El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos : descripción del proceso de selección del léxico típico del discurso económico empresarial en español

Lieve Vangehuchten

Universiteit Antwerpen – Prinsstraat 13 – 2000 Antwerpen – België
lieve.vangehuchten@ua.ac.be

Abstract

This research was triggered by my observation that the lexical content of currently existing didactic material for the purpose of the acquisition and teaching of Economic and Business Spanish is composed on the basis of merely intuitive criteria and, what is more, that it is the work of laymen, without consulting specialists (economists). This paper builds on the outcome of my PHD-research in which I designed a method to enable the objective selection of vocabulary. The method itself is based on the insights of modern terminology with regard to the identification of the terms as well as on some lexico-statistical techniques, and was tested on a sample-corpus of Applied Economics Spanish, counting 118.365 occurrences. Given the modest dimension of this corpus the results are rather indicative than representative.

Keywords : corpus linguistics, statistics for terminological purposes, terminology, lexicology, second language acquisition research.

Resumen

El punto de partida de este trabajo es la constatación de que el contenido léxico de las herramientas de aprendizaje del español económico empresarial es, en la gran mayoría de los casos, el resultado de una selección subjetiva e intuitiva. En el marco de la elaboración de una tesis doctoral, se desarrolló una metodología para la identificación y la selección científica y objetiva del léxico típico de una lengua especializada. En cuanto a la identificación, se siguió la teoría de la corriente lingüística de la Terminología que estimula la investigación semasiológica en corpus. Para el proceso de selección se ha recurrido a unas técnicas léxico-estadísticas que se han aplicado a un corpus modélico de 118.365 ocurrencias. Dado el tamaño modesto de este corpus de discurso económico empresarial, los resultados son más bien indicativos que representativos.

Palabras clave : lingüística de corpus, léxico-estadística, terminología, lexicología, lingüística aplicada a la adquisición de una lengua extranjera.

1. Introducción

El punto de partida de este trabajo es la constatación de que el contenido léxico de las herramientas de aprendizaje del español económico empresarial es, en la gran mayoría de los casos, el resultado de una selección subjetiva e intuitiva (Vangehuchten, 2003). Con vistas a la elaboración de una metodología para la selección científica y objetiva del léxico típico de una lengua especializada, se ha recurrido a algunas técnicas estadísticas que han sido aplicadas a un corpus modélico de discurso económico empresarial. En este artículo se comentarán primero la composición del corpus modélico así como su norma de lematización. Después se describirán las distintas categorías léxicas presentes en el corpus y la manera de la que la estadística contribuye a su clasificación y selección didáctica.

2. Presentación del corpus modélico

Se trata de un corpus de discurso didáctico, o sea un texto medianamente especializado, de comunicación entre un especialista y aprendices. Es un manual de introducción a la economía empresarial de Eduardo Pérez Gorostegui, catedrático de economía de la empresa Universidad Nacional de Educación a Distancia (España). Se podría reprochar a este corpus modélico el hecho de ser de un solo autor, por lo cual no sería representativo del discurso económico empresarial en general, sino del idiolecto del autor. De ahí que resulte necesario poner en claro que el corpus modélico no sirve como base para la confección de un glosario o diccionario de léxico económico empresarial, sino que desempeña la función de objeto de experimentación. Por otra parte, cabe señalar que el texto pertenece a un registro formal y escrito de lenguaje técnico sometido a un grado elevado de normalización. Asimismo, se trata de un texto de instrucción, un manual que tiene un uso generalizado en las universidades españolas. Todas estas características nos permiten decir que el carácter idioléctico de este texto está muy limitado y se manifiesta sin duda más en algunas estructuras sintácticas recurrentes del autor, que en su empleo del léxico técnico.

