Procesamiento del Lenguaje Natural, nº39 (2007), pp. 165-172 recibido 17-05-2007; aceptado 22-06-2007

Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor[1]

Gloria Corpas Pastor
Departamento de Traducción e Interpretación
Facultad de Filosofía y Letras
Universidad de Málaga
gcorpas@uma.es

Míriam Seghiri Domínguez
Departamento de Traducción e Interpretación
Facultad de Filosofía y Letras
Universidad de Málaga
seghiri@uma.es

Resumen: En las páginas que siguen a continuación vamos a describir un método[2] para calcular el umbral mínimo de representatividad de un corpus mediante el algoritmo N-Cor de análisis de la densidad léxica en función del aumento incremental del corpus. Se trata de una solución eficaz para determinar *a posteriori*, por primera vez de forma objetiva y cuantificable, el tamaño mínimo que debe alcanzar un corpus para que sea considerado representativo en términos estadísticos. Este método se ha visto implementado en la aplicación informática ReCor. Con dicha herramienta vamos a comprobar si un corpus de seguros turísticos en español que hemos compilado sería representativo para realizar estudios lingüístico-textuales y poder ser utilizarlo en traducción.
Palabras clave: Representatividad, lingüística de corpus, compilación de corpus, corpus especializado.

Abstract: In this paper we describe a method[3] to determine the representativeness threshold for any given corpus. By using the N-Cor algorithm it is possible to quantify *a posteriori* the minimum number of documents and words that should be included in a specialised language corpus, in order that it may be considered representative. This method has been implemented by means of a computer program (ReCor). This program will be used here to check whether a corpus of insurance policies in Spanish is representative enough in order to carry out text-linguistic studies and translation tasks.
Keywords: Representativeness, corpus linguistics, corpus compilation, specialised corpus**.**

1 Introducción

Hasta la fecha, mucho se ha escrito e investigado en torno la cantidad como criterio representativo así como sobre las posibles fórmulas capaces de estimar un mínimo de palabras y documentos a partir del cual un corpus especializado puede considerarse representativo sin llegar a resultados concluyentes.

Los intentos de fijar un tamaño, al menos mínimo, para los corpus especializados han sido varios. Algunos de los más significativos son los expuestos por Heaps (1978), Young-Mi (1995) y Sánchez Pérez y Cantos Gómez (1997). Según Yang et al. (2000: 21), tales propuestas presentan importantes deficiencias porque se basan en la ley de Zipf. La determinación del tamaño mínimo de un corpus sigue siendo uno de los aspectos más controvertidos en la actualidad (cf. Corpas Pastor y Seghiri Domínguez, 2007/en prensa). En este sentido, se han barajado cifras muy dispares. A modo de ilustración, diremos que Biber (1993), en uno de los trabajos más influyentes sobre corpus y representatividad, llega a afirmar que es posible representar la práctica totalidad de los elementos de un registro particular con relativamente pocos ejemplos, mil palabras, y un número reducido de textos pertenecientes a este registro, concretamente diez.

ISSN: 1135-5948

Urge, pues, resolver esta cuestión, ya que no podemos olvidar que la mayoría de estudios lingüísticos y traductológicos están utilizando corpus de reducidas dimensiones, adecuados para sus necesidades concretas de investigación, colecciones de textos que descargan directamente de fuentes de información electrónicas. La red de redes es hoy día uno de los principales proveedores de materia prima para esta lingüística de corpus "de andar por casa". Además, este tipo de corpus *ad hoc*, compilado virtualmente, ha demostrado ser tremendamente útil tanto para llevar a cabo estudios lingüísticos (cf. Haan, 1989, 1992; Kock, 1997 y 1991; Ghadessy, 2001) como para la enseñanza de segundas lenguas (Bernardini, 2000; Aston *et al*., 2004) y en traducción (Corpas Pastor, 2001, 2004, Seghiri Domínguez, 2006).

Las cifras tan dispares que se han manejado hasta la fecha, así como la poca fiabilidad que dan las propuestas para su cálculo, nos llevaron a reflexionar sobre una posible solución, que se ha visto materializada en la aplicación informática denominada *ReCor*, que pasamos a describir a continuación.

