ON BUILDING AN AUTOMATIC TEXT CLASSIFICATION MODEL WITH MINIMAL COMPUTATIONAL COSTS

NEREA MARTÍNEZ AROCA
Universidad de Murcia

ABSTRACT. *The creation of large volumes of texts and text databases in electronic form has been the result of the recent and rapid expansion of the WWW. Nevertheless, the major problem keeps on being the difficulty of accessing relevant information on a particular topic. Automated text categorisation or text classification has raised a great interest in the last decades for its applicability to Internet -as well as to other fields such as document organisation and word sense disambiguation (Sebastiani 1999: 4). It is the aim of the current paper to try and design a model of automatic text classification which allows text category discrimination as a prior step to new case assignment to previously established text categories on the basis of a series of linguistic and easily computable parameters and thus, reduced computational costs. For the purposes of the present pilot study we searched for those linguistic features which have been found to be reliable style markers, and may thus discriminate well among the categories analysed in our corpus -Cooking recipes, Ecology, Music, Oncology, Physics and Religion- and which can also be computed from unnanotated text.*

KEYWORDS: *automated text classification, discriminant analysis, classification functions.*

RESUMEN. *La creación de grandes volúmenes de textos y bases de datos en formato electrónico es resultado de la reciente y rápida expansión de Internet. El mayor problema, no obstante, reside todavía en las dificultades de acceso a información relevante en cualquier ámbito. La clasificación automática de textos nace de la necesidad imperante de organizar documentos, agruparlos, para su posterior eficaz recuperación. El objetivo del presente trabajo es diseñar un modelo de clasificación automática de textos, que permita primero diferenciar entre categorías de textos, para posteriormente asignar nuevos casos a categorías de textos preestablecidas, en base a una serie de parámetros lingüísticos y computables de fácil observación, y por lo tanto, de reducidos costes en términos de esfuerzo. Para ello, en este trabajo, presentamos un estudio piloto llevado a cabo en busca de aquellas características lingüísticas de entre las consideradas marcadores estilísticos fiables que puedan discriminar entre las categorías de textos recogidas en nuestro corpus -Recetas de cocina, Ecología, Música, Oncología, Física y Religión- y que no precisen de texto etiquetado.*

PALABRAS CLAVE: *clasificación automática de textos, análisis discriminante, funciones de clasificación.*

1. TEXT CLASSIFICATION STUDIES

Computational stylistics has widely placed its research focus on two applications of text categorisation, namely classification of texts in terms of genre and in terms of the style of authors. Aries (2005: 61) stated in relation to this that "Style […] can be approached at least from two angles: as a macro property of full texts (and/or collections of texts), something that can only be predicated of a large collection, or as a micro property that is operational in every minor linguistic choice a speaker or writer makes".

This has given rise to two major text classification areas: studies of authorship attribution (Stamatatos et al. 1999; Uzuner & Katz 2005a, b; Chaski 2005, etc.) vs. studies of genre classification (Besnier 1988; Biber 1988, 1989; Biber et al. 1994; Guinovart 2000; Karlgren & Cutting 1994; Kessler et al. 1997; Stamatatos et al. 2000b; etc.).

An early line of research in genre classification is that started by Biber (1985), followed by Besnier (1988), Biber (1988, 1989, 1995), Kim and Biber (1994) and Gómez Guinovart and Pérez Guerra (2000) among others. All these studies focus on the differences between spoken and written registers, using a multidimensional approach -MD Analysis hereafter-, whereby textual variation and linguistic relations across genres in a language are established

on the basis of co-occurrence patterns of linguistic features, interpreted as dimensions. In Biber (1988) an innovative notion of Dimension is introduced, whereby linguistic variation is described multidimensionally -under the assumption that no single dimension can capture the similarities and differences among genres.

A common core in these multi-dimensional analyses in genre classification is the set of features used, which are classified according to functional criteria, representing several major grammatical and functional characteristics. They include syntactic and lexical features which are quantitative and susceptible of being counted with the aid of a tagger. However, the main inconvenience of the features proposed by the studies following this approach, has to do with structural markers since the use of parsed or tagged text is required for automatic recognition and counting of features such as *passive counts*, *nominalization counts*, or *syntactic categories counts*. As a result, some of the features used in Biber were calculated by computational tools and the remaining were counted manually, but even the automatically obtained measures had to be checked manually.