La norma de lematización ha sido la unidad léxica (UL), por lo cual el resultado es más fino que el de un *type*/*token* análisis. Una UL es la realización de un lexema, que se puede componer de más de una sola forma y que puede sufrir variaciones gramaticales (sintácticas y morfológicas) con tal de que no afecten su significado básico. El significado de una UL es, por tanto, único e invariable, aunque permite empleos pragmáticos individuales. En total se han identificado 4.933 UL distintas sin contar los nombres propios. La lematización se ha hecho tanto automática como manualmente. Para la parte automática de la lematización del corpus, hemos recurrido a un lematizador español (*RELAX* Part-of- Speech *tagger*[1]). Ha sido imprescindible corregir los resultados manualmente, dado que este *tagger* fue desarrollado para lematizar la lengua general y no es capaz de distinguir entre significados especializados y generales, ni tampoco de reconocer términos compuestos.

El tamaño del corpus analizado sólo es de 118.365 ocurrencias (OC). No obstante, según *The Handbook of Terminology Management* un corpus de 100.000 formas léxicas ya es suficiente para sacar conclusiones significativas, dado que el léxico utilizado para tratar una temática especializada es más restringido que el léxico de un discurso no especializado :

> "*As a rule of thumb, special-language corpora already start to become useful for key terms of the domain in the tens of thousands of words, rather than the millions of words required for general-language lexicography*." (Ahmad y Rogers, 2001 : 736)[2]

Asimismo, la representatividad del corpus se puede demostrar estadísticamente al comparar la cantidad de UL nuevas por capítulo o el crecimiento del léxico a través del corpus, como se desprende del gráfico 1. Se puede ver que la evolución del léxico nuevo es bastante lógica : la tasa de UL nuevas disminuye progresivamente entre los capítulos 1 y 7, sufre altibajos entre los capítulos 7 y 10 para llegar a una situación de estabilidad relativa a partir del capítulo 11. Clijsters (1990 : 84) llama a este punto 'la saturación' del léxico :

[1] Este *tagger* fue desarrollado por el *Natural Language Processing Group* de la Universitat Politècnica de Catalunya, en colaboración con el *Computational Linguistics Laboratory* de la Universidad de Barcelona (información en http ://www.lsi.upc.es/~nlp/).

[2] En este contexto cabe citar también el estudio de Sutarsyah *et al*., (1994), quienes, para la comparación del vocabulario de la lengua económica en general – o sea, no sólo empresarial – con el vocabulario de la lengua académica, se sirven como corpus de un manual de economía general de un solo autor que cuenta 300.000 *tokens*.

> " […] *l'accroissement en vocables nouveaux tend vers un point de saturation (V n'augmente plus qu'à grand-peine)* […] "

Este punto indica que se ha empezado a agotar el léxico del asunto tratado y que habrá un momento en que el inventario de UL nuevas tocará fondo. Aún según Clijsters, este momento de inicio de agotamiento representado por el punto de saturación también es una prueba de la representatividad del corpus. Esto significa, como se desprende claramente del gráfico, que un corpus compuesto de los siete primeros capítulos del manual – 51.028 OC en total – no hubiera sido representativo, dado que después del séptimo capítulo el punto de estancamiento todavía no se alcanza. El punto de saturación parece coincidir con el capítulo 11, o sea después de 79.363 OC. Con una totalidad de 118.365 OC, el corpus supera por tanto con creces el punto de saturación.

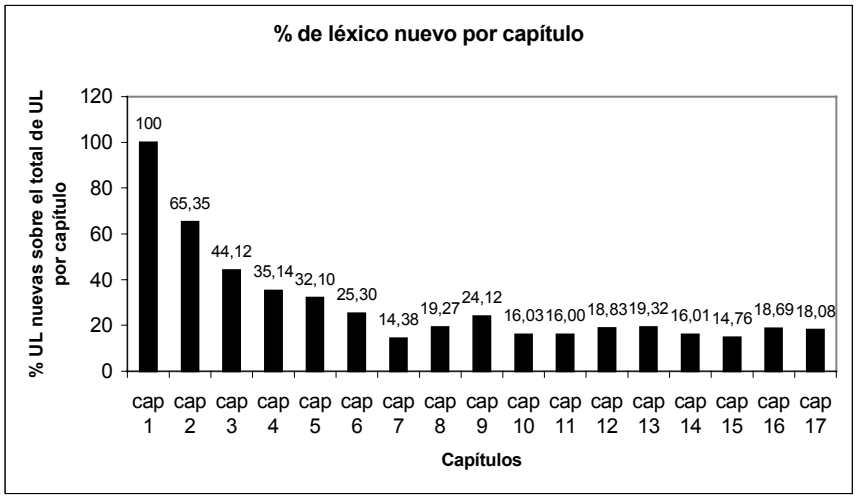

Gráfico 1.

