COMPARACIÓN DE ALGUNAS CARACTERÍSTICAS LINGÜÍSTICAS DEL DISCURSO ESPECIALIZADO FRENTE AL DISCURSO GENERAL: EL CASO DEL DISCURSO ECONÓMICO[1]

Mª TERESA CABRÉ, CARME BACH, IRIA DA CUNHA,
ALBERT MORALES, JORGE VIVALDI
Universitat Pompeu Fabra, Barcelona

RESUMEN

En este trabajo se presenta un análisis de las características lingüísticas particulares del discurso especializado frente a las características del discurso general, desde el punto de vista de la lingüística de corpus. En concreto, comparamos la frecuencia de aparición de diversos elementos en un corpus de textos especializados de economía en español y en un corpus de textos de economía aparecidos en la prensa española. El objetivo es detectar las diferencias lingüísticas entre textos que comparten un mismo tema pero difieren en cuanto a su especialización, al ser en un caso textos escritos por especialistas y, en otro, textos escritos por periodistas. Las diferencias más representativas se tomarán como criterios para la elaboración de una herramienta semiautomática de detección de textos especializados.

Palabras clave: discurso especializado, discurso general, economía, lingüística de corpus.

ABSTRACT

In this paper we present an analysis of the specific linguistic features of specialized discourse compared to the ones related to general discourse, from a corpus linguistics approach. We therefore focus our research on the contrast of the frequency rate of some lexical units in a textual corpus of Economics in Spanish, integrated by two different subcorpora. On the one hand, a set of texts published by experts on this field. On the other hand, we find articles published in Spanish in written press. This study is aimed to detect linguistic differences among multiple texts with a different degree of specialization concerning the same issue. The most representative differences may constitute a set of criteria for the future development of a semiautomatic tool for retrieving specialized texts.

Keywords: specialized discourse, general discourse, economics, corpus linguistics.

1. INTRODUCCIÓN

En este trabajo presentamos un análisis de las características lingüísticas particulares del discurso especializado en contraposición con las características del discurso general, desde el punto de vista de la lingüística de corpus. En concreto, comparamos la frecuencia de aparición de diversos elementos lingüísticos en un corpus de textos especializados de economía en español y en un corpus de textos de economía aparecidos en la prensa española. El objetivo es detectar las diferencias lingüísticas existentes entre textos que comparten un mismo tema (en este caso, la economía) pero difieren en cuanto a su especialización, al ser en un caso textos escritos por especialistas y, en otro, textos escritos por periodistas.

Durante las últimas décadas, se ha producido una explosión en la producción de documentos de texto y, en consecuencia, hay una necesidad creciente de soluciones para organizar estos documentos. La Recuperación de Información es la disciplina que se ocupa de este tema y sus campos de aplicación tradicionalmente han sido, entre otros, la categorización de noticias de agencias según temáticas preestablecidas, la clasificación de páginas web, el filtrado de correo no deseado y la identificación del autor de un documento. El trabajo de Sebastiani (2002) es una interesante reseña de la aplicación de diferentes técnicas de aprendizaje automático a estas tareas. Lamentablemente la clasificación de textos en función de su nivel de especialidad ha recibido escasa atención por parte de la comunidad científica internacional. El programa Poppins (http://www.poppinsweb.com) representa una excepción a esta tendencia general

Dicho todo esto, puede afirmarse que nuestro estudio tiene tres objetivos:

1. Objetivo general: Analizar las características lingüísticas particulares del discurso especializado en contraposición con las características del discurso general.
2. Objetivo específico: Detectar diferencias lingüísticas entre textos que comparten un mismo tema (en este caso, la economía) pero

difieren en cuanto a su especialización, al ser en un caso textos escritos por especialistas y, en otro, textos escritos por periodistas.

3. Objetivo aplicado: Emplear las diferencias más representativas como criterios para la elaboración de una herramienta semiautomática de detección de textos especializados.

