0
17	last letter (vowel=0, consonant=1)	1
18	letters A	2
19	letters B	0
20	letters C	0
21	letters D	0
22	letters E	0
23	letters F	0
24	letters G	0
25	letters H	0
26	letters I	0
27	letters J	0
28	letters K	0
29	letters L	1
30	letters M	0
31	letters N	0
32	letters O	0
33	letters P	0
34	letters R	0
35	letters S	1
36	letters T	0
37	letters U	0
38	letters V	0
39	letters W	1
40	letters Y	1
41	letters Å	0
42	letters Ä	0
43	letters Ö	0

It is necessary to emphasize that the computer was not given any other special information to improve accuracy of discrimination. For instance, no deterministic rules for identifying the language were given, i.e., nobody told the machine that a Finnish word never ends with two consonants, that the English alphabet does not contain Å, Ä, Ö, etc. Thus all the information the computer could use was restricted to 900 words described by 43 quantitative variables.

Using this information as a basis, the computer had to develop its own opinion about the differences between the languages. It can be claimed that the computer created a counterpart for that mental picture which Man would use in the same task. Let us remember, however, that Man in mastering all these three languages is superior to the computer, because he can classify most of the words by knowing their meaning. It must be mentioned further that nobody could expect the computer to recognize even its training materials, i.e., those 900 words, for Discriminant Analysis generates only a general picture of the differences between the languages, and here the effect of one single word is almost negligible.

In this case[1] the general picture was concentrated to 86 $(=2 \times 43)$ values, *discriminant loadings,* which gave the differences between English, Swedish and Finnish words in terms of two *discriminant functions.* Each discriminant function is a simple linear expression of the 43 original variables having the discriminant loadings as its coefficients. (See Table 2.)

When studying these discriminant functions, it should be realized that each of them has a very characteristic rôle. The first (stronger) of them separates Finnish from the other languages but makes almost no distinction between English and Swedish. The second discriminant function, on the other hand, separates English and Swedish from each other. These rôles are clearly revealed by the loadings of the different variables. For instance, in the first discriminant function high negative loadings indicate tendency to Finnish. Such loadings have letters H, J, K, L, M, N, P, S, T, U, V, Ä, the number of syllables with one and two letters, etc.

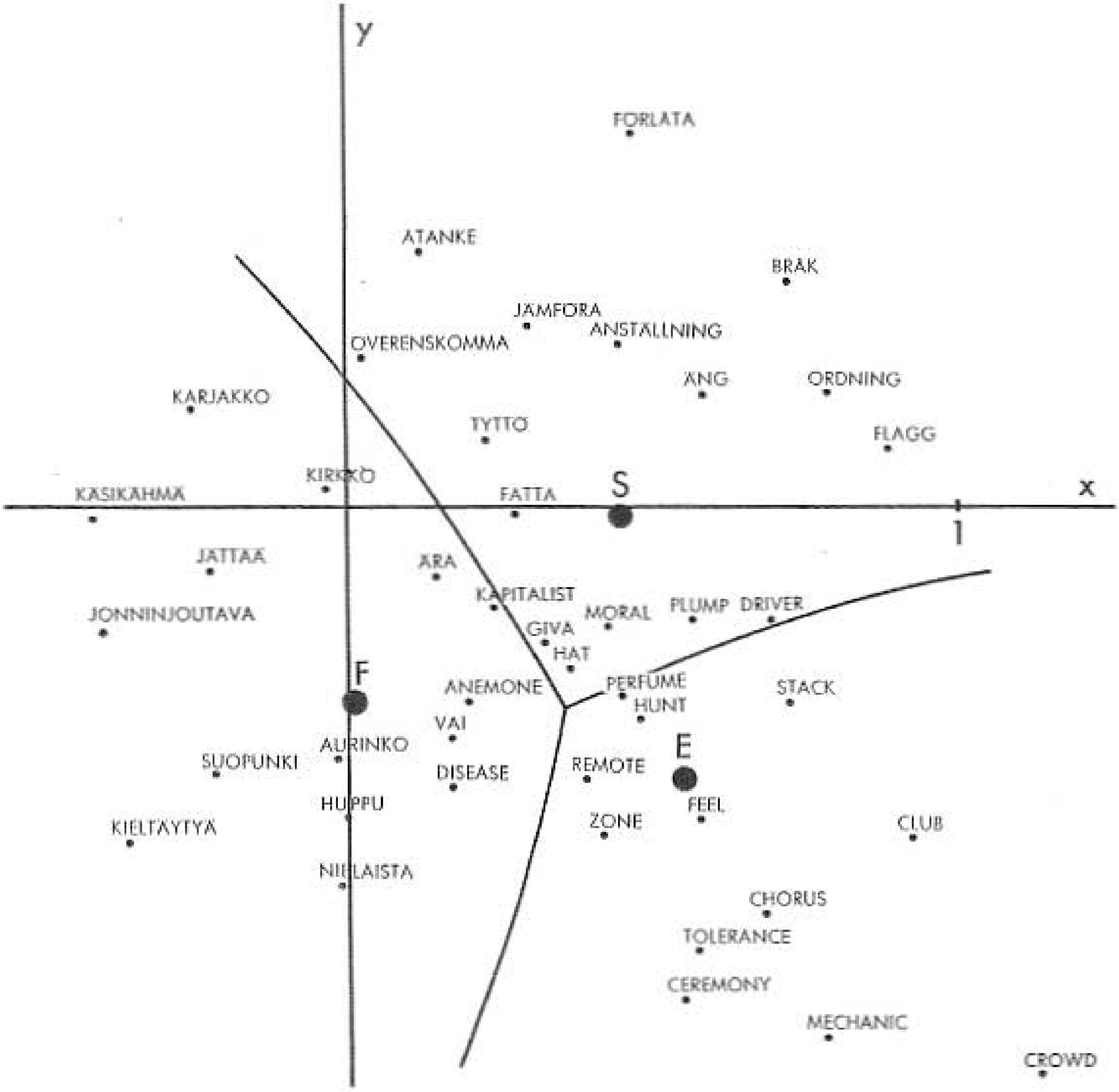
In the second discriminant function positive loadings refer to Swedish and negative to English. For instance, B, D, F, J, K, L, M, N, P, R, T, V, Å, Ö are "typical" Swedish letters and C, E, U, W, English. Also the lack of diphthongs in Swedish is indicated by a negative loading.

