

KYO KAGEURA *

MICHEL SANTANA RODRIGUEZ **

SANDRA YAMILET SANTANA **

National Center for Science Information Systems, Tokyo, Japan *

Centro Colombo Americano, Pereira, Colombia **

A quantitative morphological structure of English and Spanish technical terminology

Resumen: En este artículo, se informa del resultado del análisis estadístico de la dinámica de los elementos constitutivos de términos tanto españoles como ingleses. En los términos españoles, los elementos funcionales se utilizan mucho más frecuentemente que en los términos ingleses. Para distinguir las características conceptuales de los factores gramaticales en la estructura cuantitativa terminológica, los elementos funcionales fueron separados de los elementos que designan conceptos. El análisis muestra que el español usa menos elementos en la formación de términos que el inglés, cosa que tiene que ver con el hecho que el español utiliza más elementos funcionales. El análisis dinámico muestra que, al aumentar el tamaño de la muestra, esta tendencia continua de manera acelerada.

1. Introduction

In this paper we analyse the quantitative nature of Spanish and English technical terminology. The existing quantitative analyses of Spanish and English texts show that the same number of high frequency word types covers more tokens in Spanish than in English, although a rigid comparison is not possible. In Spanish, the first 5000 high frequency word types cover 92 % of the tokens in running texts (Juilland & Chang-Rodriguez, 1964), while in English the same number of word types cover about 86 % (Kucera & Frances, 1967).

What, then, is the quantitative structure of technical terminology? In the present study, we analyse the quantitative structure of constituent elements of terms in terminology, i.e. in the list of different terms. So the point to be observed in the present study is different from the conventional quantitative studies of lexical units. This is because, in the case of terminology, (a) the majority of terms are complex, (b) the complex term formation is considered to be carried out as a lexical process, not a syntactic process, independent of actual use. In addition, as we tend to understand the meaning of a term by means of its constituent elements, to observe the basic quantitative nature of the constituent elements of terms is an important starting point to characterise the structure of terms in the general lexical structure of a language (Miyajima, 1991).

In the following, we first compare the static structure of the terminological data in Spanish and English, and then go on to observe comparatively the dynamic tendency of the Spanish and English terminology.

2. Basic Descriptions of the Terminological Sample

2.1 The Data

For the analysis, we used the list of term entries in Kotani & Kori (1990), which covers important terms in technology in general, collected from "publications of important American and British academic institutes, the manuals and standards ..., and authoritative

works" (Kotani & Kori, 1990:viii). The Spanish equivalents are assigned then, based on various authoritative information sources. The basic quantitative information of the data is given in Table 1, where T indicates the number of different terms, N indicates the number of running constituent elements (token), and $V(N)$ the number of different constituent elements (type)¹. So N/T shows the average length of a term in terms of the number of constituent elements, and $N/V(N)$ shows the average use of a constituent element. The columns '1', '2', and '3+' indicate the number of terms, together with their ratio, with one, two and three or more constituent elements respectively. The last row indicates the ratio (values of Spanish divided by corresponding values of English) of each column.

Table 1: Basic quantities of terminology sample

	T	N	V(N)	N/T	N/V(N)	1 (%)	2 (%)	3+ (%)
Spanish	29611	74187	9729	2.51	7.63	7445 (25.1)	7803 (26.3)	14363 (48.5)
English	29611	56195	10217	1.90	5.50	8591 (29.0)	16386 (55.3)	4634 (15.7)
Ratio (S/E)	1	1.32	0.95	1.32	1.39	0.87	0.48	3.10

We can immediately notice the difference between Spanish and English, i.e. the average length of a term is 2.51 in Spanish and 1.90 in English, and the average use of a constituent element is 7.63 in Spanish but 5.50 in English. The number of terms by length also shows corresponding characteristics. This is considered to be, to a great extent, due to the functional elements in Spanish, especially "de". So they reflect the grammatical characteristics of general patterns of word or term formation in Spanish and English.

As the main function of terms is to designate specialised concepts, it is useful and important to observe the nature of terminology from the point of view of content-bearing constituent elements, as separate from the functional elements. Let us, therefore, partition the constituent elements of terminology into functional elements and content-bearing elements, and observe again the quantitative characteristics of terminology with respect to the content-bearing elements.

