Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística, editadas por Milka Villayandre Llamazares, León, Universidad de León, Dpto. de Filología Hispánica y Clásica, 2006. ISBN: 84-690-3383-2. Publicación electrónica en: http://www3.unileon.es/dp/dfh/SEL/actas.htm

FORMALIZACIÓN DE LAS CORRESPONDENCIAS ENTRE ACEPCIONES Y CONTEXTOS SINTAGMÁTICOS EN ESPAÑOL E INGLÉS[1]

AQUILINO SÁNCHEZ PÉREZ
MOISÉS ALMELA SÁNCHEZ
Universidad de Murcia

1. RELEVANCIA DE LA DESAMBIGUACIÓN LÉXICA AUTOMÁTICA

La desambiguación léxica automática, o asignación computarizada de los sentidos de una palabra en el discurso, es un instrumento útil para el acometimiento de tres tipos de tarea: primero, la aplicación de los modelos lingüísticos a la implementación de aplicaciones informáticas; segundo, el perfeccionamiento de las herramientas informáticas al servicio de la descripción lingüística; y tercero, la comprobación experimental de modelos lingüísticos.

El primer aspecto se desglosa en una serie de aplicaciones, de las cuales comentaremos algunas a continuación. Entre tales aplicaciones se encuentran las interfaces hombre-máquina. El procesamiento del

[1] Este artículo es una presentación del proyecto de investigación titulado "Definición y tipificación de las acepciones de términos léxicos polivalentes en inglés y en español, para el diseño de un prototipo de desambiguación automática". Este proyecto reúne a diez investigadores de universidades de España, Ecuador y Estados Unidos en la finalidad de desarrollar un prototipo de programa informático de desambiguación léxica automática. El proyecto, que fue concedido por el Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (HUM2004-00080/FILO), tiene prevista una duración de tres años (Dic. 2004-Dic. 2007).

lenguaje natural por parte de un ordenador alcanza resultados más afinados si los sentidos léxicos se encuentran tipificados y se dispone de un programa eficaz para la asignación automática de sentidos. Otro tipo de aplicaciones que se ven beneficiadas por la desambiguación automática son las relativas a la extracción y/o recuperación de información. La gestión eficiente de contenidos de textos electrónicos, especialmente de Internet, requiere la anotación semántico-designativa de los textos disponibles. De hecho, el programa *e*Content de la Unión Europea está basado en un informe según el cual es necesario recurrir a *metadatos* para poder llevar a cabo una extracción eficiente de la información que transita por la Red. En esencia, los metadatos textuales son, lógicamente, datos sobre los datos de un texto, es decir, información sobre el tipo de contenidos que pueden hallarse en el texto.

Más concretamente, los metadatos de tipo semántico contribuyen decisivamente a delimitar la información acerca de cuáles son los contenidos tratados en un texto. Algunas de las palabras clave de un texto pueden ser ambiguas desde el punto de vista temático. Por ejemplo, si introducimos la palabra *bisagra* en un motor de búsqueda en Internet, y nuestro objetivo es seleccionar textos relacionados con los dominios temáticos de "carpintería" o "bricolaje", es muy posible que obtengamos resultados tanto deseados como no deseados. Junto a textos que cumplirían con nuestras expectativas, hallaríamos otros que no se ajustarían a nuestro deseo inicial de buscar textos relacionados con "carpintería" o "bricolaje".

Por ejemplo, comparemos las siguientes cuatro concordancias extraídas del Corpus *Cumbre*[2], en su última versión de 40 millones de palabras. La concordancia (1) se ajustaría a nuestra intención de búsqueda, pero no así las concordancias (2)-(4). En el segundo ejemplo, lo denotado se encuadra dentro del dominio de la realidad anatómica. El tercer ejemplo se refiere a una realidad más abstracta. El sentido de *bisagra* en este caso podría parafrasearse como

[2] El Corpus *Cumbre*, propiedad de la compañía editorial SGEL, fue compilado bajo la dirección del Dr. Aquilino Sánchez. En su versión anterior de 20 millones de palabras, el Corpus sirvió como base empírica para la preparación del *Gran Diccionario de Uso del Español Actual* (*GDUESA*). La estructura del corpus, así como los criterios seguidos para su compilación, aparecen explicados en Sánchez (1995). Si bien Sánchez (1995) se refiere a la primera versión del Corpus, de 8 millones de palabras, las proporciones en la distribución de número de palabras por dialectos y tipos de texto no se han alterado en las versiones posteriores.

