INVESTIGATING TYPE-TOKEN REGRESSION AND ITS POTENTIAL FOR AUTOMATED TEXT DISCRIMINATION

Pascual Cantos Gómez
Departamento de Filología Inglesa
Universidad de Murcia

ARSI'KACT

The motivation of the present paper is based on the intuition that the sole use of data on lexical density relative to text samples of various languages, authors, linguistic domains, etc., might be a potential indicator for automated text discrimination. In order to look for a reliable and valid lexical density index, we shall review and clarify the mathematical relationship between types (word forms) and tokens (words) by discussing and constructing adequate regression models that might help to differentiate text types from each other. Additionally, we shall use multivariate statistical models (cluster analysis and discriminant function analysis) to complement the mathematical lexical density regression model (TYT-formula).

KEY WORDS: *corpus linguistics, type-token regression, text typology, automated text classification.*

RESUMEN

La motivación del presente artículo nace de la intuición de que la sola utilización de la densidad léxica de muestras textuales pertenecientes a diferentes idiomas, autores, dominios lingüísticos, etc., puede ser potencialmente válida para discriminar textos de forma automática. Con el fin de encontrar un índice de densidad léxica válido y fiable, hemos revisado y clarificado la relación matemática entre tipos (formas) y tokens (palabras), para construir modelos de regresión adecuados que nos permitan distinguir tipos de textos. Por añadidura, hemos hecho uso de modelos estadísticos multivariantes (análisis de conglomerados y análisis discriminante) con el fin de complementar y optimizar el modelo matemático de regresión para la densidad léxica (la fórmula TYT).

PALABRAS CLAVE: *lingüística del corpus, regresión de tipos (formas) y tokens (palabras), tipología de textos, clasificación automática de textos.*

1. INTRODUCTION

Simple extracts given from frequency lists only show the entries for individual words. Most frequency software also produces useful totals and sometimes offers a range of statistics based on them. The most common totals calculated for word frequency lists are usually referred to as total tokens and total types and it is important to understand the distinction between them.

In this context. a token is an individual occurrence of any word form. The paragraph:

> *"Linguists may wonder why they need statistics. **The** dominant theoretical framework in the field, that of generative grammar. has us its primary data-source judgements about the well-formedness of sentences. These judgements usually come from linguists themselves, are either-or decisions. and relate **to** the language ability of an ideal native speaker in a homogeneous speech community".*

contains altogether 56 words or tokens. but these represent only 48 different word forms or types. The frequency list shows the number of tokens found for each type. In this case. the following 5 types have more than one token:

the	4
of	ɔ
in	3
judgements	3
linguists	2

Between them. these types account for 13 of the tokens. The other 43 types occur only once and make up the overall total of 56 tokens.

The distribution oi'tokens between the types in a text can provide a useful measure of the degree of lexical variety within it. and may even provide n starting-point for examining lexical differences between different types of text. styles. authors. etc. Several statistics can be calculated from the information contained in the list. The simplest is the ratio of tokens and types. in other words. the mean frequency of each different word form. In the case of the paragraph used above. this is 56 / 43 = 1.3. This index (1.3) indicates thnt each word form or type occurs on average 1.3 times. Similarly. the reverse can be illustrated. that is. once the amount of types and tokens relative of a text sample are known. we can calculate its lexical diversity or lexical density by dividing the total number of types by the total number of tokens: 43 / 56 = 0.76. If we eventually multiply this quotient by 100. then we get the mean percentage of different types per one hundred words of the text (76% in our example). Roth indices obtained here. the token-type and the type-token ratio. are not very significant and reliable. This is obviously caiised by the smallness of the test sample (just 56 tokens or words). Longer texts. such as four of Joseph Conrand's novels

(*Nigger of the 'Narcissus'*, *Lord Jim*, *Heart of Darkness* and *The Secret Agent*), result in a token-type ratio of 15.33 and the type-token ratio of 6.56%. These figures or ratios are affected by the overall number of tokens and types in the four novels (271.056 tokens and 17.795 types).

However, the reliability of the token-type and type-token ratio as quantitative indicators of lexical diversity or lexical density are constrained because of their dependence on text size - while test length (tokens) is theoretically unlimited, the number of different words in use (types) in a language is finite (Holmes 1994: 92). That is, while any linguistic corpus increases linearly in tokens in a completely regular or stable shape, its increase in types -though close to that of the tokens at the beginning- starts declining the more the corpus grows, as it contributes fewer new types. The cumulative tokens are distributed linearly, while the cumulative types are distributed curvilinearly (Biber 1993: 350: see *Fig. 1*).