Interesting findings of these studies on genre detection and classification concern the importance of the different sets of style markers or linguistic features in relation to their contribution to the classification process. In fact, both lines offer encouraging results so as to text classification using linguistic variables more easily computed than syntactic ones. As shown in authorship attribution studies (Uzuner & Katz 2005a, b, etc.), the contribution of syntax is reduced if taken into account its computational cost.

2. METHODOLOGY

2.1. *Corpus*

The present corpus is composed of 6 text categories as shown in *Table 1.* This is as we shall refer to them hereon, following Lee (2001: 49):

> We can see that the categories to which texts have been assigned in existing corpora are sometimes genres, sometimes subgenres, sometimes 'super-genres' and sometimes something else. This is undoubtedly why the catch-all term 'text category' is used in the official documentation for the LOB and ICE-GB corpora.

For the corpus compilation, 4 written texts were collected from different websites and divided into training corpus -3 text samples of each category were included to obtain the training sample -and test corpus -composed of a text sample of each text category. The training corpus is made up of those documents that will be used to generate the model of classification. Thus, its aim is to train the text classifier on the basis of the linguistic data of such a corpus. The documents in the test corpus will be used to evaluate the accuracy of the classification model.

CODE	TEXT CATEGORY	Nº OF WORDS	Nº OF TEXTS	Training	Test
G 01	Cooking recipes	1,500	4	3	1
G 02	Ecology	1, 500	4	3	1
G 03	Music	1, 500	4	3	1
G 04	Oncology	1, 500	4	3	1
G 05	Physics	1, 500	4	3	1
G 06	Religion	1, 500	4	3	1

Table 1. *Corpus composition*

2.2. *Variables*

The variables used in the present study have been selected under the criteria that these be linguistic, quantitative and that they can be counted easily, which for our purpose means that no tagged or parsed text is required. Thus, syntactic and structural features have been kept out of the current study. The variables used for the current study belong to three groups: Punctuation variables, Lexical Distribution variables and Most frequent words of the BNC[1].

As Stamatatos et al. (2000b) show, there are cases where the frequency of occurrence of a certain punctuation mark could be used alone for predicting a certain text genre. For example, an interview is usually characterized by an uncommonly high frequency of question marks. The 8 variables of Punctuation used in the current study are: *Periods/1000 words; Commas/1000 words; Semicolons/1000 words; Colons/1000 words; Dashes/1000 words; Pairs of parentheses/1000 words; Exclamation marks/1000 words; Question marks/1000 words.*

Within the Lexical Distribution variables, we have studied 2 measures of Sentence Length -*words/sentence* and *characters/sentence-;* 5 measures of Vocabulary Richness -namely, the *Standardised Type/Token Ratio*, *Word Length in orthographic letters, Words>6 characters, Hapax Legomena* and *Hapax Dislegomena-;* and 2 measures of Readability Grades -*Automated Readability Index* and *Coleman-Liau Index.*

The frequencies of occurrence of the most frequent words is a variable commonly used in authorship attribution and genre classification studies, either considering the most frequent words of a training corpus (Burrows 1987) or the most frequent words of the entire language, as represented by the BNC in Stamatatos (2000b) under the assumption that these are more reliable discriminators of text genre.

PUNCTUATION VARIABLES								
1. PERIODS	2. COMMAS	3. SEMICOLONS	4. COLONS					
5. HYPHENS	6. PARENTHESES	7. EXCLAMATIONS	8. QUESTIONS					
LEXICAL DISTRIBUTION MEASURES								
9. WORDS/SENTENCE	10. CHARACTERS/SENTENCE	11. STANDT.TTR						
12. WORD LENGTH	13. LONG WORD COUNT	14. HAPAX LEGOMENA						
15. HAPAX DISLEGOMENA	16. AUTOMATED READ. INDEX	17. COLEMAN-LIAUINDEX						
FREQUENCY OF OCCURRENCE OF THE 30 MOST FREQUENT WORDS[2] OF THE BNC								
18. THE	19. OF	20. AND	21. A	22. IN	23. TO	24. IS	25. WAS	26. IT
27. FOR	28. WITH	29. HE	30. BE	31. ON	32. I	33. THAT	34. BY	35. AT
36. YOU	37. 'S	38. ARE	39. NOT	40. HIS	41. THIS	42. FROM	43. BUT	44. HAD
45. WHICH		46. SHE	47. THEY					

Table 2. *Variables analysed*

The statistical analysis techniques used were Cluster Analysis -hereafter CA- and Discriminant Function Analysis -hereafter DA-, within the SPSS[3] statistical package.

