Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet

Alexander Gelbukh, Grigori Sidorov, Liliana Chanona Hernández

Centro de Investigación en Computación (CIC),
Instituto Politécnico Nacional (IPN),
Av. Juan de Dios Bátiz, esq. con Miguel Othón de Mendizábal,
México D. F., Zacatenco, CP 07738, México
{gelbukh, sidorov}@cic.ipn.mx, lchanona@latinmail.com

Resumen

Para estudiar muchas propiedades de las palabras (co-ocurrencias y colocaciones, marcos de subcategorización, etc.) se emplea la investigación estadística de grandes cantidades de textos. Sin embargo, en los corpus tradicionales, aunque muy grandes, debido a la ley de Zipf unas pocas palabras tienen un número redundante de ocurrencias y ocupan la mayor parte del volumen del corpus, mientras la inmensa mayoría de las palabras tiene el número estadísticamente insuficiente de ocurrencias. La solución a este problema es el uso del corpus más grande que se ha creado por la humanidad –Internet. Los programas basados en esta idea se llaman corpus virtuales. Éstos presentan otros problemas, tales como la sobrecarga de la red, la respuesta lenta y los resultados no reproductibles: ya que el Internet se cambia constantemente, un resultado obtenido ayer no se reproduce hoy. Para combinar las ventajas de los dos tipos de corpus, se propone un corpus que proporciona los mismos resultados como un corpus virtual, pero almacenados localmente sin necesidad de descargarlos de la red. Se presenta una discusión de los problemas que surgen y se describe el sistema para la compilación automática de tal corpus.

Palabras clave: lingüística de corpus, corpus virtual, Internet, español, ley de Zipf.

Abstract

To study many properties of the words (co-occurrences and collocations, subcategorization frames, etc.), statistical study of very large corpora is necessary. However, in traditional corpora—even very large—due to the Zipf law, few words have a redundant number of occurrences and occupy the great bulk of the corpus, while the immense majority of the words have statistically insufficient number of occurrences. The solution to this problem is the use of the largest corpus created by the humanity—Internet. The programs based on this idea are called virtual corpora. However, these present other problems, such as network overload, slow response and non-reproducible results: since the Internet changes constantly, a result obtained yesterday cannot be reproduced today. To combine the advantages of both types of corpus, we suggest a corpus that provides the same results a virtual corpus but is stored locally. A discussion of the problems arising is presented, and a system for automatic compilation of such corpus is described.

Key words: corpus linguistic, virtual corpus, Internet, Spanish, Zipf law.

1. Introducción

Una de las aplicaciones principales de los corpus es proporcionar el conocimiento léxico de diferentes tipos y niveles (Biber *et al*. 1998, McEnery y Wilson 1996), usualmente las probabilidades empíricas de diferentes propiedades o situaciones en el texto. Hay dos tipos de tales probabilidades:

- Probabilidades absolutas: ¿cuál es la probabilidad de la situación *X* en el texto? Por ejemplo, ¿cuál es la frecuencia relativa de la palabra *ecuación* en el texto?
- Probabilidades condicionales: ¿cuál es la probabilidad de la situación *X* en la presencia de la situación *Y*? Es decir, la investigación estadística se aplica a un subcorpus que consiste de los contextos que contienen la situación *Y*. Por ejemplo, ¿cuál es la probabilidad de la palabra *ecuación* dado que la siguiente palabra es *diferencial*?

Aunque la primera tarea parece interesante, su importancia práctica se limita a los casos del análisis de las estructuras del texto consideradas completamente fuera del contexto, lo cual prácticamente nunca es el caso.

A diferencia de ésta, la segunda tarea tiene un sinnúmero de aplicaciones, por ejemplo:

- Aprendizaje automático de los marcos de subcategorización: ¿qué preposiciones se usan con la palabra *empezar*? (Galicia-Haro *et al*. 2001).
- Aprendizaje de las combinaciones de palabras: con la palabra *atención*, se usan las palabras *prestar*, *atraer*, *perder* (Benson *et al*., 1986; Bolshakov y Gelbukh, 2000; Bolshakov 1994).
- Aprendizaje de las propiedades sintácticas: ¿la palabra *gran* se usa antes o después de la palabra que modifica?
- Detección de los llamados malapropismos (las palabras mal escritas que parecen a las palabras existentes de la misma categoría gramatical): en la presencia de las palabras *novia*, *iglesia* y *ceremonia*, es más probable que aparezca *casar* o *cazar*?[1]

[1] *En varios dialectos del español no hay diferencia en la pronunciación de* s *y* z.