3. Las categorías léxicas

En el siguiente cuadro se presentan las categorías léxicas del corpus :

	UL	**OC**
Funcionales (F)	5,07%	55,99%
Términos (T)	28,54%	12,22%
Términos auxiliares (TA)	5,37%	1,80%
Generales (G)	61,02%	29,99%
Total	**100%**	**100%**

Cuadro 1.

3.1. *El léxico funcional*

El léxico funcional manifiesta la tasa de cobertura más importante : el 55,99% de las OC. Sin embargo, sólo constituyen el 5,07% de todas las UL, o sea que la diversidad léxica en esta categoría es muy baja. El léxico funcional, que se llama también léxico gramatical o léxico vacío por su valor semántico casi inexistente, representa una clase finita de palabras muy

frecuentes. En cuanto a la selección didáctica del léxico funcional, queda claro que este grupo finito de palabras muy frecuentes debe formar parte de manera exhaustiva del contenido léxico de un curso de español como lengua extranjera.

3.2. *Los términos y los términos auxiliares*

3.2.1. *La identificación*

La identificación se ha hecho manualmente a partir de los criterios formulados por la corriente lingüística en la Terminología (Kocourek, 1982 ; Sager, 1990 ; Cabré, 1999 ; Temmerman, 2000), ya que todavía no existe un programa de extracción terminológica que produzca un resultado lo suficientemente preciso.

3.2.2. *La selección*

Hoy en día vuelve a admitirse que la frecuencia es un criterio eficaz y fiable para seleccionar y graduar el léxico en un contexto didáctico. Para dar resultados verídicos, el criterio de la frecuencia necesita ser combinado con el de la dispersión. La estadística léxica elaboró distintas fórmulas para combinar frecuencia y dispersión, siendo, sin duda, una de las más conocidas el '*usage coefficient'* de Juilland y Chang-Rodríguez (1964 : XVIII). Según Müller (1977 : 74) es importante tomar en consideración si el corpus está dividido en partes iguales o fragmentos desiguales. En el último caso, que es el nuestro, el padre de la estadística léxica sugiere una frecuencia "corregida" : KF = $\left(\sum \sqrt{p_i f_i}\right)^2$

En esta fórmula, que está basada en la media geométrica y que fue desarrollada por Rosengren (1971 :119), p_i representa el tamaño relativo del fragmento o capítulo *i* en que figura la UL en cuestión, o sea la totalidad de OC en el fragmento *i* dividida por la totalidad de OC en el corpus. La frecuencia de la UL en el fragmento *i* corresponde a f_i. Al calcular el producto de f_i con p_i para cada capítulo en que está el lema, se obtiene una cifra que Rosengren llama la frecuencia corregida, ya que esta fórmula tiene en cuenta la dispersión de cada lema para corregir su frecuencia. Por lo tanto, los lemas que ocurren en total dos veces en dos capítulos, se estiman más importantes que los lemas con una frecuencia superior pero concentradas en un solo capítulo.

Se ha demostrado que la fórmula de la frecuencia corregida de Rosengren ofrece un criterio fiable para ordenar los términos según su importancia. No obstante, conviene hacer dos observaciones respecto del ranking así obtenido. Una primera observación concierne la utilidad de someter el resultado exhaustivo al estudiante de español económico. Hazenberg (1994 : 47) demuestra, en su intento de establecer un vocabulario fundamental neerlandés para estudiantes extranjeros que desean empezar una carrera académica, que la frecuencia y la dispersión sólo son útiles en la medida en que la tasa de nuevas palabras a aprender sea compensada por una ganancia considerable en cuanto a la tasa de cobertura. Dicho de otro modo, a partir del momento en que el aprendizaje de nuevas palabras ya no corresponde a un progreso considerable en cuanto a la tasa de cobertura, no tiene sentido continuar imponiéndolas al estudiante, ya que se entra en la zona donde las palabras manifiestan todas la misma frecuencia bajísima, por lo cual la posibilidad de toparse con ellas en un texto representativo del lenguaje examinado queda muy reducida. Claro está que esto no significa que las UL que siguen a este punto no sean importantes, sino que vale más que a partir de este momento el estudiante centre sus intereses en un subcampo determinado y que extienda sus conocimientos léxicos con el vocabulario desconocido que éste contiene.