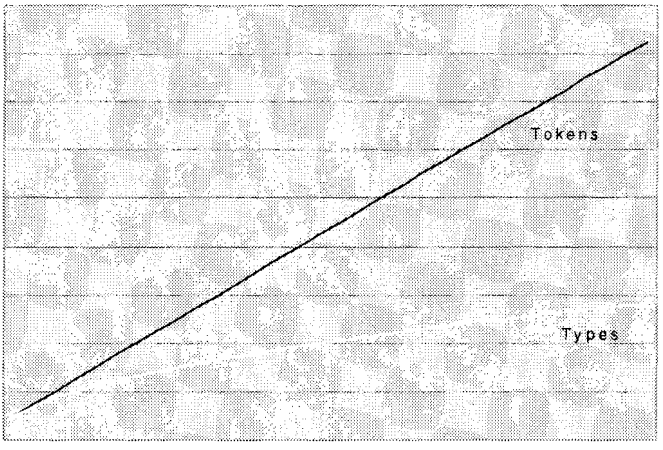

Fig. 1. Increase of tokens and types

Consequently, to overcome this reliability problem of the token-type and type-token ratio and compare tests or corpora with respect to their lexical density, the texts to be compared must be based on samples of the same size, disregarding the total length of the text or corpora. This ensures that the comparisons based on the token-type and/or token-type indices become somehow useful and relevant (Biber 1988: 238-9).

In what follows, we shall try to overcome this apparent reliability problem of the token-type relationship not by means of equalising the text samples to the same number of tokens but by means of investigating and determining the hyperbolic function of types relative to different text samples (linguistic domains, authors, etc.). We are confident that the type-token function is a positive indicator for discriminating text samples, in the belief that the non-linear growth of types is idiosyncratic and to some extent unique, depending heavily on the topics, authorship, etc., of the various text samples.

II. TYPE-TOKEN RELATIONSHIP

Our aim is to look tor a stable text independent index that determines the type-token relationship. The problem. as already outlined. is that tokens increase linearly and types do so in a curvilinear way (*Fig.* 1).

Regarding the increasing rate of word forms (types). Heaps (1978) reported that the following expression is true for a general English text of up to at least 20.000 words. where *D* (types) is related to the total number of tokens *N* by an equation relative to the way the text length increases:

$$D = kN^h \quad \text{hence. } log\, D = h\, log\, N + log\, k$$

and where k and h are constants that depend on the particular text sample. He emphasized the linear relation between *log* D and *log N* as taking common logarithms of both sides of $D = kN^h$. respectively. The purpose of his research was to create and manage index files efficiently for document retrieval. This explains why he experimented on a collection of title words of documents rather than on general English text (or corpora). Nevertheless. he did not give any explanation about how the equation was derived.

Note that Heaps just insisted that the expression above is true for general English text of "up to at least 20.000 words" rather than for texts of any size. This implies that the dependent constants or the expression itself might change as the corpus size greatly grows. In other words. even if we were to find a function that fits the given data (corpus). there is no certainty that the function would always hold.

A positive contribution to this issue can be found in Sánchez and Cantos (1997). These authors offer a detailed explanation on the type and lemma growths based on the observations of the *CUMBRE Corpus* (Corpus of Contemporary Spanish). They concluded that tokens represent a linear function ($y = ax$) and types a kind of hyperbolic function ($y = a\sqrt{x}$). The calculation of the slope a is straightforward. We just need a small sample and by means of any available concordance program get the overall tokens and types. For instance. if we assume x to be the tokens and y to be the types. we get. by simply instantiating the values obtained from. say a 250.000 token sample with 26.812 types. the following:

$$26.812 = a\sqrt{250.000}$$
$$a = 53.624$$

Now. we have to make sure that this constant value a is indeed reliable in calculating the number of types from a given number of tokens. To check this. Sánchez and Cantos defined and applied the *type-token formula* (*TYT-formula*. hereafter):

$$y = 53.624\sqrt{x}$$

where y stands for types and *x* for tokens. and compared the results with those obtained from real evidence. the *CUMBRE Corpus. Table 1* below shows eight samples chosen (1.000.000. 2.000.000. 3.000.000. 4.000.000. 5.000.000. 6.000.000. 7.000.000 and 8.000.000 tokens).

The estimation results obtained by means of our *TYT-formula* are quite close to the real ones: corpus-based. The differences between the real data and the estimations range from +4.761% to -1.353%. which translated into total figures goes from +3611 to -1778 types.

Tokens	Types (Corpus-based)	Types (Based on Estimation)	Difference (Corpus vs Estimation)	Difference in % (Corpus vs Estimation)
1.000.000	54.298	53.624	674	1.256
2.000.000	79.446	75.835	3611	4.761
3.000.000	95.764	92.879	2885	3.106
4.000.000	106.783	107.248	-465	-0.433
5.000.000	119.059	119.906	-847	-0.706
6.000.000	129.573	131.351	-1778	-1.353
7.000.000	140.283	141.875	-1592	-1.122
8.000.000	150.871	151.671	-800	-0.527

Tab. 1 Corpus-based data vs data based on estimation (CUMBRE Corpus)

It is noteworihy that the estimations are just based on a single *a-value* obtained from a subcorpus. that is only 250.000 tokens. This gives an idea of the reliability and validity of the formula. The *a-value*. though based on just 250.000 tokens. showed a great deal of accuracy in the projection of various multi-million token samples. This *a-value* is the sort of parameter that tells the function $y=\sqrt{x}$ the initial slope the curve is to have from 250.000 tokens on.

The *TYT-formula* has undergone thorough testing and severa1 more trials were undertaken. taking various samples from specific sublanguages. namely. press and general Iiction.