2 Descripción del programa ReCor

Dejando a un lado que la representatividad de un corpus depende, en primer lugar, de haber aplicado los criterios de diseño externos e internos adecuados, en la práctica, la cuantificación del tamaño mínimo que debe tener un corpus especializado aún no se ha abordado de forma objetiva. Y es que no hay consenso, como ha quedado manifiesto, sobre cuál sea el número mínimo de documentos o palabras que debe tener un determinado corpus para que sea considerado válido y representativo de la población que se desea representar. Las cifras varían, además, como hemos visto, de unos autores a otros. Pero todas estas cifras no resuelven el problema de calcular la representatividad de un corpus, dado que son cifras establecidas *a priori*, carentes de cualquier base empírica y objetivable.

Con este método pretendemos plantear una solución eficaz para determinar, por primera vez, *a posteriori* el tamaño mínimo de un corpus o colección textual, independientemente de la lengua o tipo textual de dicha colección, estableciendo, por tanto, el umbral mínimo de representatividad a partir de un algoritmo (N-Cor) de análisis de la densidad léxica en función del aumento incremental del corpus.

2.1. El algoritmo N-Cor

El presente método calcula el tamaño mínimo de un corpus mediante el análisis de la densidad léxica (*d*) en relación a los aumentos incrementales del corpus (*C*) documento a documento, según muestra la siguiente ecuación:

$$C_n = d_1 + d_2 + d_3 + ... + d_n$$

Figura 1: Ecuación base del algoritmo N-Cor

Para ello, se analizan gradualmente todos los archivos que componen el corpus, extrayendo información sobre la frecuencia de las palabras tipo (*types*) y las ocurrencias o instancias (*tokens*) de cada archivo del corpus. En esta operación se utilizan dos criterios de selección de archivos, a saber, por orden alfabético y de forma aleatoria, a fin de garantizar que el orden en el que son seleccionados los archivos no afecta al resultado. Cuando se seleccionan los documentos por orden alfabético, el algoritmo analiza el primer archivo y para éste se calculan los *tokens* y los *types*, y la densidad léxica correspondiente. Con ello ya se obtiene un punto en la representación gráfica que se pretende extraer. A continuación, siguiendo el mismo criterio de selección que en el primero, se toma el siguiente documento del corpus y se calculan de nuevo los *tokens* y los *types*, para éste, pero sumando los resultados a los *tokens* y los *types* de la iteración anterior (en este caso a los del primer documento analizado), se calcula la densidad léxica y con esto se obtiene un segundo punto para la representación gráfica. Se sigue este algoritmo hasta que se hayan tratado todos los documentos que componen el corpus que se estudia. La segunda fase del

análisis es idéntica, pero tomando los documentos en orden aleatorio.

Se emplea el mismo algoritmo para el análisis de n-gramas, esto es, la opción de realizar un análisis de la frecuencia de aparición de secuencias de palabras (2-grama, 3-grama…, n-grama). La aplicación ofrece la posibilidad de hacer el cómputo de estas secuencias considerando un rango de longitudes de secuencia (números de palabras) definido por el usuario. Al igual que se realiza con respecto a los (*tokens*), se muestra un gráfico con la información de representatividad del corpus tanto para un orden aleatorio de los ficheros como para un orden alfabético por el nombre de éstos. En el eje horizontal se mantendrá el número de ficheros consultados, y en el eje vertical el cociente (número de n-gramas distintos)/(número de n-gramas totales). A estos efectos, cada instancia de un n-grama es considerado como un *token*. Asimismo, los ficheros de salida generados indican los n-gramas.