2.2 Consideration of Functional Elements

The functional elements in terminology are identified by hand. There are some elements which are considered to be functional grammatically, but are better classified as content-bearing, due to their role in terms. For instance, in the Spanish example "montaje de espalda contra espalda", the role of "contra" is not the same as more 'neutral' functional elements such as "de". Also, the Spanish and English negative prefix "no-" takes an important semantic role in the construction of terms, as in "no-discharge operation". After examining the list of candidates for functional elements, we decided to take the conservative standpoint and classified many 'grammatical' elements as content-bearing. For our immediate analysis, this decision is justified because the ambivalent elements, not occurring frequently, do not greatly affect to the quantitative structure of terminology.

Table 2 lists the elements that are defined as functional, together with their frequencies. When they are used as a content-bearing element, e.g. "and" in "AND-OR

¹We have not so far grouped the inflexional variants of sex and number in counting different constituent elements, which will be done in the future. For our immediate purpose, it should be pointed out that, as the number of inflected elements in Spanish is much greater, it would affect positively the conclusion which is to be drawn in this paper.

circuit", they are taken as content-bearing. The frequency of "de" is notable, i.e. together with "del", it constitutes nearly 18 % of all the tokens of constituent elements in Spanish terminology.

Table 2: Functional elements in Spanish and English

Spanish	de (12090), del (1099), en (570), la (501), por (491), a (382), para (289), con (268), al (132), y (125), el (106), los (35), las (14), e (14), una (14), un (9), o (3)
English	of (472), and (59), to (24), for (22), on (16), in (13), with (9), at (8), a (7), by (6), or (4), the (2)

Table 3 shows the quantitative data of the content-bearing elements in the Spanish and English terminology. Now the number of tokens of constituent elements becomes much closer, as does the average length and the average use of constituent elements. The number of terms by length also shows quite similar patterns. The difference of the number of different constituent elements remains the same, which is natural because we have defined only a very few elements to be functional.

Table 3: Quantities of terminology sample without functional elements

	T	N	V(N)	N/T	N/V(N)	1 (%)	2 (%)	3+ (%)
Spanish	29611	58086	9717	1.96	5.98	7450 (25.2)	16863 (57.0)	5298 (17.9)
English	29611	55566	10213	1.88	5.44	8591 (29.0)	16818 (56.8)	4202 (14.2)
Ratio (S/E)	1	1.04	0.95	1.04	1.10	0.87	1.00	1.26

As far as the content bearing elements are concerned, therefore, the basic patterns of the Spanish and English terminology are much closer than they first seem to be. However, we still observe a substantial difference between Spanish and English, with respect to the number of tokens as well as the number of different constituent elements.

The quantitative characteristics revealed so far can be situated properly in a unified interpretation, i.e. as Spanish uses fewer constituent element (types) that bear content, it has to represent a concept with more content bearing elements on average than English, the grammatical construction of which is supported by the frequent use of functional elements such as "de". This corresponds to the quantitative nature of lexical elements in texts in Spanish and English.

3. Dynamic Nature of Terminology

So far we have described the basic quantitative characteristics statically, on the basis of the given sample. However, in the case of language data, most statistical measures change systematically according to the size of the sample (Tweedie & Baayen, 1998). In case of terminology, the systematic change of statistical measures can be interpreted as corresponding to the potential growth of terminology. Because what we are treating now is just a sample of terminology, the static quantitative description itself can only be treated as a snapshot of the totality of terminology, which is dynamic in nature. In this section, therefore, we observe the dynamic growth patterns of the constituent elements of terminology, on the basis of the sample.

3.1 Methodology

In order to trace the growth patterns of the constituent elements, we use the framework of binomial interpolation and extrapolation which is introduced by Good & Toulmin (1956), which requires the assumption, in the present case, that the constituent elements of terms are randomly distributed in terminology. In terminology, the order of terms can be assumed to be arbitrary, so the only deviating factor to the randomness assumption is the order of constituent elements within individual terms. Kageura (1998) shows that this does not statistically affect the randomness assumption, so we can safely rely on binomial interpolation and extrapolation. If we assume that the distribution of the length of terms does not change according to the sample size, which is not unreasonable, then the general growth patterns of the constituent elements of terms can be described with respect to the growth of the sample size in terms of T as well as of N , because T is straightforwardly derived by dividing N by the average length of a term irrespective of the sample size.