The tests were carried out by means of four 250.000 word samples from newspaper and general fiction language. The *a-vtrlzies* (press: 56.17; and general Iiction: 51.45) were obtained calculating the mean of all *a-values* of the various samples (press: 56.12 for 250.000. 56.48 for 500.000.56.34 for 750.000 and 55.74 for 1.000.000 tokens; general Iiction: 50.77 for 250.000. 51.45 for 500.000.52.29 for 750,000 and 51.32 for 1.000.000 tokens). The results for the real corpus data and the projections are given in *Tables* 2 and *3* below. The striking similarities between the real data and the estimated ones confirm once more that the formula is indeed reliable for calculating the types from a piven number of tokens. and shows that tokens and types are functionally dependent on each other. This dependency can be mathematically modeled even before compilinp any corpus.

Tokens	Corpus-based			Estimation-based			Corpus *versus* Estimation	
	Types	Increase (in Typ)	Increase (in %)	Types	Increase (in Typ)	Increase (in %)	Diff. in Types	Diff. in %
250.000	28.060	28.060	-	28.085	28.085	-	-25	-0.089
500.000	39.937	11.877	42.32	39.718	11.633	41.42	219	0.551
750.000	48.799	8.862	22.18	48.644	8.926	22.47	155	0.318
1.000.000	55.740	6.941	14.22	56.170	7.526	15.47	-430	-0.765

Tab. 2 Testing the *TYT-Formula* with press saiiiples

Tokens	Corpus-based			Estimation-based			Corpus *versus* Estimation	
	Types	Increase (in Typ)	Increase (in %)	Types	Increase (in Typ)	Increase (in %)	Diff. in Types	Diff. in %
250.000	25385	25385	-	25728	25728	-	-343	-1.333
500000	36380	10995	43.31	36380	10652	41.4	0	0
750.000	45284	8904	24.47	44557	8177	22.47	727	1631
1.000.000	51320	6036	13.32	51450	6893	15.47	-130	-0.252

Tab. 3 Testing the *TYT-Formula* with general fiction saiiiples

Similar test were performed not just for Spanish but also for English:

Tokens	Types (Corpus-based)	Types (Based on Estimation)	Difference (Corpus *vs* Estimation)	Difference in % (Corpus *vs* Estimation)
250.000	20.715	20.940	-225	-1.09%
500.000	29.202	29.613	-411	-1.41%
750.000	35.974	36.269	-295	-0.82%
1.000.000	42.130	41.880	250	0.59%
1.250.000	45.101	46.823	-1722	-3.82%
1.500.000	50.863	51.292	-429	-0.84%
1.750.000	55.653	55.402	251	0.45%
2.000.000	61.079	59.227	1852	3.03%
2.250.000	62.970	62.820	150	0.24%
2.500.000	65.190	66.218	-1028	-1.58%

2.750.000	69.234	69.450	-216	-0.31%
3.000.000	72.953	72.538	415	0.57%
3.250.000	78.312	75.500	2812	3.59%
3.500.000	78.855	78.350	505	0.64%
3.750.000	81.061	81.100	-39	-0.05%
4.000.000	84.080	83.760	320	0.38%

Tab. 4 Corpus-based data vs data based on estimation (English Corpus)

The evidence of the experimental results allows us to state that frequencies of different types are not only distributed 'curvilinearly' (Biber 1993: 250), but are distributed in a predictable way, that is, they are subject to mathematical modelling. We can still go further and say that if the relationship between types and tokens holds then we might be able to construct regression models (for a detailed discussion on the adequacy of regression models for type/lemma predicition see Yang, Cantos and Song forthcoming).

III. TYPE-TOKEN REGRESSION

In order to construct a regression model, both the information which is going to be used to make the prediction and the information which is to be predicted must be obtained from a corpus sample. The relationship between the two pieces of information is then modelled with a linear transformation. Then in the future, only the first information is necessary, and the regression model is used to transform this information into the predicted. In other words, it is necessary to have information on both variables (types and tokens) before the model can be constructed.

A notional scheme is now necessary to describe the procedure:

- *x* is the variable used to predict, and is sometimes called the independent variable. In our case, it would be the ainount of tokens.
- *y* is the observed value of the predicted variable, and is sometimes called the dependent variable. It would be the total types.
- *y'* is the predicted value of the dependent variable. It would be the predicted nuinber of types.

The goal in regression models is to create a model where the predicted *y'* and the observed *y* values of the variable to be predicted are as similar as possible. The more similar the values, the better the model.

A visual representation of the relationship between the *x* and *y* variables produces a

regression line or linear relationship between x and y. taking normally tlie form of a straight line. In general. any algebraic relation of the form

$$y = \alpha + \beta x$$

will have a graph which is a straight line. The quantity of β is called the slope or gradient *o!'* the line and α is often referred toas the intercept or intercept on the y-axis. The values of α and β remain fixed. irrespectively of the values of x and y.

If we observe. however. the type-token slopes obtained. we realise that the relationship between x and y is not linear but curvilinear. For example. see tlie type growth for English (Fig. 2).