CA is used as an exploratory technique to look for a structure of the data observed. It works so that texts under analysis are grouped into clusters in such a way that all members within each cluster are maximally similar to each other in the sharing of common properties, while each cluster is maximally distinct from the others.

DA is a multivariate analysis technique. Here, it will be used as a validation technique for CA output. It both verifies that clusters identified by CA are real and decides to which cluster a new subject observed should be assigned. DA works on a set of precategorized object-known groups and their values along a set of parameters to work out a set of discriminant functions that distinguishes between groups and allows prediction of group membership of new individuals based on their parameter scores.

CA was applied to the results obtained in the 47 variables analysed in the six text categories of our training corpus. It was carried out in seven steps: considering each group of variables, the three groups of variables at once and combining two of the three groups of variables.

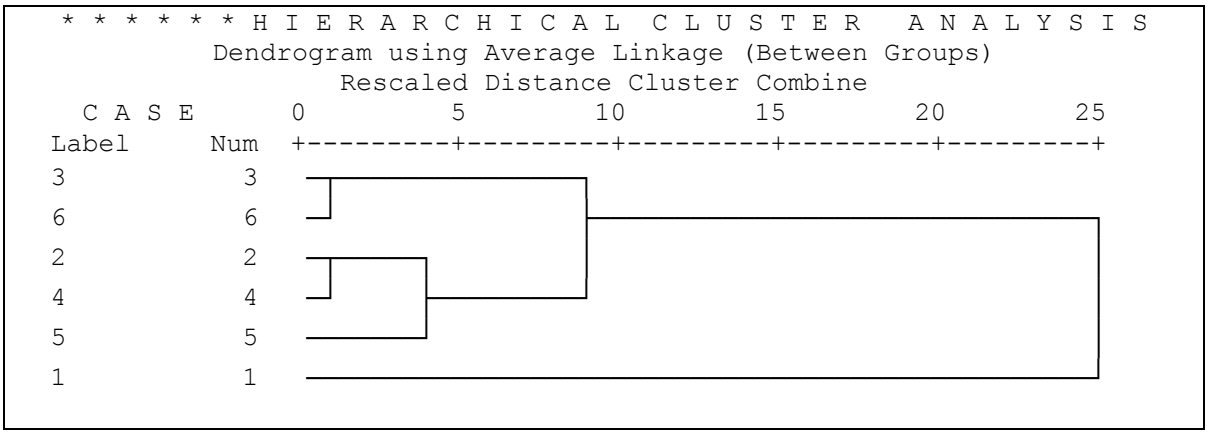
A final step in our research was to apply DA to our corpus data with the aim of assessing the relative importance of the independent variables in classifying the dependent variable.

3. ANALYSIS AND RESULTS

Although the results are quite satisfactory, the seven Cluster Analyses show problems in discriminating between cases 3 and 6. Thus, *Dendogram 1* and its corresponding Matrix of dissimilarities -*Table 4*- show that the greatest proximity, hence the greatest similarity is to be found between cases 3 and 6 with a 4645,749 Squared Euclidean Distance.

Case Number	Text Category
1	Cooking Recipes
2	Ecology
3	Music
4	Oncology
5	Physics
6	Religion

Table 3. *Correspondences of Case Numbers and Text Categories*



Dendogram 1.*Clustering of text category samples using 47 variables*

Case	Squared Euclidean Distance					
	1:1	2:2	3:3	4:4	5:5	6:6
1:1	,000	35654,676	62578,922	57998,652	29924,725	61056,351
2:2	35654,676	,000	9572,293	5133,855	7067,260	19180,701
3:3	62578,922	9572,293	,000	11467,088	24152,547	4645,749
4:4	57998,652	5133,855	11467,088	,000	15929,141	22985,064
5:5	29924,725	7067,260	24152,547	15929,141	,000	35705,489
6:6	61056,351	19180,701	4645,749	22985,064	35705,489	,000

Table 4. *Matrix of dissimilarities of the text sample categories*

Since CA did work in the current research, it was decided to use DA and apply it to our corpus to validate the results obtained with CA.