Si se aplica esta idea a los términos económicos y auxiliares del corpus, después de haberlos ordenado según su frecuencia corregida, es posible distinguir entre los términos básicos y los términos de continuación o subespecialización. A continuación se presentan las tasas de cobertura de los términos económicos y auxiliares en el corpus, teniendo en cuenta también los *hapax legomena* :

T	OC	% de cobertura	TA	OC	% de cobertura
0	0	0%	0	0	0%
200	10.803	74,68%	50	1.593	74,96%
400	12.419	85,85%	100	1.853	87,20%
600	13.213	91,34%	150	1.997	93,98%
800	13.727	94,89%	200	2.060	96,94%
1.000	14.058	97,18%	265	2.125	100,00%
1.200	14.258	98,56%			
1.408	14.466	100,00%			

Cuadro 2.

Los 200 primeros términos económicos más frecuentes y mejor repartidos, manifiestan una tasa de cobertura muy elevada del 74,68%. A partir de este punto la ganancia disminuye sin cesar : si de 200 a 400 UL todavía es de un 11,17%, baja a un 5,49% de 400 a 600 UL y a un 3,55% de 600 a 800 UL. Según Hazenberg (1994 : 48), al repartir las UL del corpus en partes equitativas por orden de frecuencia decreciente, una progresión del 3,5% aún se puede considerar sustancial en términos de comprensión potencial. Esto significa, para los términos económicos del corpus modélico, que 800 de ellos se pueden seleccionar como fundamentales en base a su frecuencia corregida. Después de este punto, el estudio de 200 UL más, sólo generaría una ganancia del 2,29%. Por consiguiente, hay 608 términos que quedan fuera de los términos seleccionados como fundamentales. Entre ellos figuran los 504 *hapax legomena*. Es necesario incluir 5 términos más por la siguiente razón. Dado que el término en la posición 800 en la lista por frecuencia corregida tiene el resultado 0.14, pero que hay 5 términos más con este mismo resultado que vienen después debido al orden alfabético, estos últimos tienen el mismo valor que el primero. Esto significa que se seleccionan, en total, 805 términos económicos. Por lo que se refiere a los términos auxiliares, la ganancia en cobertura disminuye considerablemente después de las 150 primeras formas – se reduce a un 2,96% entre las 150 y 200 UL, y a un 3,06% entre las 200 y 265 UL -, punto a partir del cual la selección en base a la frecuencia corregida deja de tener sentido.

3.3. *El léxico general*

La tasa de léxico general en el corpus es muy elevada (el 61,02% de las UL y el 29,99% de las OC). El objetivo de este apartado es examinar cómo se puede reconocer el léxico general típico del discurso económico empresarial académico. A este propósito se ha sometido el léxico general del corpus modélico a una comparación con el léxico de un corpus de español

general, el corpus Cumbre[3] (19.412.588 *tokens*). Concretamente se busca confirmación de las siguientes hipótesis :

1. la existencia de un léxico subeconómico, o sea UL con un significado orientado hacia el mundo de la economía, aunque no definidas como términos por los especialistas, por ejemplo *importe*, *suma*, *ascender*, etc.
2. la existencia de un léxico general básico, común al lenguaje general y al lenguaje económico. Este léxico general fundamental, se denomina '*core vocabulary*' en la literatura anglosajona (por ejemplo Dollerup *et al.,* 1989 : 22 ; Stubbs, 2001 : 42-43), que lo define como palabras muy frecuentes con una distribución semejante en una gran variedad de textos, estilísticamente neutras, semánticamente útiles por ser hiperónimos, e imprescindibles para comentar cualquier asunto.

