INDICE

TRADUNED Y LA DESAMBIGUACIÓN LÉXICA EN EL MARCO DE LA TRADUCCIÓN AUTOMÁTICA

José Carlos Periñán Pascual

Universidad Jaume I

Within the framework of computational lexicography, most linguistic engineers conclude that the word sense division found in every-day dictionaries is not suitable for the task of word sense disambiguation (WSD). However, TRADUNED is a prototype of knowledge-driven word sense disambiguator which relies on the idea that all the information necessary for the discrimination of either homonymous senses or polysemous ones in an input text can be found in the *Collins Cobuild English Language Dictionary* entries. Being a discourse-based system, TRADUNED automatically extracts the implicit information stated in definition texts for its application in the syntagmatic and paradigmatic levels of analysis. Moreover, our system does not only assign each word to the appropriate sense but it also describes the different decisions being made through the WSD process.

1. INTRODUCCIÓN

En el campo de la traducción automática, el principal objetivo del programador y/o lingüista es proporcionar a la máquina todo el conocimiento necesario que le permita descubrir el significado de un texto. Con el propósito de incorporar el conocimiento lingüístico a una aplicación informática tradicionalmente se ha dotado a la máquina con una gramática y un lexicón. El verdadero problema radica realmente en la integración de ambos subcomponentes de forma eficiente.

En la actualidad muchas teorías gramaticales adoptan una perspectiva lexicista, p.ej. la Teoría de Rección y Ligamiento (Chomsky 1981), la Gramática Funcional de S.C. Dik (1978), la Gramática Léxico-funcional de Kaplan y Bresnan (1982) o la Gramática del Papel y la Referencia de Foley y Van Valin (1980):

> Todos ellos concentran en el componente lexicón fenómenos que hace unos años se consideraba que formaban parte del componente sintáctico y, además, conciben el léxico como el depósito de las propiedades sintácticas de los predicados, desarrollando mecanismos de enlace que recogen la conexión entre la información almacenada en el léxico y la

ATLANTIS
Vol. XXIII Núm. 1 (Junio 2001): 135-71.
ISSN 0210-6124

> realización sintáctica de cada uno de los rasgos que conforman una entrada léxica. (Mairal 1999: 42)

Dentro de este marco general, TRADUNED se presenta como un prototipo experimental de analizador orientado al lexicón capaz de solventar de forma totalmente automática la ambigüedad léxica de las palabras lexemáticas de tres textos de entrada (*input texts*). Por medio del lexicón, componente principal de nuestro sistema, la aplicación identifica las relaciones sintagmáticas y paradigmáticas que se establecen entre las unidades léxicas del aducto, utilizando dicha información en el proceso de desambiguación léxica.

TRADUNED se basa en la idea de que toda la información que necesita un sistema de traducción automática en el proceso de desambiguación léxica de un texto de entrada puede ser extraída de las entradas léxicas del *Collins Cobuild English Language Dictionary* (1987) (de aquí en adelante, CCED), y más concretamente de sus textos definitorios. Para demostrar esta hipótesis es necesario diseñar una serie de algoritmos que permitan a la máquina explotar al máximo los textos definitorios del CCED. No obstante, no sólo estamos interesados en los resultados del proceso de desambiguación, sino además en la descripción del propio proceso. En otras palabras, nuestra aplicación permite describir las diversas decisiones que va tomando la máquina en la resolución de la ambigüedad léxica. Ésta es la razón por la cual nuestro sistema presenta en pantalla las fichas de desambiguación léxica.

Tras una descripción del problema de la ambigüedad léxica en el marco de la traducción automática (apartados 2 y 3), pasaremos a detallar la construcción del lexicón de TRADUNED (apartado 4), el proceso de desambiguación léxica tal y como ocurre en nuestra aplicación (apartado 5) y los resultados que se han obtenido (apartado 6).

2. LA AMBIGÜEDAD LÉXICA EN LA TRADUCCIÓN AUTOMÁTICA

La ambigüedad léxica[1] tiene lugar cuando un lexema está asociado a más de un sentido, siendo la homonimia y la polisemia sus fuentes más habituales.[2] Por otra parte, a partir de la tipología de Weinreich, Pustejovsky (1995: 27-28) clasifica la ambigüedad léxica en ambigüedad contrastiva (i.e. homonimia) y polisemia complementaria (i.e. polisemia), distinguiendo en esta última el subgrupo de la polisemia lógica (i.e. polisemia intracategorial). La ambigüedad léxica es muy frecuente en el lenguaje natural, aunque pasa desapercibida porque se trata de un proceso que ocurre sin acceso consciente. A pesar de que la polivalencia semántica de un lexema fomenta la economía del lenguaje natural, esa misma polivalencia presenta un grave problema en la disciplina de la traducción automática, ya que el

[1] Russell y Norvig (1996: 718-19) hacen hincapié en la importancia de distinguir entre ambigüedad léxica y ambigüedad semántica. Esta última tiene lugar como producto de la ambigüedad léxica y sintáctica, produciéndose un error en la interpretación. Por ejemplo, la expresión *coast road* se refiere tanto a un camino que bordea la costa como a un camino cuyo destino es la costa.

[2] Aunque Raskin (1987: 46) identifica los homógrafos como una fuente adicional, nosotros los consideramos, junto con los homófonos, una subclase de los homónimos (Lázaro Carreter 1984: 226).

ordenador debe recorrer diligentemente todos los caminos posibles que se vayan trazando durante el proceso de desambiguación. En la construcción de un lexicón automático, no es pertinente realizar la diferenciación entre polisemia y homonimia porque sólo nos interesa el estudio sincrónico del significado.[3] El hecho de que dos sentidos estén relacionados históricamente o que se trate realmente de un accidente ortográfico no tiene en absoluto repercusiones gramaticales. Por consiguiente, el único problema que debemos tratar a este respecto es que un mismo significante tenga asociados en el lexicón varios significados, independientemente de que se trate de un caso de polisemia u homonimia.

Al igual que el hablante humano utiliza el contexto[4] para resolver esta ambigüedad léxica, debemos utilizar una estrategia que permita a la máquina obtener la mayor cantidad de información posible a partir del contexto local y global.[5] A nivel computacional, es precisamente en la fase del análisis del aducto donde tiene lugar la desambiguación léxica, la cual se fundamenta en la cohesión sintagmática y paradigmática, utilizando así tanto el contexto intraoracional como interoracional (Laffling 1991: 2). Un análisis aislado del contexto local, el cual ha sido aplicado con mayor o menor profundidad en innumerables sistemas operativos de traducción automática, como p.ej. TAUM-METEO, METAL, etc. (Little 1990; Tucker 1987: 30-33), no es con frecuencia suficiente para proporcionar al sistema la información necesaria que le permita resolver con éxito la ambigüedad léxica intracategorial,[6] principalmente si se trata de una palabra polisémica cuyos sentidos son extremadamente parecidos:

> Selection restrictions work best in discriminating homonymous senses of words ... where the meanings are unrelated and therefore distinct, rather than polysemous senses, where the meanings are related and thus often

[3] Escobedo Rodríguez (1994: 41) llega incluso a afirmar que la homonimia es un caso de polisemia sincrónica.

[4] A lo largo de este artículo emplearemos el término "contexto" en el sentido de "contexto intralingüístico" o "co-texto": i.e. aquellas palabras que acompañan a una determinada palabra en el texto (Albaladejo Mayordomo y García Berrio 1983: 219).

[5] De especial interés es mencionar los experimentos psicolingüísticos realizados por Estévez y De Vega (1999), quienes estudian el proceso de resolución del significado de las palabras ambiguas que aparecen en contextos lingüísticos extensos. Este estudio supone una novedad frente a la mayoría de los estudios psicolingüísticos sobre el procesamiento de palabras ambiguas, los cuales se basan en la presentación de homógrafos aislados o bien incluidos en contextos muy breves (i.e. una oración). Los resultados de este estudio muestran que la resolución de la ambigüedad léxica no debe centrarse únicamente en el contexto local, corroborando así la existencia de un trazo temático o discursivo a lo largo del texto, que se concibe como una representación del mundo más que como una representación del propio texto. Estos autores afirman además que es necesario que el lector sea capaz de establecer relaciones de coherencia global, por lo que se alejan del enfoque minimalista, el cual considera que el lector se ocupa primero de procesar la coherencia local del texto y sólo excepcionalmente la coherencia global cuando existe una ruptura de la coherencia local.

[6] La ambigüedad intercategorial es mucho más fácil de solucionar si nuestro sistema va provisto de un subcomponente que posibilite la asignación de las categorías gramaticales de las palabras que configuran el aducto. En TRADUNED, este analizador se implementa a nivel de diseño en el procedimiento *ActivarParser*, aplicado a los textos definitorios del lexicón, y en *Etiquetar1*, *Etiquetar2* y *ReglasContext1* para el texto introducido por el usuario. Véase el apéndice 1 para una descripción detallada de la arquitectura semántica de esta aplicación.