'vínculo intermediario entre dos cosas opuestas al menos aparentemente'. Resulta difícil adscribir este sentido a un dominio temático en particular. Por último, la concordancia (4) ejemplifica el uso de la palabra *bisagra* como término político.

(1) Además, para mantener un alto grado de protección en una colisión lateral, han diseñado una puerta posterior, que se articula sobre una única *bisagra* de gran tamaño y que dispone de fuertes anclajes en el arco del techo y en el larguero.

(2) Mi cuerpo se había vuelto de plomo con remaches en todas las *bisagras* de mis piernas y brazos.

(3) En segundo lugar, el triunfo de Luiz Inacio Lula da Silva y del Partido de los Trabajadores, PT, es un acontecimiento que puede transformarse en sí mismo en una *bisagra* histórica.

(4) Y quizás este partido sea una *bisagra* y hasta nos haya servido la derrota para saber que no somos invencibles.

Parece cierto que un mismo significado lingüístico de fondo subyace a los cuatro usos arriba mencionados. El significado de 'vínculo intermediario entre dos cosas' parece ser aplicado en cada caso a dominios distintos de la realidad. Así, en el ejemplo (1), *bisagra* denota la unión de dos superficies separadas, una de ellas fija y la otra móvil; en (2), *bisagra* se refiere al punto de unión entre huesos del cuerpo, es decir, a las articulaciones; en (3), se refiere a la relación de continuidad entre dos etapas históricas, y finalmente en (4), lo denotado es la función de un partido político como intermediario entre otros partidos y/o ideologías más opuestos entre sí.

En cualquier caso, la hipotética unidad semántica subyacente a todos los usos de la palabra poco importa a los propósitos de la extracción de información de textos electrónicos, dado que la unidad y especificidad designativo-conceptual es prioritaria para la gestión eficiente de datos, y dichas cualidades no emanan del posible significado unitario de la palabra, sino de los usos en entornos textuales estereotípicos. Por ello, la asignación de un sentido, o uso designativo-conceptual, a los distintos usos textuales de la palabra es

necesario para poder restringir los dominios temáticos que son objeto de búsqueda en cada caso.

Además, los programas de desambiguación léxica automática contribuyen a la mejora de las herramientas computacionales de apoyo a la investigación lingüística básica (no aplicada). En este sentido, la contribución puede ser directa o indirecta. Un ejemplo de contribución directa sería la búsqueda –pongamos por caso– de esquemas valenciales específicos de una acepción del lexema verbal, o bien de un grupo de acepciones relacionadas. Por ejemplo, si quisiéramos conocer con precisión las características morfo-sintácticas y léxico-semánticas de los entornos oracionales del verbo *saltar*, sabiendo que las características de cada grupo de contextos correlacionan con las acepciones de dicho verbo, deberíamos examinar las pertinentes concordancias de un corpus agrupadas en función de los sentidos de *saltar*. En un verbo de uso tan frecuente como éste, la discriminación manual de los sentidos en cada concordancia constituiría un proceso arduo y lento. Por otra parte, si prescindiéramos de la agrupación de concordancias por acepciones del verbo, la identificación de los esquemas valenciales sería una tarea extremadamente farragosa, ya que contextos de muy diversas características se hallarían entremezclados. La mejor solución sería, pues, contar con un programa de desambiguación léxica que permitiera clasificar las concordancias de forma automática según los sentidos del verbo.

Un ejemplo de contribución indirecta al perfeccionamiento de las herramientas informáticas para la investigación lingüística es la implementación de los etiquetadores gramaticales. El uso de corpus computacionales para determinados fines descriptivos ofrece mejores posibilidades con etiquetación gramatical que sin ella. Hoy en día, existen programas de etiquetación gramatical que alcanzan un grado de eficacia bastante alto y satisfactorio, pero no hay ninguno que logre el cien por cien de precisión en la asignación de categorías gramaticales. La información de tipo léxico-semántico a menudo se solapa con categorías gramaticales. Por ello, un programa de desambiguación léxica que no dependa de la anotación gramatical puede servir para complementar los propios etiquetadores gramaticales, aumentando la eficacia de los mismos. Es decir, los mecanismos disponibles para asignar una acepción a la aparición de un lexema podrían servir para afinar los mecanismos empleados para

la asignación de una categoría gramatical, por ejemplo, la etiquetación morfológica en términos de clase de palabras.