Sin embargo, los corpus tradicionales –grandes colecciones de textos aleatoriamente seleccionados de entre los textos del cierto género, tema, etc., véase (Biber, 1993)– son más apropiados para la primera tarea, más específico, para determinar si la frecuencia de una cierta palabra (construcción, etc.) es alta o baja.

En cualquier corpus de este tipo, unas pocas palabras ocurren muchísimas veces, ocupando la mayor parte de su volumen y tomando la mayor parte del tiempo de procesamiento del corpus. Por otro lado, la inmensa mayoría de las palabras de lenguaje tiene la representación muy poca o ninguna en el corpus específico. Este fenómeno se conoce como la ley de Zipf: la palabra con el rango estadístico n tiene aproximadamente la frecuencia C / n, donde C es una constante, véase, por ejemplo, (Gelbukh y Sidorov 2001).

Esto no presenta mayor problema para la tarea del aprendizaje de las probabilidades absolutas (primera tarea). Sin embargo, la utilidad de los corpus del tamaño razonable para el aprendizaje de las probabilidades condicionales (segunda tarea) es limitada debido a que para casi todas las palabras del lenguaje el número de sus ocurrencias es estadísticamente insuficiente para hallar los resultados confiables. Por otro lado, la inmensa mayoría de las ocurrencias de las palabras en el corpus son las repeticiones de las mismas palabras, redundantes desde el punto de vista estadístico: bastaría un número mucho menor de las ocurrencias de cada una de éstas.

Una solución a este problema es el uso del corpus más grande que se ha creado por la humanidad, lo que es Internet. En un corpus tan enorme, hay información suficiente para aprender las propiedades de un gran número de palabras.

Los sistemas que utilizan Internet como un corpus se conocen como los corpus virtual (Kilgariff, 2001) y se utilizan de la siguiente manera: el usuario presenta su petición especificando la palabra en cuestión (la *Y* en los términos de la segunda tarea descrita al inicio de esta sección), el programa busca en Internet un cierto número de los contextos relevantes (no todos ya que pueden ser demasiados) y presenta al usuario un subcorpus construido de esta manera para su investigación estadística. Un ejemplo de tal herramienta es, digamos, WebCorp (www.webcorp.org.uk).

A pesar de la indudable utilidad de los corpus virtuales, los corpus "reales" –en forma de un archivo– tienen unas ventajas importantes de que los corpus virtuales carecen:

- Respuesta rápida que no depende del tráfico de la red.
- Uso local de los recursos. No sobrecarga la red.
- La posibilidad de revisión, control de calidad y limpieza manual, marcado manual (digamos, marcado sintáctico o morfológico) y toda la clase de la preparación que distingue un corpus de una colección de textos no preparados.
- Resultados estables y reproducibles en tiempo y espacio. A la misma petición hecha por diferentes usuarios en diferentes días o años se presenta la misma respuesta. Lo que facilita la utilización del corpus para la comparación de diferentes métodos y sistemas, para el trabajo paralelo de diferentes grupos de desarrollo o pruebas, etc. Internet, al contrario, se cambia constantemente: cada segundo aparecen nuevos sitios y páginas, se encienden o se apagan los servidores, las máquinas de búsqueda indexan nuevas páginas y quitan otras, etc.

Para combinar las ventajas de los corpus virtuales y los localmente almacenados, nosotros proponemos la construcción, a través de Internet, de un diccionario grande de contextos de palabras.

Los diccionarios de contextos se conocen como las concordancias de tipo KWIC (*key words in context*). En nuestro caso, el diccionario en este formato se compila a través del Internet, lo que permite hacerlo muy grande (dar la información para un número grande de las palabras) y emplearlo para las tareas para las cuales normalmente se emplean los corpus.

Específicamente, se trata de ejecutar las peticiones al Internet sobre cada palabra, guardando la respuesta. La recuperación posterior de estas respuestas simula el resultado de la petición a un corpus virtual con toda su riqueza léxica, pero sin necesidad de conectarse a la red y con todas las ventajas descritas arriba (la posibilidad de control y marcado manual, resultados estables, etc.).

Es decir, nuestro sistema es virtualmente un corpus virtual (ya que simula su comportamiento) sin serlo físicamente (sin necesidad de conectarse a la red). Gracias al número limitado de los contextos que se obtienen y se guardan para cada palabra, no se trata de guardar todo el Internet en el disco local, sino de un fichero relativamente pequeño que cabe en un CD.