Now, the expected number of different constituent elements $E[V_{N_0}(N)]$ in the sample size N , and the number of different constituent elements that appear m times, $E[V_{N_0}(m, N)]$, in the sample of size N , both given the original sample of size N_0 , can be defined as follows:

Using these formulae, we can observe the growth patterns of constituent elements of terminology for changing sizes of terminology, up to about twice the original sample size².

$$E[V_{N_0}(N)] = V(N_0) + \sum_{m=1}^{N_0} (-1)^m V(m, N_0) \left(\frac{N}{N_0} - 1 \right)^m \quad (1)$$

$$E[V_{N_0}(m, N)] = \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0} \right)^m \left(1 - \frac{N}{N_0} \right)^{k-m} \quad (2)$$

As terms are very real elements recognised at the level of *la parole*, observation within this range is expected to be very useful.

3.2 Descriptive Indices

To observe the dynamic characteristics of terminology, we observe the three indices of (i) the number of different constituent elements, (ii) the average use, or frequency of occurrence, of a constituent element, and (iii) the growth rate of the constituent elements. Each of the transitions of these indices can be calculated by the formulae (1) and (2).

The third index needs further explanation. The growth rate $P(N)$ here means the probability, in the mathematical sense, of encountering a new constituent element type, when the sample size is increased, and is defined by the following formula (Baayen, 1989):

$$P(N) = \frac{E[V(1, N)]}{N}$$

²Beyond that size, the values begin to diverge, due to the terms $(N/N_0 - 1)^m$ and $(1 - N/N_0)^m$ in the formula.

Note the denominator is N , so the sample size is counted in terms of constituent elements. It can be derived from the theoretical model on the basis of the binomial assumption, though we do not give the procedure of derivation here.

3.3 Results and Interpretations

Figures 1 to 3 show the transitions of the number of different constituent elements, of the average use, and of the growth rate, respectively, up to $N = 1.9 \cdot N_0$. Note that X-axis is the number of terms, not the number of constituent elements. In Figure 1, the transition of the number of hapax legomena, i.e. the elements that occur only once, is also indicated, as this is used in the calculation of the growth rate³. In each figure, the panel on the left indicates the transitions of the actual values of the index, while the panel on the right shows the ratio of the values in Spanish and English, i.e. the values in Spanish divided by the values in English. In the panels on the left, the dots show the values for all the elements, and the lines show the values for content-bearing elements only. Spanish values are indicated by white dots and dashed lines, while English values are indicated by black dots and solid lines. In the panels on the right, the dots correspondingly show the values for all the elements and the lines show the values for content-bearing elements.

From Figure 1, we can observe that, when the sample size is smaller, the number of different constituent elements in Spanish and English does not differ greatly. In fact, the panel on the right shows that, at first (up to about $T = 2000$), the number of different constituent elements in Spanish is expected to be greater than in English, as indicated by the fact that the values exceed 1 at the beginning. This can be explained by the fact that, as the average length of the Spanish term is greater than that of the English terms, more tokens of constituent elements are introduced when a new term is added. That the value exceeds 1 at the beginning, therefore, does not mean that Spanish re-uses the existing elements less than English at first. This point will be discussed again shortly in relation to the growth rate. The value of the ratio then declines, sharply at first and then more slowly. From the transition of the ratio, we can say that, although the ratio perhaps converges at some point to some value, which corresponds to the ratio of population numbers of the constituent element types in Spanish and English, we may very well observe that the number of different English constituent elements becomes relatively greater than that of Spanish elements in a realistic terminology size range.

Figure 2 shows the prominent influence of the functional elements on the average use of constituent elements in Spanish. In English, on the other hand, the functional elements are almost negligible. Without functional elements, the Spanish and English become much closer to each other. However, the ratio of average frequencies in the two languages becomes higher as the sample size is increased, though the line becomes more flat, which may well indicate, again that the ratio converges at a certain point⁴.