Pongamos por caso una investigación sobre el adjetivo *bienvenido*. La misma forma *bienvenido* puede funcionar no sólo como adjetivo, sino también como nombre propio[3], e incluso algunos modelos gramaticales podrían observar un uso con categoría de interjección. La acepción adjetival no es idéntica al uso de *bienvenido* en una interjección, y por supuesto no comparte significado con el uso de la misma forma como nombre propio. Por ello, la discriminación de sentidos se solapa con la asignación de categoría gramatical. De ahí que si desarrollamos un programa de desambiguación léxica que no dependa en principio de la información gramatical, los resultados de la desambiguación léxica pueden ser de utilidad para confirmar o, en su caso, corregir los resultados de la etiquetación gramatical. Incluso sería posible incorporar el proceso de desambiguación léxica como un modelo complementario integrado en el propio programa de etiquetación gramatical.

(5) Hasta un extraño es *bienvenido* a recibir su contagiosa alegría.

(6) El resto del elenco también logró excelentes interpretaciones; *Bienvenido* Martínez como el productor Eduardo, Laura Newman como la radioescucha madre de Mauricio, Christine King como la medio boba y coqueta secretaria, y como una radioescucha también, muy buena actuación en ambos personajes.

(7) Tenemos a su autor: Jusep Samperas, escritor. ¡*Bienvenido*! Es autor de este libro, junto a Antonio Ortí.

Finalmente, el tercer aspecto por el cual la desambiguación léxica es relevante repercute en la propia teoría lingüística que genera el modelo de asignación de acepciones. Estos programas requieren la formulación de algoritmos basados en conocimiento lingüístico-

[3] Obviamente, hay muchos casos en los que la grafía mayúscula o minúscula no servirá para determinar la categoría gramatical. Tales casos comprenden los contextos en posición inicial de oración, los errores ortográficos –por ejemplo, nombre propio escrito en minúscula– y, de manera más evidente, las transcripciones erráticas del lenguaje oral.

científico, por ejemplo, la identificación de factores contextuales (distancia entre colocados, fijación sintagmática, etc.) que tienen una influencia en la discriminación de acepciones. Por tanto, la obtención de resultados satisfactorios podría considerarse como una prueba de validez experimental del modelo de análisis, ya que significaría que el modelo en cuestión es formalizable, y que dicha formalización es conducente a buenos resultados.

Así pues, si conseguimos que el ordenador clasifique las concordancias de *bisagra* en cuatro grupos correspondientes a los cuatro sentidos mencionados arriba, ello significará que el algoritmo generado por el modelo de análisis está basado en un conocimiento lingüístico válido empíricamente. Del mismo modo, el cuestionamiento de los resultados habría de redundar en una revisión del modelo de análisis, ya que si somos incapaces de hacer que el ordenador clasifique automáticamente las concordancias según los sentidos de *bisagra*, cabría interpretar este fracaso como sintomático de una teoría inadecuada sobre la correlación entre acepciones y entornos textuales. En suma, el desarrollo de programas de desambiguación léxica automática puede verse como una especie de test para el correspondiente modelo lingüístico de análisis polisémico.

2. MECANISMOS DE DESAMBIGUACIÓN

Hasta la fecha, los principales métodos utilizados en la desambiguación léxica automática abarcan desde las implementaciones en el campo de la Inteligencia Artificial hasta el empleo de técnicas estadísticas basadas en datos de corpus, pasando por el recurso de las bases de datos computacionales, como los diccionarios electrónicos. Un buen resumen de los distintos métodos es el que proporcionan Ide y Véronis (1998).

