

SIMPLE EXTRACTOR: VALIDACIÓN DE LA HERRAMIENTA Y APLICACIÓN EN EL ANÁLISIS LINGÜÍSTICO

CRISTINA TEJEDOR MARTÍNEZ
LAURA MARTÍN-PÉREZ GONZÁLEZ
NATALIA MUÑOZ PARDO
Universidad de Alcalá, Alcalá de Henares

RESUMEN

La herramienta SimpleExtractor ofrece unas prestaciones para estudios lingüísticos que es necesario validar dado que se trata de un nuevo producto. Para ello, describiremos las características de la herramienta para entender su funcionamiento. A continuación, probaremos la herramienta para el análisis lingüístico del corpus y analizaremos los resultados. El estudio consistirá en analizar un corpus compilado seleccionando los discursos de los galardonados con el Premio Cervantes. El uso del SimpleExtractor nos ofrecerá listas del léxico utilizado en ambos subcorpus; el léxico será clasificado en categorías semánticas y se procederá a comparar ambas listas y clasificaciones. Este estudio permitirá determinar una aplicación real en el ámbito de la lingüística de dicha herramienta. De hecho, una de las conclusiones de este trabajo es que el desarrollo tecnológico ayuda a la automatización de tareas de análisis lingüístico.

Palabras clave: Lingüística de corpus, extractor terminológico, lexicología y semántica

ABSTRACT

SimpleExtractor tool has some features to be used for Linguistic research and has to be validated as it is a new software application. Therefore, we will describe briefly its characteristics in order to understand how it works. After that, the tool will be applied for the linguistic analysis of our corpus and the results will be analysed. This study deals with the analysis of the corpus compiled from the delivered speeches by the prize-winners with the Premio

Cervantes. The use of SimpleExtractor will offer us lists of the vocabulary used in both subcorpora; the vocabulary will be classified in semantic categories and the lists and classification will be compared. This study will allow us determine a real application of the tool in the linguistic field. In fact, one of the conclusions of our work is that technological development helps in the automation of tasks of linguistic analysis.

Keywords: corpus linguistics, terminological extractor, lexicology and semantics

1. INTRODUCCIÓN

Los estudios de carácter lingüístico realizados contando con corpus de textos se han generalizado, especialmente a partir del último cuarto del siglo XX; además, cabe señalar que “the trend towards the corpus-driven approach is growing, as people realize that the corpus can tell them more than they could ever imagine if they were to rely mainly on their own intuition” (Pearson, 1998: 50). Uno de los ejemplos destacables es la creación y desarrollo del corpus COBUILD liderado por el profesor John Sinclair. Este corpus ha tenido gran repercusión en la elaboración de materiales diversos y trabajos lexicográficos, así como en la preparación de otros ejemplos. La compilación de un corpus se basa en una serie de criterios (Renouf, 1987: 1-40) que buscan proporcionar representatividad para el estudio que se plantea tras su análisis; entre estos criterios se pueden destacar, por ejemplo, el tamaño, el tipo de textos, la/s lengua/s de los mismos, el periodo de tiempo que se cubre, etc. Otros autores como Leech (1991), Biber (1993) y McEnery, Xiao y Ono (2006), entre otros, describen los criterios para buscar la representatividad. Estos últimos autores (2006: 15) distinguen entre “two broad types of corpora in terms of the range of text categories represented in the corpus: *general* and *specialized* corpora”, nuestro trabajo se enmarca en el segundo grupo.

