# Hacia la verificación de diccionarios explicativos asistidos por computadora

*Alexander Gelbukh*
*Grigori Sidorov*
Laboratorio de Lenguaje Natural
Centro de Investigación en Computación, IPN

*An explanatory dictionary is a complex system with numerous relations between the elements located in different places in its text, as well as between the definitions and the live usage of the words in language. This makes it very difficult to manually detect certain types of defects in the dictionary, such as vicious circles in the system of the definitions, an inconsistent inventory of the words used in the definitions (defining vocabulary), inconsistent or insufficient definitions, incorrect subdivision of the entries into word senses, inconsistent synonymy and antonymy marks, etc. In this paper we explain how computational algorithms can help in the quality control of the dictionary and its interactive development, as well as present the corresponding software tool.*

*Un diccionario explicativo es un sistema complejo con numerosas relaciones tanto entre los elementos localizados en diferentes lugares en su texto, como entre las definiciones y el vivo uso de las palabras en el lenguaje. Debido a esta complejidad se hace muy difícil la detección manual de ciertos tipos de defectos en el diccionario, tales como círculos viciosos en el sistema de definiciones, un inventario inconsistente de las palabras usadas en las definiciones (vocabulario definidor), definiciones inconsistentes o insuficientes, división incorrecta de los artículos en los sentidos específicos, marcas inconsistentes de sinonimia y antonimia, etc. En este artículo explicamos cómo los algoritmos computacionales pueden ayudar al control de calidad del diccionario y en el desarrollo interactivo del mismo, y presentamos la herramienta computacional correspondiente.*

Palabras clave: *lexicografía computacional, diccionarios explicativos, herramientas de desarrollo, lingüística computacional, lingüística de corpus.*
Fecha de recepción del manuscrito: febrero del 2003

**Alexander Gelbukh y Grigori Sidorov**
Laboratorio de Lenguaje Natural, Centro de Investigación en Computación, IPN
Av. Juan de Dios Bátiz 07738, Zacatenco, México D. F.
correos electrónicos: gelbukh@cic.ipn.mx, sidorov@cic.ipn.mx;

## 1. Introducción

Los diccionarios explicativos son el corazón de la descripción lexicográfica de un lenguaje, la máxima autoridad que rige el correcto y preciso uso y comprensión de sus palabras, el acervo de la sabiduría de todo un pueblo. Son elaborados con gran esmero por equipos de profesionales durante muchos años, para garantizar su impecable calidad.

Sin embargo —como se verá en los ejemplos que presentaremos— es muy difícil garantizar esta calidad con los métodos tradicionales. Esta dificultad se debe a que un diccionario es un sistema complejo de elementos interrelacionados entre sí y al vivo uso del lenguaje. Como en el caso de cualquier sistema complejo, su calidad no se puede evaluar, ni mucho menos garantizar, observando y analizando sus elementos (los vocablos) aislados, uno por uno; dicha evaluación se puede llevar a cabo sólo si se toman en cuenta las relaciones entre los elementos que se encuentran en muy diferentes lugares del diccionario completo.

Por ejemplo, en el diccionario más popular del idioma ruso (Ozhegov, 1990), la *gallina* (en ruso, *kúritsa*) se define como *la hembra del gallo* y, a cientos de hojas de esta definición, encontramos otra para el *gallo* (en ruso, *petukh*) que lo define como *el macho de la gallina*. Aunque ambas definiciones son igualmente correctas y válidas, obviamente no son compatibles dentro del mismo sistema lógico, ya que a quien no sabe que son *kúritsa* y *petukh*, no le proporcionan esta información.

Para los humanos es muy difícil, por no decir imposible, detectar manualmente los problemas de esta naturaleza en un diccionario grande. Pero es allí donde podemos obtener una ayuda indispensable de la computadora, la infatigable ayudante capaz de pasar, sin desfallecer, palabra por palabra, comparando las definiciones esparcidas entre cientos de diferentes hojas, calculando estadísticas, verificando exhaustivamente todos los pormenores. Obviamente la máquina no puede sustituir al experto humano, pero sí puede atraer su atención a las anomalías encontradas y presentarle la información que le facilite tomar una decisión más informada y mejor fundamentada.

Dentro de la lingüística computacional ya se están desarrollando los métodos para el análisis automático del léxico, los cuales permiten automatizar algunos tipos de análisis; véanse, por ejemplo, Saint-Dizier y Viegas, (1995); Vossen, (2001). Esos métodos pueden ayudar al lexicógrafo en el desarrollo de las definiciones y en la evaluación formal de los diccionarios explicativos.En este artículo presentamos varias ideas que llevan a la creación de una herramienta computacional que ayudaría al lexicógrafo a detectar los defectos en la estructura del diccionario y proponer los posibles cambios, específicamente en los casos donde se trata de inconsistencias entre vocablos distantes en el texto.

Aquí sólo abordamos dos problemas relacionados con la calidad de los diccionarios explicativos:

1. Relaciones entre las definiciones en el diccionario,
2. División de los vocablos en sentidos, en los casos de polisemia.

El primer tipo trata de todo un conjunto de problemas que van desde la selección de las palabras a través de las cuales se tienen que definir otras palabras hasta la lógica propia de las definiciones. Estos temas se han tratado en la literatura. Las palabras apropiadas para su uso en las definiciones conforman lo que se conoce como vocabulario definidor (LDOCE, OALD) o, en el contexto más teórico, primitivos semánticos (Wierzbicka, 1996). Tales palabras no son únicas, hay muchas posibles maneras de elegirlas. Sin embargo, según nuestro conocimiento hasta ahora se eligen para cada diccionario de forma artesanal, por prueba y error, sin criterios bien definidos. En la sección 2, presentamos un método que da al lexicógrafo la información necesaria para formar un mejor vocabulario definidor.

Acerca de la construcción de definiciones, en la literatura de la lexicografía tradicional (Hartmann, 2001; Landau, 2001; Singleton, 2000) normalmente sólo se dan recomendaciones de carácter muy general de cómo hay que escribirlas. El principio básico es la idea de Aristóteles de que la definición debe contener el género y las diferencias. Otras ideas más específicas se basan en el trabajo clásico de Zgusta (1971): no definir palabras más simples a través de palabras más complejas (más difíciles de entender) ; definir, a su vez, todas las palabras empleadas en la definición; evitar el uso de la palabra o sus derivadas en su propia definición o en la definición de las palabras explicadas a través de ésta (como en nuestro ejemplo con *gallina* y *gallo*); empezar la definición con la parte más importante; hacer las definiciones simples y breves, etc.; véase, por ejemplo, Landau (2001: 156–171). En las secciones 2 y 3, presentaremos los métodos para verificar automáticamente algunos de estos requerimientos y demostramos que éstos, en efecto, no son completamente compatibles.