The growth rate in Figure 3 again indicates that the Spanish without functional elements shows a pattern much closer to English, which is in accordance with all the observations so far. The actual transition of the values and the ratio of the growth rate of

³The upward transition observed in the right tails of the ratios of $E[V(1,N)]$ and the growth rate should be interpreted as divergence caused by numerical computation, and therefore unreliable.

⁴If we assume that the number of different constituent elements is finite, then for $N \rightarrow \infty$ the average frequency approaches infinity as well, which in turn means the ratio becomes 1. What is meant here by convergence, therefore, should be understood as the turning point, where the value of the ratio stops increasing.

Spanish and English shows, however, that, the more terms are created, the greater the difference in the growth rate, i.e. the speed that the new constituent elements are introduced in English is accelerated relative to Spanish. Although the speed of acceleration itself becomes slower, it is not clear at what point the acceleration stops. Note also that even at the beginning, the ratio is less than 1, i.e. the growth rate in English is higher. This fact reinforces the discussion above that the value of the ratio over 1 in the panel on the right in Figure 1 is due to the different average length of Spanish and English terms. The growth rate is calculated by means of N .

All in all, the transitions of the three indices show that, although there may be a point of convergence where the difference between Spanish and English is no longer accelerated (which is implied by the fact that the **rate of increase or decrease of the ratio** of the indices becomes lower), within a realistic range of terminology the basic difference observed between Spanish and English terminology in the static analysis not only continues but can also be expected to continue with acceleration.

4. Conclusions

We have seen that the dynamic analysis of the quantitative nature of terms can be used successfully to reinforce the static quantitative analysis. The direction has good potential but the present study only addresses a preliminary part of what should be done in the quantitative study of terminology. The present study should be extended, with at least the following points being taken into consideration:

- (1) Finer classification of the constituent elements. At the grammatical level, we have not yet treated the inflexional variants by sex and number. To observe the quantitative nature of terminology, we believe it is crucial to introduce proper qualitative categories, e.g. head and modifier, semantic categories, etc.
- (2) A more powerful method of extrapolation. The framework by Good & Toulmin (1956) is theoretically proper but has a practical limitation in that it cannot cope with extrapolation to a very large sample size. We are currently examining the applicability of LNRE models introduced by Chitashvili & Baayen (1993).
- (3) A proper way of relating the statistical model, which is naturally defined over constituent elements of terms, to a model that can possibly be defined over terms. Currently there is a significant gap to be filled at the stage of interpreting the quantitative dynamics observed in the constituent elements at the level of terms and terminology. To fill the gap, such factors as the distribution of the length of terms should be more properly taken into account.

Bibliography

- Baayen, R. H.** (1989) *A Corpus-Based Approach to Morphological Productivity*. PhD Thesis, Free University of Amsterdam.
- Chitashvili, R. J. and Baayen, R. H.** (1993) "Word Frequency Distributions", In: Hrebicek, L. and Altmann, G. (eds.) *Quantitative Text Analysis*. Trier: Wissenschaftlicher Verlag. p. 54-135.
- Good, I. J. and Toulmin, G. H.** (1956) "The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased", *Biometrika*. 43(1), p. 45-63.
- Juilland, A. and Chang-Rodriguez, E.** (1964) *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

Kageura, K. (1998) "A Statistical Analysis of Morphemes in Japanese Terminology", COLING-ACL'98. 10-14 August, Montreal, Canada.

Kotani, T. and Kori, A. (1990) *Dictionary of Technical Terms*. Tokyo: Kenkyusha.