> overlap. Hence, it is no coincidence that the verbs and nouns selected for many studies of lexical ambiguity resolution in linguistics and NLP have either homonymous senses or distinct polysemous ones. (Fass 1993: 268)

En los textos utilizados por TRADUNED para su desambiguación léxica se presentan casos de polisemia que sólo pueden ser resueltos con un análisis paradigmático.[7] Esto implicará el diseño de un componente semántico fuerte que permita capturar las relaciones conceptuales que se formen entre las palabras del aducto.

3. TRADUNED Y LOS MÉTODOS DE DESAMBIGÜEDAD LÉXICA

En este apartado, describiremos aquellos modelos dentro del marco de la ingeniería lingüística que han influido en cierta medida en el diseño de TRADUNED, además de las principales aportaciones de nuestra aplicación en la resolución de la ambigüedad léxica.

En los años 60 y 70 los psicolingüistas destacaron el importante papel que desempeña en la desambiguación humana el *priming* semántico, o proceso por el cual la introducción de un determinado concepto facilita el procesamiento de conceptos semánticamente relacionados que serán introducidos posteriormente. Esta idea se implementó a través del modelo de las redes semánticas: es decir, los conceptos están organizados en la memoria en forma de una red conceptual formada por nodos conectados por distintios tipos de vínculos asociativos, donde cada nodo representa un sentido de un lexema. Los conceptos en una red semántica son activados por el uso, extendiéndose dicha activación a los nodos vecinos. Las redes semánticas se diseñaron en principio para demostrar cómo es posible construir la estructura semántica de la mente humana y procesar algorítmicamente dicha macroestructura por medio de un ordenador; por lo tanto, se puede considerar como un modelo psicológico y computacional de la memoria humana. La red semántica del psicolingüista M. Ross Quillian (1968) es una de las primeras implementaciones de una red de activación utilizada para la desambiguación léxica. Quillian creó esta red a partir de las definiciones del diccionario, pero la mejoró manualmente con su conocimiento como usuario de la lengua.

A este respecto, TRADUNED puede crear una red de motivación conceptual que interconecte los distintos sentidos del lexicón partiendo únicamente de sus definiciones lexicográficas. En el proceso de desambiguación léxica nuestra aplicación construye de forma totalmente automática una macrorred virtual de relaciones asociativas a partir de las listas de densidad semántica (de aquí en adelante, LDS) de las palabras que configuran el aducto.[8]

---

[7] Véase el apéndice 2.

[8] Una LDS es una lista de los diversos sentidos asociados conceptualmente a un determinado sentido de una palabra; dicha lista se construye automáticamente a partir de los textos definitorios del CCED. En el apartado 4.3 detallaremos la fundamentación lingüística y la implementación computacional de estas LDS.

La Semántica de Preferencias de Yorick Wilks (1975) es uno de los primeros enfoques específicamente diseñados para tratar el problema de la ambigüedad léxica. La Semántica de Preferencias resalta la idea de que los predicados verbales tienen preferencia por determinadas clases semánticas de argumentos, y los modificadores tienen preferencia por determinadas clases semánticas de núcleo, pero siempre teniendo presente que cualquiera de estas preferencias puede ser violada, por lo que no se imponen como restricciones de selección. Cada uno de los sentidos de las palabras registradas en el lexicón se representa en forma de una estructura arbórea binaria formada por primitivos semánticos, la cual se considera una expresión formal de la definición que nos podemos encontrar en un diccionario monolingüe convencional. El primer paso en el proceso de desambiguación léxica consiste en construir una plantilla semántica formada únicamente por los núcleos de las fórmulas semánticas de las palabras que configuran el texto de entrada. A continuación, se busca esta plantilla en la lista de plantillas prediseñadas del sistema. Si tras este proceso de reconocimiento, sólo se encuentra una coincidencia, entonces el proceso de desambiguación se da por concluido. En caso contrario, cada plantilla se enriquece con la expansión de las preferencias de selección por medio de los primitivos semánticos. La plantilla que satisface más preferencias selectivas es la que se asigna como interpretación semántica del texto de entrada.

La influencia de este modelo computacional en TRADUNED se puede observar en las relaciones sintagmáticas que se establecen en el aducto a partir de las preferencias de colocación de los nombres y adjetivos y las preferencias de los argumentos del marco predicativo verbal. Además, la interpretación semántica del aducto se determina por medio de un sistema de puntuación basado en el número de coincidencias de preferencias selectivas que se establecen al relacionar los textos definitorios y el aducto. No obstante, nuestro sistema se distancia del presente enfoque en ciertos aspectos que hemos intentado mejorar. El más evidente de todos implica la laboriosa elaboración manual de las fórmulas semánticas de cada uno de los sentidos del lexicón en el modelo de Wilks; en nuestro sistema el proceso de identificación de las preferencias sintáctico-semánticas de cada lexema está totalmente automatizado. Por otra parte, la utilización de plantillas semánticas para la desambiguación léxica exige la implementación de un analizador local, el cual no podrá hacer uso de la información textual en el proceso de desambiguación. En cambio, nuestro sistema aprovecha el contexto local y textual para dicho proceso, consiguiendo así resultados óptimos.

El método de desambiguación léxica utilizado en el sistema KT (Dahlgren, McDowell y Stabler 1989) es un enfoque "local combinado" por dos motivos. Por una parte, el método es básicamente local, pero cuando las fuentes de información intraoracional no son suficientes, entonces se tienen en cuenta otras oraciones en el texto de entrada. Por otra parte, el algoritmo también es combinado porque emplea hasta un máximo de tres tipos diferentes de información: las locuciones idiomáticas, la información sintáctica (principalmente, restricciones de selección) y el razonamiento del sentido común o "semántica ingenua". La semántica ingenua especifica un nivel de conocimiento que es general y común para muchos hablantes de un lenguaje natural, siendo la base de la explicación de la interpretación textual.

Su objetivo no es encontrar un conjunto mínimo de primitivos necesarios para distinguir los conceptos entre sí, sino representar por medio de otras palabras el concepto cognitivo asociado a cada palabra a partir de estudios psicolingüísticos de los conceptos. El equipo de Dahlgren descubrió que la mitad de la desambiguación se lleva a cabo gracias a las locuciones idiomáticas y la información sintáctica, mientras que la otra mitad se resuelve utilizando el razonamiento del sentido común.

TRADUNED realiza tanto un análisis sintagmático como paradigmático de un texto de entrada, si bien asigna un mayor peso específico a cada coincidencia que ocurra durante el análisis sintagmático.[9] No obstante, debemos tener presente que un análisis aislado del contexto local no aporta a menudo la información necesaria que permita resolver la ambigüedad léxica intracategorial de palabras polisémicas cuyos sentidos son extremadamente parecidos, razón por la cual abogamos por una complementariedad de ambos planos. También compartimos con el modelo de Dahlgren la posibilidad de ubicar los lexemas en una dimensión cognitiva teniendo en cuenta las palabras que se asocian al sentido de ese lexema. Mientras el enfoque de Dahlgren identifica esas palabras intermedias por medios psicolingüísticos, nuestra aplicación lo hará automáticamente a través de las definiciones lexicográficas.

La elaboración manual de las fuentes de conocimiento de los sistemas del procesamiento del lenguaje natural (de aquí en adelante, PLN) suele dar lugar a implementaciones con un lexicón *ad hoc* que sólo cubre un pequeñísima fracción del vocabulario de una lengua. Aunque teóricamente correctas, el interés práctico que pueden provocar estas aplicaciones es inexistente. No obstante, a mediados de los años 80 el trabajo en torno a la desambiguación léxica alcanzó un punto decisivo cuando recursos léxicos a gran escala, como los diccionarios, los tesauros y los córpora, empezaron a ser bastante accesibles. El interés computacional en los diccionarios surgió en el mismo momento en que algunos editores empezaron a elaborar formatos lexicográficos más formalizados en cuanto a la sintaxis y el vocabulario utilizados en los textos definitorios de las palabras. Diccionarios como *Longman Dictionary of Contemporary English, Oxford Advanced Learner's Dictionary of Contemporary English* o *Collins' COBUILD Dictionary of the English Language* parecen casi haber sido diseñados para la investigación en lingüística computacional, a pesar del hecho de que sus formatos radicalmente nuevos fueron diseñados para los estudiantes de inglés. Casi todos coinciden en que los diccionarios contienen una gran cantidad de conocimiento sobre el mundo. La cuestión es si la información que puede ser extraída de los diccionarios es suficiente para el PLN. Por ejemplo, es difícil extraer automáticamente relaciones tan simples como la hiperonimia, debido en gran parte a las inconsistencias que muestran los diccionarios elaborados manualmente, además de los errores de omisión y comisión (Fontenelle 1992: 91). De hecho, la única base de conocimiento léxico disponible a gran escala (*WordNet*[10]) se creó manualmente. La explicación que suelen dar algunos

[9] El plano sintagmático también recibe un tratamiento preferente en aquellos casos en que la mayor puntuación final de una palabra ambigua es compartida por dos o más sentidos; la máquina elige finalmente aquel sentido que posee más puntos en su análisis sintagmático.