En nuestro Proyecto, hemos optado por las técnicas basadas en el tratamiento cuantitativo de datos de corpus. La hipótesis es que, por lo general, toda la información necesaria para interpretar el sentido de una palabra se encuentra de un modo u otro codificada en la superficie textual. Además, en el caso de las acepciones convencionales o estereotipadas de una palabra –es decir, todas aquellas que no proceden del uso creativo de la lengua–, los datos

co-textuales necesarios para la asignación de sentido son limitados y conmensurables. Por tanto, son formalizables, tipificables y aplicables a un programa informático. A continuación, comentaremos algunos de los parámetros que tenemos en cuenta en nuestro Proyecto de Investigación para el desarrollo de un prototipo de programa de desambiguación.

2.1. *Contexto sintagmático*

La información contenida en la superficie textual se expresa en estructuras de diverso tipo, y en la fase actual del Proyecto hemos preferido centrarnos sólo en una, al menos provisionalmente. Hasta el momento, hemos concentrado nuestros esfuerzos en el tratamiento de los *colocados léxicos*, por considerar que la optimización del tratamiento de este tipo de información plantea más dificultades teóricas y metodológicas que otros tipos de datos textuales y co-textuales.

En la tradición del Contextualismo Británico y la actual Lingüística de Corpus de corte neo-firthiano, el concepto de "colocado" se define como coocurrencia habitual en el entorno próximo de una palabra. Este concepto deja por tanto abierta a la discusión la cuestión acerca de los límites de lo "habitual" y de la proximidad en el texto. En efecto, dos de las cuestiones de estudio más controvertidas en la Lingüística de Corpus son las concernientes a las mediciones estadísticas de asociación léxica (*cf*. Stubbs 1995; Barnbrook 1996) y a la búsqueda de la ventana colocacional óptima (*cf*. Mason 2000).

A continuación, pondremos un ejemplo muy sencillo que ilustra la utilidad de los colocados léxicos como mecanismos de desambiguación léxica. A continuación, citamos todas las concordancias del Corpus *Cumbre* (40 millones) en las que *tabla* coocurre con *cálculo*[4] (ejemplos 8-22). Se observa que el colocado *cálculo* tiene una gran influencia sobre la selección de acepción en *tabla*. A tenor de lo que muestran los ejemplos de abajo, la colocación con *cálculo* excluye la práctica totalidad de las acepciones de *tabla*, excepto la que podría parafrasearse como "recuadro de texto y/o cifras". Es importante observar que esta correlación entre

[4] Adviértase que el gestor del Corpus *Cumbre* incluye en una sola concordancia todo el fragmento de texto que abarca de punto a punto.

colocado léxico y acepción de la palabra analizada no impone ninguna restricción aparente sobre aspectos sintagmáticos tales como la posición relativa de los colocados o la función sintáctica. Incluso la distancia que separa los dos colocados en el texto llega a ser considerable en algunos casos, y, sin embargo, el vínculo entre el colocado y la acepción parece mantenerse a pesar de ello.

(8) Es interesante el poder resaltar de ellos sus *cálculos* astronómicos sobre los eclipses contenidos en el Códice de Dresde, (Introducción a la edición del Códice de Dresde de la Biblioteca Anglosajona de Dresde, Helmut Deckert, traducción en Antropología Centramericana, Antología, David Luna Desola, EDUCA, 1977), así como las *tablas* de multiplicación para las conjunciones del planeta Venus y las correcciones para la órbita solar; también están las *tablas* de ascensos y ocasos de Venus durante un período de 312 años, y las de los eclipses de Sol y Luna.

(9) Microsoft Access 2000: Bases de datos - Las *tablas* - Las consultas - Relaciones - Hacer *cálculos* dentro de las consultas - Informes - Encabezados y pies - Informes con totales - Importar y exportar datos - Macros.

(10) El análisis: Como extraer mas información de sus planillas - *Tablas* dinámicas - Aplicar filtros - Obtención de sub *tablas* - Análisis con dos variables - *Cálculos* imposibles. Si pincha sobre el servicio que le interesa, dispondrá de una *tabla* para realizar el *cálculo* de honorarios.

(11) Pero en tanto uno aporta la ciencia práctica y el conocimiento directo, el otro las especulaciones teóricas, los *cálculos*, las *tablas* y los mapas.