A continuación, en primer lugar, explicamos la metodología empleada en nuestro trabajo. En segundo lugar, detallamos cómo hemos conformado el corpus de análisis. En tercer lugar, delimitamos las unidades de análisis, partiendo de los trabajos de Cabré et al. (2007) y Cabré (2007). En cuarto lugar, detallamos el método de búsqueda y análisis de las unidades, y mostramos los resultados obtenidos. En quinto lugar, establecemos las conclusiones en base a los resultados obtenidos. Finalmente, listamos la bibliografía consultada.

2. METODOLOGÍA

La metodología que hemos seguido en este estudio pasa por varias fases. En primer lugar, conformamos el corpus, dividido en un subcorpus de textos especializados de economía y en un subcorpus de textos de economía de la prensa general, ambos en español.

En segundo lugar, delimitamos los rasgos lingüísticos que nos interesa analizar. Para ello, partimos de los trabajos de Cabré (2007) y Cabré et al. (2007), donde se presentaron algunos avances sobre la detección de aspectos lingüísticos específicos del discurso especializado con relación al discurso general, a partir del marco de la Teoría Comunicativa de la Terminología (TCT) de Cabré (1999).

En tercer lugar, seleccionamos las herramientas informáticas mediante las que realizamos las búsquedas, efectuamos dichas búsquedas y obtenemos resultados cuantitativos, que convertimos en porcentajes de frecuencia de los rasgos gramaticales buscados.

Finalmente, en base a los resultados obtenidos y teniendo en cuenta los dos trabajos ya mencionados, establecemos una serie de conclusiones, siempre de cara a establecer criterios para la elaboración de una herramienta automática de detección de textos especializados.

3. CORPUS

Como ya hemos comentado, el corpus de análisis está dividido en un subcorpus de textos especializados de economía y en un subcorpus de textos de economía de la prensa general (en concreto, de *El País* y de *La Vanguardia*), ambos en español. Cada uno de ellos consta de un millón de palabras.

Los textos se han extraído del Corpus Técnico del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF). Este corpus está formado por textos especializados de seis ámbitos (derecho, economía, medicina, medioambiente, informática y prensa) en cinco lenguas (español, catalán, inglés, francés y alemán). Todos los textos que contiene este corpus fueron seleccionados por especialistas de cada uno de los ámbitos y se agruparon a partir de una clasificación temática propuesta por los mismos especialistas. Posteriormente, estos textos se marcaron de acuerdo con el estándar SGML y con las directrices marcadas por el *Corpus Encoding Standard* (CES) de la iniciativa EAGLES.

La única limitación práctica con la que nos encontramos al conformar el corpus fue el hecho de que las noticias de economía de prensa que incluye el Corpus Técnico son fragmentos de documentos más amplios (que incluyen fragmentos de noticias de otros ámbitos temáticos), por lo que fue necesario crear un programa para extraer automáticamente estos fragmentos y poder formar el subcorpus de noticias de economía de la prensa.

4. UNIDADES DE ANÁLISIS

Para delimitar las unidades de análisis partimos, como ya se ha mencionado, de los resultados obtenidos en los trabajos de Cabré (2007) y Cabré et al. (2007).

En Cabré et al. (2007) se analizó la frecuencia de aparición de sustantivos, verbos y adjetivos, en sus diversas variantes, en un corpus de textos especializados (de derecho, economía, informática, medioambiente y medicina) y en un corpus de textos del ámbito general (noticias de prensa de diversos temas). Cada corpus incluía cinco millones de palabras.

Nosotros tomamos los rasgos lingüísticos que los autores consideraron representativos, es decir, las unidades que obtuvieron como resultado una diferencia de más del 20% en cuanto a su frecuencia de aparición en cada uno de los corpus. Los rasgos lingüísticos seleccionados son los siguientes:

- nombres propios
- nombre + adjetivo calificativo
- nombres acabados en –ción
- verbos en pasado
- verbos en 1ª persona singular
- verbos en 2ª persona plural

Los resultados de Cabré et al. (2007) indican que los nombres propios, los verbos en pasado, los verbos en 1ª persona singular y los verbos en 2ª persona del plural son más frecuentes en el discurso general, mientras que las secuencias "nombre + adjetivo calificativo" y los nombres acabados en –ción son más frecuentes en el discurso especializado.