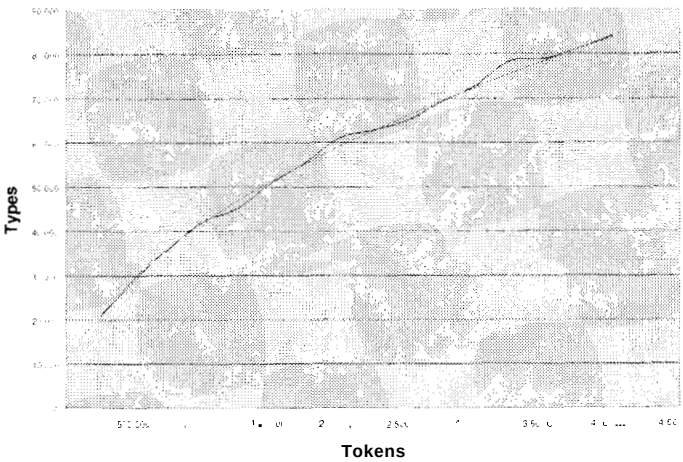

Fig.2 Real vs estimated type-growth

What we need to do here is to linearize the type-token relationship by means of transforming the data. This mathematical transformation allows the dala to fit better to simple repression models. *Figure* 2 shows that the relationship between the two variables x (tokens) and y (types) is clearly not a straight line. It is similar in shapr to curves which can be expressed by an equation of the form:

$$Y = AX^{b}$$

where A and B are constants or parameters. Now instead of y consider its logarithm. $log\ Y$:

$$log\ Y = log\ (AX^{b})$$
$$log\ Y = log\ A + b\ log\ X$$

If we write

$$W = log\ Y$$
$$Z = log\ X,\ and$$
$$a = log\ A$$

tlie equation can be written:

$$W = a + bZ$$

u-hich is exactly the form of the simple linear regression model. *Figure* 3 shows a graph of W ($log\ Y$) against Z ($log\ X$) for English and Spanish type growths and indicates a much more linear relationship than u-as apparent in the previous figure (*Fig.* 2). A linear regression could then be safely fitted to the logarithm of the original scores.

Fig. 3 Spanish versus English transformed type-token regression

The evidence of the experimental results allows us to state that frequencies of'different types are not only distrihuted 'curvilinearly' (Biber 1993: 230). but are distributed in a predictahle way. that is. they are subject to mathematical niodelling.

The analytic technique for predicting types applied by Sánchez and Cantos (1997) is simple and straightforward and tlie resulting formula

$$TYPES = K\sqrt{TOKENS}$$

is easy to use. flexible and can be applied quickly to any corpora or language samples. The practicality of this formula relies on its simplicity which -and this is important- goes hand in hand with its effectiveness and transparency. In particular. the *TYT-Formula* due to its thorough

testing on various text samples of various sizes. seems very reliable with a more than acceptable error niargin of ±5%. and this speaks eloquently of its validity.

The most positive contributions of the *TYT-Formula* can be summarised in the following points:

- It is a stable indicator of lexical diversity and lexical density.
- It overcomes the reliability flaw of both the token-type ratio and type-token one as it is not constrained or dependent on text lengtli.
- It can be used as a predictive tool to account for thr total amount of word forms (types) and lemmas any hypothetical corpus might contain (see Sanchez and Cantos 1998).

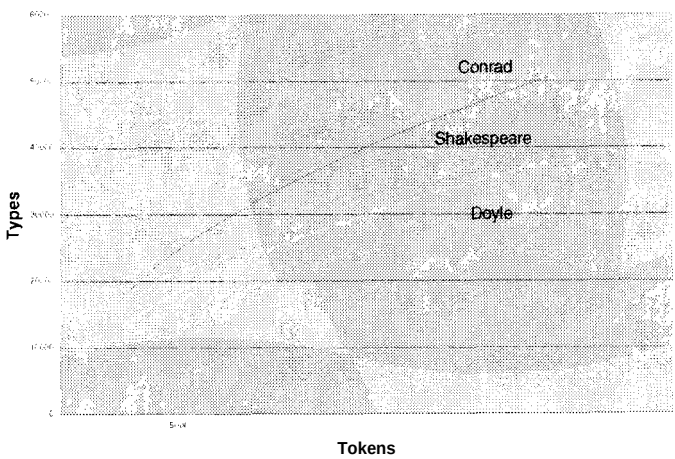

Fig. 4 Compared type slopes for Conrad, Shakespeare and Doyle

A revealing issue is that the application of the *TYT-Formula* on different text samples yields, giving idiosyncratic. unique and distinctive slopes. The contrastive graph above (*Fig. 4*) clearly reveals that. for example. Conrad's lexical density is superior to Doyle's and Shakespeare's. And this is further evidenced by their correspondent linear regression transformation niodels (*Fig. 5*).

This evidence suggests that the *TYT-Formula* might also be valid for text. author and language classifications. among others. In u-hat follows we shall experiment on this issue using the *CUMBRE Corpus* (Corpus of contemporary Spanish).