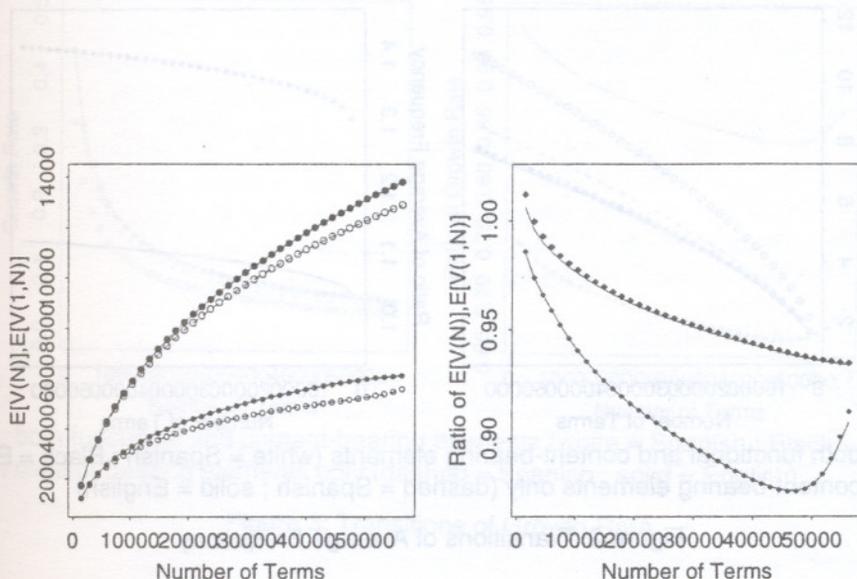
Kucera, H. and Frances, W. N. (eds.) (1967) *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Miyajima, T. (1981) *Problems of Technical Language*. Tokyo: Syuei Syuppan. [in Japanese]

Tweedie, F. J. and Baayen, R. H. (1998) "How Variable May a Constant be? Measures of Lexical Richness in Perspective", *Computers and the Humanities*. (forthcoming)

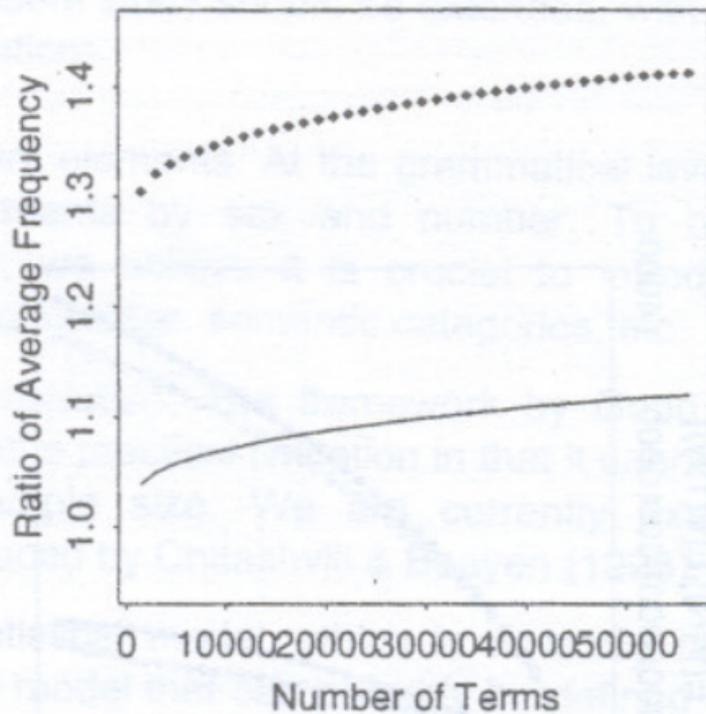
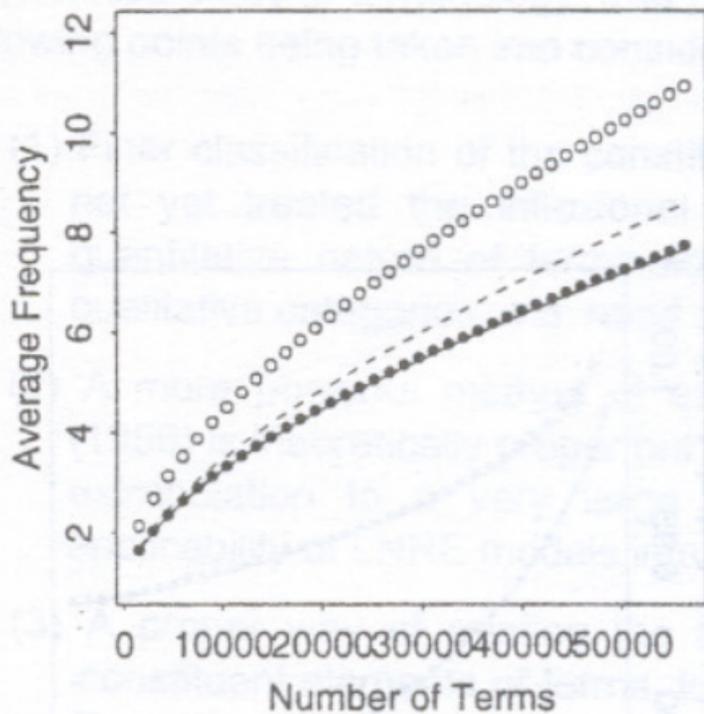
Acknowledgement

We would like to thank Ms. Stephanie Coop and Ms. Ariadna Font Llitjós for reading the draft and cheking some expressions.