For the construction of such a model, the test corpus samples were included. *Table 5* shows the results of DA applied to the 12 text samples[4] -both training and test corpus- considering all the independent variables identified.

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
group	Squared Mahalanobis Distance	2,97	2,97	2,99	2,63	1,63	1,8	2,97	2,97	2,99	2,63	1,63	1,8
2nd Highest	Group	3	4	6	2	3	3	3	4	6	2	2	3
	Squared Mahalanobis Distance	115,6	14,93	6,2	8,22	23,68	5,66	99,25	11,62	15,33	17,64	21,11	13,48
Discriminant scores	Function 1	11,29	-4,05	1,29	-4,61	-2,91	1,48	10,55	-5,75	,59	-3,1	-4,85	,06
	Function 2	-1,08	-2,25	,143	-2,99	1,918	1,301	-1,07	-2,25	1,196	-,317	2,22	3,2
	Function 3	-,308	-1,37	1,128	1,465	-1,73	1,196	-,534	-,650	-2,05	2,096	-1,21	1,986
	Function 4	-1,20	1,685	1,126	-,309	-1,33	,018	,513	-,955	,665	-,604	-,087	,488
	Function 5	1,395	,263	,352	,651	-,393	-,784	-1,49	-,968	,487	-,102	,526	,065

Table 5. *Discriminant function analysis considering Punctuation Variables, Lexical Distribution measures and Frequency of occurrence of the 30 most frequent words from the BNC*

The most outstanding results of DA are that whether we consider only one, the three or combinations of two out of the three groups of variables, the new six cases introduced are successfully assigned to their real group, showing thus, variables are reliable classifiers (*Tables 6-11* below*).*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual group		1	2	3	4	5	6	1	2	3	4	5	6
Highest group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis Distance	2,996	2,798	2,500	2,669	1,037	2,999	2,996	2,798	2,500	2,669	1,037	2,999
2nd Highest	Group	6	4	4	3	2	3	6	4	4	3	2	3
	Squared Mahalanobis Distance	11140	212,6	47,22	37,87	725,3	355,9	11251	275,7	41,05	50,73	707,3	319,6
Discriminant scores	Function 1	110,5	-29,81	-13,50	-14,05	-56,77	5,607	111,	-31,77	-12,59	-16,55	-56,75	4,663
	Function 2	-1,572	,693	4,489	-,646	-5,520	2,120	-3,036	-,973	4,858	-,400	-4,584	4,572
	Function 3	1,060	-1,882	2,650	-1,488	2,334	,958	-,676	-2,721	,115	-1,784	1,262	,171
	Function 4	1,325	-,435	-,258	,184	-,403	-,007	-1,153	,309	1,350	,296	,291	-1,499
	Function 5	,417	-,515	,006	,549	,663	-,860	-,271	1,313	,189	-1,515	-,617	,640

Table 6. *Casewise statistics considering Punctuation Variables*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis	2,975	1,885	2,281	2,492	2,768	2,598	2,975	1,885	2,281	2,492	2,768	2,598
2nd Highest	Group	6	5	5	2	2	1	6	5	5	2	2	1
	Squared Mahalanobis	93,27	9,634	15,30	14,30	7,915	140,7	152,3	8,713	23,50	38,24	12,19	104,0
Discri minant	Function 1	-9,922	4,398	3,540	7,384	3,194	-7,847	-12,15	4,158	4,296	9,333	1,593	-7,968
Scores	Function 2	-3,704	1,023	-,580	-,695	-1,017	6,514	-5,900	1,358	-,866	-1,717	,748	4,835
	Function 3	,613	,536	-2,197	1,699	1,942	-,228	-,026	1,713	-3,417	-,237	-,285	-,112
	Function 4	-,458	-,302	,336	,303	-,870	1,646	,784	,319	-,582	1,447	-1,532	-1,090
	Function 5	-,222	-1,634	-1,306	,382	,199	,231	,136	,732	1,172	,115	,209	-,016