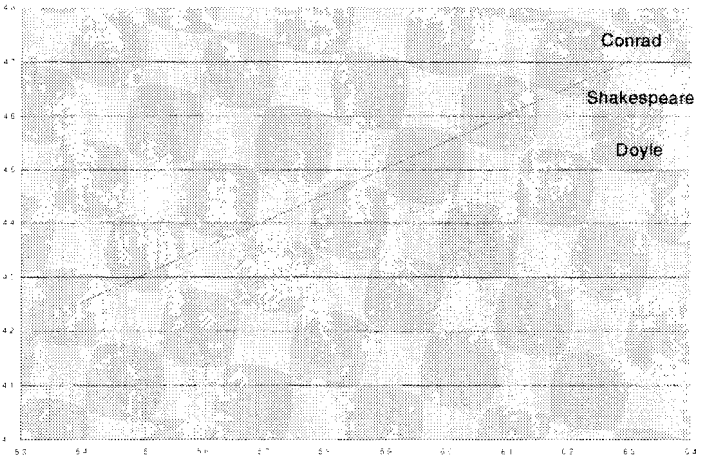

Fig. 5 Type-token regressions: Conrad, Shakespeare and Doyle

IV. COMPARING TYPE-TOKEN REGRESSIONS

In this experiment, we (a)extracted (from the *CUMBRE Corpus*) 11 different text samples from textbooks and manuals for secondary education and university level relative to various subjects or linguistic domains. (b) obtained their total amounts of tokens and types, and (c) calculated their *K-values* (constant value; see *TYT-Formula*). The results are illustrated below in *Table 5.*

Sample	Tokens	Types	*K-value*
Architecture	64431	11225	44.22
Chemistry	22539	2771	18.46
Computing	18822	2344	17.09
Geography	48544	7341	33.32
History	29711	5671	32.90
Mathematics	18700	1907	13.95
Medicine	39639	5228	26.26
Natural Sciences	41650	5982	29.31
Philosophy	20385	3344	23.42
Physics	15233	2378	19.27
Sociology	75149	11522	42.03

Tab. 5 Tokens, types and k-values relative to eleven linguistic domains (CUMBRE Corpus)

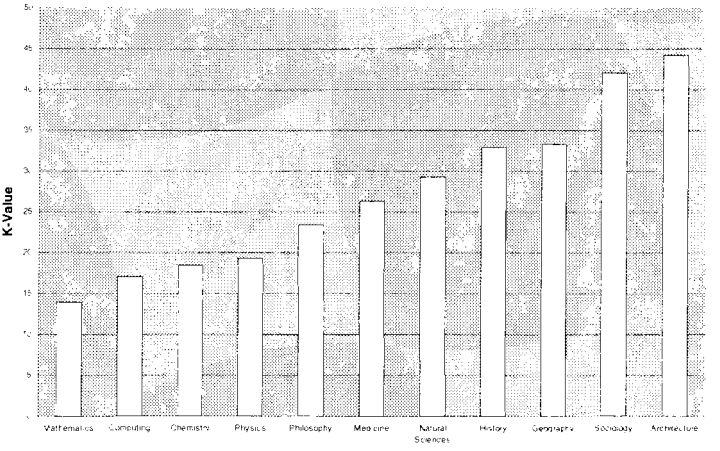

Fig. 6 Text types and lexical density (K-values)

The mean *K-value* for the 11 sample is 27.29 and its standard deviation 9.43. Comparing these figures with the individual *K-values* from the table above (*Tab. 5*) reveals a great deal of variability or dispersion among the various text samples. The sample on *physics* compared with the *sociology* one indicates huge differences in lexical density, not to say the relation between *mathematics* and *architecture*. However, *geography* and *history* seem to have a very similar lexical density. The outstanding lexical density of *architecture* can be explained on the basis that it might contain many proper names (artists, architects, places, etc.) and specific terms, whereas the very low density of *mathematics* niight probably rely on its high proportion of figures and formulaic expressions in substitution of word forins. The histogram (*Fig.* 6) displays graphically the various text types ordered according to their lexical densities (*K-values*). Interesting, here is the fact how the lexical density scale moves smoothly from pure science subjects (*mathematics, computing, chemistry*, etc.) to more arts and humanistic content texts. Additionally, neighbourhood on the histogram niight suggest subject relatedness: the more dissimilar the lexical density indices (*K-values*) the less the subjects relate to each other.

The *K-values* suggest that discrimination between *chemistry* (18.46) and *sociology* (42.03) texts niight indeed be possible as both figures diverge significantly. However, a sole *K-value* based distinction between *chemistry* (18.46) and *physics* (19.27) seems less reliable, due to its closeness. Intuitively, it seenis as if a really fine grained classification is not viable.

To carry on exploring the extent and potential of our mathematical regression model, we proceeded in constructing a purely statistical model. We started experimenting with a descriptive, non-inferential statitistical technique: cluster anal>sis.

To put it succinctly, cluster analysis classifies a set ofobservations into two or more mutually exclusive groups based on the combination of interval variables. The purpose of cluster analysis is to discover a system of organizing observations into groups, where members

o!' the group share coninion properties. Cluster analysis classifies unknown groups while discriminant function analysis classifies known groups. A common approach to doing a cluster analysis is to iirst create a table or niatrix of relative similarities or differences between all objects and second to use this information to combine objects into groups. The table oi'relative similarities is called a proximity or dissimilarity matrix. *Table 6* displays the dissimilarity niatrix (note that both proximity matrices are symmetrical. Symmetrical means that row and column entries can be interchanged or that the numbers are thr same on each half of the niatrix defined by a diagonal running from top left to bottom right). The distance measure used is the *squared Euclidean distance.*