[10] *WordNet* (Miller 1995; Fellbaum 1998) es una gran base de datos léxica *on-line* que permite construir

investigadores reticentes a la transducción de un diccionario convencional en un lexicón tratable por la máquina es que los diccionarios se crean para el uso humano y no para ser explotados por una máquina. En este sentido, Kilgariff (1997: 100) expone diversas razones por las que el mismo conjunto de sentidos que se presenta en una obra lexicográfica no puede ser relevante para una aplicación del PLN, entre las cuales destaca el subjetivismo del propio lexicógrafo al seguir sus propias intuiciones en la discriminación de los sentidos de una palabra.[11] Incluso los propios lexicógrafos son conscientes de la ausencia de acuerdo en definir los sentidos y en las divisiones de los sentidos. No obstante, la mayoría de los investigadores en desambiguación léxica se apoyan actualmente en las distinciones de sentido proporcionadas por los diccionarios por resultar fácilmente disponibles.

Lesk (1986) fue uno de los primeros que sugirió la utilización de un diccionario legible por la máquina para la resolución de la ambigüedad léxica de forma totalmente automática. Lesk creó una base de conocimiento que asociaba a cada sentido del *Longman Dictionary of Contemporary English* una lista de las palabras que aparecían en la definición de ese sentido. Se parte del principio de que existe una gran probabilidad de que los sentidos de las palabras que ocurren en una misma oración estén relacionados semánticamente. Esto es debido a que la mayoría de las oraciones se comportan como un todo coherente, donde cada término está implicado en alguna relación y cada par de términos se une por alguna cadena de relaciones. En este modelo, la desambiguación se realiza al seleccionar el sentido de la palabra cuya lista contenga el mayor número de solapamientos con las listas de las palabras vecinas en su contexto. El método de Lesk ha servido de base para posteriores trabajos de desambiguación con diccionarios legibles por la máquina, aunque intentando mejorar el modelo original en tres aspectos fundamentales: la resolución de la ambigüedad léxica de las palabras que configuran la definición de un término titular, un tratamiento adecuado de las formas flexionadas y la construcción de una *stop list* apropiada. Por ejemplo, Wilks et alii (1993) mejoraron el algoritmo de Lesk a través de un método que explora la coocurrencia de las palabras en el *Longman Dictionary of Contemporary English*. Los datos de coocurrencias pueden proporcionar automáticamente una medida del grado de relación semántica de las palabras. Este sistema se utiliza posteriormente con un método de vectores que relaciona cada palabra y su contexto. Utilizando funciones de relación podemos expandir los contextos y las entradas léxicas para incluir palabras relacionadas. De esta forma, resulta más fiable la técnica de buscar la máxima coincidencia.[12]

---

una red semántica con los lexemas de la lengua inglesa. *WordNet* es hoy en día el recurso más conocido y más utilizado para la desambiguación léxica en inglés, contribuyendo a ello el hecho de que se pueda conseguir gratuitamente a través de Internet.

[11] Este subjetivismo se reduce considerablemente al basar el trabajo lexicográfico en los cópora lingüísticos computerizados. A este respecto, el CCED marcó un hito en la lexicografía al hacer corresponder a cada distinción de sentido una distinción de contexto identificada a partir de un corpus actual.

[12] No obstante, esta técnica de desambiguación léxica podría clasificarse como un "método débil". Si se combinase con métodos que utilizasen la información sintáctica y etiquetasen las palabras con su correspondiente parte del discurso, podría resultar un enfoque bastante adecuado para el procesamiento de

A pesar de que se han solucionado la mayoría de las deficiencias que presentaba el modelo inicial de Lesk, no resulta en absoluto un método de desambiguación perfecto, principalmente por tratarse de una técnica basada en el contexto local: la entrada léxica de un determinado sentido y la oración en la que aparece un uso de ese sentido no tienen a menudo palabras en común. Es indudable la influencia que ha recibido TRADUNED del modelo de Lesk y otras aplicaciones basadas en éste, pero nuestro sistema se diferencia de éstos por ofrecer la posibilidad de capturar las relaciones conceptuales que se establecen a nivel textual entre las palabras del aducto, tomando evidentemente como fuente de conocimiento un diccionario convencional.

4. EL LEXICÓN DE TRADUNED

4.1. Las bases de datos

El componente léxico de TRADUNED está constituido por un lexicón principal y un diccionario auxiliar. Mientras el lexicón principal guarda las entradas léxicas de las palabras titulares que contiene el CCED, el diccionario auxiliar permite al usuario registrar nuevos términos.[13] Desde la perspectiva del diseño computacional, este componente léxico se implementa a través de las bases relacionales *lexicon.mdb* (lexicón principal) y *lexaux.mdb* (diccionario auxiliar), las cuales se podrían describir esencialmente como tablas de atributo-valor a gran escala (Schubert 1992: 71).[14] Por ejemplo, la base de datos *lexicon.mdb* está formada por una tabla padre (*Principal*) y cuatro tablas hijas (*Definición, Traducción, Red* y *Semántica*), todas ellas interrelacionadas por el campo clave de la referencia (*Ref*), como se muestra más abajo (fig. 1).

Cada uno de los campos de una tabla representa una relación atributo-valor. Mientras la tabla padre almacena la información más recurrente a lo largo del procesamiento, o sea, la realización morfológica del término titular (*Palabra*) y su categoría gramatical (*Categoría*), las tablas hijas aportan información sobre el texto definitorio (*Definición*), el equivalente de traducción (*Traducción*), la lista de densidad semántica (*Relación*) y, en caso de tratarse de un verbo, el tipo semántico al que pertenece (*Tipo*). Con referencia a la base de datos *lexaux.mdb*, el sistema sólo dispone de una única tabla en la que se guarda la realización morfológica del término titular, su equivalente de traducción y su categoría gramatical.[15]

---

las entradas del *Longman Dictionary of Contemporary English* o del lenguaje en general.

[13] La diferencia terminológica entre "lexicón" y "diccionario" se basaría en el hecho de que el primero puede explorar la competencia semántica del hablante (Leech 1974: 207), a pesar de que en nuestro caso no representamos formalmente el sentido de las palabras.

[14] Muchas teorías lingüísticas actuales, p.ej. la Gramática Léxico-funcional de Kaplan y Bresnan (1982) o la Gramática de la Estructura Sintagmática Generalizada de Gazdar, Klein, Pullum y Sag (1985), utilizan una descripción funcional del conocimiento en forma de pares atributo-valor.

[15] Esta versión de TRADUNED utiliza un procedimiento asistido en la fase de inserción de entradas léxicas nuevas: i.e. a través de una interfaz agradable para el usuario no experto en lingüística, la máquina nos permite "copiar" la información que contienen las entradas léxicas del CCED. Podemos fácilmente imaginar que si tuviéramos este diccionario en un formato tratable por la máquina, o incluso utilizáramos un programa convencional de OCR y escaneáramos el diccionario para obtenerlo en formato ASCII, nuestra aplicación podría llevar a cabo los mismos procesos constructivos de forma totalmente automática

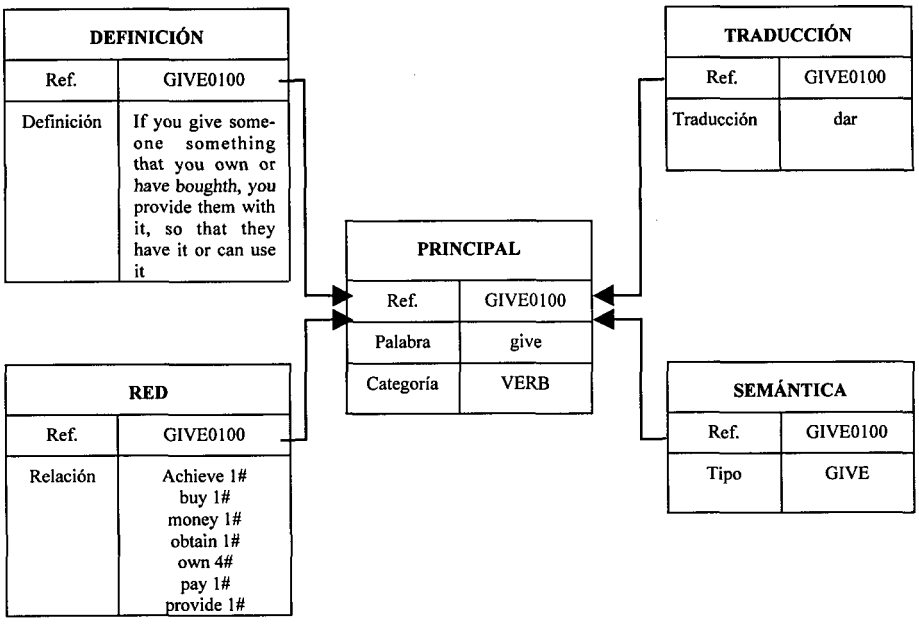


Figura 1

La razón de crear diversas tablas y no una base de datos monolítica se encuentra en el hecho de que las tablas extensas resultan poco eficientes para trabajar y son difíciles de modificar. Un lexicón legible por la máquina a gran escala requiere una enorme cantidad de memoria operativa, por lo que ésta tiene que gestionarse eficientemente:

> For reasons of system performance and memory management, the subdivision of lexical entries into several parts is even more important for a machine-readable lexicon than for the mental lexicon ... One essential demand on the implementation of a machine-readable lexicon is the most economical use of a machine's working memory. (Handke 1995: 237)

En relación con la rapidez en el procesamiento, existe una necesidad de eliminar la información redundante de las bases de datos. Si comparamos las tablas asociadas al sentido **give 1** (fig. 1) con la correspondiente entrada léxica que se presenta en la pantalla (fig. 2), observamos que no toda la información de la entrada léxica está explícitamente almacenada en la base de datos.

sin implicar grandes modificaciones.