Cuando ya se han determinado los criterios, en función de la investigación que se quiere realizar, el siguiente paso es proceder al análisis de los datos contenidos en ese corpus (Sinclair, 1991 y 2004). Para facilitar esta labor, se han desarrollado diversas herramientas que nos ayudan en la extracción de información. Por ejemplo, una

herramienta utilizada por los investigadores es WordSmith, como se puede comprobar en la información disponible en la página web: http://lexically.net/wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm

El objetivo de este estudio es comprobar la utilidad de otra herramienta *SimpleExtractor* en la extracción del léxico de un corpus dividido en dos subcorpus para realizar un estudio lingüístico. La finalidad es extraer las listas de sustantivos en ambos subcorpus que corresponden a dos momentos históricos diferentes para comparar la carga semántica de esas unidades léxicas y señalar esas diferencias tras el estudio de la clasificación semántica del léxico utilizado por los galardonados. Se trata de poder aplicar este estudio a la enseñanza de lenguas. La decisión de trabajar con el género del discurso se debe a que no se utilizan frecuentemente en el aula y también porque queríamos comprobar si a través del estudio de los sustantivos se puede analizar y contrastar la temática en dos etapas distintas.

2. SIMPLEEXTRACTOR Y ANÁLISIS DEL CORPUS

2.1 *SimpleExtractor*: descripción de la herramienta

*SimpleExtractor*¹ es un extractor terminológico muy sencillo de manejo, pero potente en sus prestaciones, y que da prioridad a la gestión del mismo por su usuario. Es decir, los usuarios podemos configurar la capacidad de extracción, el cálculo de frecuencias, y la selección de términos que consideremos válidos, de hasta siete palabras, gestionando las listas de palabras vacías (palabras a ignorar) en varias lenguas: español, portugués, ruso, inglés. Permite además extraer términos desde ficheros en diferentes formatos y ordenar los resultados por frecuencia o alfabéticamente los términos extraídos y según número de palabras.

El diseño de *SimpleExtractor* está orientado al usuario, primando la facilidad y la sencillez de su uso, con interfaces intuitivos y un flujo de trabajo claro donde el usuario sabe siempre cuál es el próximo paso. Asimismo, *SimpleExtractor* se ha desarrollado como una aplicación de escritorio, de forma que se garantiza la confidencialidad de los textos y materiales de trabajo del usuario. Además, para poder

trabajar más cómodamente con la lista de términos extraídos, *SimpleExtractor* ofrece la posibilidad de realizar un informe que recoja, en diferentes formatos (txt, Word, PDF o CSV), dichos términos. Desde el punto de vista de las prestaciones, *SimpleExtractor*² se caracteriza por su alta velocidad de extracción y su capacidad de extracción.

2.2. Desarrollo del estudio

El estudio que se ha realizado tiene como objetivo la comparación del léxico utilizado en los discursos de los galardonados con el Premio Cervantes en dos etapas: la primera etapa cuando se instituyó el premio y prácticamente la última etapa del mismo. Para llevar a cabo el trabajo se ha procedido a la compilación del corpus de textos seleccionados que se subdivide en dos subcorpus: el primer subcorpus consistirá en los primeros 10 discursos impartidos entre 1976 y 1984 (8 años, porque hay dos galardonados en la edición de 1979); y el segundo subcorpus constará de los últimos diez discursos entre 2003 y 2012 (9 años).

El primer paso consistió en la extracción de términos simples de los diez primeros discursos con el programa *SimpleExtractor*, limitando la misma a una frecuencia de aparición igual o mayor a 10. De la misma manera se ha procedido con los discursos de la segunda etapa analizada. Una vez obtenida esta lista de términos frecuentes, se han seleccionado de forma manual los sustantivos, por ser las palabras que tienen mayor carga semántica, dado que el objetivo de nuestro estudio es comparar aspectos semánticos a través del léxico utilizado en los mismos. La herramienta proporciona un informe que nos ha permitido trabajar con la lista de términos en un archivo Word.