Sin embargo, lo llamamos diccionario porque es estructurado: proporciona los datos para cada palabra encabezado.

Obviamente, el costo de estas comodidades es la pérdida de la flexibilidad de los corpus virtuales reales. Los parámetros del corpus (tales como el umbral del número de los contextos para cada palabra) se deben determinar de antemano. El tipo de la petición (contextos de una palabra versus contextos que contienen dos palabras específicas, etc.) también se fija en el tiempo de compilación. Para cambiar los parámetros sólo se tiene que repetir el proceso automático de compilación del corpus.

En el resto del artículo se presenta con mayor detalle el diccionario compilado, se describe el método que hemos empleado para su compilación y los parámetros variables, los resultados experimentales y el trabajo futuro.

2. El diccionario de contextos

El diccionario presenta, para cada palabra encabezado, un cierto número N de los contextos de su uso. Los contextos deben ser lingüísticamente válidos, es decir, ser oraciones completas (o unas ventanas de texto) en español que representan el uso de la palabra en el lenguaje.

El diccionario de este tipo también se puede llamar un corpus representativo (Gelbukh *et al*. 2002) ya que cada palabra para la cual hacemos la búsqueda tiene en él la representación estadísticamente significativa.

El número N se determina con las formulas correspondientes de la estadística matemática basándose en el porcentaje requerido de la confiabilidad de los resultados de la investigación estadística. O bien, se puede determinar simplemente como lo suficientemente grande basándose en la experiencia del investigador. En nuestros experimentos usamos $N = 50$.

Para algunas palabras no se puede obtener del Internet el número requerido N de contextos válidos. Estas palabras, en el sentido estricto no pertenecen al vocabulario del diccionario (porque no cumplen los requisitos) pero sus contextos se guardan y se presentan al usuario con la advertencia que la

información puede ser estadísticamente menos confiable.

Las ventajas de nuestro diccionario (un corpus virtual, virtual) en comparación con el corpus tradicional (un texto muy largo) son:

- Este corpus ocupa mucho menos espacio y entonces es más fácil de manejar porque no contiene el número inmenso de las palabras frecuentes.
- Por otro lado, este corpus contiene los contextos de las palabras raras, de las cuales el corpus tradicional contiene pocos o ningunas ocurrencias.

Entonces, con este diccionario se puede obtener las estadísticas del uso de las palabras de baja frecuencia, las cuales de hecho son la mayoría de las palabras del lenguaje.

Por otro lado, las ventajas de nuestro diccionario en comparación con un corpus virtual se han descritas en la sección anterior.

Los corpus virtuales también tienen sus problemas. Uno de éstos es la baja calidad de los contextos encontrados:

- La palabra se puede usar como nombre propio de una persona, empresa, producto, etc. O bien, puede coincidir con alguna palabra extranjera, etc.
- En algunos casos los contextos no son lingüísticos sino son nombres de campos de formularios o los botones, rótulos de imágenes, direcciones de Internet o nombres de ficheros, tablas, etc.
- En otros casos el contexto no es del uso de la palabra sino, por ejemplo, un artículo del diccionario que la explica o traduce.
- Finalmente, el lenguaje usado por los autores en Internet puede ser poco culto o gramáticamente mal formado.

En cuanto al último problema no se puede hacer mucho al respecto. Además, no es claro si es un problema del corpus o una propiedad que refleja el uso real (a diferencia del uso prescrito) del lenguaje.
En cuanto a los demás problemas, los hemos solucionado a cierto grado empleando los filtros que detectan esos defectos en un contexto dado y previenen que el contexto entre en el diccionario.
Otra complicación es el manejo de las formas gramaticales de las palabras. Obviamente, el lexema a buscar se realiza en los textos y se debe representar en los contextos del diccionario como una forma de palabra específica (*pienso*, *pensaríamos* versus *pensar*). Hay tres estrategias posibles para calcular el porcentaje deseado de las formas correspondientes en el diccionario. Se puede representar las formas:

- Uniformemente: si la palabra pensar tiene $K = 65$ formas, buscar N/K contextos para *pensar*, N/K para *pienso*, N/K para *pensaríamos*, etc. Con esta estrategia, cada lexema se representa con N contextos.
- Independientemente: N contextos para *pensar*, N para *pienso*, N para *pensaríamos*, etc. Con esta estrategia, cada lexema se representa con $K \times N$ contextos, es decir, los lexemas de verbos serán mejor representados que los de sustantivos.
- Proporcionalmente a su uso en los textos: los N contextos para el lexema se dividen de tal manera que a las formas más usadas les corresponde un mayor número de los contextos.