[1] Discriminant Analysis has been carried out according to the technique reported in Cooley-Lohnes (2) pp. 116–145. The computations have been performed with an Elliott 803 computer.

TABLE 2.	Discriminant functions:	I	II
1	different letters	.021	—.024
2	different vowels	.028	.067
3	syllables	.414	—.209
4	1-letter syllables	—.467	.176
5	2-letter syllables	—.232	.073
6	3-letter syllables	—.011	.058
7	letters in first syllable	.056	—.020
8	letters in last syllable	.078	—.049
9	AB- type syllables	.076	—.073
10	BA- type syllables	—.038	0.17
11	ALB-type syllables	—.084	.167
12	BAI-type syllables	—.160	.137
13	BAL-type syllables	—.161	—.009
14	letter twins	—.033	—.004
15	diphthongs	—.247	—.134
16	first letter	—.043	—.120
17	last letter	.071	.054
18	A	—.055	—.033
19	B	.009	.187
20	C	.076	—.163
21	D	—.073	.173
22	E	—.001	—.122
23	F	—.078	.283
24	G	—.011	.184
25	H	—.215	.032
26	I	—.089	—.090
27	J	—.224	.208
28	K	—.261	.230
29	L	—.142	.126
30	M	—.127	.191
31	N	—.167	.143
32	O	—.071	—.093
33	P	—.146	.132
34	R	—.099	.127
35	S	—.164	.149
36	T	—.132	.171
37	U	—.107	—.121
38	V	—.175	.155
39	W	.060	—.299
40	Y	—.095	—.074
41	Å	—.025	.408
42	Ä	—.126	—.002
43	Ö	—.043	.166

The rôles of discriminant functions need not generally be characteristic. It often happens that rôles are mixed and no sensible interpretation is possible. Since, however, Finnish in this investigation was fairly different from English and Swedish, it is quite obvious that discriminant functions were characteristic.

The rôles of discriminant functions become also evident from Fig. 3 where x-coordinate corresponds to the first one and y-coordinate to the second one. In the figure, the mean points of each language and the borders separating languages have been drawn. The figure can be enlarged to a "map of languages" by locating the words in their own places by the aid of the discriminant loadings. The words located in the West on the map are typically Finnish and words in the Nort-East and the South-East are Swedish and English, respectively.



The main interest in this investigation was concentrated on the second task of Discriminant Analysis. We wanted to know what was the effect of teaching, i.e., how wise the computer now was in classification.