Tanto en el análisis por orden alfabético como en el aleatorio de n-gramas llegará un momento en el que un determinado documento no aporte apenas *types* al corpus, lo cual indicará que se ha llegado a un tamaño adecuado, es decir, que el corpus analizado ya se puede considerar una muestra representativa de la población en términos estadísticos. En una representación gráfica estaríamos en el punto en el que las líneas de *types* y *tokens* se estabilizan y se aproximan al cero. Si el corpus es realmente representativo la gráfica tenderá a descender exponencialmente porque los *tokens* crecerán en cada iteración mucho más que los *types*, debido a que, en teoría, cada vez irán apareciendo menos palabras nuevas que no estén almacenadas en las estructuras de datos que utiliza el programa. Así pues, podremos afirmar que el corpus es representativo cuando la gráfica sea constante en valores cercanos a cero, pues los documentos siempre van a contener variables del tipo números o nombres propios, por ejemplo, que tenderán a constituir instancias de *hapax legomena* y, por tanto, aumentarán el grado de variabilidad léxica del corpus. Una posible solución podría ser el empleo de expresiones regulares y técnicas de análisis superficial (*shallow parsing*) para la detección de nombres propios. En cualquier caso, conviene señalar que, en la práctica, es imposible alcanzar la incorporación de cero *types* en el corpus, aunque, por el contrario, sí que irán presentado una tasa muy baja de incorporación, como permite predecir la ley Heaps.

2.1.2. Especificaciones del programa

ReCor es una aplicación informática creada con objeto de poder estimar la representatividad de los corpus en función de su tamaño y que se caracteriza, ante todo, por la sencillez de su interfaz de usuario (cf. Figura 2), frente a la carga eminentemente matemática y de formulación que abundan en este tipo de trabajos.

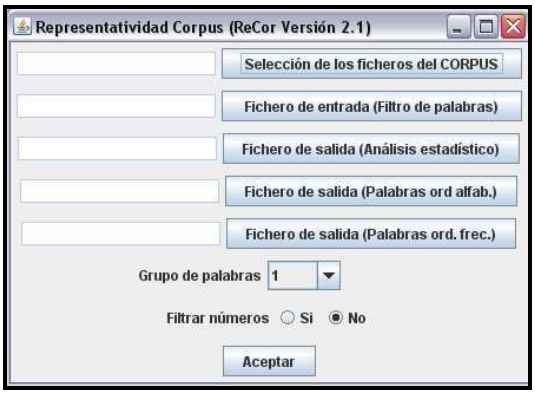

Figura 2: Interfaz de ReCor (versión 2.1)

Hasta el momento se han implementado tres versiones del programa *ReCor*: 1.0, 2.0 y 2.1. El funcionamiento es básicamente similar y corresponde a la descripción genérica que ofrecemos a continuación. Ahora bien, la versión 2.0 difiere de la versión 1.0 en que permite a) seleccionar automáticamente un directorio completo de documentos (en vez de tener que pulsar la tecla *Shift* como en la versión anterior) y b) permite seleccionar un número de n-gramas para el cálculo, donde $n \geq 1$ y $n \leq 10$. Ambas versiones (1.0 y 2.0) generan archivos estadísticos en texto plano (.txt). La versión 2.1. difiere de su predecesora en que presenta los archivos estadísticos simultáneamente en formato .txt y en forma de tablas en Excel.

3 Funcionamiento del programa

En este apartado mostraremos el programa *ReCor* en funcionamiento (versión 2.1.). Para la ilustración del funcionamiento del programa hemos compilado un corpus de seguros turísticos en español. Este corpus, por su diseño[4] —es monolingüe[5], comparable[6], textual[7] y especializado[8]—, responde a los parámetros de creación de corpus, por lo que estaría en condiciones de ser utilizado de forma independiente para la realización de estudios lingüísticos y traductológicos sobre los elementos formales de este tipo contractual.