(12) Para el *cálculo* de la extensión, un método simple que puede utilizarse es la "regla de los nueve de Wallace", mediante la cual se considera que las distintas regiones anatómicas corporales representan un 9% cada una o un múltiplo de 9% de la superficie corporal total (*tabla* 1).

(13) Se pueden registrar los métodos de *cálculo* y *tablas* de componentes en memoria fija, o transferirlos a un disco flexible de 1,44 MB. El apéndice A se encuentra conformado por *tablas* simplificadas donde se puede consultar el tamaño de la muestra requerido sin necesidad de efectuar los *cálculos* matemáticos.

(14) Posteriormente se presentan conceptos acerca del tamaño de la muestra, los principios y elementos que se necesitan para el *cálculo* de la misma, las fórmulas empleadas para los diferentes diseños, acompañado de algunos ejemplos y luego algunas *tablas* para consulta directa de la muestra requerida.

(15) Supongamos que X sale embarazada en el mes de abril del año 2001 (para este *cálculo* la cuenta es por semana, con una *tabla* donde aparecen las mismas).

(16) Los *cálculos* están resumidos en la *tabla* II.

(17) En el caso de las fuentes móviles tenemos que el *cálculo* de la concentración de monóxido de carbono en horas pico aplicando la fórmula 2 para diferentes tramos de algunas de las principales avenidas del Municipio arrojó los resultados que se muestran en la *tabla* No III.

(18) Permitieron obtener las alturas de la ola junto al muro de contención a lo largo de la costa, resumiéndose en la *Tabla* 2 el valor de este parámetro frente a cada uno de los puntos topográficos, empleados en el *cálculo* del rebase del oleaje.

(19) Los resultados de este *cálculo* se muestran en la *Tabla* 4.

(20) Además, tienen programas "gratis" para *cálculo* de casi todas las áreas de Ingeniería Mecánica, desarrollados por ingenieros mecánicos y afines (*tablas* de vapor, *cálculo* de deformaciones en elementos mecánicos, *cálculo* de esfuerzos de diferentes elementos y situaciones, etc).

(21) Se pueden crear informes que incorporen *cálculos* basados en los datos de las *tablas* para mostrar resultados totales o promedios o bien para generar catálogos.

(22) En ese momento se actualizará los *cálculos* de la *tabla* dinámica, mostrando las Unidades Equivalentes de la(s) planta(s) correspondiente(s) al análisis.

Según resultados preliminares, los colocados léxicos en tanto elementos discriminadores de acepción se muestran igualmente eficaces en la lengua inglesa. Así, la coocurrencia del sustantivo *sibling* con *father* tiende a excluir todas las acepciones de *father* menos una, a saber, la de "progenitor masculino". A continuación

(ejemplos 23-28), citamos todos los ejemplos extraídos del Corpus *LACELL*[5]. Como puede observarse, el vínculo entre la acepción y el colocado se preserva incluso en distancias textuales relativamente largas, como en (27), donde la ventana colocacional supera las 50 palabras.

(23) Nils's wife Christine suffered a debilitating illness and Maria had assumed the responsibility for looking after her younger *siblings* and her *father*.

(24) Focusing on the female X chromosome that men inherit from their mother (they also get a male Y from their *father*), the researchers found that two-thirds of the gay *siblings* shared a distinctive pattern along a segment of their X chromosome.

(25) My *father* had eight brothers and sisters and only three of them survived the famine in Russia at the turn of the century; my mother had fifteen *siblings* and only three survived.

(26) The son of a rotten *father*, Russ/O'Brian became a rotten *father* himself, cutting off all contact with his son, granddaughters, and even *siblings*.

(27) Utilizing original interviews, archival video, photographs, newspapers, and home movies, Curran introduces his *siblings* (brothers Desmond and Gavin, the undaunted striped bass fishermen and his sister Maeve, a fiercely competitive bodybuilder); his mother Mary Jane, a stoic Catholic who has relied on her faith to cope with tragedy; and himself as the insightful narrator, filmmaker Tom III, who questions and probes, striving to fulfill his *father's* legacy.

(28) So, although I am genetically related to my mother's brother's, or my mother's *siblings'* children or my *father's siblings'* children, the fact is, the *siblings* of both groups are genetically related to each other.