En Cabré (2007) se empleó el mismo corpus de análisis para realizar nuevas búsquedas y se llegó a nuevas conclusiones. Por un lado, se observó que tienen una representatividad significativa en el discurso general:

- pronombres tónicos de 1ª persona
- conjunciones *pero*, *porque* y *ni*

Por otro lado, se constató que hay otros elementos más representativos en el discurso especializado, como son:

- conjunción *o*
- marcadores metadiscursivos
- voz pasiva
- pronombre *uno*

Asimismo, se observó que en los textos de especialidad se usa escasamente la 1ª persona del singular (*yo*) a favor de la 1ª persona

plural (*nosotros*) y que no se emplea la 2ª persona del singular y del plural.

En nuestro estudio analizamos la frecuencia de aparición de todos estos rasgos en nuestros dos subcorpus, de cara a corroborar la representatividad detectada en los dos trabajos mencionados. También buscamos otros rasgos, como son:

- conjunciones
- conjunción *y*
- adverbios
- adjetivos
- preposiciones
- pronombres
- pronombre *se*
- pronombres átonos
- pronombres tónicos
- pronombres tónicos de 2ª persona
- nombre + preposición + (artículo) + nombre
- subjuntivo
- infinitivo
- gerundio
- participio
- verbos en 1ª persona plural
- verbos en 2º persona singular

5. ANÁLISIS Y RESULTADOS

Una vez delimitadas las unidades de análisis, realizamos las búsquedas. Para ello empleamos BwanaNet, la herramienta de explotación del Corpus Técnico del IULA, que permite realizar búsquedas complejas (por lema, por forma, por secuencias, etc.) sobre los textos seleccionados de este corpus (véase Figura 1). En nuestro caso, las búsquedas se realizan sobre textos pertenecientes al subcorpus especializado de economía y sobre textos pertenecientes al subcorpus de prensa (en concreto sobre noticias de economía). Puede accederse a esta herramienta de explotación del corpus desde la siguiente dirección: http://bwananet.iula.upf.edu/.

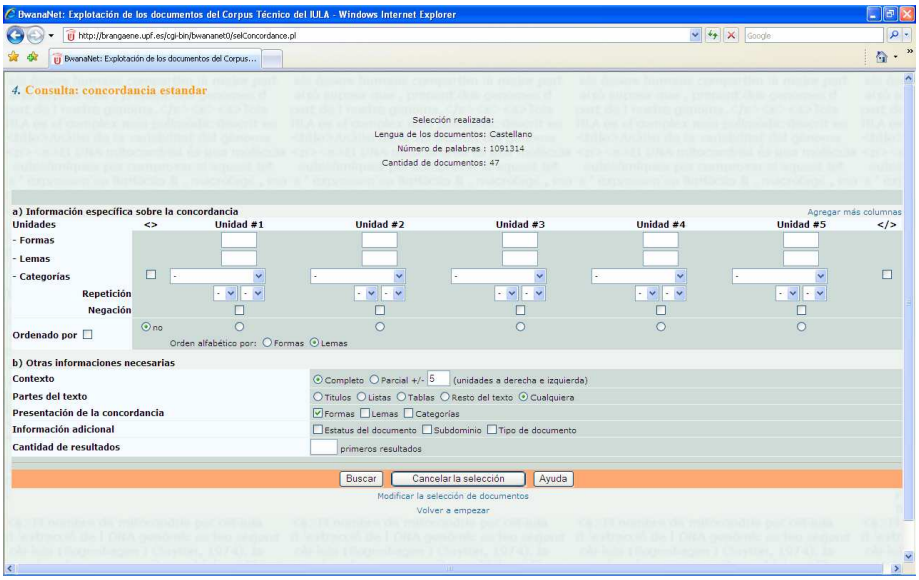

Figura 1. Pantalla de búsqueda de BwanaNet

Mediante el análisis con BwanaNet obtenemos la frecuencia de aparición de las unidades buscadas. A continuación calculamos el porcentaje de ocurrencias de cada una de ellas en los textos de economía especializados y en los textos de economía de la prensa. Consideramos que un rasgo lingüístico es relevante cuando su frecuencia de aparición en textos especializados y textos de prensa difiere en más de un 20%.