Para la clasificación del léxico seleccionado de ambos subcorpus se ha seguido la propuesta hecha por MacArthur (1981) en su diccionario semántico del vocabulario de la lengua inglesa, *Longman Lexicon of Contemporary English*. Este lexicón distribuye el vocabulario de la lengua inglesa a través de 14 categorías de tema general que son matizadas incorporando unas subcategorías de carácter más específico. Estas categorías son: (1) Life and Living Things; (2) The Body: Its Functions and Welfare; (3) People and the

Family; (4) Buildings, houses, the home, clothes, belongings, and personal care; (5) Food, drink, and farming; (6) Feelings, emotions, attitudes, and sensations; (7) Thought and communication, language and grammar; (8) Substances, Materials, Objects, and Equipment; (9) Arts and crafts, science and technology, industry and education; (10) Numbers, measurement, money, and commerce; (11) Entertainment, sports, and games; (12) Space and Time; (13) Movement, Location, Travel and transport; (14) General and Abstract Terms. Como es sabido, los nombres propios no se recogen generalmente en este tipo de obras de referencia, pero son de especial importancia en nuestro análisis de los discursos, por ello se ha añadido la categoría (15) Proper noun.

Para clasificar los elementos léxicos se ha tenido en cuenta el contexto de uso de las palabras que nos ofrece siempre el *SimpleExtractor*. Por ejemplo, la palabra *razón* puede hacer referencia a una facultad de los seres humanos, pero también es sinónimo de *causa* o *motivo*. Por esta razón, se ha revisado en la misma extracción el contexto en el que aparece, porque va a ser determinante para encuadrar esta palabra en una u otra categoría o en ambas.

Los datos obtenidos se muestran en las siguientes tres tablas:

	Nº de palabras	Nº de sustantivos	Porcentaje
Subcorpus 1	4642	1371	33.9%
Subcorpus 2	6030	2149	35.6%
Total	10672	3520	32.9%

Tabla 1. Resultados globales

Al analizar los datos de la tabla 1, se observa un porcentaje similar en cuanto al número de sustantivos utilizados en ambos subcorpus, solamente aumenta un 1.7%. Podemos señalar que al estudiar la clasificación de los sustantivos, en el primer periodo se han contabilizado 61 tipos de sustantivos y en el segundo periodo 91. Se ha comprobado que 30 tipos de sustantivos aparecen utilizados en ambos subcorpus.

	CATEGORÍAS SEMÁNTICAS: subcorpus 1	Nº sustantivos	Tipos de sustantivos
1	LIVE AND LIVING THINGS	67	2
2	THE BODY, ITS FUNCTIONS AND WELFARE	59	5

3	PEOPLE AND FAMILY	70	5
4	BUILDINGS, HOUSES, THE HOME, CLOTHES, BELONGINS AND PERSONAL CARE	10	1
5	FOOD, DRINK AND FARMING	0	0
6	FEELINGS, EMOTIONS, ATTITUDES AND SENSATIONS	57	2
7	THOUGHT AND COMMUNICATION, LANGUAGE AND GRAMMAR	305	14
8	SUBSTANCES, MATERIALS, OBJECTS AND EQUIPMENT	48	4
9	ARTS AND CRAFTS, SCIENCE AND TECHNOLOGY, INDUSTRY AND EDUCATION	49	2
10	NUMBERS, MEASUREMENT, MONEY AND COMMERCE	63	2
11	ENTERTAINMENT, SPORTS AND GAMES	0	0
12	SPACE AND TIME	270	12
13	MOVEMENT, LOCATION, TRAVEL AND TRANSPORT	0	0
14	GENERAL AND ABSTRACT TERMS	52	3
15	PROPER NOUN	302	9
		1371	61

Tabla 2. Clasificación semántica del léxico: primer subcorpus

	CATEGORÍAS SEMÁNTICAS: subcorpus 2	N° sustantivos	Tipos de sustantivos
1	LIVE AND LIVING THINGS	95	3
2	THE BODY, ITS FUNCTIONS AND WELFARE	34	2
3	PEOPLE AND FAMILY	183	8
4	BUILDINGS, HOUSES, THE HOME, CLOTHES, BELONGINS AND PERSONAL CARE	20	1
5	FOOD, DRINK AND FARMING	0	0
6	FEELINGS, EMOTIONS, ATTITUDES AND SENSATIONS	70	4
7	THOUGHT AND COMMUNICATION, LANGUAGE AND GRAMMAR	777	32
8	SUBSTANCES, MATERIALS, OBJECTS AND EQUIPMENT	38	2
9	ARTS AND CRAFTS, SCIENCE AND TECHNOLOGY, INDUSTRY AND EDUCATION	97	6
10	NUMBERS, MEASUREMENT, MONEY AND COMMERCE	0	0
11	ENTERTAINMENT, SPORTS AND GAMES	78	3
12	SPACE AND TIME	322	10
13	MOVEMENT, LOCATION, TRAVEL AND	41	3