El segundo tipo de problemas, el tratamiento de homonimia y polisemia, es aún más difícil de manejar de modo uniforme y consistente, ya que cada palabra presenta sus propias peculiaridades y, por otro lado, cada lexicógrafo tiene sus propios gustos y experiencia sobre el uso de las palabras específicas. Efectivamente, los pocos ejemplos del uso de cada palabra que una persona puede escuchar o leer en su vida, no dan una información estadísticamente significativa de todos sus usos y menos de los matices sutiles de su significado. Aquí, el análisis automático de grandes cantidades de texto, tan grandes que no podría una persona leerlos en toda su vida, es una ayuda indispensable.

Algunas consideraciones acerca del problema de la división de palabras en sentidos se presentan en la sección 3. En esta sección también analizamos en breve cómo reflejar en el diccionario la polisemia regular (Apresjan, 1974). Otros tipos de verificación automática del diccionario, tales como la verificación de ortografía y la verificación del sistema de marcas de sinonimia y antonimia, se presentan en la sección 4.

Finalmente, en la sección 5 describimos las funciones de la herramienta «ayudante del lexicógrafo», la cual está bajo desarrollo en el Laboratorio de Lenguaje Natural del CIC, IPN. En la sección 6 presentamos las conclusiones.

## 2. Relaciones entre las definiciones

Para el análisis formal de los diccionarios comúnmente se usa la representación del diccionario como una red semántica o lo que en las matemáticas se llama un grafo dirigido, ya que los métodos de tal análisis son orientados a la estadística o a la teoría de grafos (Kozima and Furugori, 1993; Evens, 1988; Gelbukh y Sidorov, 2002). Está fuera del alcance de este artículo discutir esos métodos a detalle; sólo mencionaremos algunas ideas fundamentales de este tipo de análisis necesarias para entender sus aplicaciones prácticas.

Un diccionario para el usuario humano tiene como propósito explicar la palabra, maximizando la probabilidad de que la definición contenga las palabras que el usuario ya conoce. Nótese que con esto, las parafrases sinonímicas en la definición aumentan la probabilidad de que el usuario entienda por lo menos una variante (mientras que «para la computadora» son confusas e inútiles). Otro modo de aumentar la probabilidad de comprensión es usar, en las definiciones, sólo un número restringido de las palabras más simples y conocidas (vocabulario definidor). En la práctica es recomendable que sólo se usen alrededor de 2 mil palabras, como, por ejemplo, en los diccionarios de inglés de Longman (LDOCE) o de Oxford (OALD).

Para maximizar la probabilidad de que el usuario entienda la definición, no debe haber círculos viciosos cortos en el sistema de definiciones. Por ejemplo, el diccionario Anaya (Grupo Anaya, 1996) da, efectivamente, las siguientes definiciones:

1. *abeja: insecto que segrega miel*
   *miel: sustancia que producen las abeja.*
2. *convenio: pacto, acuerdo*
   *acuerdo: pacto, tratado*
   *tratado: convenio*

En el primer caso, una palabra se define a través de otra y aquélla a través de la primera, así que un usuario que no sabe qué son *abeja* y *miel* —y consulta al diccionario para saberlo— no tiene ninguna manera de entender las dos definiciones. En el segundo caso el círculo es de longitud 3: *convenio* — *acuerdo* — *tratado* — y nuevamente *convenio*; una persona que no sabe de antemano ninguna de estas tres palabras, no entenderá las definiciones. Sin embargo, si el círculo es bastante largo, la probabilidad de que el usuario no sepa ninguna de sus palabras es baja y entonces los círculos largos —a diferencia de los cortos— no son problemáticos para el uso tradicional del diccionario explicativo del léxico general.

No sucede así con los diccionarios explicativos terminológicos, de términos especiales o técnicos, donde es altamente probable que el usuario no sepa ninguna de las palabras en una cadena de los términos explicados uno a través de otro. Así se desarrolla la exposición de la geometría escolar: todos los términos se construyen, aunque indirectamente, de los tres términos «básicos» —*punto*, *recta*, *pertenecer*— los cuales no se definen sino se ilustran con dibujos o ejemplos. Nótese que para no crear círculos viciosos,

algunas de las palabras usadas en las explicaciones no deben tener explicación (ya que en un grafo donde cada nodo tiene vínculos salientes, necesariamente hay ciclos); es decir, las recomendaciones de la lexicografía tradicional de no formar ciclos y explicar cada palabra usada, son contradictorias.

El concepto de palabras «básicas» es acorde con la tradición lexicográfica donde se pretende definir (aunque indirectamente) todas las palabras a través de un conjunto muy restringido de los llamados *primitivos semánticos* (Wierzbicka, 1996). La diferencia entre el vocabulario definidor y los primitivos semánticos es que las palabras del vocabulario definidor son las únicas palabras que pueden aparecer en las definiciones, pero no importa si algunas de éstas se definen a través de otras. En cambio, los primitivos semánticos son independientes: no se puede definir unas a través de otras. Lo que significa que su conjunto es mínimo: no se puede remover de él ninguna palabra (primitiva semántica) sin perder la posibilidad de definir todas las demás palabras en el diccionario a través de este conjunto. Representando el diccionario como un grafo dirigido (Gelbukh y Sidorov, 2002), la diferencia es que las palabras del vocabulario definidor deben ser accesibles en exactamente un paso por los vínculos del grafo, mientras que las primitivas semánticas pueden ser accesibles en varios pasos. Eso se debe al hecho de que las palabras del vocabulario definidor están presentes físicamente en las definiciones de las palabras (por eso el nombre de vocabulario definidor); a diferencia de los primitivos semánticos que se presentan virtualmente en las definiciones, por el hecho de ser accesibles en el grafo pasando, tal vez, por varios nodos.

Existe una aplicación muy importante, aunque menos tradicional, de los diccionarios, en la cual —al igual que en los diccionarios terminológicos— los círculos, no importa qué tan largos sean, están prohibidos. A saber, aparte de su uso tradicional como fuente de referencia para los usuarios humanos, los diccionarios se pueden aplicar como fuente de información sobre el lenguaje y el mundo real en los sistemas computacionales de inteligencia artificial basados en inferencia lógica. En esta aplicación, no se espera que el sistema experto sepa de antemano palabra alguna ya que su única fuente de conocimiento sobre el lenguaje es el mismo diccionario; además, los círculos viciosos destruyen el sistema de razonamiento lógico ya que entra en ciclos infinitos. En el uso del diccionario explicativo «para las computadoras», es necesario seleccionar algunas palabras como primitivas semánticas, eliminar sus definiciones (para romper los círculos) y definirlas por medios de programación, no de explicación, semejante al trato de los términos *punto*, *recta*, *pertenecer* en la geometría escolar.

Entonces, en el análisis y la evaluación de la calidad de los diccionarios con respecto a las relaciones entre las definiciones surgen los siguientes problemas:

– ¿Cómo escoger las palabras usadas en las definiciones (vocabulario definidor)?
– ¿Cómo escoger los primitivos semánticos (para el uso computacional)?
– ¿Cómo evitar los círculos viciosos (cortos) en las definiciones?