Gracias a una sencilla interfaz, *ReCor* resulta de fácil manejo. Así, procedemos a la selección de los archivos que conforman el subcorpus de seguros turísticos en español mediante el botón «Selección de los ficheros del corpus». Una vez seleccionados los archivos que integran el corpus en español, podremos incorporar, si se desea, un «filtro de palabras». En nuestro caso, hemos incluido un filtro que contiene numeración romana. Además, el programa genera tres ficheros de salida (*Análisis estadístico, Palabras ord. alf.* y *Palabras ord. frec*.) que se crearán por defecto en la ubicación que determine la aplicación. Si se desea otra localización de los archivos de salida generados, puede indicarse una nueva ruta. El primero, «Análisis estadístico», recoge los resultados de dos análisis distintos; de un lado, los ficheros ordenados alfabéticamente por nombre; de otro, para los ficheros ordenados en orden aleatorio. El documento aparecerá estructurado en cinco columnas, a saber, muestra de *types*, *tokens*, cociente entre palabras distintas y totales (*types*/*tokes*), número de palabras con una parición (V1) y número de palabras con dos apariciones (V2). El segundo, «Palabras ord. alfa.», generará dos columnas en la que aparecerán las palabras ordenadas por orden alfabético, de una parte, y sus correspondientes ocurrencias, de otra. En tercer lugar, «Palabras ord. frec.», presenta la misma información que el fichero de salida anterior, pero esta vez las palabras se ordenan en función de su frecuencia, es decir, por rango.

Por último, procederemos a especificar «Grupo de palabras», esto es, los n-gramas. Escogemos, para una primera ilustración, uno (cf. Figura 3). Asimismo, indicaremos «sí» en la opción «Filtrar números».

3.1. Representaciones gráficas

Una vez se han seguido los pasos descritos más arriba, la aplicación está lista para realizar el análisis, cuyo resultado se expresa en forma de representaciones gráficas y ficheros de salida en .txt con datos estadísticos exportables a tablas y tablas en Excel. Para generar las representaciones gráficas A y B, pulsamos «Aceptar». *ReCor* creará, además de los ficheros de salida, las representaciones gráficas A y B, que serán las que nos permitan determinar si, efectivamente, nuestra colección es representativa. (cf. Figura 3). El tiempo que tarde el programa en generar las representaciones gráficas y los archivos de análisis dependerá del número de n-gramas seleccionados para el cálculo, del tamaño del corpus analizado y de la versión utilizada.

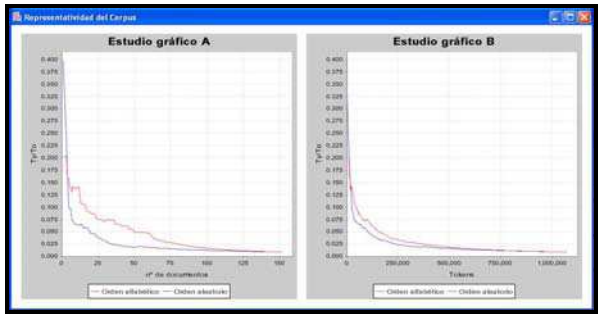

Figura 3: Representatividad del corpus de seguros turísticos (1-grama)

A partir de los datos arrojados por *ReCor*, podemos concluir que el corpus español de contratación de seguros turísticos (cf. Figura 3) es representativo a partir de 140 documentos y 1,0 millón de palabras.

Si deseamos ver los resultados para dos o más gramas, repetiremos los pasos anteriormente expuestos y especificaremos la cifra en «Grupo de palabras». A continuación, mostramos los resultados arrojados por ReCor para 2-gramas.

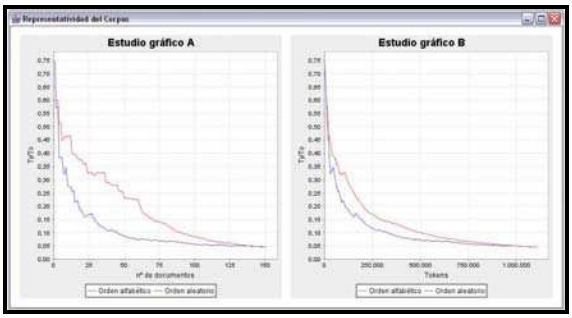

Figura 4: Representatividad del corpus de seguros turísticos (2-gramas)

De este modo, a partir de los datos que nos ofrece el programa para 2-gramas, se desprende que el corpus español de contratación de seguros turísticos (cf. Figura 4) es representativo a partir de 150 documentos y 1,25 millones de palabras.