	TRANSPORT		
14	GENERAL AND ABSTRACT TERMS	80	5
15	PROPER NOUN	314	12
		2149	91

Tabla 3. Clasificación semántica del léxico: segundo subcorpus

Atendiendo a la distribución de los términos, más del 60% de los sustantivos extraídos en ambos subcorpus se aglutinan en tres categorías: *Thought and communication, language and grammar; Space and time* y *Proper noun*. Sin embargo, se aprecian diferencias en el orden y distribución de las palabras en ambos subcorpus (Figura 1). En el primer subcorpus, tres categorías carecen de elementos; en el segundo, solamente dos. Cabe resaltar la importancia que la categoría *Proper noun* en ambos casos, pero especialmente en el primer subcorpus, ya que es la categoría con más ítems. Se detecta una forma de ritual dentro de los discursos, ya que los mismos nombres propios son repetidos tanto en el primer periodo como en el segundo, por ejemplo, Cervantes, el Quijote, España y referencia a América.

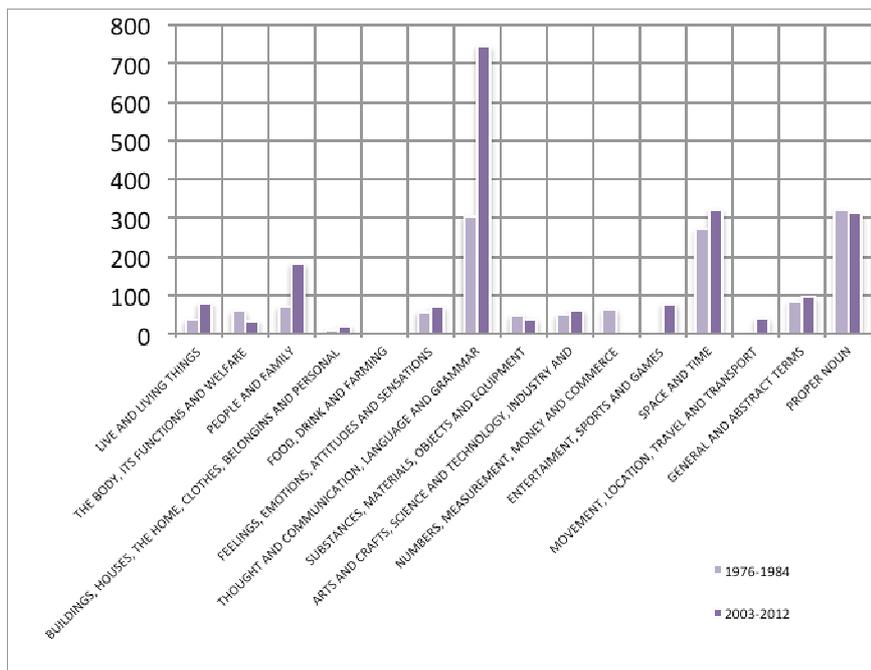


Figura 1. Clasificación semántica del corpus

La diferencia más notable entre los dos periodos es que se centran más en el tema *Thought and communication, language and grammar* en el segundo periodo (35.80%) que en el primero (22.2%). Así, los discursos del segundo periodo parecen centrarse en aspectos metalingüísticos, dando gran importancia a la lengua, libros, novelas, cuentos, poesías, a la lectura, palabras y al escritor; sin embargo, en el primer periodo se reparten los ítems de manera similar entre tres categorías *Thought and communication, language and grammar*; *Space and time* y *Proper noun*, aunque las referencias a la lengua, la novela, las palabras y los escritores aparecen con frecuencia en los discursos. En el primer subcorpus se menciona la palabra ‘crisis’, no en el segundo, cuando ya era una realidad en la sociedad; en cambio en este segundo periodo aparecen las palabras ‘muerte’, ‘pobreza’, ‘seriedad’, ‘trabajo’ y ‘guerra’.