Case	Arch	Chem	Comp	Geo	Hist	Math	Med	Nat	Phil	Phys	Soc
Arch		663.58	736.04	118.81	128.14	916.27	322.56	222.31	432.64	622.50	4.80
Chem	663.58		1.88	220.82	208.51	20.34	60.84	117.72	24.60	0.66	555.55
Comp	736.04	1.88		263.41	249.96	9.86	84.09	149.33	40.07	4.75	622.00
Geo	118.81	220.82	263.41		0.18	375.20	49.84	16.08	98.01	197.40	75.86
Hist	128.14	208.51	249.96	0.18		359.10	44.09	12.89	89.87	185.78	83.36
Math	916.27	20.34	9.86	375.20	359.10		151.54	235.93	89.68	28.30	788.49
Med	322.56	60.84	84.09	49.84	44.09	151.54		9.30	8.07	48.86	248.69
Nat	222.31	117.72	149.33	16.08	12.89	235.93	9.30		34.69	100.80	161.80
Phil	432.64	24.60	40.07	98.01	89.87	89.68	8.07	34.69		17.22	346.33
Phys	622.50	0.66	4.75	197.40	185.78	28.30	48.86	100.80	17.22		518.02
Soc	4.80	555.55	622.00	75.86	83.36	788.49	248.69	161.80	346.33	518.02	

Tab. 6 Matrix of dissimilarity of the text sample subjects

Looking at the matrix we find that the least dissimilarity or closest similarity of all is 0.18. between the *history* text sample and the *geography* onr. We could say that these seem to form the pair that is most alike. *Physics* and *chemistry* have a very low dissimilarity index (0.66)and could be grouped. too. Since *history* is related to *geography* we could say that these form a cluster. On the opposite scale. we find the hugest difference between *mathematics* and *architecture* (016.27).

After the distances between tlie text types have been found. tlie next step in the cluster analysis procedui-e is to divide the text types into groups based on the distances. The results of the application of the clustering technique are best described using a dendogram or binary tree. Tlie objects are represented as nodes in the dendogram and the branches illustrate when the cluster method joins subgroups containing that object. The length of the branch indicates the distance between the subgroups when they are joined.

```
* * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * *

Dendrogram using Average Linkage (Between Groups)

                        Rescaled Distance Cluster Combine

   C A S E          0         5        10        15        20        25
 Label      Num    +---------+---------+---------+---------+---------+
 Geo          4    ••
 Hist         5    ••••••••
 Nat          8    ••      ••••••••••••
 Med          7    ••••••••           •
 Phil         9    ••                 •••••••••••••••••••••••••••••••
 Chem         2    ••                 •                              •
 Phys        10    ••••               •                              •
 Comp         3    ••  ••••••••••••••••                              •
 Math         6    ••••                                              •
 Arch         1    ••••••••••••••••••••••••••••••••••••••••••••••••••
 Soc         11    ••
```

The interpretation of the dendogram is fairly straightforward. For example, *Geo/His/Nat* form a group. *Chem/Phys/Comp/Math* form another group and *Arch/Soc* is called a "runt" because they do not enter any group until near the end of the procedure. A dendogram that clearly differentiates groups of objects will have small distances in the far branches of the tree and large differences in the near branches. The dendogram above illustrates 1 cluster or solution at distance 25. 2 clusters at distance 10. 3 at 4. 4 at 3. 5 at 1 and 11 at 0. This results into 6 possible solutions or groupings (see *Tab.* 7).

Cluster analysis methods always produce a grouping. The grouping produced by the cluster analysis may or may not prove useful for classifying objects. To validate these cluster analysis outputs we shall use them in conjunction with discriminant function analysis on the resulting groups (solutions) to discover the linear structure of either the measures used in the cluster analysis and/or different measures.

Solution	Rescaled Distance	Clusters
1	0	11: *Geo* *His* *Mat* *Med* *Phil* *Chem* *Phys* *Comp* *Math* *Arch* *Soc*
2	1	5: *Geo His Nat* *Med/Phil* *Chem/Phys/Comp* *Math* *Arch Soc*
?	?	4: *Geo/His Nat* *Med Phil* *Chem/Phys Comp Math* *Arch Soc*
4	4	?. *Geo/His Nat/Med Phil* *Chem/Phys/Comp Math* *Arch Soc*
5	10	?. *Geo His Nat Med/Phil Chem/Phys Comp Math* *Arch Soc*
6	25	1: *Geo/His Nat/Med Phil Chem/Phys Comp Math Arch Soc*

Tab. 7 Possible text type clustering

Obviously, the best solution is 1 (the posibility of discriminating all 11 text types). whereas 6 is clearly the worst one (unable to differenciate any text type).

Cluster analysis is a positive exploratory tool for clustering possible groupinp solutions and for constructing at a later stage a group membership predictive model by neans of the discriminant Iunction analysis. This later multivariate technique is based on a linear combination of the interval variables (*K-values*). It begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a rnodel that allows prediction of membership when only the inteval variables are known. A second purpose of discriminant function analysis is an understanding of the data set. as a careful examination of the prediction model that results Iiom the procedurecan give insight into the relationship between group membership and the variables used to predict group nienibership.