Las diferentes estrategias corresponden a diferentes tipos de aplicación del diccionario. Ya que nuestra motivación era el estudio de las propiedades combinatorias de las palabras que poco dependen de las formas gramaticales, en nuestros experimentos empleamos la última estrategia.

3. Compilación del diccionario a través del Internet

Para la compilación del diccionario se ejecutaron los siguientes pasos, todos salvo el primero siendo completamente automáticos:

1. Se compiló la lista de las palabras que debieron ser representadas en el diccionario.
2. Se generaron todas las formas gramáticas (morfológicas) de estas palabras.
3. Para cada palabra, se calculó la proporción (en por cientos) de sus diferentes formas gramáticas en la cual éstas deben ser representadas en el diccionario. Esta proporción corresponde al uso de las formas en Internet. Por ejemplo, si la forma singular de la palabra ocurre en 400 mil documentos en Internet y la forma plural en 100 mil, entonces en sus N = 50 contextos en el diccionario va a ocurrir 40 veces en singular y 10 en plural. Para los verbos, algunas formas (menos usadas para

el verbo específico) no obtuvieron ninguna representación en el diccionario.

4. Para cada forma morfológica de la palabra (digamos, singular y plural), se buscaron en Internet los documentos que la contienen haciendo una petición a un buscador de Internet (se experimentó con *Google* y *Altavista*) y analizando su respuesta con el fin de obtener el URL del documento.
5. Para cada URL obtenido, se bajó el documento, se analizó su estructura HTML y se extrajeron los contextos de la palabra en cuestión.
6. Para cada contexto, se aplicaron los filtros heurísticos para rechazar los contextos no válidos (más cortos que $n = 8$ palabras sin traspasar las marcas HTML, o donde la palabra es un nombre propio, una dirección de Internet, etc.).

El paso 5 se repitió hasta encontrar, de ser posible, el número requerido (determinado en el paso 3) de los contextos válidos de la forma morfológica dada de la palabra.

Para el paso 2 se empleó un sistema de análisis y generación morfológico (Gelbukh y Sidorov, 2002). Si para el paso 1 se usara un corpus, las palabras se lematizarían empleando el mismo sistema. Sin embargo, nosotros utilizamos la lista de palabras extraída de un diccionario que ya estuvieron en la primera forma gramatical (singular, infinitivo, etc.).

En el paso 3, para cada forma morfológica de la palabra se ejecutaba la misma petición a las máquinas de búsqueda que en el paso 4, pero sólo se analizaba el número total de documentos que contienen esta palabra (los buscadores proporcionan este número).

La interacción técnica con Internet en los pasos 4 y 5 se realizó con el componente THTTP del Borland C++ Builder 5.0. Empezando con el paso 2, el programa funciona totalmente automático sin intervención humana alguna.

Se asegura que los textos serán en el lenguaje español con la configuración del buscador usado (todos los buscadores principales de Internet permitan especificar el lenguaje de los documentos a buscar).

En el paso 6 se requirió un tamaño mínimo del contexto porque los contextos cortos no contienen suficiente información lingüística. Además, frecuentemente no son expresiones de lenguaje natural sino otros tipos de datos (rótulos de imágenes, nombres de campos, botones o ficheros, etc.).

También, los experimentos han mostrado que una palabra puede tener muy alta frecuencia pero como apellido y no como palabra normal. Por ejemplo, muchos contextos encontrados para la palabra *abad* referían al apellido. Por eso no se consideraron válidos los contextos donde la palabra en cuestión empieza con mayúscula en el medio de la oración.

Nótese que con rechazar algún contexto no se pierde mucho ya que usualmente el número de contextos disponibles es muy grande, a reserva de que no se rechazan de modo sistemático que afecta sus propiedades estadísticas.

Obviamente, uno de los filtros verifica que los contextos no se repitan literalmente (lo cual significaría que se trata de copias múltiples del mismo documento).

4. Resultados experimentales

El programa que aplica el algoritmo descrito se realizó de tal manera que los siguientes parámetros:

- Los umbrales numéricos tales como N y n,
- El conjunto de los filtros a aplicar,
- El nombre del fichero que contiene la lista de palabras encabezado y
- El lenguaje para el cual se compilará el diccionario

se especifican a través de la interfaz de usuario. Una vez especificados los parámetros, el programa funciona en el modo automático hasta que se genera el diccionario necesario.