En este sentido, los criterios para mejorar el diccionario serían:

– Tener el menor número de palabras en el vocabulario definidor
– Tener el menor número de círculos viciosos cortos (en el caso de que las palabras del vocabulario definidor también sean definidas)

## 3. Separación de los significados en sentidos

En la tarea de separación de sentidos de palabras hay tres posibles problemas:

– El diccionario no contiene algún sentido presente en el texto
– Varios sentidos del diccionario corresponden a un solo sentido en el texto (y no se trata de neutralización de algunas características)
– Un sentido del diccionario corresponde a varios sentidos en los textos

Esos casos se analizan en las siguientes subsecciones.

### *3.1. Falta de sentidos específicos*

Uno de los problemas del diccionario se presenta cuando éste no contiene algún sentido específico de una palabra. Por ejemplo, para la palabra *gato* se dan sentidos correspondientes al

*1) animal doméstico que maúlla*
*2) animal felino*

Pero no a la

*3) herramienta mecánica para la reparación de carros*

Este tipo de problemas, a diferencia de algunos otros, no se puede detectar automáticamente con tan sólo analizar el diccionario sino que es necesario comparar el diccionario con el uso real del lenguaje. Aparte de la introspección del lexicógrafo (la que no discutimos en este artículo), el método más adecuado es verificar si todas las palabras en un corpus grande corresponden a algún significado específico en el diccionario. Dicha verificación se puede hacer de dos maneras: manual y automática. Como siempre, la ventaja de la verificación manual es la calidad y la ventaja de la verificación automática es la rapidez.

Para la verificación manual, en una selección grande de ejemplos del uso de la palabra, cada ocurrencia de la misma se marca, manualmente, con uno de los sentidos seleccionados del diccionario; el hecho de que el anotador no encuentre ningún sentido adecuado (como sería con el ejemplo de *gato* arriba mencionado y con el texto *Para reparar su carro Juan tuvo que comprar un gato neumático*) indica el problema en el sistema de significación.

Para facilitar la anotación manual, en nuestro Laboratorio fue desarrollada una herramienta computacional (Ledo-Mezquita *et al*., 2003) que automáticamente selecciona cada palabra significativa del texto, una por una (pasando por alto las palabras funcionales como preposiciones), y presenta al usuario la lista de posibles significados de la palabra previstos en el diccionario, de entre los cuales el lexicógrafo puede escoger uno o, en su caso, marcar la palabra como la que tiene un sentido no previsto en el diccionario. La herramienta facilita la labor del anotador usando los métodos de la lingüística computacional para seleccionar automáticamente el sentido más probable en el contexto dado, el cual el anotador, en la amplia mayoría de los casos, puede simplemente confirmar. Si la preselección automática fue errónea, el programa ofrece al usuario el siguiente sentido más probable (según las heurísticas computacionales usadas), etc.

Ya que tal labor manual es costosa y aburrida, la manera más económica —aunque no más simple técnicamente— es la verificación puramente automática. En este caso sólo se verifica que las heurísticas usadas para elegir el sentido de cada palabra lo puedan hacer con cierto nivel mínimo de certeza. Los métodos correspondientes, del estado del arte actual, cometen una cantidad significativa de errores de dos tipos. Por un lado, a una palabra se le puede, erróneamente, asignar un sentido no pertinente en el contexto dado, con lo cual puede quedarse sin detectar una verdadera falta del sentido correcto en el diccionario. Por otro lado, en algunos casos, el error se puede reportar no debido a un problema real en el diccionario sino al fallo de las heurísticas o bien debido a que el contexto no presenta la información suficiente para la selección del sentido. Sin embargo, la ventaja de los métodos automáticos es la posibilidad de procesar una gran cantidad de textos prácticamente sin costo alguno. Sólo de esta manera es factible encontrar y considerar los sentidos de frecuencia baja y muy baja.

Entre los métodos para la selección automática de los sentidos de palabras en el contexto se pueden mencionar diferentes variantes del método de Lesk (1986). La idea básica de este método es buscar automáticamente en el contexto inmediato de la palabra, las palabras usadas en su definición. Por ejemplo, en el contexto *mi gato maúlla cuando ve al perro* está presente una palabra de la definición del primer sentido de nuestro ejemplo. Pero el contexto *Juan no pudo reparar su coche sin un gato* sólo es compatible con el tercer sentido del mismo ejemplo. Existen modificaciones de este método que usan diccionarios de sinónimos (Banerjee y Pedersen, 2002; Sidorov y Gelbukh, 2001) y métodos lingüísticos para la comparación de las palabras; por ejemplo, en el último contexto *coche* es sinónimo de *carro* y *reparar* es una derivación de *reparación*.

Sea la verificación manual o automática, es importante que en el corpus aparezca un número suficiente de ejemplos del uso de la palabra en cuestión, lo que es muy difícil de lograr para la mayoría de las palabras del diccionario. Efectivamente, según la famosa ley de Zipf, en cualquier texto unas cuantas palabras se repiten muchas veces, mientras que la mitad de las palabras que aparecen en el texto, aparecen en él sólo una vez. Con eso podemos deducir que incluso en un corpus muy grande, casi todas las palabras de un diccionario lo suficientemente completo, aparecen muy pocas veces o

ninguna. Entonces, el aplicar los métodos descritos arriba a un gran corpus tradicional parece un gran desperdicio de esfuerzo: se procesan muchísimas ocurrencias de unas pocas palabras del diccionario y muy pocas de casi todas las demás palabras.

Este problema se puede resolver con un corpus de un tipo específico, que llamamos un corpus representativo respecto al vocabulario dado (Gelbukh *et al*., 2002a), equivalente a una concordancia de palabras en contexto. Este tipo de corpus se colecciona automáticamente de Internet, el repositorio de textos más grande creado hasta ahora por el ser humano. Para cada palabra del diccionario, se colecciona un cierto número de contextos. Así, incluso para las palabras más raras que conocemos encontramos en este corpus el número de contextos suficiente para su investigación estadística. Lo que soluciona el problema que la ley de Zipf presenta para toda investigación basada en corpus.

### *3.2. Sistema de sentidos demasiado detallado*

Otro posible problema se presenta cuando los sentidos son demasiado finos, es decir, a un sentido del texto pueden corresponder varios sentidos del diccionario e incluso un humano tiene dificultades de escoger el sentido correcto.

Es importante decir que a veces la imposibilidad de escoger un sentido predeterminado está relacionada con la neutralización de algunas características semánticas, cuando el contexto no tiene la suficiente información para elegir un sentido predeterminado.

Por ejemplo, en el diccionario Anaya tenemos dos sentidos de la palabra *ventana*:

1. *Abertura, vano en un muro para iluminar y ventilar*.
2. *Armazón, marco con cristales para cerrarla*.