Es preciso resaltar también la escasa importancia de la referencias a términos abstractos en los discursos, en contraposición a los términos que se refieren al espacio y al tiempo. Por último, resulta llamativo que en el segundo periodo se reduce totalmente la alusión a *Numbers, measurement, money, and commerce*, es decir, que es un tema a evitar.

3. CONCLUSIONES

El estudio ha consistido en extraer y contabilizar los sustantivos utilizados con una frecuencia de aparición de al menos diez veces en cada subcorpus en los discursos impartidos por los galardonados con el Premio Cervantes en dos periodos históricos; posteriormente se clasificaron atendiendo a lista de categorías utilizada previamente con el vocabulario de la lengua inglesa. Como pretendíamos comprobar, el extractor terminológico *SimpleExtractor* ha facilitado la tarea de recuperar los términos y su frecuencia en ambos subcorpus, así como mostrarnos el contexto original de cada ítem para realizar la clasificación en las categorías seleccionadas. Podemos concluir que se trata de una herramienta de fácil uso que permite diversas opciones para realizar investigaciones lingüísticas de corpus porque el usuario puede adaptar las opciones de búsqueda, adecuándolas a sus necesidades.

Los resultados del análisis de la clasificación de los sustantivos en categorías semánticas nos han permitido determinar la importancia que en los discursos tiene el vocabulario relacionado con el lenguaje, la gramática, el pensamiento y la comunicación, así como el uso casi establecido de un número de nombres propios indudablemente esenciales por el contexto de los discursos: recoger el Premio Miguel de Cervantes de literatura en lengua española. Este trabajo puede tener aplicaciones didácticas, de hecho la herramienta se puede utilizar con los estudiantes para realizar trabajos que conduzcan al aprendizaje del léxico y combinaciones léxicas, junto a la mejora de la comprensión y análisis de textos.

Tras la realización de este trabajo, se ha sugerido una mejora de la herramienta para las próximas versiones: que el extractor esté vinculado a un diccionario de forma que clasifique automáticamente las diferentes categorías gramaticales para facilitar futuras investigaciones.

NOTAS

¹ El diseño de esta herramienta ha sido llevado a cabo por el Grupo de investigación en Validación y Aplicaciones Industriales de la Universidad Politécnica de Madrid. La empresa DAIL Software SL ha desarrollado el prototipo.

² Se puede acceder al enlace <https://www.dail-software.com/es/> para más información sobre las características de la herramienta.

REFERENCIAS BIBLIOGRÁFICAS

- Biber, D. 1993. "Representativeness in corpus design", *Literary and Linguistic Computing*, 8/4: 243-257.
- Leech, G. (1991), "The state of an art in corpus linguistics", en Aijmer, K. & B. Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 8-29.
- MacArthur, T. 1981. *Longman Lexicon of Contemporary English*. Harlow: Longman.

- McEnery, T., Xiao, R., Ono, Y. 2006. *Corpus-based Language Studies. An Advanced Resource Book*. Oxon: Routledge.
- Pearson, J. 1998. *Terms in Context*. Amsterdam: John Benjamins.
- Renouf, A. 1987. "Corpus development", J.M. Sinclair (ed.). *Looking Up. An account of the COBUILD Project in lexical computing*. London: HarperCollins Publishers, 1-40.
- Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: University Press.
- Sinclair, J. 1994. "Corpus typology", *EAGLES DOCUMENT EAG-CSG/IR-T1.1*, Commission of the European Communities.
- Sinclair, J. 2004. *Trust the text: language, corpus and discourse*. London: Routledge.
- Scott, M. 2008. "Developing WordSmith", *IJES*, 8/1: 95-106.