[5] Este corpus genérico del inglés fue compilado en la Universidad de Murcia, Dpto. de Filología Inglesa, por miembros del Grupo de Investigación LACELL. El corpus tiene un tamaño de 20 millones de palabras y una estructura comparable a la del Corpus *Cumbre*. El binomio formado por estos dos corpus constituye nuestra principal herramienta de estudios contrastivos inglés-español.

Otro ejemplo en inglés es el de *trial* y *conviction*, cuya coaparición tiende a seleccionar acepciones en una y otra palabra. La colocación de ambas parece excluir todos los sentidos de *trial* (p. ej. "experimento", "periodo de prueba laboral", etc.) e inducir el sentido "proceso judicial". En cuanto a *conviction*, su coaparición con *trial* tiende a excluir la acepción "convicción" y a primar el sentido "sentencia condenatoria". En los ejemplos 29-32 (también del Corpus *LACELL*), se observa cómo, al igual que en ejemplos anteriores, el vínculo entre el colocado y la acepción es resistente a multitud de variaciones sintagmáticas (distancia, funciones sintácticas, posición relativa).

(29) Equally, common sense demands that the operated transsexual should not be able to avoid prosecution and *conviction* for soliciting or importuning, as the case may be, by suddenly adopting for the duration of the *trial* the prior and now abandoned sex.

(30) Paul Coggins, a Dallas criminal lawyer who represents Mr McBirney, complains that the *trials* can be compared to drug-dealing *trials* in the 1960s when winning a *conviction* was as easy as dropping the dope on the courtroom table.

(31) The investigation, raid, hearing, *trial* and *conviction* all took place swiftly, between February and November 1959.

(32) When it does, it tends to be a reaction to perceived injustice, such as internment without *trial*, or the *conviction* of a son by a sole judge in a *trial* held in total secrecy and on the evidence of unseen witnesses, or a simple case of one's house being badly mauled by careless soldiers searching for arms.

Además de los colocados léxicos, la información de tipo fraseológico y semi-fraseológico es evidentemente muy útil para los programas de desambiguación léxica. Volviendo al ejemplo de *tabla*, siempre que este sustantivo aparece como miembro de la locución *tener tablas*, el sentido de *tabla* está fijado: "soltura, experiencia en una actividad". Igualmente, colocaciones semi-fraseológicas como *tabla de multiplicar* o *tabla de planchar* restringen la interpretación de *tabla*. La extracción automática de construcciones de este tipo –es decir, con cierta fijación fraseológica– debe complementar la información proveniente de los colocados léxicos en el sentido más

laxo del término. En cualquier caso, los colocados en sentido laxo proporcionan una información muy valiosa para desambiguar la palabra cuando esta no se utiliza como miembro de una construcción (semi-)fraseológica.

2.2. *Frecuencia de las acepciones*

Además de las colocaciones, la frecuencia de uso de cada acepción es un parámetro relevante para poder precisar la predicción automática del sentido. Pongamos por caso la palabra *abuela*. Miembros del equipo de investigación de nuestro Proyecto clasificaron las concordancias de *abuela* en la versión de 20 millones de palabras del Corpus *Cumbre*. El procedimiento fue semi-automático. Se diseñó un programa informático para que cada concordancia abriera una pestaña con los cinco usos que el GDUEsA (Sánchez 2001) registra en la entrada *abuela*, de los cuales había que seleccionar manualmente una acepción por concordancia. Este proceso hizo posible no sólo la computación de colocados distribuidos por acepciones, sino también la determinación de las proporciones de uso de cada sentido. En este último aspecto, los resultados fueron contundentes. La acepción "madre del padre o de la madre de una persona" copa el 92% de las ocurrencias de *abuela* en el corpus.