Ahora vamos a ver los siguientes ejemplos:

1. *Juan saltó por la ventana* $_{1\ (\text{pero no } 2)}$
2. *Juan rompió la ventana* $_{2\ (\text{pero no } 1)}$
3. *Juan saltó por la ventana* $_{2\ (\text{pero no } 1)}$ *y la rompió*
4. *Juan está mirando a través de la ventana* $_{1\ o\ 2}$
5. *Juan abrió la ventana* $_{1\ o\ 2}$

En los dos primeros ejemplos está muy claro de qué ventana —en el primer sentido o en el segundo sentido— se está hablando, mientras que en el caso del cuarto y quinto ejemplo, cualquiera de los dos sentidos es aceptable. Digamos, en el ejemplo cinco, puede ser que se abrió el espacio o que se movió el marco. Como vemos, la interpretación exacta depende del enfoque del hablante u oyente; en este caso el contexto no contiene la suficiente información para elegir. Sin embargo, eso no significa que no existan los dos sentidos, porque sí hay otros contextos donde ambos se distinguen.

Lo interesante del tercer ejemplo está relacionado con el hecho de que el contexto que se encuentra antes de la palabra *ventana* es igual al ejemplo 1, sin embargo, el sentido de la palabra es diferente. Es así, porque la segunda parte del contexto contiene la restricción para elegir el sentido –la ventana se rompe, lo que es aplicable solamente a la $ventana_2$. Es decir, el contexto contiene los datos (el conocimiento del mundo) que solamente son compatibles con $ventana_2$.

Ahora bien ¿cómo un humano escoge el sentido que corresponde al contexto? Se analiza el contexto y se aprovecha el conocimiento del mundo. Si hay algo que sólo es compatible con uno de los sentidos, se puede escoger este sentido, en caso de que esta información no está disponible, se neutraliza la diferencia entre los sentidos. Es decir, en el ejemplo 1, el conocimiento es que se puede saltar por algún espacio. En el ejemplo 2 se sabe que normalmente las ventanas se hacen de algún material como vidrio que se puede romper y que es parte del marco que cubre las ventanas. En el ejemplo 3, se sabe que se puede romper la ventana, por lo tanto se abrirá el espacio, y se puede saltar por el espacio abierto. Tal vez, el ejemplo 3 es el uso metafórico del sentido 2 en lugar del sentido 1. En los ejemplos 4 y 5 no hay información adicional, entonces no se puede elegir uno de los sentidos. De hecho, no está claro que tan relevante es la diferencia en el caso de esos ejemplos.

Veamos otro ejemplo. La palabra *agobiarse* tiene dos sentidos en el diccionario Anaya:

1. *Causar molestia o fatiga*
2. *Causar angustia o abatimiento*

Sin embargo, en el contexto *Él se agobió*, obviamente, no hay la posibilidad de escoger uno de esos dos sentidos.

En el caso de *ventana* existe una polisemia regular (Apresjan, 1974). Cuando los objetos son «un espacio plano limitado de los lados», como *puerta*, *esclusa*, etc., puede uno referirse a este objeto como a un espacio, y al mismo tiempo como a un objeto que cubre este espacio. Es decir, de un sentido siempre se puede inferir el otro. En el caso de *agobiarse* no existe el fenómeno de polisemia regular.

Proponemos que la solución al problema de sentidos demasiado finos (y la neutralización de sus diferencias) puede ser la representación del sentido como una jerarquía –en los niveles altos, se definen los sentidos más generales, y en los niveles más profundos, se especifican los sentidos más a detalle.

En el caso de polisemia regular el nivel más alto es la unión de los sentidos de nivel más bajo. Los sentidos en este caso son muy diferentes, por lo tanto, no tienen un sentido generalizado. También la definición debe tener la referencia que contiene el fenómeno de polisemia regular.

En el caso de *agobiarse* es necesario generalizar los dos sentidos, como, por ejemplo:

*Causar una sensación desagradable en el cuerpo humano*

Nótese que no se especifica si el sentimiento está relacionado con el estado físico o el estado psicológico. En el nivel más bajo, se dan las definiciones como están en el diccionario. La profundidad posible de la jerarquía es el objeto de investigaciones futuras.

Entonces, en algunos contextos se puede determinar cuál sentido de nivel más bajo se usa, y si no se puede, se hace la referencia al sentido generalizado.

Es importante mencionar que el fenómeno que tratamos no es el caso de falta de precisión (*vagueness*, en inglés). La diferencia entre la ambigüedad (en nuestro caso, de sentidos) y de la falta de precisión es que, en el caso de la ambigüedad, algo puede tener varios sentidos, y lo que no está claro es cuál sentido se usa en el contexto. Mientras que la falta de precisión se refiere al hecho de que un concepto no está bien definido. Los casos que vimos se trataban de ambigüedad.

Ahora bien, ¿cómo se puede aplicar el análisis automático para ayudar al lexicógrafo a detectar las situaciones de sentidos potencialmente similares, los cuales pueden ser tanto los casos de polisemia regular como requerir la generalización? Recordemos que es el lexicógrafo quien toma las decisiones y el sistema sólo trata de ayudarle.

Hemos desarrollado un método que permite calcular la similitud entre los sentidos de la misma palabra (Gelbukh *et al*., 2003). Brevemente, la idea es calcular la similitud de los sentidos usando la medida de semejanza entre las definiciones, muy parecida a la medida de similitud de los textos conocida como el coeficiente de *Dice*, véase, por ejemplo (Rasmussen, 1992). El coeficiente de *Dice* representa la intersección normalizada de las palabras en los textos. Es decir, se toma la intersección textual medida en palabras de dos textos y se divide entre la suma de las palabras en los textos. De preferencia las palabras deben estar normalizadas, por ejemplo, *trabajabas*, *trabajar*, y *trabajaron* se refieren a la misma palabra (lema) *trabajar*.

La medida modificada toma en cuenta adicionalmente los sinónimos de las palabras, porque por definición los sinónimos expresan los mismos conceptos y para algunas tareas se puede ignorar los matices de sentidos que normalmente tienen los sinónimos. La medida propuesta es como sigue:

$$S(t_1, t_2) = \frac{|W_1 \cap W_2| + |W_1 \circ W_2|}{\max(|W_1|, |W_2|)}$$

donde $W_1$ y $W_2$ son conjuntos de palabras en los textos $t_1$ y $t_2$, $|W_1 \cap W_2|$ significa que se calcula el número de las palabras (por lemas; recordemos que aplicamos la normalización morfológica automática) que se encuentran en definiciones de ambos sentidos de la palabra y $|W_1 \circ W_2|$ representa el número de intersecciones usando los sinónimos. Es decir, para cada palabra se toma su lista de sinónimos y cada sinónimo de esta lista se busca en el otro texto. En caso que este sinónimo se encuentre allá, se aumenta el número de intersecciones. El algoritmo está diseñado de tal manera que cuenta cada intersección sólo una vez —si la palabra o su sinónimo ya se encontró, no se buscan más sinónimos de esta palabra. Eso significa que el número de intersecciones no puede ser mayor que el

número máximo de las palabras en uno de los textos (el que contiene más palabras) —el valor que aparece en el denominador. El denominador sirve para una normalización, que significa que el resultado no depende del tamaño del texto.