Con el análisis realizado mediante BwanaNet se obtienen los resultados mostrados en la Tabla 1. Se observa que los rasgos representativos en cada subcorpus son los siguientes:[2]

Textos especializados:

- conjunción *ni*
- conjunción *o*
- nombres en -ción
- pronombres átonos
- pronombres tónicos
- subjuntivo
- verbos en 1ª persona del plural
- marcadores metadiscursivos
- voz pasiva
- nombre + adjetivo calificativo

Textos generales:

- nombres propios
- pronombres de 2ª persona
- tiempos en pasado
- verbos en 1ª persona singular
- verbos en 2º persona singular

Unidades	Corpus de prensa		Corpus especializado	
	nº unidades	% unidades	nº unidades	% unidades
conjunciones	40320	44,91%	49446	55,09%
y	20809	48,05%	22495	51,94%
ni	375	39,26%	580	60,63%
o	1799	27,17%	4822	72,82%
pero	1294	50,23%	1282	49,76%
porque	686	48,72%	722	51,27%
adverbios	39548	44,4%	49500	55,6%
adjetivo	72386	46,72%	82517	53,27%
nombres propios	57335	70,37%	24135	29,62%
nombres en –ción	16142	38,7%	25582	61,3%
nombre + adjetivo	32314	40%	47404	60%
nombre + prep + (art) + nombre	58402	48,7%	61388	51,3%
preposiciones	182237	50,6%	177833	49,4%
pronombres	42633	46%	49812	54%
se	13320	50,8%	12856	49,2%
pronombres átonos	4024	34%	7783	66%
pronombres tónicos	1393	36,4%	2429	63,6%
pron. tónicos 1ª per.	177	50%	174	50%
pron. tónicos 2ª per.	63	92,6%	5	7,4%
uno	21233	51%	20336	49%
indicativo	54261	43,7%	69810	56,3%
gerundio	2738	43%	3639	57%
infinitivo	22711	47,1%	25460	52,9%
subjuntivo	1457	14,6%	8495	85,4%
pasado	22377	61,8%	13831	38,2%
verbos en 1º persona (sg y pl)	3708	41,8%	5157	58,2%
verbos en 1º sg	2276	77,1%	676	22,9%
verbos en 1º pl	1432	30,1%	4481	69,9%
verbos en 2º persona (sg y pl)	915	86,6%	142	13,4%
verbos en 2º sg	905	86,9%	136	13,1%
verbos en 2º pl	10	62,5%	6	37,5%
pasiva	500	32,8%	1022	67,2%
marcadores metadiscursivos	503	23,8%	1613	76,2%

Tabla 1. Resultados obtenidos mediante BwanaNet

6. CONCLUSIONES

Una vez obtenidos los resultados del análisis realizado, hemos determinado cuáles son las unidades lingüísticas más representativas del discurso especializado en contraposición con el discurso general en textos que traten del mismo tema, en este caso la economía. Se confirman o no algunas de las conclusiones de Cabré (2007) y Cabré et al. (2007) y se extraen algunas otras.

Mediante el análisis realizado con BwanaNet se obtienen diversas conclusiones. En primer lugar, en este trabajo se confirma la predominancia de ciertos rasgos lingüísticos (ya detectados en los dos estudios anteriores) en el discurso especializado, como son los nombres acabados en -ción, las secuencias nombre + adjetivo, la conjunción *o*, la voz pasiva y los marcadores metadiscursivos (más presentes en el discurso especializado por la necesidad de retomar e introducir términos propios de cada ámbito, en este caso la economía). A su vez, se confirma también la predominancia de otros rasgos en el discurso general, como son los nombres propios, los tiempos en pasado y los verbos en 1ª persona del singular.