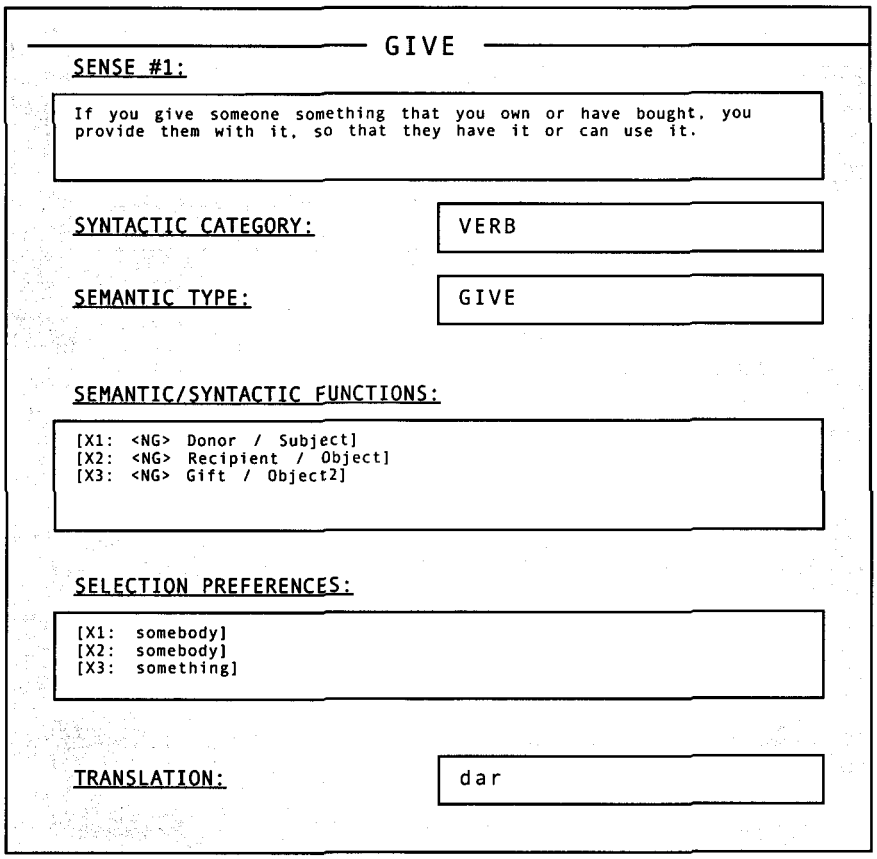


Figura 2

Flickinger (1992: 66), basándose en las propuestas de Ritchie (1987: 232-33), presenta tres objetivos que deben estar presentes durante el proceso de implementación de un diccionario en un sistema computacional:

> (i) the minimizing of redundancy, to satisfy theoretical commitments, enable sustained development of a system, and save space; (ii) the maximizing of efficiency in lexical access and in parsing; and (iii) the effectiveness of the semantic link to applications of the system.

Con el propósito de reducir el coste de almacenamiento, el diseño de nuestro lexicón se basó en la idea de que un sistema de traducción automática ha de ser capaz de obtener toda la información necesaria para el procesamiento del aducto a partir de unas bases de datos que permitan el uso de generalizaciones y mecanismos de extracción de la información implícita. Por un lado, los diccionarios no presentan

ningún tipo de generalización, por lo que resulta preciso utilizar una fuente auxiliar al componente léxico que permita, por ejemplo, deducir el comportamiento sintáctico de una palabra a partir de su contenido semántico. En nuestro caso, la gramática de enfoque semántico de R.M.W. Dixon (1991) permite al sistema identificar las funciones semánticas y sintácticas de los argumentos de un marco predicativo verbal construido a partir del texto definitorio del CCED. Por otra parte, los propios textos definitorios del CCED contienen mucha información implícita que es imprescindible para la resolución de la ambigüedad léxica: p.ej. el término *genus* de un nombre, las preferencias de colocación de los nombres y los adjetivos, o las preferencias selectivas de los argumentos del marco predicativo verbal. No será necesario almacenar toda esta información en la base de datos ya que el sistema dispone de ciertos mecanismos que le permiten obtener dicha información de forma totalmente automática durante el propio proceso de desambiguación o cuando el usuario consulta una entrada léxica.[16] En cambio, cuando este prototipo de desambiguador léxico se implemente en un proyecto operativo y funcional de traducción automática, p.ej. en un modelo de traducción indirecta basado en la transferencia, recomendaremos entonces, en aras a una reducción del tiempo de procesamiento, no guardar el texto definitorio y en su lugar almacenar toda esa información que es utilizada por el sistema durante el proceso de desambiguación léxica.

Nos gustaría destacar que una de las razones por la cual hemos optado por el CCED como fuente de información para la construcción de nuestro lexicón se encuentra en su propio método lexicográfico. A diferencia de otros diccionarios semasiológicos, el sentido de una palabra del CCED se define formalmente con la ecuación $\alpha = \beta$ donde α representa aquella parte de la definición donde aparece el *definiendum* en su estructura más típica, β representa el *definiens* y el operador de igualdad se formaliza principalmente en una coma para los textos definitorios verbales:

> It is this unique property of the Cobuild explanations which makes them more useful as a subject for automatic analysis than conventional definitions. (Barnbrook 1993: 315)

Esta estructuración del texto definitorio posibilita enormemente una rápida focalización de la información que se desea extraer. Por ejemplo, la máquina sólo utiliza la primera parte del texto definitorio como aducto para la construcción del marco predicativo, lo que implica una mayor rapidez en el procesamiento. Este

[16] Debemos hacer hincapié en el hecho de que la presente versión de TRADUNED es en definitiva un proyecto investigador y docente, razón por la cual la información sobre los términos *genus* del nombre, las preferencias de colocación de los nombres y los adjetivos, y el marco predicativo verbal que utiliza nuestra aplicación en el nivel sintagmático del análisis no aparece explícitamente registrada en su componente léxico, sino que se obtiene en tiempo real durante el propio proceso de desambiguación. Al permitir al usuario almacenar los textos definitorios completos del CCED en nuestro lexicón, por el principio de minimización de la redundancia informativa nos vemos obligados a no incluir en las bases de datos toda esa información que puede ser inferida a partir del texto definitorio de las entradas léxicas.

estrategia de analizar lo estrictamente necesario se aplica también en otros sistemas del PLN que utilizan diccionarios como fuente de conocimiento, como ocurre con PATR-II:

> The analysis process is intended to extract the most important information from definitions without necessarily having to produce a complete analysis of the whole of a particular definition text since attempting to produce complete analyses would be difficult for many LDOCE definition texts. (Alshawi *et al* 1985: 177)

4.2. La influencia de la Gramática Funcional de Dik

La Gramática Funcional de S.C. Dik (1978, 1989, 1997) ejerce igualmente una notable influencia en el lexicón de TRADUNED. Por una parte, tanto el modelo funcional de Dik ($Give_v$ $(x_1$:<anim>$)_{Ag}(x_2)_{Go}(x_3$:<anim>$)_{Rec}$) como TRADUNED (fig. 2) presentan un marco predicativo verbal que especifica la forma del predicado, su categoría sintáctica, las valencias cuantitativa y cualitativa, y las preferencias de selección de cada uno de sus argumentos.

El diseño de un modelo computacional basado en la teoría de la Gramática Funcional de S.C. Dik[17] no sólo tendría cierta similitud con TRADUNED por la construcción de un marco predicativo verbal, sino además por la naturaleza del mismo. Es decir, ambos adoptan un enfoque relacional en la descripción del significado, donde no se recurre a un metalenguaje, como la utilización de primitivos semánticos o predicados abstractos, sino a unidades léxicas de la propia lengua (Dik 1997: 83). No obstante, el marco predicativo de la presente versión de TRADUNED difiere del propuesto por Dik al incluir la realización morfológica y la función sintáctica de los argumentos. De esta forma, dirigimos nuestra aplicación a un usuario más general y no únicamente a un experto en lingüística. Además, nuestro patrón de subcategorización es más genérico, apareciendo con bastante frecuencia *somebody* o *something* como preferencias selectivas.[18]

Por otra parte, el proceso de identificación de los términos *genus* de un nombre está influido por el principio de la "descomposición léxica gradual" de la teoría de Dik. Este principio permite establecer un sistema de interrelación de las entradas léxicas en el cual el predicado del *definiens* de un postulado de significado puede convertirse en *definiendum* de otro postulado de significado (Dik 1989: 85). El principal problema radica en el hecho de encontrar el sentido que posee el término *genus* en un texto definitorio. En caso de que se trate de un término léxicamente polivalente, nuestro sistema activará un procedimiento para la resolución de la ambigüedad léxica basado en los mismos principios que guían la construcción de las LDS.