3.2. Datos estadísticos

Además de las representaciones gráficas A y B, el programa también genera de forma simultánea tres tipos de archivos de salida, cuyo formato (.txt y Excel) depende de la versión utilizada. El primero de ellos, presenta un «Análisis estadístico» del corpus, tanto por orden alfabético como aleatorio, estructurado en cinco columnas: *types, tokens*, cociente entre palabras distintas y totales (*types/tokens*), número de palabras con una aparición (V1) y número de palabras con dos apariciones (V2):

Archivo de salida 1 - Bloc de notas

Archivo Edición Formato Ver Ayuda

8388.0	246362.0	0.03404746	2245	1462
8390.0	255585.0	0.03282665	2194	1400
8390.0	264530.0	0.03171663	1999	1509
8403.0	270783.0	0.031032229	1976	1528
8467.0	278014.0	0.0304553	2016	1540
8512.0	289467.0	0.02940577	2032	1482
8514.0	300924.0	0.028292857	1993	1496
8771.0	305357.0	0.028723756	2165	1483
8772.0	315121.0	0.027836926	2150	1465
8961.0	321784.0	0.027847873	2280	1485
8975.0	329563.0	0.027233033	2131	1587
8977.0	337116.0	0.026628817	2122	1438
8979.0	345078.0	0.026020205	2116	1435
9226.0	349633.0	0.02638767	2291	1458
9226.0	363022.0	0.025414437	2291	1266
9265.0	377700.0	0.02453005	2269	1306
9265.0	392012.0	0.02363448	2269	1291
9265.0	395688.0	0.023414914	2236	1317
9265.0	410645.0	0.022562066	2216	1329
9265.0	426865.0	0.021704754	2216	1317
9265.0	441856.0	0.02096837	2216	1317
9265.0	453700.0	0.020420983	2209	1275
9265.0	462646.0	0.02002611	2207	1138
9265.0	471875.0	0.019634437	2207	1118
9265.0	481098.0	0.01925803	2205	1120
9265.0	490043.0	0.018906504	2205	1120
9265.0	496296.0	0.018668294	2203	1121
9265.0	503527.0	0.018400205	2158	1152
9265.0	514980.0	0.01799099	2158	1114
9265.0	526437.0	0.017599447	2156	1116
9265.0	530870.0	0.017452484	1965	1253
9265.0	540634.0	0.017137287	1964	1239

Figura 5: Fichero de salida (Análisis estadístico)-Español (v. 2.1)

A partir de este análisis estadístico, se puede observar cómo los *types* (primera columna) no incrementan y se mantienen estables —9265.0— a pesar de que el volumen del corpus —*tokens*— sigue en aumento tal y como ilustra la segunda columna (de 392012.0 a 540634.0). De este modo, se comprueba, efectivamente que el corpus ya es representativo para este campo de especialidad y que la inclusión de nuevos textos apenas incorporará novedades significativas al corpus.

En segundo tipo de archivo, «Palabras ord. alf.», nos muestra las palabras que contiene el corpus ordenadas por orden alfabético (primera columna) acompañadas de su frecuencia de aparición (segunda columna):

Archivos de salida 2 - Bloc de notas

Archivo Edición Formato Ver Ayuda

a78393188	2
ab	4
abajo	21
abandono	51
abdomen	28
abierto	9
abintestato	8
ablaciÃ³n	1
ablación	29
abogado	36
abogados	7
abolición	12
abona	18
abonada	46
abonadas	42
abonado	12
abonadopara	1
abonados	17
abonar	74
abonara	2
abonarse	8
abonará	299
abonarán	4
abonará	5
abone	33
abono	102
abonos	8
abonándole	1
aborto	8
abortos	4
abre	1
abril	165

Figura 6: Ficheros de salida (Palabras ord. alf.) de los corpus de seguros turísticos (español)

Por último, el tercer fichero de salida «Palabrar ord. frec» presenta las palabras del corpus ordenadas (primera columna) en función de su frecuencia (segunda columna):