Table 7. *Casewise statistics considering Lexical Distribution measures*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual Group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis Distance	2,970	2,973	2,992	2,632	1,632	1,800	2,970	2,973	2,992	2,632	1,632	1,800
2nd Highest	Group	3	4	6	2	3	3	3	4	6	2	2	3
	Squared Mahalanobis Distance	115,6	14,93	6,200	8,229	23,68	5,668	99,25	11,62	15,33	17,64	21,11	13,48
Discrim inant	Function 1	11,29	-4,051	1,294	-4,616	-2,914	1,485	10,55	-5,750	,593	-3,104	-4,853	,061
	Function 2	-1,081	-2,258	,143	-2,999	1,918	1,301	-1,076	-2,254	1,196	-,317	2,227	3,201
	Function 3	-,308	-1,373	1,128	1,465	-1,732	1,196	-,534	-,650	-2,056	2,096	-1,218	1,986
	Function 4	-1,204	1,685	1,126	-,309	-1,336	,018	,513	-,955	,665	-,604	-,087	,488
	Function 5	1,395	,263	,352	,651	-,393	-,784	-1,491	-,968	,487	-,102	,526	,065

Table 8. *Casewise statistics considering Frequency of occurrence of the 30 most frequent words from BNC*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis distance	2,996	2,798	2,500	2,669	1,037	2,999	2,996	2,798	2,500	2,669	1,037	2,999
2nd Highest	Group	6	4	4	3	2	3	6	4	4	3	2	3
	Squared Mahalanobis distance	11140	212,6	47,22	37,87	725,3	355,9	11251	275,7	41,05	50,73	707,3	319,6
Discriminant	Function 1	110,5	-29,81	-13,50	-14,05	-56,77	5,607	111,1	-31,77	-12,59	-16,55	-56,75	4,663
	Function 2	-1,572	,693	4,489	-,646	-5,520	2,120	-3,036	-,973	4,858	-,400	-4,584	4,572
	Function 3	1,060	-1,882	2,650	-1,488	2,334	,958	-,676	-2,721	,115	-1,784	1,262	,171
	Function 4	1,325	-,435	-,258	,184	-,403	-,007	-1,153	,309	1,350	,296	,291	-1,499
	Function 5	,417	-,515	,006	,549	,663	-,860	-,271	1,313	,189	-1,515	-,617	,640

Table 9. *Casewise statistics considering Punctuation Variables and Lexical Distribution measures*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis Distance	2,996	2,798	2,500	2,669	1,037	2,999	2,996	2,798	2,500	2,669	1,037	2,999
2nd Highest	Group	6	4	4	3	2	3	6	4	4	3	2	3
	Squared Mahalanobis Distance	11140	212,6	47,22	37,87	725,3	355,9	11251	275,7	41,05	50,73	707,3	319,6
Discriminant Scores	Function 1	110,5	-29,81	-13,50	-14,05	-56,77	5,607	111,0	-31,77	-12,59	-16,55	-56,75	4,663
	Function 2	-1,572	,693	4,489	-,646	-5,520	2,120	-3,036	-,973	4,858	-,400	-4,584	4,572
	Function 3	1,060	-1,882	2,650	-1,488	2,334	,958	-,676	-2,721	,115	-1,784	1,262	,171
	Function 4	1,325	-,435	-,258	,184	-,403	-,007	-1,153	,309	1,350	,296	,291	-1,499
	Function 5	,417	-,515	,006	,549	,663	-,860	-,271	1,313	,189	-1,515	-,617	,640

Table 10. *Casewise statistics considering Punctuation Variables and Frequency of occurrence of the 30 most frequent words from the BNC*