[17] La propuesta más ambiciosa la presenta el propio Dik en el proyecto *Functional Grammar Computational Model of the Natural Language User* (1990).

[18] La razón de esta generalidad se encuentra precisamente en el hecho de que en nuestro proceso de construcción del patrón de subcategorización no existe ningún tipo de interacción humana, interviniendo tan sólo la información presente en las definiciones del CCED.

4.3. Las listas de densidad semántica

4.3.1. La fundamentación lingüística

TRADUNED es capaz de construir automáticamente una lista de los diversos sentidos asociados conceptualmente a cada sentido de una palabra titular tomando como punto de partida el propio texto definitorio.[19] La cuestión que se plantea es cómo construir estos campos asociativos a partir de las definiciones del CCED. La respuesta la podemos encontrar en las siguientes palabras de Moon (1987: 87):

> A particular word is unlikely to be ambiguous at the moment of utterance, irrespective of how many different senses for it are recorded in a dictionary.... Context restricts interpretation and thereby resolves ambiguity. Meaning is the product of context.

Podemos afirmar que las relaciones semánticas que se establecen entre los lexemas de un campo léxico (i.e. sinonimia, hiperonimia, meronimia, etc.) proporcionan una guía útil para la desambiguación léxica:

> Lexical sets and fields may be used in disambiguation, and this criterion can be seen as halfway between the formal and the semantic, since it presupposes the existence of sets that are, arguably, signalled by coincidence of collocation. (1987: 97)

Como en el CCED el término titular de una entrada léxica aparece contextualizado en su propia definición, la LDS de cada uno de los sentidos del lexicón se formará a partir de las relaciones cohesivas que se establezcan entre el término titular y las palabras lexemáticas de su texto definitorio. La idea básica es que las relaciones cohesivas léxicas[20] que se establezcan a nivel lexicológico se proyecten sobre los textos definitorios de cada uno de los sentidos que intervengan en el aducto, formando todo ello una macroestructura semántica.

Uno de los principios básicos en la construcción de nuestras LDS es la "implicación semántica": dos palabras lexemáticas están léxicamente asociadas si una aparece en el texto definitorio de la otra (i.e. implicación explícita) o si ambas comparten al menos una palabra lexemática en sus textos definitorios (i.e. implicación putativa). Por ejemplo, *heat, hot* y *temperature* son palabras léxicamente asociadas:

---

[19] Desde una perspectiva psicolingüística, Belinchón et alii (1992: 372) proponen un modelo hipotético de lexicón mental en el que existiría una lista de términos o conceptos asociados a la entrada léxica por su significado.

[20] Las relaciones cohesivas fueron clasificadas por Halliday y Hasan (1976) en cohesión gramatical, con la subclasificación de conjunción, referencia, sustitución y elipsis, y cohesión léxica, siendo esta última la más importante de todas: "Lexical cohesion is the only type of cohesion that regularly forms multiple relationships (though occasionally reference does so too). If this is taken into account, lexical cohesion becomes the dominant mode of creating texture. In other words, the study of the greater part of cohesion is the study of lexis, and the study of cohesion in text is to a great degree the study of patterns of lexis in text" (Hoey 1991: 10).

heat
1 When you heat something, you raise its temperature, for example by using a flame or a special piece of equipment.

hot
1 Something that is hot has a high temperature.

Heat y *temperature* están léxicamente asociadas porque *temperature* aparece en la definición de *heat*, mientras que la explicación para el caso de *heat* y *hot* se encuentra en el hecho de que ambas palabras comparten el término "mediador" *temperature*. También podemos decir que *temperature* está léxicamente asociada tanto a *heat* como a *hot*.

La relación de implicación semántica se puede definir como una relación de equivalencia, es decir, una relación binaria reflexiva, simétrica y transitiva:

> Todo elemento está relacionado consigo mismo (propiedad reflexiva); si un elemento está relacionado con otro, éste también está relacionado con el primero (propiedad simétrica); por último, si un elemento está relacionado con otro, y éste con un tercero, el primero está relacionado con el tercero. (Garrido 1988: 33)

La propiedad transitiva de la relación de cohesión léxica se ilustra perfectamente en el diagrama siguiente, inspirado en Hoey (1991).

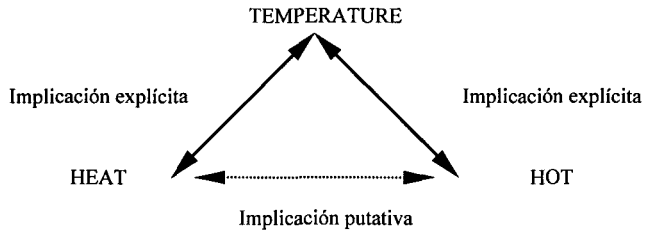


Figura 3

Aunque el principio de implicación semántica se acaba de definir como una relación binaria, es frecuente observar que, como ocurre en *garden-dig* o *sunshine-cloud*, las relaciones se proyecten sobre otras palabras del léxico gracias a su transitividad formando así una auténtica constelación de relaciones cohesivas:

GARDEN-DIG
garden
1 A garden is a piece of land next to someone's house where they grow flowers, vegetables, or other plants.

land
1 Land is an area of ground, especially one that is used for a particular purpose such as farming or building.

dig
1 When people or animals dig, they make a hole in the ground or in a pile of earth, stones, or debris.

SUNSHINE-CLOUD

sunshine
1 Sunshine is the light and heat that comes from the sun.

sun
1 The sun is the ball of fire in the sky that the Earth goes round, and that gives us heat and light.

cloud
1 A cloud is a mass of water vapour that floats in the sky.

De esta forma, no nos limitamos al texto definitorio de una palabra, sino que navegamos por el lexicón en busca de las definiciones asociadas a los conceptos que aparecen referenciados en el texto definitorio inicial. Por ejemplo, si explotamos algunas de las conexiones que se pueden establecer entre *hospital, doctor* e *ill* a partir del CCED, el grafo resultante, donde los nodos representan conceptos y las aristas establecen relaciones asociativas, tendría la forma de la figura 5:

hospital
1 A hospital is a place where people who are ill are looked after by nurses and doctors.

doctor
1 A doctor is someone who is qualified in medicine and treats people who are ill.

ill
1 Someone who is ill is suffering from a disease or a health problem.

medicine
1 Medicine is the treatment of illness and injuries by doctors and nurses.

disease
1 A disease is an illness which affects people, animals, or plants, for example one which is caused by bacteria or infection.

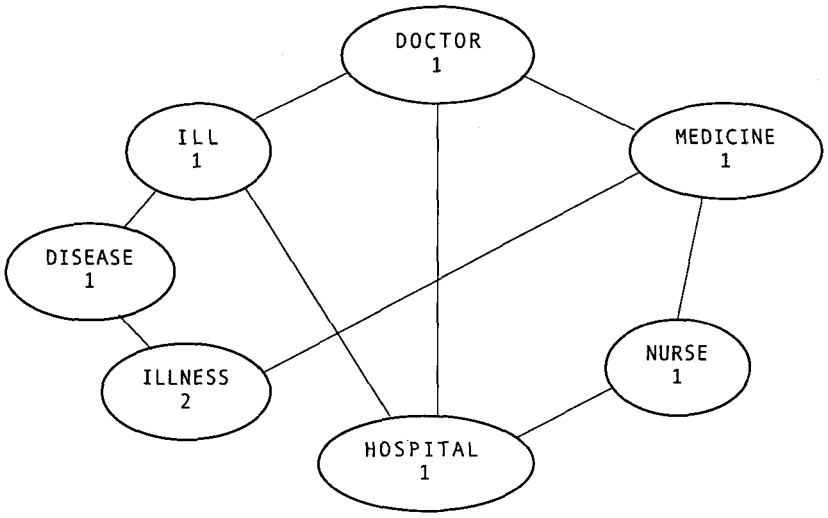


Figura 4

Sería conveniente destacar el hecho de que los miembros de las LDS no son unidades ortográficas, sino unidades semánticas: al igual que la mente humana, la computadora construye una red de asociaciones entre sentidos o conceptos,[21] razón por la cual cada nodo del grafo de la figura 4 hace referencia a un vocablo titular y un número de acepción.

4.3.2. La implementación computacional

Desde una perspectiva computacional, la LDS de un determinado sentido se construye tomando como punto de partida el texto definitorio correspondiente a esa acepción, el cual hemos denominado "definición nuclear". A modo de ilustración, vamos a suponer que un usuario desea conocer los sentidos que se encuentran léxicamente relacionados con **body 1** a partir de su texto definitorio:

> **body**
> **1** Your body is all your physical parts, including your head, arms and legs.