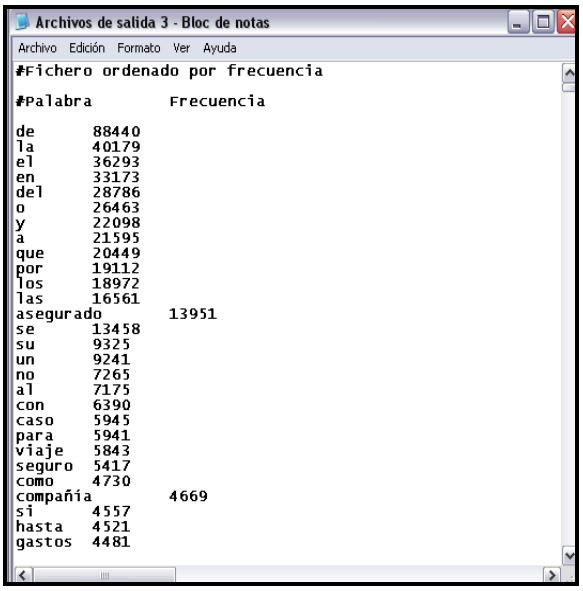

Figura 7: Ficheros de salida (Palabras ord. frec.) de los corpus de seguros turísticos (español)

Finalmente, la versión 2.1. genera simultáneamente, además los anteriores resultados en .txt, tablas de Excel. La Fig. 8 ilustra una tabla en Excel de 2-gramas, ordenados por frecuencia, que ha generado la versión 2.1. para el corpus español.

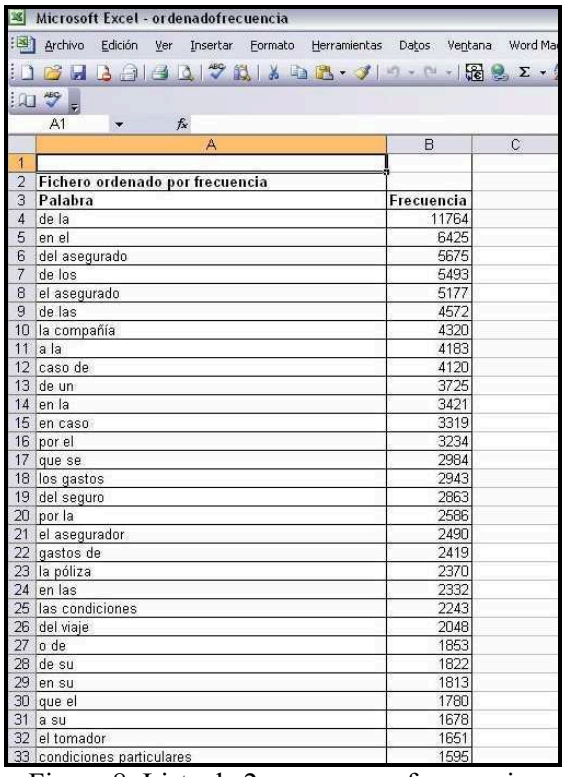

Figura 8: Lista de 2-gramas por frecuencia-Español (v. 2.1.)

4 Conclusiones

Una de las características principales de los corpus virtuales o *ad hoc* es que suelen ser eminentemente desequilibrados, puesto que su tamaño y composición finales vienen determinados, normalmente, sobre todo en los lenguajes de especialidad, por la disponibilidad (Giouli y Piperidis, 2002) y, por consiguiente, es imprescindible contar con herramientas que nos aseguren su representatividad. Sin embargo, el problema estriba en que no existe acuerdo sobre el tamaño que debe tener un corpus para que sea considerado «representativo», a pesar de que la «representatividad» sea el concepto clave que diferencia a un corpus de otros tipos de colecciones y repertorios textuales. Sin embargo, las propuestas realizadas hasta la fecha para el cálculo de la representatividad no resultan fiables, como ya hemos señalado. Conscientes de estas deficiencias, Yang et al. (2000) intentaron superarlas con una nueva propuesta, una formulación matemática capaz de predecir la relación entre los *types* de un corpus y el tamaño de éste (*tokens*). Sin embargo, los autores, al concluir su trabajo admiten que su enfoque presenta serias limitaciones y entre ellas, destacan la siguiente: «the critical problem is, however, how to determine the value of tolerance error for positive predictions» (Yang et al. 2000: 30).