A fin de comparar las frecuencias de UL de dos *corpora* de tamaño distinto, la estadística ha elaborado varias fórmulas. La más conocida es, sin duda, el *método del χ cuadrado o de Pearson*, pero éste requiere para el objetivo que nos proponemos frecuencias absolutas superiores a 5, lo que no es el caso de todas las UL generales en Econ. Una fórmula que no exige una frecuencia mínima es el denominado '*log-likelihood ratio*', aplicado a la comparación léxica de dos *corpora* por Rayson y Garside (2000). Este método parte de las frecuencias absolutas u observadas (O) de una palabra en dos corpus para calcular en primer lugar sus frecuencias teóricas o esperadas (E). Para ello se necesita conocer el número de formas léxicas en cada corpus (N) :

$$E_i = \frac{N_{ii} \sum_i O_i}{\sum_i N_i}$$

Las frecuencias esperadas de la palabra i corresponden al producto de N_i o el total de OC en el corpus concernido, con la suma de las frecuencias observadas dividido por la suma de la totalidad de OC en los dos corpus. Una vez que se conocen las frecuencias esperadas de la palabra, el '*log-likelihood ratio*' se calcula según la siguiente fórmula :

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

El '*log-likelihood ratio*' (LL) se obtiene al multiplicar por dos la suma del producto de cada frecuencia observada con el logaritmo de la división de esta misma frecuencia observada por su frecuencia esperada. El valor obtenido expresa la diferencia entre la frecuencia en el corpus de base y la del corpus de referencia. Cuanto más alta la cifra, más significativa la diferencia. Los resultados obtenidos mediante la fórmula de '*log-likelihood ratio'* son estadísticamente significativos a partir de 3,84 cuando alpha = 0.05 y sólo a partir de 6.63 cuando alpha = 0.01 (χ^2 α[1] = 3,84/6,63 para un test unilateral, Kanji 1993 : 168).

Al aplicar la fórmula de *log-likelihood* a las frecuencias absolutas y esperadas de las UL generales se obtiene una cifra para las UL generales que figuran en ambos corpus. Dado que en lingüística se suele adoptar un umbral de probabilidad de error de 1 sobre 20 o del 5%,

[3] El corpus, compuesto por A. Sánchez *et al.* y cuyo léxico fue puesto a nuestra disposición gracias a su amabilidad, cuenta con 54.737 *types* que se realizan 19.412.588 veces y "es ampliamente representativo de la lengua española en España e Hispanoamérica (sin olvidar las áreas hispanohablantes de Estados Unidos), en su variedad escrita y oral, y en géneros y ámbitos variados" (2001 : 8).

incluimos entre las UL características del lenguaje económico empresarial académico todas las palabras cuya frecuencia relativa en Econ es superior a la que aparece en Cumbre y que tienen un resultado '*log-likelihood*' superior a 3,84. En total se trata de 482 UL que ocurren juntas 18.999 veces en Econ (el 16,05% de todas las OC) y 750.955 veces en Cumbre (el 3,87% de las OC). Consideramos a estas 482 UL el léxico subeconómico del corpus Econ, ya que parece que está al servicio del léxico económico.

Después de haber encontrado un método para determinar qué léxico general es típico de una lengua especializada – en este caso la económica -, está por resolver la cuestión de encontrar un método para detectar el vocabulario general básico. Para 1.856 UL en Econ el resultado obtenido con el '*log-likelihood ratio*' no se puede considerar estadísticamente significativo. Esto significa que son UL iguales de frecuentes en los dos corpus. Ahora bien, para poder contar estas UL como pertenecientes al '*core vocabulary*' de la lengua española, es necesario ordenarlas por su frecuencia corregida y examinar a partir de qué punto cesa el progreso sustancial de cobertura en el corpus. Este punto se introduce a partir de las 1.025 UL, cuando la progresión de la tasa de cobertura cae por debajo del umbral del 3,5% propuesto por Hazenberg**.**

Finalmente, la aplicación del '*log-likelihood ratio*' a las UL generales de Econ indica 318 UL como significativamente más frecuentes en Cumbre que en Econ. No obstante, al mirar estas UL, muchas de ellas – las más frecuentes – parecen intuitivamente fundamentales para la lengua en general, como se puede constatar en la siguiente lista que presenta las 10 primeras ordenadas según su frecuencia corregida : *hacer, decir, sólo, dar, saber, sí, quedar, menos, vez* y *crear*. De ahí que se haya aplicado una vez más el criterio propuesto por Hazenberg, por lo cual quedan seleccionadas las 180 UL más frecuentes y mejor repartidas de este grupo.