Aplicamos este algoritmo al diccionario Anaya, comparando los pares de sentidos de cada palabra, y obtuvimos que cerca de 1% de todos los pares de sentidos son muy parecidos (contienen más de 50% de los mismos conceptos) y cerca de 10% de los pares son sustancialmente parecidos (contienen más de 25% de los mismos conceptos). Consideramos, que, por lo menos, para ese 10% de los sentidos parecidos, el lexicógrafo evaluará si sus definiciones son válidas.

### *3.3. Sentidos demasiado generales*

Un tercer posible problema se presenta cuando el mismo sentido del diccionario cubre usos claramnente diferentes de las palabras. Por ejemplo, la definición de *llave* como *un objeto que se usa para abrir o cerrar algo* cubre tanto el contexto *Juan sacó la llave de su bolsillo y abrió la puerta* como *Juan entró al baño y abrió la llave del agua caliente*. Sin embargo, los hablantes tendemos a considerar «la llave para la puerta» y «la llave para el agua» como cosas muy diferentes, al grado de que el uso de la misma palabra para cosas tan diferentes parece ser pura coincidencia.

El procedimiento descrito en la sección 3.1 no detectará ningún problema con esta definición, ya que en ambos contextos a la palabra en cuestión se le asignará un sentido del diccionario. Tampoco es simple detectar el problema manualmente comparando los contextos en los cuales a la palabra se le asignó el mismo sentido, ya que están en lugares distantes en el corpus, además del alto costo de la gran labor manual necesaria para tal comparación.

Se pueden usar varios algoritmos para la verificación automática de la homogeneidad del conjunto de los contextos en los cuales a la palabra se le marcó con el mismo sentido. Aquí discutiremos dos métodos basados en el agrupamiento automático (clusterización, de la palabra inglesa *clustering*) de los contextos. Por contexto de una palabra entendemos las palabras que la rodean en el texto; este concepto se puede precisar de diferentes maneras, desde la oración que la contiene hasta las palabras que están dentro de una cierta distancia desde la palabra dada. Los dos métodos tratan de analizar diferentes sentidos de una palabra dependiendo de su contexto.

En el primer método, para cada sentido de la palabra se seleccionan los contextos y se agrupan, según una medida de semejanza entre los textos (Alexandrov y Gelbukh, 1999; Alexandrov *et al*., 2000), en dos grupos de tal manera que la distancia entre los elementos dentro de cada grupo se minimiza y la distancia entre los dos grupos se maximiza. Esta última distancia da una medida de la calidad de la definición. En el caso de una definición mala los contextos se dividirán claramente en dos o más grupos no parecidos entre sí. En el caso de nuestro ejemplo con la palabra *llave*, un grupo se caracterizará por las palabras *puerta*, *llavero*, *bolsillo*, *insertar*, *olvidar*, mientras que el otro por las palabras *agua*, *caliente*,

*fría*, *baño*, *lavar*. Nótese que nuestro método no penaliza indiscriminadamente los sentidos generales: aunque la palabra *objeto* (en el sentido de *cualquier cosa*) es muy general y en consecuencia los contextos de su uso son muy diversos, éstos no se dividen en grupos claramente distinguibles sino llenan uniformemente un área amplia.

Otro método (Jiménez-Salazar, 2003) ayuda a verificar todo el conjunto de los sentidos de una palabra en el diccionario. Los contextos de la palabra dada encontrados en el corpus se agrupan automáticamente, también usando alguna medida de semejanza entre dos contextos, por ejemplo, el número de palabras que ambos contextos compartan. La hipótesis del método es que diferentes sentidos de la palabra se usan en diferentes contextos, entonces los grupos de contextos tales que los contextos son parecidos dentro de cada grupo y diferentes entre grupos diferentes, representan los sentidos diferentes de la palabra. Usando los métodos descritos más arriba, tales como las distintas modificaciones del método de Lesk, se puede incluso asociar los sentidos presentes en el diccionario para la palabra dada con los grupos de contextos detectados en el corpus. La buena correspondencia indica que el sistema de los sentidos está bien hecho mientras la mala es una alarma. Nótese que en este caso el procedimiento de evaluación es puramente automático, mientras que la resolución de los problemas encontrados necesitan la intervención del lexicógrafo.

Resumiendo, el primer método usa las técnicas de clasificación automática y sólo se analiza un sentido de palabra a la vez para precisar si la definición del sentido es buena o no. Se supone de antemano que todos los contextos corresponden al mismo sentido. El segundo método usa las técnicas de desambiguación de sentidos de palabras y trata de asociar cada contexto con algún sentido. En caso de no encontrar un sentido apropiado se reporta un posible problema.

## 4. Otros tipos de verificación formal

Aunque no lo discutimos a detalle en este artículo, hay muchos otros aspectos del diccionario explicativo que se pueden verificar automáticamente. La base de tal verificación son las propiedades formales (parecidas a lo que en el contexto de las gramáticas formales o bases de datos se llaman *restricciones*) que demuestran las relaciones entre los elementos de este sistema tan complejo que es el diccionario explicativo. Aquí sólo damos unos pocos ejemplos.

### *4.1. Verificación de la ortografía y la estructura de los artículos*

A diferencia de otros tipos de verificación que discutimos en este artículo, en esta subsección mencionamos brevemente dos tipos de verificación local, que no involucra ninguna comparación de los elementos distantes en el texto del diccionario: la verificación de la ortografía y la verificación de la estructura.

La verificación de ortografía y gramática se aplica a cualquier texto, sin ser excepción un diccionario explicativo. Existe una vasta cantidad de literatura y una gran variedad de métodos y heurísticas utilizados para la verificación de este tipo (Kukich, 1992). Incluso cualquier procesador de palabras moderno (como Microsoft Word™) contiene herramientas de esta naturaleza. Por esta razón no dedicaremos más espacio en este artículo a la presentación de los métodos de verificación de ortografía y gramática.

Sin embargo, haremos notar que debido a la gran importancia de la perfección de los diccionarios, tiene sentido aplicar métodos que garantizan mayor calidad de verificación que los tradicionales, es decir, verificación más exhaustiva. Aquí el punto clave es el balance entre el número de errores omitidos y las alarmas falsas (lo que en la literatura especializada se llama la relación entre especificidad (*recall*, en inglés*)* y precisión. Los métodos de verificación que producen un número demasiado alto de alarmas falsas (de baja precisión); es decir, los que reportan un posible error que la verificación manual no confirma, que es muy característico de los métodos de verificación exhaustiva —de alta especificidad (*recall*)— no son prácticos en el uso cotidiano; sin embargo, pueden ser de gran utilidad en la verificación de diccionarios y otros textos importantes.