Como la lista de las palabras encabezado usamos el vocabulario del diccionario Anaya de la lengua española, el cual contiene 33 mil de las palabras más comunes del español. Ya que estas palabras ya están en la forma morfológica principal, no usamos el lematizador en este paso. Obviamente, la lista de palabras se puede ampliar sin la necesidad de repetir el proceso para las palabras actuales.

Para estas 33 mil palabras, la ejecución automática del programa tomó alrededor de tres semanas, ejecutándose en un computador PC Pentium IV conectado a la red del área local de velocidad mediana y variable según el tráfico. Durante estas tres semanas, para el total de 100 mil formas

morfológicas de las palabras de la lista se ejecutaron alrededor de 200 mil peticiones (pasos 3 y 4 del algoritmo) al buscador de Internet y se bajaron alrededor de $33 \times N = 1650$ mil documentos. Para la compilación del diccionario se usó el buscador *Google* ya que su tiempo de respuesta es muy bueno.

El archivo de resultado tiene el tamaño de 221 MB, es decir, una tercera parte de un disco compacto.

Para la mayoría de las palabras de la lista inicial el diccionario contiene los 50 contextos esperados. Sin embargo, para unos 10% de las palabras el número de contextos válidos encontrados es menor. En el momento actual no es claro si esto se debe a la baja frecuencia de su uso en Internet o a algún problema técnico del programa o de la implementación de los filtros.

5. Conclusiones y trabajo futuro

Se propuso el diccionario que combina las ventajas de los corpus virtuales:

- Menor número de palabras redundantes,
- El numero suficiente de los contextos de la mayoría de las palabras para el aprendizaje estadísticamente confiable

así como las de los corpus tradicionales:

- Tamaño razonablemente pequeño,
- Manejo local de los recursos con la respuesta rápida y sin sobrecarga de la red,
- La posibilidad de clasificación, limpieza y marcado manual,
- Estabilidad y reproducibilidad de los resultados en el tiempo y espacio.

El método también herede algunas desventajas de los corpus virtuales:

- La calidad inferior de los textos debido al lenguaje cotidiano de Internet (depende mucho de la calidad de los filtros empleados),
- Imposibilidad de elegir el género, tópico, autor, etc.,
- Imposibilidad (sin aplicar las herramientas correspondientes) de resolver la homonimia de partes de oración (*¿trabajo* es verbo o sustantivo?), lo que aumenta el ruido en el corpus obtenido; y distinguir los sentidos de palabras diferentes, lo que hace que un sentido frecuente es representado mucho más amplio que los demás.

así como las de los corpus tradicionales en comparación con los virtuales:

- Menor flexibilidad de las peticiones,
- Los parámetros del corpus se fijan en el tiempo de compilación.

El primer grupo de restricciones es inherente para los corpus basados en Internet y puede limitar su uso. El último grupo no presenta mayor problema ya que con el programa que hemos desarrollado, la obtención de un corpus con otros parámetros es la cuestión de unos clics (y varias semanas de trabajo automático del computador). Véase también el trabajo futuro más adelante.

Basándonos en esta idea, realizamos un programa para la compilación automática de tales diccionarios, con el cual compilamos un diccionario que para unos 33 mil de palabras españolas da unos 50 contextos para cada palabra, divididos entre 100 mil formas gramaticales en total ponderadas según sus frecuencias del uso. Lo que resultó en un fichero de tan sólo 221 MB –una tercera parte de un CD.

El diccionario obtenido se usará para el aprendizaje automático de los diccionarios estadísticos de diferentes tipos para el español, tales como el diccionario de los marcos de subcategorización y el diccionario de combinaciones de palabras (atracción léxica, o colocaciones). Este último diccionario será la base para la compilación manual del diccionario de las funciones léxicas de español.

Se pueden mencionar las siguientes direcciones del trabajo futuro:

- Mejorar los filtros heurísticos para los contextos.
- Experimentar con diferentes valores de los parámetros, tales N y n.
- Experimentar con diferentes maneras de ponderación de las formas gramaticales.
- Probar otros tipos de peticiones. Por ejemplo, compilar un diccionario que para cada par de palabras de cierta lista (Bolshakov y Gelbukh 2000) dé ejemplos de sus coocurrencias. También, un diccionario donde la unidad de la petición no es una palabra sino una cierta combinación de características gramaticales (por ejemplo, pretérito plural de segunda persona): encontrar N contextos para cada

combinación de las características. El algoritmo correspondiente es obvio y la modificación del programa existente será mínima.