Tras etiquetar gramaticalmente las palabras que configuran la definición nuclear, se identificarán aquellas palabras lexemáticas que no resulten semánticamente relevantes para esa definición, es decir, aquellas que pertenezcan a la *stop list* de nuestro sistema:

> A *stop list* usually means taking out certain commonly occurring closed-class words on the assumption that, since they often appear in both the

[21] Al estar configurada una LDS por sentidos específicos de palabras, evitamos así el problema de la ambigüedad léxica de sus miembros.

> context and the sense definition, their appearance contributes nothing. (Wilks et alii 1996: 191)

La *stop list* de TRADUNED[22] contiene aquellas palabras lexemáticas que suelen aparecer con mucha frecuencia en las definiciones del CCED, pero cuya aportación a la descripción del sentido de la palabra titular resulta poco determinativa (p.ej. *be, example, happen, have, involve, mean, part, particular, piece, refer* o *way* entre otras muchas). Los miembros de esta lista no se ajustan a nuestro principio de "relevancia semántica", ya que no son candidatos válidos para el establecimiento de relaciones cohesivas por servir únicamente como armazón de la explicación lexicográfica. Por consiguiente, nuestro principio de la implicación semántica puede reformularse como sigue: dos palabras lexemáticas están léxicamente asociadas si una aparece en el texto definitorio de la otra o si ambas comparten al menos una palabra lexemática en sus textos definitorios, siempre y cuando ninguna de esas palabras pertenezca a la *stop list* del sistema.

Volviendo al ejemplo de **body 1**, mostramos de forma visual como sería el resultado de esta búsqueda, donde las palabras rechazadas por el sistema se marcan con una tachadura, mientras que las palabras aceptadas son doblemente subrayadas:

> **body**
> **1** Your ~~body~~ is all your physical ~~parts~~, including your head, arms and legs.

Las palabras *body* y *parts* son rechazadas por coincidir con la palabra titular y pertenecer a la *stop list* respectivamente. Por otra parte, *physical, head, arms* y *legs* logran pasar con éxito la criba.

El siguiente paso consistirá en buscar el texto definitorio de cada uno de los nombres y verbos identificados como miembros de la LDS para así poder identificar a su vez otras palabras semánticamente relevantes en estas definiciones. En el caso del adjetivo, sólo se aceptan como semánticamente relevantes aquellos que se encuentren en la definición nuclear, ya que un adjetivo que aparezca en otro nivel del análisis suele especificar una cualidad sólo atribuible al nombre al que califica y no al vocablo titular de la definición nuclear. Por este motivo, sólo se profundizará en el análisis de las definiciones nominales y verbales. En nuestro ejemplo, primero se buscará la definición de *head* para analizarla en busca de otras palabras semánticamente relevantes. El mismo proceso se repetirá con *arm* y finalmente con *leg*:

[22] Esta *stop list* se ha confeccionado partiendo de la información estadística que nos ofrece la lista de frecuencia de las formas léxicas utilizadas en los 1228 textos definitorios correspondientes a las 777 palabras titulares de nuestro lexicón. Cuanto más precisa sea la selección de los artículos léxicos que formen parte de la *stop list*, más fiable será nuestro sistema.

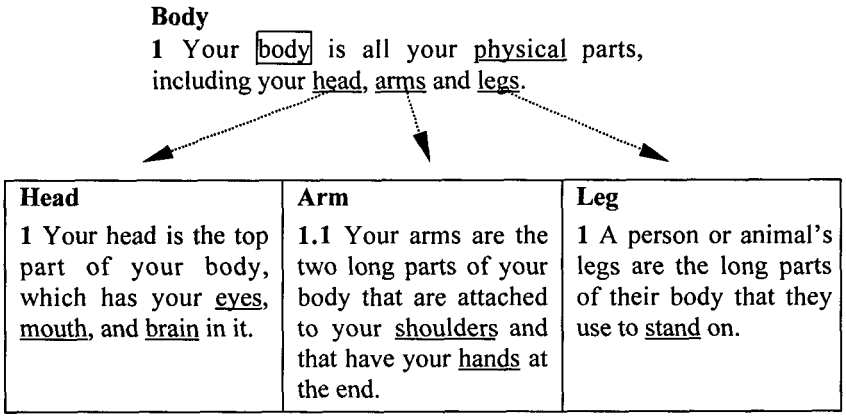


Figura 5

Uno de los problemas del proceso de identificación del sentido de una palabra dentro de una definición es precisamente la ambigüedad léxica. Para ello TRADUNED va provisto de un desambiguador léxico a nivel lexicográfico que se activa cuando una palabra lexemática de un texto definitorio está asociada a más de una entrada léxica.

A continuación comentaremos el resultado del proceso de identificación de las palabras semánticamente relevantes en cada uno de los textos definitorios correspondientes a los sentidos asignados a las palabras *head, arm* y *leg* de la definición nuclear:

head
1 Your ~~head~~ is the ~~top~~ ~~part~~ of your ~~body~~, which has your eyes, mouth, and brain in it.

arm
1.1 Your ~~arms~~ are the two ~~long~~ ~~parts~~ of your ~~body~~ that are ~~attached~~ to your shoulders and that ~~have~~ your hands at the ~~end~~.

leg
1 A ~~person~~ or ~~animal~~'s ~~legs~~ are the ~~long~~ ~~parts~~ of their ~~body~~ that they ~~use~~ to stand on.

En **head 1**, la palabra *head* es rechazada por tratarse del término titular de la definición, mientras *body* es rechazada por tratarse del término titular de la definición nuclear. No se aceptan *top* y *part* por encontrarse en nuestra *stop list*. En **arm 1.1**, *arms* coincide con la palabra titular de la definición, *body* con la palabra titular de la definición nuclear, y *parts, have* y *end* son miembros de la *stop list*. *Long* se descarta igualmente por tratarse de un adjetivo fuera de la definición nuclear y *attached* es rechazado al no ser identificado como vocablo titular de una entrada léxica del lexicón principal de TRADUNED. En **leg 1**, las palabras *person, animal,*

parts y *use* se encuentran en nuestra *stop list*, *legs* coincide con la palabra titular de la definición, *body* es la palabra titular de la definición nuclear y *long* es tratado de forma idéntica a la definición anterior. La LDS que TRADUNED ha construido de forma totalmente automática para **body 1** tendría finalmente la siguiente forma:

BODY 1: arm 1.1, brain 1, eye 1, hand 1, head 1, leg 1, mouth 1, physical 1, shoulder 1, stand 1

4.4. Una base de datos léxica enriquecida

Según Nirenburg y Levin (1992: 6), existen dos grandes tradiciones en el estudio de la semántica léxica que tuvieron importantes repercusiones en la lingüística computacional: por una parte, un enfoque que busca descubrir las propiedades semánticas de los artículos léxicos a partir de las cuales se pueda predecir el comportamiento sintáctico, y por otra parte, un enfoque que intenta establecer el significado de los textos con la ayuda de un modelo del mundo u "ontología" que explique las relaciones entre las entidades del mundo más que las relaciones entre las unidades léxicas. Mientras el enfoque sintáctico tomaría la forma de una base de datos léxica, el enfoque ontológico se implementaría en una base de conocimiento. Sin embargo, la utilización de las LDS en el proceso de desambiguación léxica nos proporciona un punto de vista alternativo.

Una base de conocimiento incorpora toda la información de una base de datos léxica, además de la instalación de una red de motivación conceptual que interconecte las distintas palabras del lexicón. La inferencia se entenderá por tanto como la navegación del sistema por una red semántica, por lo que será necesario construir un modelo del mundo sobre el que se proyecten las unidades léxicas del texto de entrada durante el proceso de análisis:

> The purpose of the lexicon is to facilitate the above linking. Of course, the lexicon contains not only mappings into the ontology. It also contains information about morphological, syntactic, pragmatic, and collocational properties of the lexical unit —since all and any of these components serve as clues at various stages of the process of mapping text into representation or vice versa. But the semantics of most of the open-class lexical items is described in terms of ther mappings into instances of ontological concepts. (Onyshkevych and Nirenburg 1992: 290)

Nirenburg y Levin (1992: 10) nos indican cómo construir ese modelo ontológico computacional:

> One way of constructing a model is to come up with a set of properties describing things in the world, define value sets for these properties and then describe each concept in the world as a set of particular property-value pairs.

Generalmente los conceptos se representarán por medio de primitivos semánticos, por lo que se plantea la cuestión de cómo expresar los matices de

significado con un grupo restringido de estos primitivos. Además nos encontramos con el doble problema de construir un modelo teórico del mundo a gran escala y su posterior implementación computacional, lo que requerirá un trabajo humano muy laborioso.