In order to construct a model using discriminant function analysis. we added 11 more test samples. one for each text type. as the data available was insufficient. Next. using the exploratory cluster analysis data. we constructed the first model. taking solution 1. that is. a model subject to discriminate11 different text types (namely. *Geography. History. Natural Sciences. Medicine. Philosophy. Chemistry. Physics. Computer Science. Mathematics. Architecture* and *Sociology*). The case number. actual group. group assignments (*Highest Group* and *2nd Highest Group*) and discriminant scores are given below (*Table* S; note that wrong group assignment in *Highest Group* is marked with "**").

Case Number	Actual Group	Highest Group	2nd Highest Group	Discrim Scores
1	1	1	11	23219
2	2	2	3	-12,424
3	3	3	2	-14,320
4	4	4	5	8,137
5	5	5	4	7,556
6	6	6	3	-18,664
7	7	7	9	-1,631
8	8	8	7	2,589
9	9	9	7	-5,561
10	10	2**	10	-11,303
11	11	11	1	20,189
12	1	1	11	21697
13	2	10**	2	-11,234
14	3	3	2	-13005
15	4	5**	4	7,418
16	5	4**	5	7971
17	6	6	3	-17,045
18	7	7	8	-,165
19	8	8	7	,984
20	9	9	7	-3,209
21	10	10	2	-10127
22	11	11	1	18930

Tab. S Discriminant function analysis for solution 1 (11 clusters/groups)

The discriminant niodel for 11 text types revealed a success rate (correct group assignment) of 81.81% (it failed in correctly assigning cases 10. 13. *13* and *16.* which were. however. correctly classified in the second choice *-2"" Highest Group*).

The next discriminant niodel based on solutioii 3 *(5* text types) resulted into a very promising 95.5% success rate. It just failed in classifying correctly case 19 (*Nat* test) which was groiiped to the *Med/Phil* cluster.

Case Number	Actual Group	Highest Group	2nd Highest Group	Discrim Scores
1	1	1	5	10196
2	10	10	6	-5455
3	10	10	6	-6,288
4	5	5	7	3573
5	5	5	7	3318
6	6	6	10	-8,196
7	7	7	5	-716
8	5	5	7	1137
9	7	7	10	-2442
10	10	10	6	-4,963
11	1	1	5	8,865
12	1	1	5	9527
13	10	10	6	-4933
14	10	10	6	-5711
15	5	5	7	3257
16	5	5	7	3500
17	6	6	10	-7485
18	7	7	5	-,072
19	5	7**	5	432
20	7	7	10	-1,409
21	10	10	7	-4447
22	1	1	5	8312

Tab. 9 Discriminant function analysis for solution 2 (5 clusters/groups)

The next solution (3 with 4 clusters/groups) differs from solution 2 in that it groups *Math* within the *Chem/Phys/Comp* cluster. without solving tlie wrong groiip assigment of solution 2 (case 19). This is only solved within solution 4 (where *Nat* is grouped within *Med/Phil*, resiilting into just 3 clusters: (1) *Geo/Hist/Nat/Med/Phil*, (2)*Chem/Phys/Comp/Math* and *(3)Arch/Soc*). with a success rate of 100%. However. this solution has a serious flaw: its niinimal accuracy and discriminaton power (*Tab. 10* displays a summary oi'all 6 solutions).

Solution	Clusters	Success Rate
1	11	81.81%
2	5	95.5%
3	4	95.5%
4	3	100%
5	2	100%
6	1	100%

Tab.10 Summary of the various solutions, cluster divisions and associated success rates

The previous analyses are very revealing and it is now up to the reader to choose or decide which is the best solution. depending on his/her researcli goals (recall that discriniinant function analysis is not *inferential*). Nevertheless. it is our opinion that the hest model is solution 2. because oi'its reasonable discrimation power (it is able to discriminate 5 different text types: (1) *Geo/Hist/Nat. (2)Med/Phil. (3)Chem/Phys/Comp.* (4) *Math* and (5) *Arch/Soc*) and its accuracy (95.5%).

Another positive contribution of discriminant function analysis is that once the groups (interval variables) are known we can construct a model tliat allows prediction of membership. This is done by means of the resulting discriminant function coefficients The coefficients for solution 2 are:

		TEXTYPE			
	Areh/Soc	Geo/Hist/Nat	Math	Med/Phil	Chem/Phys/ Comp
K VALUE	15.734	11.670	5.365	9.434	6,909
(Constant)	-336.914	-186069	-40603	-121909	-66267

Tab.11Coefficients

To illustrate its prediction power. take. for example. a text with a *K-value* = 14.01

*TEXTTYPE = Constant + (K_VALUE * 14.01)*

We just need to niaximize the five coefiicients:

*Arch/Soc = -336.914 + (15.734 * 14.01) = -116.48*
*Geo/Hist Nat = -186.069 + (11.67 * 14.01) = -22.57*
*Math = -40.603 + (5.365 * 14.01) = 34.56*
*Med/Phil = -121.909 + (9.424 * 14.01) = 10.12*
*Chem/Phys/Comp = -66.267 + (6.909 * 14.01) = 30.52*

This results in that a hypothetical text with a *K-value* = 14.01 is most likely to he classified in first choice as being a *mathematics* text. as *Math* is tlie highest resulting coefficient (34.56); and in second choice. it would be classified as *Chem/Phys/Comp* (30.52). Siniilarly. the least likely group membership would be *Arch/Soc* (-116.48).