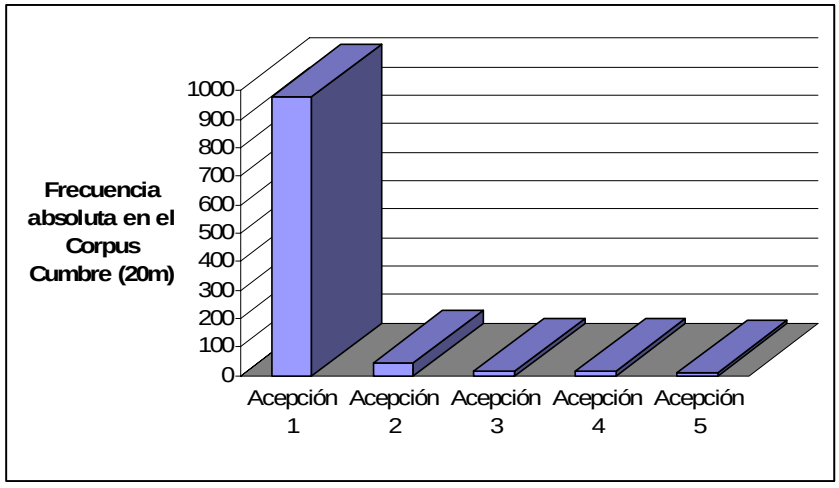

Tabla 1. *Frecuencia de las acepciones de "abuela".*

En la medida en que el corpus sea representativo del uso de la lengua española, los datos indican que la inmensa mayoría de los usos de *abuela* expresan dicha acepción. Esto significa que, antes de analizar la información colocacional, sabemos de entrada que, en principio, hay alrededor de un 92% de probabilidades de que el uso de *abuela* active la acepción "madre del padre o de la madre". Así pues, en caso de no hallarse colocados léxicos lo suficientemente discriminatorios del significado, y de que la palabra en cuestión no aparezca dentro de ninguna construcción fraseológica tipificada, la información sobre la frecuencia de uso de las acepciones inclinaría la balanza del lado de la primera acepción de *abuela*. En general, hay más posibilidades de acertar en la desambiguación de palabras con una distribución muy desigual de la frecuencia por acepciones. Por ello, es relevante tener en cuenta este parámetro.

2.3. *Tipología textual*

Lógicamente, determinados tipos de texto pueden inducir unas acepciones y reprimir la activación de otras. Por ejemplo, en un manual para el usuario de *hardware* informático, las probabilidades de que *ratón* active el sentido "dispositivo para mover el cursor sobre una pantalla" serán notablemente mayores que en un libro de texto de ciencias naturales. Por casos como éste, concluimos que la información acerca del tipo textual puede contribuir a predecir la acepción. Los métodos para identificar automáticamente el tipo de texto son potencialmente integrables en un prototipo de desambiguación léxica automática, y la información que aportan es significativa.

3. LIMITACIONES Y RETOS

La relación con el ámbito de la lexicografía plantea uno de los principales retos que deben afrontar los programas de desambiguación léxica automática en un futuro próximo. La mayoría de estos programas parten necesariamente de una lista de acepciones provista por una entrada de diccionario, electrónico o no, ya sea dicha entrada compilada para la ocasión u obtenida de uno o varios

diccionarios comercializados. Esto plantea un problema teórico-metodológico: no existen, a día de hoy, criterios formalizables para la discriminación de sentidos, ni parece que alguien pueda desarrollar tales criterios en un futuro cercano. Gran parte de esta tarea depende de procesos intuitivos. De ello se colige que el punto de partida de estos programas de desambiguación es de por sí dudoso, al estar asentado sobre unas bases tan inestables como la capacidad perceptiva de diferencias semánticas.

Además, de la dependencia del diccionario se podría desprender un problema práctico, ya que la compatibilidad entre el programa de desambiguación y el *input* de las entradas léxicas podría llegar a ser muy limitada. Hemos de tener en cuenta que la discriminación de acepciones en diccionarios distintos puede ser divergente, y de hecho lo es en numerosos casos. Si desarrollamos un programa basándonos en las entradas de un diccionario y lo evaluamos de acuerdo con el mismo modelo de polisemia, no podemos estar seguros de que el mismo programa dé resultados satisfactorios introduciendo las entradas léxicas de otros modelos polisémicos, por ejemplo de otros diccionarios. Se plantea, pues, la pregunta de hasta qué punto la validez de un programa de desambiguación automática está limitada sólo a entradas léxicas que siguen un determinado modelo de polisemia. Lógicamente, el diseño del algoritmo debe aspirar a alcanzar la máxima compatibilidad posible: será más práctico aquel programa que esté capacitado para funcionar recibiendo *inputs* desde entradas léxicas de diccionarios con características muy distintas en el tratamiento de la polisemia. En concreto, uno de los retos más difíciles es superar las discrepancias en torno a lo que se viene denominando la "granularidad". Un mismo grupo de concordancias puede abarcar una única acepción en un diccionario, pero varias acepciones en otro.