En segundo lugar, hay ciertos resultados obtenidos en los dos estudios anteriores que no concuerdan con los resultados de este trabajo. En Cabré (2007) se observa que los pronombre tónicos de 1ª persona y las conjunciones *pero* y *porque* tienen una frecuencia más representativa en el discurso general, y que el pronombre *uno* a su vez es más frecuente en el discurso especializado. Estos resultados no se reflejan en el presente estudio, en el que estas unidades tienen una frecuencia similar en ambos corpus (general vs. especializado). Se observa también una divergencia relevante, ya que en este estudio la conjunción *ni* tiene una frecuencia muy representativa en el discurso especializado, mientras que en Cabré (2007) dicha conjunción predomina en el discurso general.

La presencia significativa de las conjunciones *ni* y *o* frente a la conjunción *y* en el discurso especializado de la economía puede ser un indicio de que en este tipo de discurso se presentan muchas más disyunciones que en el discurso especializado en general. De todos modos, habrá que demostrar esta afirmación con estudios posteriores más detallados.

Finalmente, el trabajo aquí presentado refleja nuevos resultados relevantes. Por un lado, se observa en el discurso general una predominancia representativa de los pronombres tónicos de 2ª persona y de los verbos en 2ª persona singular. Por otro lado, los resultados indican que en el discurso especializado predominan los pronombres tanto tónicos como átonos, los tiempos en subjuntivo y los verbos en 1ª persona del plural.

El análisis de especificidades de Lexico3, por su parte, refleja que el uso de algunas de las formas analizadas (la conjunción *o* en especial) está por encima de la media esperada estadísticamente y que, por tanto, podrían ser rasgos lingüísticos que ayuden a discriminar el discurso especializado. De todas maneras, en este artículo presentamos unas conclusiones muy primerizas en este sentido, que deberían constatarse con estudios futuros.

La aparición o no de las unidades lingüísticas representativas detectadas será uno de los criterios en los que se basará la futura herramienta semiautomática de detección de textos especializados que pretendemos desarrollar.

NOTAS

[1] Este trabajo se enmarca en el proyecto "TEXTERM III. Fundamentos, estrategias y herramientas para el procesamiento, extracción y representación de información especializada", financiado por el Ministerio de Educación y Ciencia (HUM2006-09458), y a su vez en el grupo de investigación consolidado IULATERM del Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra, Barcelona).

[2] Para contrastar los resultados obtenidos en cuanto a las conjunciones *y*, *o*, *ni*, *pero* y *porque* empleamos un programa estadístico de lexicometría: Lexico3 (Lamalle et al. 2003). Este programa busca formas concretas de las unidades léxicas y ofrece su "nivel de especificidad". Mediante este análisis se obtiene que, de las cinco formas analizadas, cuatro (*y*, *o*, *ni*, *pero*) presentan especificidad positiva en el bloque de textos especializados. La forma *o* es la que, según el análisis estadístico de Lexico3, presenta un índice de especificidad positiva más alto en dicho subcorpus, por lo que su elevada presencia podría ser un rasgo que permitiese caracterizarlo como especializado. El resto de formas (*y*, *ni*, *pero*) también se presentan como específicas de dicho bloque de textos, aunque su grado de especificidad es inferior.

REFERENCIAS BIBLIOGRÁFICAS

Cabré, M. T. (1999): *La terminología. Representación y comunicación*. Barcelona: IULA-UPF.

Cabré, M. T. (2007): "Constituir un corpus de textos de especialidad: condiciones y posibilidades". En Ballard, M.; Pineira-Tresmontant, C. (eds.). *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université. 89-106.

Cabré, M. T.; Bach, C.; Castellà, J. M.; Martí, J. (2007): "La caracterización lingüística del discurso especializado". En Mairal, R. et. al. (eds.). *Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva: perspectivas aplicadas entre disciplinas. Actas del XXIV Congreso Internacional de AESLA*. Madrid: UNED-AESLA. 851-857.

Lamalle, C. et al. (2003): *Manuel d'utilisation. Lexico3 (Version 3.41 - Février 2003).* París: SYLED - CLA^{2}T. Université de la Sorbonne nouvelle - Paris 3.

Sebastiani, F. (2002): "Machine Learning in Automated Text Categorization". *ACM Computing Surveys* 34 (1).1-47.