V. DISCUSSION AND CONCLUSIONS

From the evidence above. me are confident tliat the *K-value* is indeed a stable and robust lexical density indicator conipared to the *type-token ratio* and/or *token-type ratio*. This constant value

seems not just a reliable lexical density indicator but also a decisive index for type/lemma prediction in an *x* token corpus. Distinct to tlie *type-token/token-type ratio*. the *K-value* is text length independent and stays unaltered throughout. This reliability and validity results in a useful lexical density indicator.

It is precisely its robustness that has niotivated the preseiit study, on the assuniption that different test types/samples relative to distinct linguistic doniains are likely to exhibit unique *K-values*. The experimental data. as well as previous experinients. have revealed that different languages. authors or linguistic domains. etc., differ from each other. among niany other things. in their lexical density. that is. in the relation of distinct word forms (types) to tlie text/corpus word size. This enables us. for instance. to distinguish (a) languages: general Spanish has a *K-value* of 54.29. whereas general English 41.43; (b) text types: Spanish fiction: 50.77 and Spanish press: 56.12: (c) authors: Conrad: 35.75. Shakespeare: 37.96 and Doyle: 26.78: and (4) linguistic domains: archicture (Spanisli): 44.22. chemistry (Spanish): 18.46. computing (Spanish): 17.09. geography (Spanish): 33.32. history (Spanish): 32.9. mathematics (Spanisli): 13.95. medicine (Spanisli): 36.26. natural sciences (Spanish): 29.31. philosophy (Spanish): 23.42. physics (Spanish): 19.27 and sociology (Spanish): 42.03. Clearly. distinct *K-values* indicate different text types. authors. linguistic doniains. etc.

If we concentrate on the examined linguistic domains. we can appreciate a huge variation between *architecture* (44.22) and *mathematics* (13.95). for example. This sugpests that discrimiating these two dornains u-ould not be too difficult. However. distinguisliing between *geography* (33.32) and *history* (32.9) seenis nearly impossible.

Interesting in this sense are *Figure* 6. the cluster analysis and the discriminant function analysis. *Figure* 6 represcnts visually the *K-value* ordered linguistic doniains. where we can appreciate a logical and smooth text type transition. that goes froni pure science (*mathematics*) to clear humanity contents (*sociology/architecture*). This stratification is based on a single lexical density feature: the *K-value*. Complementary. the cluster analysis offers an exploratory grouped hierarchical structure of the text types. highlighting the major flaw of the *K-value*: incapacity of distinguishing between closely nearby *K-values*. as little dissimilar lexical densities are grouped into single clusters. Clearly. the *K-value* fails to distinguish between (a) *geography*, *history* and *natural sciences*; (b) *medicine* and *philosophy*; (c) *chemistry*, *physics* and *computing*; and (d) *sociology* and *architecture*. However. the final modelling of the data by nieans of the discriminant Iunction analysis reveals that the *K-value* is valid and reliable to successfully differentiate (a) *geography history/natural sciences*. (b) *medicine philosophy*. (c) *chemistry physics computing*. (d) *sociology/architecture* and (e) *mathematics* from each other.

Though a potential text discriminator using *K-value* does not. in principle. produce a very specific classification. it does not invalidate the use of lexical density for text differentiation. The resulting text classification froni the experiment is far from being erroneous or exaggeratedly generic. On the contrary. it discrimates clearly distintive text type clusters:

(a) *mathematics*. (b) *chemistry/physics/computing*. (c) *medicine/philosophy*. (d) *chemistry/physics computing* and (5) *sociology/architecture*. with an accuracy rate of 95.5%.

In sum. we are coniident of the usefulness of the lexical density for autoniated text classification. if a reliable and valid lexical density index such as the *K-value* is used. Tlie conjunction of the *K-value* with multivariate statitistical tecliniques (cluster analysis and discriniinant function analysis) has resulted into very positive and promising data niodels. where the potential preciseness of the text typifation has been much more specific than one might espect at first sight. It needs to be recalled that neither linguistic knowledge. linguistic paradipnis nor linguistic feature data were used. just a single index specifying the relationsliip between words (tokens) and word forms (types) relative to each text sample.

REFERENCES

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

Biber. D. (1993) "Representativeness in Corpus Design". *Literary and Linguistic Computing*. 8(4): 243-257.

Heaps. H. S. (1978) *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Holnies. D. I. (1994) "Authorship Attribution". *Computers and the Humanities*. 38: 87-106.

Sánchez. A. and P. Cantos (1997) "Predictability of Word Fornis (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish". *International Journal of Corpus Linguistics*. 2(2): 259-280.

Sánchez. A. and P. Cantos (1998) "El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras. de las lenguas inglesa y española y en cinco autores de ambas lenguas". *ATLANTIS*. XIX (1): 205-223.

Yang. D.-H.. P. Cantos and M. Song (*forthcoming*) "An Algorithm Sor Predicting the Relationship between Lemmas and Corpus Size". *ETRI* (Electrnnics and Telecommunications Research Institute) *Journal*.