4. Conclusión

La eficacia de las técnicas utilizadas ha sido puesta a prueba de la siguiente manera. Respecto de la tasa de cobertura necesaria para la comprensión de un texto, los resultados de la investigación experimental (por ejemplo Laufer, 1992 ; Nation, 1990, 2001) sugieren que un estudiante medio necesita entender entre el 95% y el 98% de las OC en un texto escrito para que su comprensión de dicho texto sea aceptable, o sea para que obtenga una nota de aprobado. Hazenberg (1994) añade que a partir de un 95% de comprensión, el alumno es capaz de inferir el significado del léxico desconocido en base al contexto.

Ahora bien, juntas, las OC de las UL seleccionadas cubren el 98,22% del corpus Econ, una tasa que supera con creces el 95% requerido según la bibliografía para asegurar una comprensión mínima e, incluso, el umbral del 98% que garantiza una lectura fluida y agradable sin asistencia necesaria. Este resultado se restringe, por supuesto, al corpus examinado. A fin de componer el léxico que cubra por lo menos el 95%, o mejor aun, el 98% de cualquier texto de discurso económico empresarial, se necesita ampliar el presente corpus con otros textos más.

Referencias

Ahmad K. y Rogers M. (2001). Corpus Linguistics and Terminology Extraction. In Wright S. y Budin G. (Eds), *Handbook of Terminology Management,* vol. (2). John Benjamins Publishing Company : 725-760.

Cabré M.T. (1999). *La terminología. Representación y comunicación*. IULA.

Clijsters W. (1990). *Mille lettres d'affaires en chiffres*. Champion/Slatkine.

Dollerup C. *et al*. (1989). Vocabularies in the reading process. *AILA Review*, vol. (6) : 1-33.

Hazenberg S. (1994). *Een keur van woorden.* Ridderprint.

Juilland A. y Chang-Rodríguez E. (1964). *Frequency dictionary of Spanish words.* Mouton&Co.

Kocourek R. (1982). *La langue française de la technique et de la science*. Oscar Brandstetter Verlag.

Laufer B. (1992). How much lexis is necessary for reading comprehension? In Béjoint H. y Arnaud P. (Eds), *Vocabulary and Applied Linguistics*. MacMillan : 126-132.

Müller Ch. (1977). *Principes et méthodes de statistique lexicale*. Hachette.

Nation I.S.P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Nation I.S.P. (1990). *Teaching and learning vocabulary.* Newbury House.

Pérez Gorostegui E. (1997). *Introducción a la administración de empresas.*Editorial Centro de Estudios Ramón Areces

Rayson P. y Garside R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38 annual meeting of the Association for Computational Linguistics (ACL 2000), 1-8 October 2000, Hong Kong, 1-6, en línea : http://www.comp.lancs.ac.uk/computing/users/paul/publications/rg_acl2000.pdf

Rosengren I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Etudes de Linguistique Appliquée*, vol. (1) : 103-127.

Sager J.C. (1990). *A practical course in terminology processing.* John Benjamins Publishing Company.

Sánchez A. *et al*. (2001). *Gran diccionario de uso del español actual*. SGEL.

Stubbs M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Blackwell Publishers.

Sutarsyah C. *et al*. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal*, vol. (25/2) : 34-50.

Temmerman R. (2000). *Towards new ways of terminology description. The sociocognitive approach*. John Benjamins Publishing Company.

Vangehuchten L. (2003). *El léxico del discurso económico empresarial : elaboración de una metodología con vistas a su descripción y análisis en ELE*. Tesis inédita.