Por otra parte, hay pocas alternativas a la dependencia del diccionario. En los sistemas de desambiguación basados en datos de corpus, la única alternativa parece ser la extracción automática de grupos de colocaciones que recibirían una interpretación semántica a posteriori. Esta opción presentaría, en principio, dos ventajas. En primer lugar, se parte de criterios formalizables, ya que el algoritmo opera directamente sobre la agrupación y separación de formas lingüísticas, y no presupone el establecimiento intuitivo de significados o sentidos (estos constituirían más bien el punto de llegada). En segundo lugar, la compatibilidad con distintos modelos

de polisemia aumentaría. Distintas colocaciones (en el sentido laxo del término) podrían agruparse o separarse en distintas acepciones, en función de las necesidades específicas de cada tarea. Este planteamiento coincide con las recomendaciones de Kilgarriff (1997). Según este autor, conviene que los programas de desambiguación automática traten los listados de acepciones como perspectivas planteadas por la tarea, más que como representaciones de una realidad lingüística objetiva.

En cualquier caso, la extracción automática de agrupaciones de palabras en función de su relevancia para la polisemia es todavía un objetivo utópico y lejano. Para ello, necesitaríamos contar con técnicas capaces de relacionar el vínculo estadístico (de coocurrencia) con el vínculo semántico-designativo entre palabras. Merece la pena dedicar esfuerzos a la consecución de dicho objetivo. Mientras tanto, la desambiguación automática seguirá siendo esencialmente una prolongación de las técnicas lexicográficas de análisis semántico, en vez de moldear a éstas.

REFERENCIAS BIBLIOGRÁFICAS

BARNBROOK, G. (1996): *Language and Computers. A Practical Introduction to the Computer Analysis of Language*, Edinburgh: Edinburgh University Press.

CANTOS, P. y SÁNCHEZ, A. (2001): "Lexical constellations: What collocates fail to tell", *International Journal of Corpus Linguistics*, 6(2), 199-228.

IDE, N. y VÉRONIS, J. (1998): "Word sense disambiguation: The state of the art", *Computational Linguistics*, 24(1), 1-41.

JONES, S. y SINCLAIR, J. (1974): "English lexical collocations. A study in computational linguistics", *Cahiers de lexicologie*, 24, 15-61.

KILGARRIFF, A. (1997): "'I don't believe in word senses'", *Computers and the Humanities*, 31(2), 91-113.

LAVID, J. (2005): *Lengua y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid: Cátedra.

MASON, O. (2000): "Parameters of collocation: The word in the centre of gravity", en J. M. Kirk (ed.), *Corpora Galore. Analyses and Techniques in Describing English*, Amsterdam/Atlanta, Georgia: Rodopi, 267-280.

RAVIN, Y. y LEACOCK, C. (eds.) (2000): *Polysemy. Theoretical and Computational Approaches*, Oxford: Oxford University Press.

SÁNCHEZ PÉREZ, A. (1995): "Organización del Corpus *Cumbre*", en A. Sánchez Pérez *et al*. (eds.), *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*, Madrid: SGEL, 25-37.

SÁNCHEZ PÉREZ, A. (ed.) (2001): *Gran Diccionario de Uso del Español Actual*, Madrid: SGEL.

SÁNCHEZ PÉREZ, A. y ALMELA SÁNCHEZ, M. (2004): "Polysemy and sense discrimination in Lexicography", en J. M. Bravo (ed.), *A New Spectrum of Translation Studies*, Valladolid: Universidad de Valladolid, 141-174

SINCLAIR, J. (1991): *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

STUBBS, M. (1995): "Collocations and semantic profiles. On the cause of trouble with quantitative studies", *Functions of Language*, 2(1), 23-55.