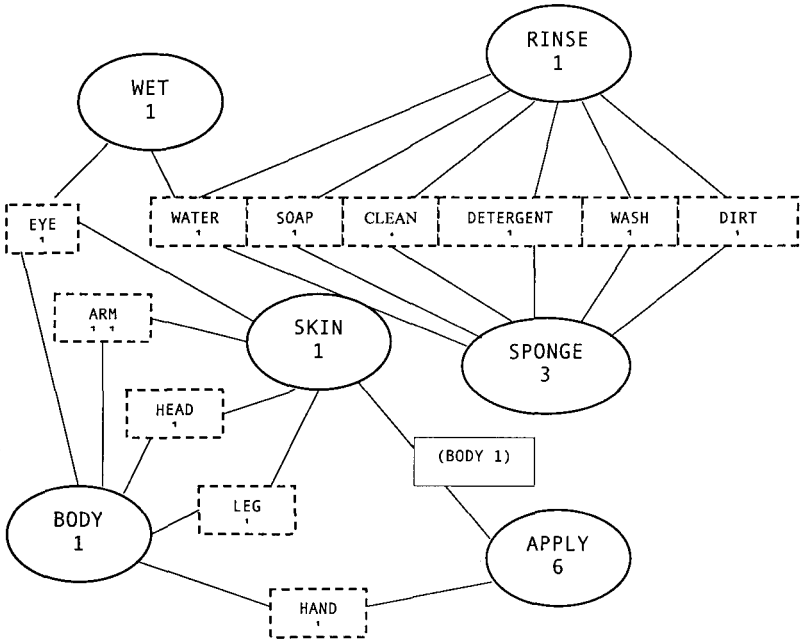


Figura 6

En el marco de la lexicografía computacional, la aportación del lexicón de TRADUNED con sus LDS se podría definir como la construcción de un componente semántico fuerte implementado en una base de datos léxica enriquecida. En otras palabras, el sistema dispone de un componente léxico más potente que una simple base de datos léxica, al permitir la interrelación de los diversos sentidos que contiene el lexicón por medio de las LDS, pero además el diseño de su modelo de lexicón es menos complejo que el de una base de conocimiento, ya que no es necesario construir un modelo ontológico ni definir un grupo restringido de primitivos semánticos que funcionen como conceptos ontológicos. Al igual que en una base de conocimiento, ya hemos visto cómo TRADUNED puede crear una red de motivación conceptual que interconecte los distintos sentidos del lexicón. Como afirman Meijs y Vossen (1992: 144), las definiciones lexicográficas permiten trazar conexiones entre las palabras con el fin de crear una enorme red multidimensional.[23]

[23] Por ejemplo, el proyecto LINKS (Meijs 1989, 1992; Vossen 1989, 1994; Meijs y Vossen 1992) utiliza las definiciones del *Longman's Dictionary of Contemporary English* para establecer una macroestructura de relaciones dentro del léxico, principalmente entre lexemas nominales. El significado de una palabra sería una función de la posición que ocupa en esa red y de sus conexiones con todas las palabras de la red con las que se encuentre más estrechamente relacionada. TRADUNED, no obstante, emplea tanto los

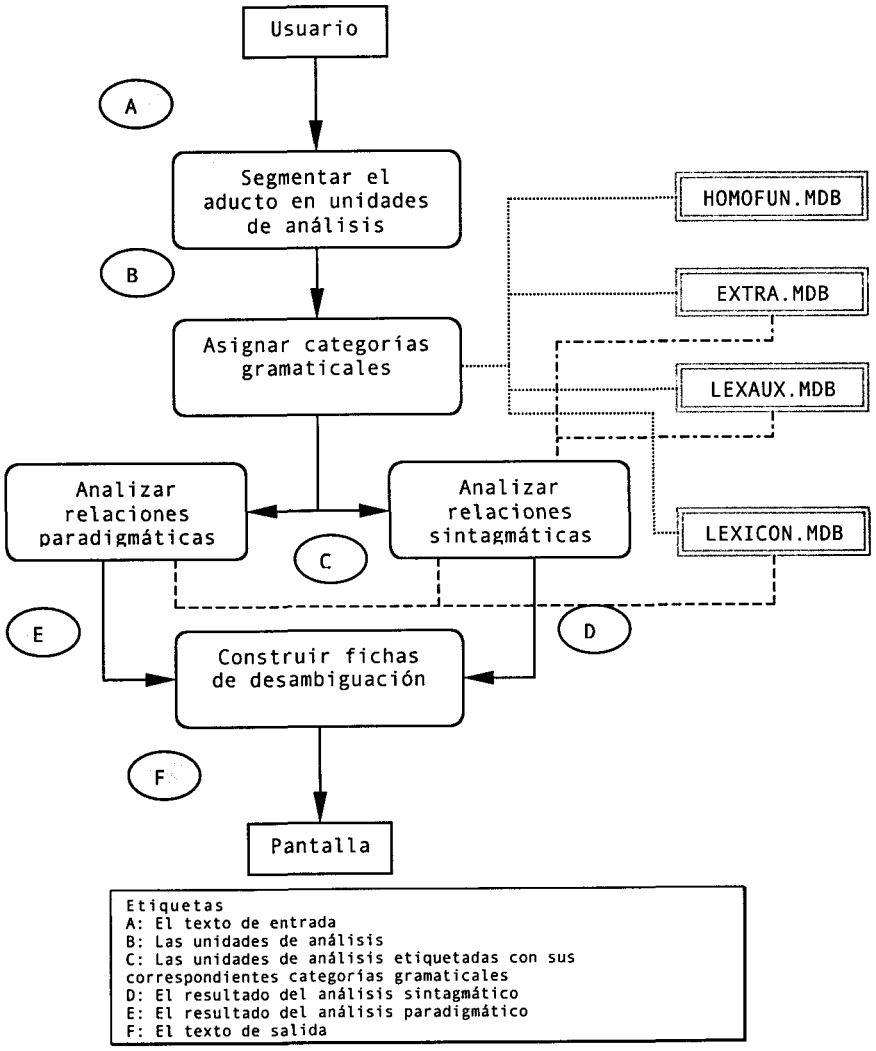


Figura 7

En el proceso de desambiguación léxica nuestra aplicación construye de forma totalmente automática una macrorred virtual de relaciones asociativas a partir de las LDS de las palabras que configuran el aducto. Por ejemplo, la figura 6 muestra la macroestructura que se formaría en la memoria del sistema para un texto de entrada como "With body sponge, apply all over wet skin. Rinse thoroughly".

componentes *genus* como *differentiae* para interrelacionar los predicados del lexicón, por lo que a este respecto se parece más a *EuroWordNet* (Vossen 1998).

5. EL PROCESO DE DESAMBIGUACIÓN LÉXICA EN TRADUNED

Las características generales del diseño de TRADUNED, aplicación implementada en *Visual Basic*,[24] Además, podríamos añadir que, gracias a la facilidad de comprensión de su sintaxis, *Visual Basic* se convierte en un excelente lenguaje introductorio para la formación de investigadores, permitiendo a su vez el desarrollo de programas complejos. se adecuan a las recomendaciones propuestas por Johnson (1987: 259-260): un sistema de traducción automática debe ser ampliable, modular y transparente. En este apartado describiremos la interacción de los diversos componentes y subcomponentes que se activan en TRADUNED durante el proceso de la desambiguación léxica del aducto, y cuya arquitectura del *software* se describe en el apéndice 1. En aras a facilitar su comprensión, presentamos a través del siguiente diagrama de flujo de datos las diferentes fases en que se divide dicho proceso (fig. 7).[25]

5.1. La segmentación del aducto

El desambiguador léxico de TRADUNED tiene el aspecto de un procesador de textos con una serie de menús y una barra de botones (fig. 8). Si deseamos aplicar un proceso completo de desambiguación léxica a un texto que se visualice en pantalla, lo primero que debemos hacer es seleccionar dicho texto con el ratón y ejecutar seguidamente el comando *Syntagmatic and Paradigmatic Levels* del submenú *Lexical Ambiguity Resolution* en el menú *Tools*.

El proceso de desambiguación se inicia con el mecanismo fundamental de la segmentación del aducto (*SegmentarTexto*),[26] el cual va precedido de una breve fase de preprocesamiento en la que caracteres como los retornos manual y automático del carro y el apóstrofe reciben un tratamiento especial (*PrepararTextoEntrada*). Este proceso de segmentación consiste en la identificación de las palabras ortográficas del aducto como unidades de análisis. Como apunta Sinclair (1991: 28 y 41), en esta fase del análisis es habitual definir la palabra como toda subcadena que va seguida de un espacio en blanco, con la única salvedad de que si el último carácter de esa subcadena es un signo de puntuación, dicha subcadena se divide en dos unidades de análisis (*Puntuación*).

[24] *Visual Basic* presenta numerosas ventajas como herramienta de programación para el desarrollo de aplicaciones, ya sea en el contexto de la ingeniería lingüística o en la enseñanza de lenguas asistida por ordenador. Hickman y Langdon (1996) destacan las siguientes: a) Variedad en la presentación del texto; b) Posibilidad de manipulación de los datos lingüísticos; c) Flexibilidad de los mecanismos de respuesta al aducto; d) Facilidad de navegación para el usuario; e) Transparencia de las técnicas de programación; f) Extensibilidad en los componentes de *software*.