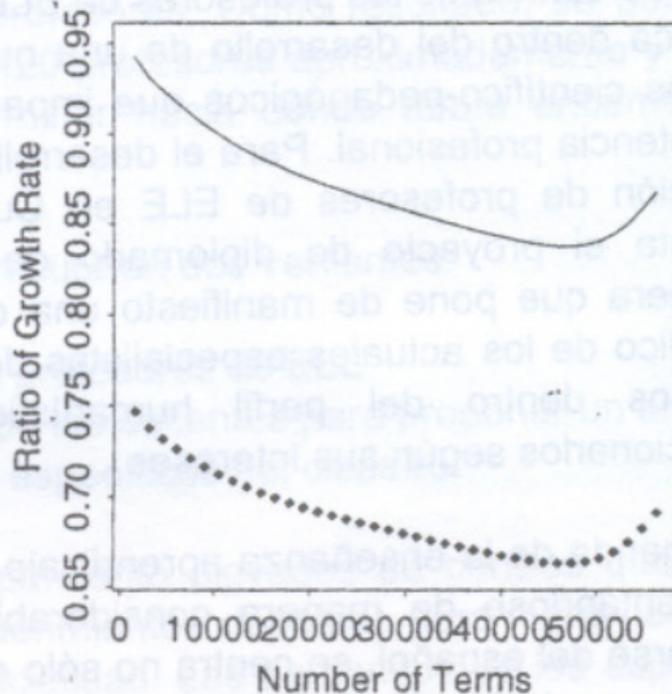
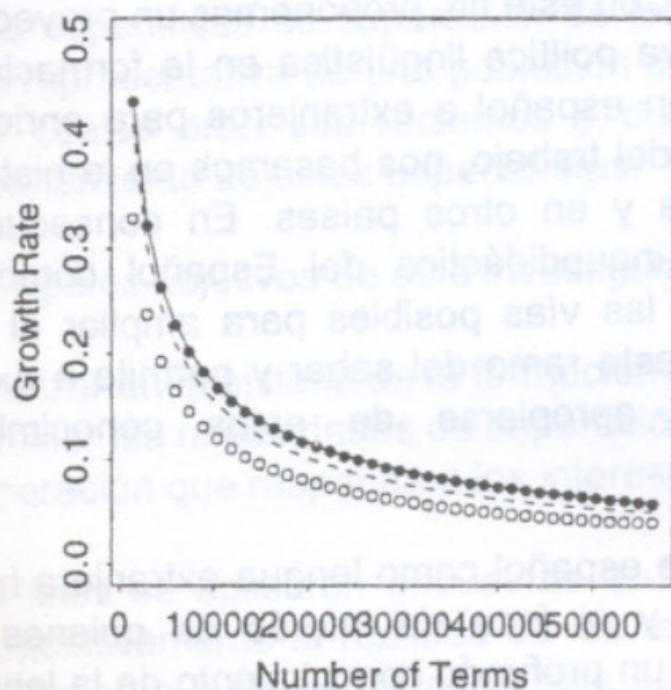
Dots : both functional and content-bearing elements (white = Spanish ; Black = English)
Lines : content-bearing elements only (dashed = Spanish ; solid = English)
(small dots/thin lines indicate $E[V(1, M)]$)

Figure 1: Growth of Constituent Elements



- Dots : both functional and content-bearing elements (white = Spanish ; Black = English)
- Lines : content-bearing elements only (dashed = Spanish ; solid = English)

Figure 2: Transitions of Average Frequency



- Dots : both functional and content-bearing elements (white = Spanish ; Black = English)
 Lines : content-bearing elements only (dashed = Spanish ; solid = English)

Figure 3: Transitions of Growth Rate