Nuestra propuesta supera a las anteriores en tanto no necesita determinar la constante C (=tamaño del corpus) para sobre ello intentar calcular su representatividad (algo, por otra parte, casi tautológico), como es habitual en los enfoques basados en la ley de Zipf. Tampoco necesita determinar el valor del error máximo de tolerancia, que es la principal deficiencia del enfoque de Biber (1993) y del de Yang *et a*l. (2000). El algoritmo N-Cor permite establecer *a posteriori*, sin tener que establecer valores prefijados, el umbral de representatividad de un corpus bien construido, es decir, compilado conforme a criterios de diseño cualitativos (externos e internos). Concretamente, se parte de la idea de que el cociente entre las palabras reales de un texto y las totales —*types/tokens*—, que da cuenta de la densidad o riqueza léxica de un texto, no aumenta proporcionalmente a partir de un número de textos determinado. Lo mismo ocurre cuando la representatividad se calcula en

función de la densidad léxica a partir secuencias de palabras (n-gramas).

Sobre esta base teórica, se ha implementado un programa (*ReCor*), que permite ilustrar gráficamente el punto a partir del cual un corpus que ha sido compilado según criterios cualitativos comienza a ser representativo en términos cuantitativos. La representación gráfica, a partir de dos líneas —documentos incluidos alfabéticamente y aleatoriamente—, que se estabilizan a medida que se aproximan al valor cero, muestra el tamaño mínimo de la colección para ser considerada representativa.

En el caso de los corpus especializados de tamaño reducido de ámbitos concretos, no es posible determinar *a priori*, exactamente, un número óptimo de palabras o de documentos, puesto que estará en función de las restricciones propias del campo de especialidad, de cada país y lengua. Nuestro método permite realizar dicha estimación a posteriori, esto es, una vez que se ha terminado de compilar el corpus, durante la compilación o durante la fase de análisis y verificación.

Hasta el momento esta metodología se ha probado con éxito para corpus especializados de seguros turísticos y condiciones generales de contratos de viaje combinado en inglés, español, alemán e italiano (cf. Corpas Pastor y Seghiri Domínguez, 2007/en prensa). También se ha utilizado para comprobar la representatividad del corpus multilingüe utilizado por la Agencia Catalana de Noticias para alimentar su sistema de traducción automática español-inglés-francés-catalán-aranés (occitano).

Actualmente estamos trabajando en una nueva versión (ReCor 3.0) que esté optimizada para trabajar con múltiples ficheros o con archivos de gran extensión de forma rápida y, al mismo tiempo, permita extraer unidades fraseológicas a partir del análisis en n-gramas ($n \geq 1$ y $n \leq 10$) del corpus.

Bibliografía

Aston, G., S. Bernardini y D. Stewart.. 2004. *Corpora and Language Learners.* Amsterdam y Filadelfia: John Benjamins.

Bernardini, S. 2000. *Competence, capacity, corpora*. Bolonia: Cooperativa Libraria Universitaria Editrice.

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

Corpas Pastor, G. 2001. «Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada». *TRANS: revista de traductología*. 5. 155-184.

Corpas Pastor, G. 2004. «Localización de recursos y compilación de corpus vía Internet: Aplicaciones para la didáctica de la traducción médica especializada». En Consuelo Gonzalo García y Valentín García Yebra (eds.). *Manual de documentación y terminología para la traducción especializada.* Madrid: Arco/Libros. 223-257.

Corpas Pastor, G.; Seghiri Domínguez, S. 2007/en prensa. *El concepto de representatividad en lingüística de corpus: aproximaciones teóricas y consecuencias para la traducción.* Málaga: Servicio de Publicaciones de la Universidad.

Ghadessy, M.., A. Henry, R. L. Roseberry (eds.). 2001. *Small corpus studies and ELT: theory and practice*. Ámsterdam y Filadelfia: John Benjamins.