Entre los métodos de este tipo podemos mencionar la detección de malapropismos. El malapropismo es un tipo de error de la palabra existente en un lenguaje (*real-word errors* en inglés), el cual consiste en sustituir, por accidente, una palabra con otra igual de correcta y válida en el mismo lenguaje. Lo que en algunos casos resulta en una palabra de una categoría gramatical distinta, tales casos son simples de detectar con un análisis puramente gramatical, por ejemplo: *este articulo es interesante* (en vez de *artículo*). Sin embargo, en otros casos —éstos se llaman malapropismos— sólo las consideraciones semánticas permiten detectar el error, por ejemplo: *centro histérico de la ciudad*, *en la reserva la casa de venados está prohibida* / *mi caza tiene tres pisos y está pintada de blanco*. Los métodos existentes de detección de malapropismos (Hirst y Budanitsky, 2003; Bolshakov y Gelbukh, 2003) demuestran usualmente muy baja precisión cuando están configurados para una especificidad (*recall*) razonablemente alta. Eso limita su uso en los procesadores de palabras comunes, pero todavía pueden ser útiles para una verificación más exhaustiva de los diccionarios.

Otro tipo de verificación local es el análisis de la estructura de los artículos. Por ejemplo, verificar que cada palabra significativa (no funcional) usada en el texto del diccionario tenga definición en éste, y en su caso proporcionar al lexicógrafo la lista de palabras usadas sin ser definidas (lo que en la sección 2 hemos llamado el vocabulario definidor). También se puede verificar que cada artículo contenga las partes obligatorias, por ejemplo, pronunciación, etimología, explicación y ejemplos. Igualmente se puede observar la numeración correcta de los sentidos y subsentidos, el orden de los elementos del artículo, el orden alfabético de los artículos, las fuentes tipográficas correspondientes a diferentes elementos del artículo, etc.

### *4.2. Verificación de las marcas de sinonimia y antonimia*

Usualmente los diccionarios explicativos marcan las relaciones básicas entre palabras, tales como sinonimia y antonimia, y en algunos casos —como, por ejemplo, WordNet (Fellbaum, 1998)— otras relaciones tales como meronimia, etc. En el sistema de estas relaciones existen ciertas propiedades (restricciones), por ejemplo:

- Simetría: si la palabra *A* es sinónima de la palabra *B* entonces normalmente *B* es sinónima de *A*
- Transitividad: si la palabra *A* es sinónima de la palabra *B* y *B* es sinónima de *C* entonces es probable (aunque en muchos casos no cierto) que *A* sea sinónima de *C*

Como en otros casos de las propiedades de las relaciones entre las palabras colocadas distantemente en el texto del diccionario es muy difícil (o por lo menos laborioso) verificar tales restricciones manualmente, pero es más fácil hacer que un programa los verifique y atraiga la atención del lexicógrafo a los posibles problemas detectados. Nótese que se pueden tratar de manera semejante otras relaciones, tales como antonimia, meronimia, etc. Incluso se pueden combinar las verificaciones que involucran relaciones diferentes: por ejemplo, un antónimo de una palabra normalmente no debe ser su merónimo, ni su sinónimo, ni un sinónimo de su sinónimo, etc.

Uno puede argumentar que el autor del diccionario, en su sano juicio, no puede marcar la palabra *A* como sinónima de la *B* y a la vez marcar la *B* como antónima de la *A*, y que entonces no tiene caso aplicar en práctica las heurísticas que aquí discutimos. Sin embargo, la aplicación de tales heurísticas no le sirve al programa para argüir con el autor del diccionario sobre los asuntos lingüísticos, sino para detectar posibles errores mecanográficos o incluso errores puramente ortográficos, de la manera semejante a la detección de malapropismos. Por ejemplo:

*cuerdo* < ... >. Antónimo: *poco*

en vez de *loco*. Aquí, el error probablemente ocurrió debido a que el dedo tocó la tecla *p* en lugar de la cercana *l*, lo que puede suceder en el proceso de preparación de texto. Sin embargo, la única manera que podemos imaginar para detectar automáticamente este error no es la verificación de la ortografía, por muy exhaustiva que esta sea, sino atraer la atención del lexicógrafo que en la definición de la palabra *poco* no se indica, como se esperaba, que tenga un antónimo *cuerdo*.

Otra posible técnica para la verificación de las marcas de sinonimia o antonimia es la comparación de las definiciones. En este caso, más bien se trata de determinar automáticamente qué palabras son sinónimas y verificar si así están marcadas en el diccionario. La hipótesis que aquí se verifica es que las palabras cuyas definiciones son semejantes deben ser marcadas como sinónimas (o antónimas, ya que es difícil interpretar

las negaciones automáticamente) y ningunas otras deben ser así marcadas. El incumplimiento de esta hipótesis para un par dado de palabras puede significar, o bien la marca de sinonimia mal puesta, o bien (mucho más probable) algún problema de las definiciones. Por ejemplo, si las palabras marcadas como sinónimas se definen de manera muy diferente, eso puede indicar inconsistencia en las definiciones. Por otro lado, si dos palabras no marcadas como sinónimas se definen de modo demasiado semejante, eso puede indicar que las definiciones son demasiado generales para reflejar el significado específico de estas palabras.

Como medida de semejanza se puede usar el número de palabras compartidas entre las dos definiciones o variantes de este método, como se describe en la sección 3.2 más arriba. Para obtener una medida más estricta, se puede considerar también el orden de las palabras compartidas, es decir, alguna medida derivada de la distancia de Levenshtein (1966).

Otra posible fuente de información sobre la sinonimia es un corpus grande de oraciones. Aquí la hipótesis a verificar es que los sinónimos se usan en contextos iguales o muy parecidos. Sean las dos palabras en cuestión $p_1$ y $p_2$ y sea que aparecen en los dos oraciones (digamos, oraciones) $C_1$ y $C_2$, respectivamente. ¿Cómo podemos saber que las textos $C_1$ y $C_2$ se parecen? No basta con identificar que ambas cadenas son iguales o muy parecidas y que sólo difieren en que en $C_1$ se usa $p_1$ y en $C_2$ se usa $p_2$ (en vez de $p_1$), lo difícil es saber si significan lo mismo; por ejemplo, aunque las palabras *vaca* y *cabra* pueden aparecer en contextos iguales —*la leche de vaca* (*cabra*) *es sabrosa y nutritiva*— eso no significa que son sinónimas ya que el significado de estos textos no es idéntico. Una de las maneras en que podemos saber si el significado de dos textos, cortos pero diferentes, resulta idéntico es con la comparación de diccionarios explicativos diferentes, sobre todo terminológicos, ya que en éstos se reduce la ambigüedad (Sierra y McNaught, 2003; Sierra y Alarcón, 2002). Por ejemplo, supongamos que tres diccionarios diferentes dan las siguientes definiciones:

- Diccionario 1: *velocímetro: dispositivo para medir la velocidad de movimiento*
- Diccionario 2: *velocímetro: dispositivo para determinar la velocidad de movimiento*
- Diccionario 3: *velocímetro: aparato que se usa para determinar la rapidez de moción de algo*

Comparando la definición en el diccionario 1 con la del diccionario 2 es simple notar que la palabra *determinar* se usa en vez de *medir*; nótese que el hecho de que ambos textos definan la misma palabra *velocímetro* garantiza que el significado de los mismos es idéntico. En la práctica es más común el caso que se presenta en la comparación de las definiciones de los diccionarios 1 y 3: en este caso no es tan simple detectar automáticamente la semejanza entre los dos textos, sin embargo existen técnicas para hacerlo (Sierra y McNaught, 2000).