We took a *new* random sample of 300 words, 100 words from each language, and let the computer predict the correct language of each of them. First the values of the discriminant functions were computed for every new word, and hence its position on the "map of languages" was determined. Classification was then based on the "distances"[2] between the new word and the mean points of the languages; the computer always chose the language nearest to the new word. Furthermore, the computer could tell us how sure it was about the prediction by giving a probability for every language. (For instance, the probability of REMOTE being an English word was .56. The probabilities for Swedish and Finnish were .23 and . 21, respectively.)

Table 4 indicates the results of classification when no "Don't know" answers were allowed to the computer.

When examining the results one must remember that complete accuracy of classification is hardly achieved in applications like this where the different groups (as languages here) are not disjoint but have a great many members (words) which are, or at least could be, common to several of those groups.

If this method of classification is compared with deterministic methods in which a great number of strict rules form the basis of classification, it is obvious that our method, being so blind in many details, cannot contest

TABLE **4.**

	Classification	E	S	F
Correct language	E	59	28 (19)	13 (11)
	S	11	79	10 (8)
	F	1	8	91

The numbers in brackets are valid after eliminating words common to two languages.

[2] A special measure of distance (Mahalanobis' D^2) is used in which the interplay of variables has been counted.

If 70 % certainty (one of the probabilities over .70) is required, the results are

TABLE 5.
Classification with 70 % certainty

	English	Swedish	Finnish	Don't know
English	45	AGOG (.87) DEMAND (.75) FLAME (.77) GUARD (.86) SKILL (.71) 5	ANEMONE (.75) DISEASE (.86) HAVE (.75) LEAVE (.79) POTATO (.76) SANITARY (.80) (SO) (.72) 7	43
Swedish	BEGE (.77) MEDARBETARE (.97) SCEN (.99) 3	41	ANA (.85) KÄLLA (.87) HUVUD (.78) HÄNDELSE (.70) HÖJA (.74) (LEKA) (.72) (MALARIA) (.94) ÄRA (.71) 8	48
Finnish	MIES (.85) 1	TYTTO (.76) 1	79	19

The numbers in brackets are probabilities corresponding to misclassified words.

To sum up, on different certainty levels the results were

TABLE 6.

Classification	correct	wrong	don't know
70 % certainty	55 %	8 %	37 %
60 % certainty	62 %	13 %	25 %
50 % certainty	72 %	19 %	9 %
all classified	76 %	24 %	—

with those in accuracy. Nevertheless it seems to have some advantages worthy of mention.

Deterministic classification is not very interesting, since it hardly can reveal anything new about the subject. All the information about the rules of classification must be given by the investigator. In the method based on Discriminant Analysis no rules for classification are given in advance. The method itself has to generate those rules within a frame of a sophisticated statistical model. Then there is always a *possibility* that something new pertaining to the relations between the groups will be brought to light.

In our example speaking about such a possibility may be nonsense, but let us suppose that we should have some related dialects instead of quite different languages. An interesting question might be: "What is the effect of geographical relations (distances, etc.) between different areas to the linguistic relations between them?"

The simplest way to answer that question by the aid of Discriminant Analysis is to compare the "map of languages" to the geographical map. If the main differences can be explained by geographical relations, these two maps have to be similar to some extent. It is also valuable that this technique makes it possible to compare dialects as a whole and not only using a few key words.

The method based on Discriminant Analysis is also superior to deterministic methods in its "sense of humour". For instance, it is not very sensitive to "printing errors" and other slight modifications in word structure. Let us take as an example the word ALWAYS. In our investigation ALWAYS was English with probability one; but if it is written in the form ÅLWAYS, it is still English with probability .89, and probability for Swedish is only .11. A deterministic method having rules like "Å will appear only in Swedish", "Diphthong AY does not appear in Swedish" would have more trouble in such cases.

References

[1] Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*. New York, John Wiley and Sons, Chapter 6.

[2] Cooley, W.W-Lohnes, P.R. (1962) *Multivariate Procedures for the Behavioral Sciences*. New York, John Wiley and Sons, Chapter 6 and 7.

[3] Kendall, M.G. (1957) *A Course in Multivariate Analysis*. London, Charles Griffin and Co., pp. 105-116.

[4] Rao, C.R. (1952) *Advanced Statistical Methodes in Biometric Research*. New York, John Wiley and Sons, Chapter 8.

[5] Sebestyen, G.S. (1962) *Decision-Making Processes in Pattern Recognition*. New York, The Macmillan Company.