		Cases											
		1	2	3	4	5	6	7	8	9	10	11	12
Actual Group		1	2	3	4	5	6	1	2	3	4	5	6
Highest Group	Predicted Group	1	2	3	4	5	6	1	2	3	4	5	6
	Squared Mahalanobis Distance	2,970	2,973	2,992	2,632	1,632	1,800	2,970	2,973	2,992	2,632	1,632	1,800
2nd Highest	Group	3	4	6	2	3	3	3	4	6	2	2	3
	Squared Mahalanobis Distance	115,6	14,93	6,200	8,229	23,68	5,668	99,25	11,62	15,33	17,64	21,11	13,48
Discriminant scores	Function 1	11,29	-4,051	1,294	-4,616	-2,914	1,485	10,55	-5,750	,593	-3,104	-4,853	,061
	Function 2	-1,081	-2,258	,143	-2,999	1,918	1,301	-1,076	-2,254	1,196	-,317	2,227	3,201
	Function 3	-,308	-1,373	1,128	1,465	-1,732	1,196	-,534	-,650	-2,056	2,096	-1,218	1,986
	Function 4	-1,204	1,685	1,126	-,309	-1,336	,018	,513	-,955	,665	-,604	-,087	,488
	Function 5	1,395	,263	,352	,651	-,393	-,784	-1,491	-,968	,487	-,102	,526	,065

Table 11. *Casewise statistics considering Lexical Distribution measures and Frequency of occurrence of the 30 most frequent words from the BNC*

On account of achieving 100% accuracy rate, we decided to go a bit further and look into the feature reduction field, with the aim of finding a subset of the features analysed which would allow the classification task at target in the most optimum way.

Stepwise DA was used to select those variables with a greater discriminant capacity to assign cases to a priori defined groups and to generate a predictive discriminant model to classify new cases thanks to the classification functions. Stepwise DA reduced the number of variables from the original 47 variables used to a subset of only 6 variables -*Table 12*-.

			Step					
			1	2	3	4	5	6
Introduced			Automated Readability Index	Question marks/1000 words	THEY	FROM	AT	Hyphens/1000 words
Wilkins Lambda	Statistic		,082	,007	,000	,000	,000	,000
	gl1		1	2	3	4	5	6
	gl2		5	5	5	5	5	5
	gl3		6,000	6,000	6,000	6,000	6,000	6,000
	Exact F	Statistic	13,503	11,029				
		gl1	5	10				
		gl2	6,000	10,000				
		Sig.	,003	,000				
	Approximate F	Statistic			17,718	23,789	34,474	38,303
		gl1			15	20	25	30
		gl2			11,444	10,900	8,932	6,000
		Sig.			,000	,000	,000	,000

Table 12. *Feature Subset Selection*

As for the classification functions, these allow computation of classification scores for some new observations. Once we have computed the classification scores for a case, we decide to classify it in general as belonging to the group for which it has the highest classification score.

These classification functions have the following formula:

$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + ... + w_{im}*x_m$

	Corpus					
	1	2	3	4	5	6
Hyphens/1000 words	-22,571	-35,184	11,647	-14,630	-9,271	-85,460
Question marks/1000 words	-565,143	-343,892	7496,632	965,894	1299,538	-5724,540
Automated Readability Index	345,959	600,819	643,899	388,236	338,109	874,021
AT	-382,028	-636,455	-369,194	-337,577	-277,929	-1176,966
FROM	682,873	1157,162	921,669	669,045	566,576	1945,606
THEY	714,537	1113,508	-339,787	451,799	306,758	2699,965
	2222,906	6602,755	11917,5	2926,466	2314,242	15826,9

Table 13. *Coefficients of the classification functions*

4. DISCUSSION AND CONCLUSIONS

Bergo (2001: 8) argues the main task of text categorisation is "how can documents be assigned to a category with a highest possible chance of being correct without assigning too many incorrect categories, and at acceptable computational costs".

With this pilot study we have searched for a model of classification based on a feature selection under the criterion of lowering computational costs while maintaining high accuracy rates in case assignment to group belonging. Results have shown that the set of linguistic variables proposed, in addition to being easily identified and computed -in contrast to other linguistic features used in research studies in the area that require manual supervision because of ambiguity (see Biber 1995: 85)-, can accurately discriminate among text categories as seen through CA. Furthermore, DA offered a 100% accurate classification of text samples into the categories analysed. Interestingly enough, all groups of variables achieved a 100% accuracy rate with no cases misclassified.

As for the set of variables, our major finding was the arrival at a classification model based on the combination of the highest discriminating variables which provided us with 6 discriminating functions, one for each text category analysed.