## 5. Herramienta ayudante de lexicógrafo

Las ideas presentadas en las secciones anteriores nos llevaron al desarrollo de una herramienta que permita al lexicógrafo investigar la estructura del diccionario con el fin de detectar y corregir varios tipos de defectos en la estructura del diccionario. La herramienta analiza el texto del diccionario y atrae la atención del lexicógrafo a los problemas encontrados, según lo expuesto en las secciones 3 y 4.

Además, la herramienta proporciona la interfaz interactiva para el desarrollo o la modificación del diccionario. Este software está diseñado para proporcionar al lexicógrafo la siguiente información:

- Visualiza el diccionario en una interfaz gráfica amigable, en un formato tabular, claramente distinguiendo diferentes elementos de cada definición, tales como la pronunciación, etimología, sentidos, subsentidos, ejemplos, relaciones con otras palabras, etc.
- Para la palabra elegida, muestra varias características de la misma, tales como su frecuencia en las definiciones del diccionario, el tamaño de su propia definición, el largo mínimo del ciclo en el sistema de definiciones en que está involucrada (se refiere a las definiciones como *gallina es hembra del gallo* y *gallo es macho de la gallina*), etc.
- También, proporciona la información sobre el uso de la palabra en el gran corpus de textos y en Internet [1], tal como la frecuencia, los contextos del uso, los contextos agrupados, un árbol del agrupamiento de los contextos —desde la división *grosso modo* hasta los matices finos— lo que se usa para facilitar la división del artículo en sentidos, etc. Aquí, la herramienta permite al usuario elegir los sentidos para las ocurrencias de las palabras en el corpus (véase más abajo).
- Permite buscar las palabras por sus definiciones, por ejemplo: *¿cómo se llama un dispositivo para medir la velocidad de movimiento?* En esto se aplican los métodos de búsqueda inteligente usando sinonimia entre las palabras de la petición y el texto (Sierra y McNaught, 2003; Gelbukh *et al*., 2002b).
- Construye la lista de las palabras usadas en el corpus con una frecuencia considerable pero ausentes al vocabulario del diccionario. Para esto se emplea la normalización morfológica (lematización, cf. *stemming* en inglés) –para que el programa no reporte todas las formas morfológicas de las palabras (por ejemplo, *piensas*) como ausentes al vocabulario (que sólo contiene *pensar*).

En cuanto a los últimos puntos, la herramienta proporciona la interfaz gráfica para el estudio y marcaje del corpus (Ledo-Mezquita *et al*., 2003), permitiendo al usuario elegir los sentidos específicos, de entre los que el diccionario proporciona, para cada ocurren-

[1] En caso de Internet, la frecuencia aproximada se calcula usando de las máquinas de búsqueda existentes, tales como *Google*, las cuales determinan el número de los documentos donde se encuentra la palabra.

cia de cada palabra significativa, como se describe en la sección 3.1 y con los fines de desarrollar un corpus marcado con sentidos necesario para la aplicación de los algoritmos mencionados en la sección 3.3.

Otro módulo de la herramienta ayuda al lexicógrafo a construir un mejor conjunto de las palabras primitivas, según lo expuesto en la sección 2; por ejemplo, el lexicógrafo debe considerar que el conjunto definidor no debe tener muchas palabras de frecuencia baja. Para esto la herramienta:

- Genera diferentes conjuntos definidores mínimos permitiéndole al usuario controlar varios parámetros del algoritmo de su generación. Muestra los conjuntos generados junto con la información (tal como la frecuencia) sobre cada palabra incluida en el conjunto.
- Permite al lexicógrafo cambiar manualmente el conjunto definidor generado y verifica que el conjunto cambiado todavía es un conjunto definidor y que es mínimo.
- Permite al lexicógrafo cambiar las definiciones de las palabras e inmediatamente muestra el impacto en los conjuntos definidores que se generan.
- Dada una lista de las palabras que el lexicógrafo quiere que sean no primitivas,verifica si existe algún conjunto definidor que no las contiene. Éste existe siempre y cuando las palabras elegidas no formen círculos viciosos. Si así es, genera una o varias variantes de tal conjunto. Si no es así, muestra los círculos, lo que ayuda a eliminar de la lista las palabras que los causan.
- Dada una lista de las palabras que el lexicógrafo quiere que sí sean definidoras, genera uno o varios conjuntos definidores que contengan estas palabras. Si tal conjunto definidor no puede ser mínimo, sugiere eliminar ciertas palabras de la lista.
- Dado un conjunto definidor mínimo, la herramienta puede:
- Para una palabra no primitiva, mostrar su definición expandida a las palabras definidoras, es decir, la que consiste sólo de las palabras definidoras.
- Para una palabra primitiva, mostrar los ciclos (más cortos o todos) que su definición actual causa en el diccionario.

En la actualidad no todos los módulos de la herramienta están completamente implementados, aunque disponemos de los algoritmos necesarios y planeamos incorporarlos en la herramienta. Los módulos de la herramienta más desarrollados hasta la fecha son los del marcaje del corpus y la selección del vocabulario definidor.

## 6. Conclusiones y trabajo futuro

Un diccionario explicativo es un sistema complejo con numerosas relaciones entre sus elementos y con diferentes restricciones (requerimientos) que tales relaciones deben satisfacer para garantizar la integridad y consistencia del diccionario. La verificación de tales requerimientos involucra el análisis no local, es decir, la consideración de los elementos localizados en diferentes lugares en su texto, lo que es casi imposible de hacer manualmente, pero que se facilita en gran medida con el uso de computadoras y la aplicación de algoritmos correspondientes, de diferente grado de complejidad e inteligencia.

La verificación automática no sustituye al lexicógrafo sino atrae su atención a posibles problemas y le proporciona la información necesaria para tomar una decisión informada y consciente, sea ésta el hacer modificaciones al texto del diccionario o dejarlo tal cual. Más allá de la verificación, las herramientas computacionales permiten el desarrollo interactivo del diccionario proporcionándole al lexicógrafo la información sobre las relaciones entre la palabra actual y las palabras relacionadas con ésta («cercanas» a ella en la estructura lógica), aunque distantes en el texto plano del diccionario.

Aún más allá, las técnicas computacionales permiten la construcción puramente automática de muchos de los elementos del diccionario —desde el vocabulario y la información estadística hasta la división de los artículos en sentidos con los ejemplos correspondientes, y la detección de sinonimia entre las palabras—, en la mayoría de los casos a partir del análisis de una gran cantidad de textos es decir, un corpus. En este artículo sólo hemos considerado tales posibilidades con el único fin de comparar los datos obtenidos automáticamente con los presentes en el diccionario. Otro uso de estos métodos que no hemos discutido, es la construcción automática de un borrador del diccionario completo, para su perfección manual posterior.