- Tratar de desarrollar los filtros temáticos o de género con el fin de la compilación de un diccionario para un cierto género o tema específico. El algoritmo será basado en la clasificación temática (Gelbukh *et al*. 1999) y búsqueda con enriquecimiento de la petición (Gelbukh 2000).
- Realizar el esquema de autoenriquecimiento del corpus: usar los archivos bajados de Internet para descubrir nuevas palabras que no están en la lista inicial, e incluirlas en el diccionario (buscar los N contextos para ellas).

Para dicho autoenriquecimiento resultará útil generar las formas gramaticales de las palabras descubiertas. Para eso se usará un analizador y generador morfológico heurístico (Gelbukh y Sidorov, 2002). Para comprobar sus hipótesis sobre la categoría gramatical (sustantivo, verbo, etc.) y el tipo de inclinación o conjugación de la nueva palabra, el analizador también usará el Internet. Por ejemplo, para la palabra *internetizados* se generarán las formas hipotéticas *internetizar*, *internetizarán*, etc.; su presencia en Internet comprobaría que la palabra inicial es el verbo del tipo de conjugación predicho.

Agradecimientos

Este trabajo fue realizado con el apoyo parcial del gobierno de México (CONACyT, SNI), del Instituto Politécnico Nacional, México (CGEPI-IPN, COFAA, PIFI) y la red RITOS-2 del Subprograma VII de CYTED.

Referencias

Benson, M., E. Benson, and R. Ilson. *The BBI Combinatory dictionary of English*. John Benjamins Publishing Co., 1986. (1986)

Biber, D. Representativeness in corpus design. *Literary and linguistic computing*, 8:243-257. (1993)

Biber, D., S. Conrad, and D. Reppen. *Corpus linguistics. Investigating language structure and use*. Cambridge University Press, Cambridge, 1998. (1998).

Bolshakov, I. A. Multifunction thesaurus for Russian word processing. Proceedings of *4th Conference on Applied Natural language Processing*, Stuttgart, October 13-15, 1994, p. 200-202. (1994)

Bolshakov, I. A. and A. Gelbukh. A Very Large Database of Collocations and Semantic Links. *NLDB'2000: Applications of Natural Language to Information Systems*. Lecture Notes in Computer Science N 1959 Springer-Verlag, 2000, pp. 103–114. (2000).

Galicia-Haro, S., A. Gelbukh, I.A. Bolshakov. Acquiring syntactic information for a government pattern dictionary from large text corpora. *IEEE SMC-2001: Systems, Man, And Cybernetics*. Tucson, USA, 2001, IEEE, pp. 536–542. (2001).

Gelbukh, A. Lazy Query Enrichment: A Simple Method of Indexing Large Specialized Document Bases. *DEXA-2000: Database and Expert Systems Applications*. Lecture Notes in Computer Science N 1873, Springer-Verlag, 2000, pp. 526–535. (2000).

Gelbukh, A. and G. Sidorov. Zipf and Heaps Laws' Coefficients Depend on Language. *CICLing-2001, Intelligent Text Processing and Computational Linguistics*. Lecture Notes in Computer Science N 2004, Springer-Verlag, 2001, pp. 330–333. (2001).

Gelbukh, A. and G. Sidorov. Morphological Analysis of Inflective Languages through Generation. Journal *Procesamiento de Lenguaje Natural*, Vol. 29, Spain, 2002, pp 105-112. (2002).

Gelbukh, A., G. Sidorov, and L. Chanona-Hernández. Compilation of a Spanish representative corpus. *CICLing-2002*, *Intelligent Text Processing and Computational Linguistics*. Lecture Notes in Computer Science N 2276, Springer-Verlag, 2002, pp. 285–288. (2002).

Gelbukh, A., G. Sidorov, A. Guzman-Arenas. Use of a weighted topic hierarchy for document classification. *TSD-99: Text, Speech and Dialogue*. Lecture Notes in Artificial Intelligence N 1692, Springer-Verlag, 1999, pp. 130–135. (1999).

Kilgariff, A. Web as corpus. In: Proc. of *Corpus Linguistics 2001 conference*, University center for computer corpus research on language, technical papers vol. 13, Lancaster University, 2001, pp 342-344. (2001).

McEnery, T. and A. Wilson. *Corpus linguistics*. Edinburg University Press, 1996. (1996).