Given the reduced applicability of the model because of the size of our training corpus, in future research we will aim at increasing our training set so as to allow profile extraction of the text categories during the learning phase. We will therefore try to increase the number of text categories so as to try to cater for the greatest text classification range. For such an aim, terminological revision will have to be approached so as to make quite clear the basis of future groups in text categorisation. We are aware that not only do groups need increasing but group representation as well, that is, more work needs to be done in the analysis of as many samples from each text category as possible with the aim of enhancing feature reduction.

NOTAS

1. The BNC is the British National Corpus, a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. http://www.natcorp.ox.ac.uk/
2. This list was taken from a non-lemmatised list of the most frequent words of the BNC, retrieved from http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html
3. Statistical Package for the Social Sciences
4. Attention must be drawn to the fact that we have basically compared just two parallel sets of text samples, i.e., there has only been one test sample of each text category for DA to classify against the values of another single text sample -the training sample-, though the training sample was made up of three written texts.

REFERENCES

Aires, R., Aluísio, S. & Santos, D. (2005). "User-aware page classification in a search engine". [Document available on the Internet at http://eprints.sics.se/26/01/style2005.pdf]

Bergo, A. (2001). "Text categorization and prototypes". [Document available on the Internet at http://www.illc.uva.nl/Publications/ResearchReports/ MoL-2001-08.text.pdf]

Besnier, N. (1988). "The linguistic relationships of spoken and written Nukulaelae registers". *Language* 64: 707-36.

Biber, D. (1985). "Investigating macroscopic textual variation through multifeature/ multidimensional analyses". *Linguistics* 23: 337-360.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1989). "A typology of English texts". *Linguistics* 27: 3-43.

Biber, D. (1995). *Dimensions of register Variation: A cross linguistic comparison*. Cambridge: Cambridge University Press.

Burrows, J. F. (1987). "Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style". *Literary and Linguistic Computing* 2:2: 61-70.

Cantos, P. (2000). "Investigating Type-token Regression and its Potential for Automated Text-Discrimination". *Cuadernos de Filología Inglesa* 9:1: 71-92.

Chaski, C. (2005). "Computational Stylistics in Forensic Author Identification". [Document available on the Internet at http://eprints.sics.se/26/01/style2005.pdf]

Gómez Guinovart, X. & Pérez Guerra, J. (2000). "A multidimensional corpus-based analysis of English spoken and written-to-be-spoken discourse". *Cuadernos de Filología Inglesa* 9.1: 39-70.

Karlgren, J. & Cutting, D. (1994). "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis". [Document available on the Internet at http://eprints.sics.se/56/01/cmplglixcol.pdf]

Kessler, B., Nunberg G. & Schutze, H. (1997). "Automatic Detection of Text Genre". [Document available on the Internet at http://acl.ldc.upenn.edu/P/P97/P97-1005.pdf]

Lee, D. (2001). "Genres, Registers, Text-Types, Domains and Styles: Clarifying the Concept and Navigating a Path through the BNC Jungle". *Language Learning & Technology* 5.3: 37-72.

Quirk, R., Greenbaum, S., Leech, G. & Svartviket, J. (1985). *A Comprehensive Grammar of the English Language*. Longman: London.

Sebastiani, F. (2005). "Text categorization". *Text Mining and its Applications*. Ed. Alessandro Zanasi. Southampton: WIT Press. 109-129.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (1999). "Automatic Authorship Attribution". [Document available on the Internet at http://www.cs.mu.oz.au/acl/E/E99/E99-1021.pdf]

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2000a). “Automatic Text Categorisation in Terms of Genre and Author”. *Computational Linguistics* 26.4: 471-495.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2000b). “Text genre detection using common word frequencies”. [Document available on the Internet at http://www.cs.mu.oz.au/acl/C/C00/C00-2117.pdf]

Uzuner, Ö. & Katz, B. (2005a). “A Comparative Study of Language Models for Book and Author Recognition”. [Document available on the Internet at http://people.csail.mit.edu/ozlem/ijcnlp-05-UzunerO.pdf]

Uzuner, Ö. & Katz, B. (2005b). “Style vs. Expression in Literary Narratives”. [Document available on the Internet at http://people.csail.mit.edu/ozlem/sigir-05-cc-UzunerO-cr.pdf]