Estas consideraciones llevaron al desarrollo en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del CIC-IPN, de una herramienta computacional que proporcione estos servicios al lexicógrafo, junto con las facilidades para el marcaje semiautomático de los sentidos de palabras en el corpus. A la fecha, se han desarrollado los algoritmos principales de tal herramienta y se están integrando con la interfaz gráfica amigable al usuario.

## Referencias

ALEXANDROV, M., GELBUKH, A. (1999) “Measures for determining thematic structure of documents with domain dictionaries”. Proc. *Text mining workshop at 16th International joint conference on artificial intelligence* (*IJCAI’99*), Stockholm, Sweden: 10–12.

ALEXANDROV, M., GELBUKH, A., MAKAGONOV, P. (2000) “On metrics for keyword-based document selection and classification”. Proc. *CICLing-2000, International conference on intelligent text processing and computational linguistics*, Mexico City, February: 373–389.

APRESJAN, J. D. (1974) “Regular polysemy”, *Linguistics*, No. 142: 5–32.

BANERJEE, SATANJEEV, and PEDERSEN, T. (2002) "An adapted lesk algorithm for word sense disambiguation using WordNet". Proc. *CICLing-2002, Computational linguistics and intelligent text processing*. Lecture notes in computer science N 2276, Springer-Verlag: 136–145.

BOLSHAKOV, I. A., and A. GELBUKH (2003) "On detection of malapropisms by multistage collocation testing". Proc. *NLDB-2003, 8th International workshop on applications of natural language to information systems*, Lecture notes in computer science, Springer-Verlag (to appear).

GELBUKH, A., SIDOROV, G. (2002) "Selección automática del vocabulario definidor en un diccionario explicativo". España: *Procesamiento de lenguaje natural,* SEPLN: 29: 55–64.

GELBUKH, A., SIDOROV, G., and CHANONA-HERNANDEZ. L. (2003) "Automatic evaluation of the quality of an explanatory dictionary by comparison of word senses". Proc. *5th Conference on perspectives of informatics systems*. Lecture notes in computer science, Springer-Verlag, to appear.

GELBUKH, A., SIDOROV, G., and CHANONA-HERNÁNDEZ, L. (2002a) "Compilation of a Spanish representative corpus". Proc. *CICLing-2003, Computational linguistics and intelligent text processing*. Lecture notes in computer science N 2588, Springer-Verlag: 285–288.

GELBUKH, A., G. SIDOROV, A., and GUZMÁN-ARENAS (2002b) "Relational data model in document hierarchical indexing". In: E. Ranchhold, N. J. Mamede (Eds.). Proc. *PorTAL-2002, Advances in natural language processing*. Lecture notes in computer science, N 2389, Springer-Verlag: 259–262.

Grupo Anaya (1996) *Diccionario de la lengua española*: www.anaya.es.

EVENS, M. N. (ed.) (1988) *Relational models of lexicon: Representing knowledge in semantic network*. Cambridge: Cambridge University Press.

FELLBAUM, C. (ed.) (1998) *WordNet: an electronic lexical database*, Cambridge Mass: MIT Press.

HARTMANN, R.R.K. (2001) *Teaching and researching lexicography*. Pearson Education Limited.

HIRST, G., and BUDANITSKY, A. (2003) "Correcting real-word spelling errors by restoring lexical cohesion". *Computational linguistics* (to appear).

JIMÉNEZ-SALAZAR, H. (2003) "A method of automatic detection of lexical relationships using a raw corpus". Proc. *CICLing-2003, Computational linguistics and intelligent text processing*. Lecture notes in computer science N 2588, Springer-Verlag: 325–328.

KOZIMA, H. And FURUGORI, T. (1993) "Similarity between words computed by spreading activation on an English dictionary". *Proc. 6th conf. of the european chapter of ACL*: 232–239.

KUKICH, K. (1992) "Techniques for automatically correcting words in texts", *ACM Computing surveys*, N 24 (4): 377–439.

LANDAU, S. (2001) *Dictionaries: the art and craft of lexicography*. Cambridge: Cambridge University Press.

*LDOCE* (*Longman dictionary of contemporary English*). Longman: www.longman.com/ dictionaries/ which_dict/ldocenew.html.

LEDO-MEZQUITA, Y., SIDOROV, G., GELBUKH, A. (2003) "Tool for computer-aided Spanish word sense disambiguation". Proc. *CICLing-2003, Computational linguistics and intelligent text processing*. Lecture notes in computer science N 2588, Springer-Verlag: 277–280.

LESK, M. (1986) "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone". Toronto, Canada: Proc. of *ACM SIGDOC Conference*: 24-26.

LEVENSHTEIN, V. I. (1966) "Binary codes capable of correcting deletions, insertions, and reversals". *Cybernetics and control theory*, 10 (8): 707–710.

*OALD* (*Oxford advanced learner's dictionary*). Oxford University Press, www1.oup.co.uk/ elt/oald.

OZHEGOV, S. I. (1990). *Diccionario explicativo del idioma ruso* (en ruso), Moscú, Rusia. Edición 22ª.

SAINT-DIZIER, P. and VIEGAS, E. (eds.) (1995) *Computational lexical semantics*. Cambridge: Cambridge University Press.

SIDOROV, G., and A. GELBUKH (2001) "Word sense disambiguation in a Spanish explanatory dictionary". Tours, France: *Proc. TALN-2001*: 398–402.

SIERRA, G. and ALARCÓN, R. (2002) "Recurrent patterns in definitory context". Proc. *CICLing-2002, Computational Linguistics and intelligent text processing*. Lecture notes in computer science N 2276, Springer-Verlag: 438–440.

SIERRA, G. and MCNAUGHT, J. (2000) "Analogy-based method for semantic clustering". Proc. *CICLing-2000, International conference on intelligent text processing and computational linguistics*, Mexico City, February.

SIERRA, G. and MCNAUGHT, J. (2003) "Natural language system for terminological information retrieval". Proc. *CICLing-2003, computational linguistics and intelligent text processing*. Lecture notes in computer science N 2588, Springer-Verlag: 543–554.

SINGLETON, D. (2000) *Language and the lexicon: an introduction*. Arnold Publishers.

VOSSEN, P. (2001) "Condensed meaning in EuroWordNet". In: P. Boillon and F. Busa, *The language of word meaning*. Cambridge: Cambridge University Press: 363-383.

WIERZBICKA, A. (1996) *Semantics: primes and universals*. Oxford: Oxford University Press.

ZGUSTA, L. (1971). *Manual of lexicography*. Hague: Mouton, Prague: Academia.

## Agradecimientos

El trabajo ha sido realizado con el apoyo parcial del Gobierno de México (CONACyT y SNI) y el Instituto Politécnico Nacional, México (CGEPI, COFAA, COTEPABE). Expresamos nuestro más cordial agradecimiento a la Dra. Sofía Galicia Haro por sus útiles consejos.