Giouli, V. y S. Piperidis. 2002. *Corpora and HLT. Current trends in corpus processing and annotation*. Bulagaria: Insitute for Language and Speech Processing. S. pag. <http://www.larflast.bas.bg/balric/eng_files/corpora1.php> [Consulta: 18/05/2007].

Haan, P. 1989. *Postmodifying clauses in the English noun phrase. A corpus-based study*. Amsterdam: Rodopi.

Haan, P. 1992. «The optimum corpus sample size?». En Gerhard Leitner (ed.). *New dimensions in English language corpora. Methodology, results, software development*. Berlín y Nueva York: Mouton de Gruyter. 3-19.

Heaps, H. S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Nueva York: Academic Press.

Kock, J. 1997. «Gramática y corpus: los pronombres demostrativos». *Revista de filología románica*. 14 (1): 291-298. <http://www.ucm.es/BUCM/revistas/fll/0212999x/articulos/RFRM9797120291A.PDF> [Consulta: 18/05/2007].

Kock, J. 2001. «Un corpus informatizado para la enseñanza de la lengua española. Punto de partido y término». *Hispanica Polonorum*. 3: 60-86. <http://hispanismo.cervantes.es/documentos/kock.pdf> [Consulta: 18/05/2007].

Sánchez Pérez, A. y P. Cantos Gómez. 1997. «Predictability of Word Forms (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.

Seghiri Domínguez, M. 2006. *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Tesis doctoral Málaga: Universidad de Málaga.

Yang, D., P. Cantos Gómez y M. Song. 2000. «An Algorithm for Predicting the Relationship between Lemmas and Corpus Size». *ETRI Journal*. 22 (2) : 20-31. <http://etrij.etri.re.kr/Cyber/servlet/GetFile?fileid=SPF-1042453354988> [Consulta: 18/05/2007].

Young-Mi, J. 1995. «Statistical Characteristics of Korean Vocabulary and Its Application». *Lexicographic Study*. 5 (6): 134-163.

---

[1] El presente trabajo ha sido realizado en el seno del proyecto *La contratación turística electrónica multilingüe como mediación intercultural: aspectos legales, traductológicos y terminológicos* (Ref. nº HUM-892, 2006-2009. Proyecto de Excelencia, Junta de Andalucía).

[2] La metodología descrita en este trabajo ha recibido el *Premio de Investigación en Tecnologías de la Traducción* (III convocatoria) concedido por el Observatorio de Tecnologías de la Traducción. Para más información, véase <http://www.uem.es/web/ott/>.

[3] This method has been awarded the *Translation Technologies Research Award* (*Premio de Investigación en Tecnologías de la Traducción)* by the Translation Technologies Watch (Observatorio de Tecnologías de la Traducción). Further information at the URL: <http://www.uem.es/web/ott/>.

[4] Para una visión más amplia acerca del protocolo de compilación de corpus especializados, véase Seghiri Domínguez (2006).

[5] Aunque es un corpus monolingüe (español), se encuentra delimitado diatópicamente. De este modo, los textos que integran el corpus de seguros turísticos son elementos formales del contrato que hayan sido redactados exclusivamente en España.

[6] Se trata de un corpus comparable pues está integrado por textos originales para la contratación turística, concretamente, elementos formales del contrato y legislación.

[7] El corpus de seguros turísticos compilado incluye documentos completos ya que este tipo de corpus es el que permite llevar a cabo investigaciones lingüísticas léxicas y de análisis del discurso, a la par que posibilita la creación de un subcorpus, o un componente, a partir de la selección de fragmentos más pequeños (Sinclair, 1991). De hecho, Sinclair (1991) y Alvar Ezquerra et al. (1994) han puesto de manifiesto la necesidad de incluir textos enteros porque, de este modo, se elimina la discusión en torno a la representatividad de las distintas partes de un texto así como a la validez de las técnicas de muestreo.

[8] Los textos que integran el corpus de seguros turísticos son, específicamente, elementos formales del contrato, a saber, solicitudes de seguro, propuestas, cartas de garantía y pólizas.