[25] Con respecto a la notación gráfica utilizada en este diagrama de burbujas, el rectángulo representa una entidad externa al sistema *software*, el círculo indica un proceso o transformación de los datos, las flechas etiquetadas indican el flujo de los datos y su dirección, y la línea doble representa un almacén de datos, cuyo acceso se indica mediante una línea punteada.

[26] Entre paréntesis se indica el procedimiento activado por TRADUNED para la consecución del objetivo en cuestión. El apéndice 1 nos permite ver el lugar que ocupa cada procedimiento en la arquitectura de la aplicación para el proceso de desambiguación léxica.

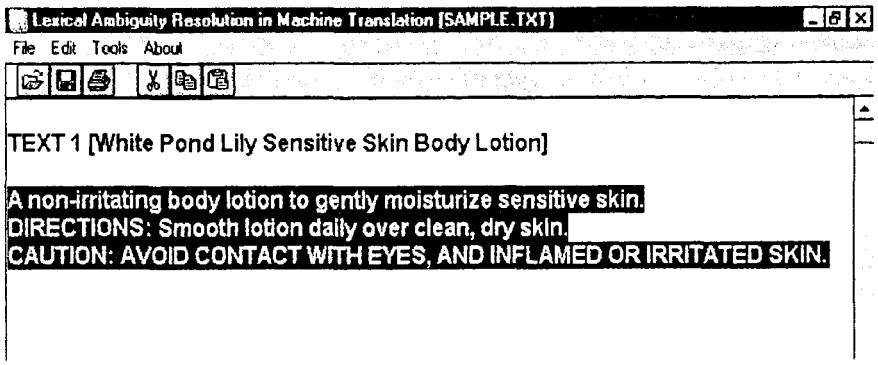


Figura 8

5.2. El etiquetado gramatical

En una segunda fase, tiene lugar la asignación de las categorías gramaticales de cada una de las unidades lingüísticas del aducto, identificando primero la categoría de la mayoría de las palabras funcionales simples o complejas, los participios irregulares y cualquiera de las formas verbales de *be, do* o *have* (*Etiquetar1*). Como en este último punto se aplica un breve análisis de la morfología flexiva, es necesario consultar una lista de palabras cuyas terminaciones puedan confundirse con un morfema flexivo verbal, como en el caso de *red, wing,* etc. (*ExcepAnaMorf*). Posteriormente, el sistema identifica la categoría gramatical de las palabras lexemáticas del aducto; en primer lugar, las que se encuentran registradas en su lexicón principal (*Etiquetar2*) o en el diccionario auxiliar (*ConsultarDicAuxiliar*), para luego identificar la categoría de las restantes palabras por medio de un análisis morfológico y la aplicación de una serie de reglas sintagmáticas (*ReglasContext1*).[27]

5.3. El análisis sintagmático

En una tercera fase, se establecen las relaciones sintagmáticas entre las palabras del aducto (*Sintagmático*), para lo cual, dependiendo de la categoría de cada palabra, la máquina emprende unas acciones u otras. En el caso del adjetivo y el nombre, se identifican las preferencias de colocación tras haber etiquetado las palabras que configuran sus correspondientes textos definitorios (*IdentificarColRest1* y *IdentificarColRest2*, respectivamente). Dichas preferencias de colocación se cotejan con su contexto local (*SelPrefDesamb1* y *SelPrefDesamb2*, respectivamente), siendo necesario identificar el término *genus* del núcleo nominal al que modifican (*IdentificarGenus*).

En el caso del verbo, el sistema determina el número de argumentos del marco predicativo junto con las preferencias selectivas de cada uno de estos argumentos

[27] El grado de eficacia de este etiquetador gramatical repercute notablemente en el desarrollo correcto del proceso de desambiguación léxica: "The availability of very good automatic part-of-speech tagging programs allows some lexical ambiguity to be resolved using the parts of speech" (Wilks et alii 1996: 187).

(*AnalizarSintDef*), asignando posteriormente los papeles temáticos correspondientes (*IdentificarPapeles*). El marco predicativo resultante se aplica al contexto local del aducto para el establecimiento de las relaciones sintagmáticas (*ArgumentosVerb*).

La máquina asigna tres puntos a cada una de las coincidencias que se establecen en este nivel del análisis.

5.4. El análisis paradigmático

En una cuarta fase, TRADUNED identifica las relaciones paradigmáticas que se establecen entre los lexemas que configuran el texto de entrada a partir de sus correspondientes LDS (*Paradigmático*). En esta fase del procesamiento, primero se activa un procedimiento que permite identificar qué palabras lexemáticas no son miembros de la *stop list*, siendo éstas las únicas que pueden intervenir en las posibles relaciones paradigmáticas que establezca nuestro sistema. Seguidamente se buscan todos los sentidos asociados a cada una de esas palabras, siempre y cuando su palabra titular tenga la misma categoría gramatical que la palabra del aducto.

Las relaciones paradigmáticas se construyen a partir de los diferentes sentidos que configuran las LDS de todos los sentidos seleccionados para las palabras lexemáticas que intervienen en el aducto. El sistema recorre todo el texto de entrada deteniéndose en cada una de las palabras ambiguas y coteja cada uno de los sentidos de la LDS correspondiente a cada sentido de la palabra ambigua con el resto de las LDS. Nuestra teoría se basa en la idea de que entre dos palabras se establece una relación paradigmática siempre que exista alguna coincidencia entre sus LDS. Por ello, todas las coincidencias que se van encontrando son tratadas como relaciones paradigmáticas entre un determinado sentido de una palabra del aducto y los sentidos de algunas de las palabras que configuran su contexto. La máquina asigna un punto por cada una de estas coincidencias dentro de una misma relación paradigmática.

5.5. Las fichas de desambiguación léxica

Finalmente tiene lugar una fase de posprocesamiento del aducto (*PresentarDesamb*), la cual permite visualizar en pantalla el texto que el usuario seleccionó para su desambiguación, además de marcar con un subrayado simple todas aquellas palabras que han resultado ser intercategorial o intracategorialmente ambiguas (fig. 9). Haciendo un clic sobre cualquiera de estas palabras se presenta en primer plano la ficha de desambiguación léxica del sentido que el sistema ha asignado a dicha palabra (fig. 10),[28] por lo que podemos decir que es aquí donde finaliza el proceso de desambiguación léxica.

[28] Cuando se visualiza una ficha de desambiguación léxica, en el título de la pantalla se muestra la palabra titular, el número de acepción en el CCED, el equivalente de traducción y la puntuación obtenida durante el proceso de desambiguación.

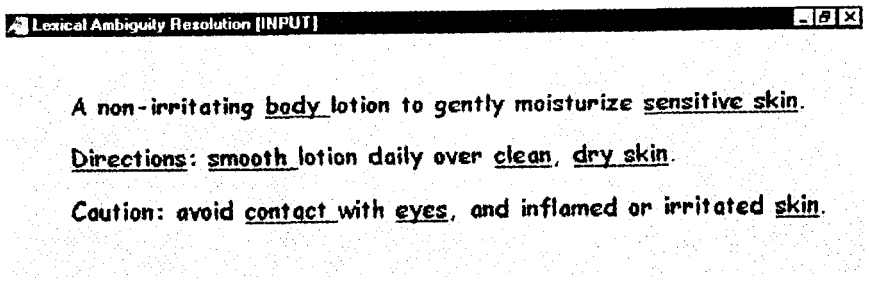


Figura 9

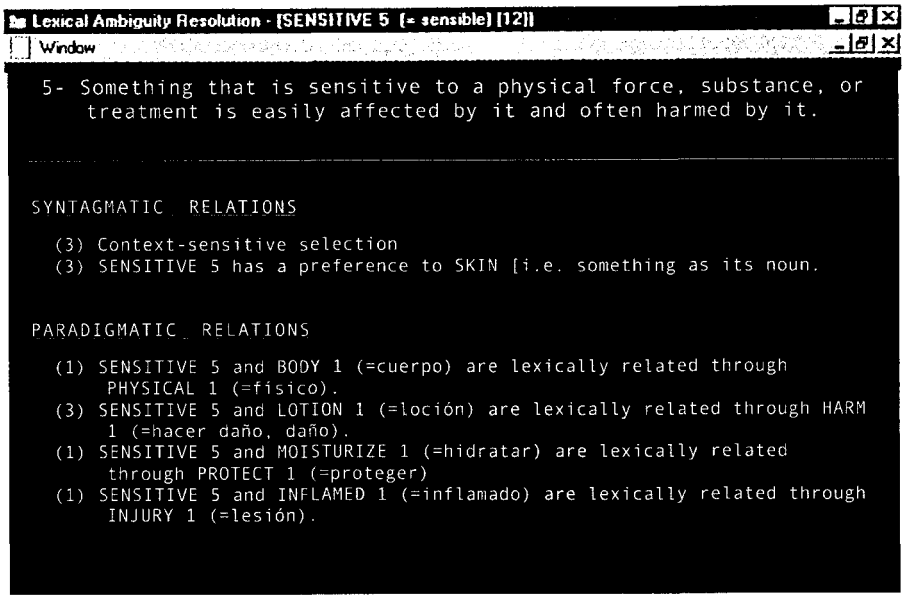


Figura 10

La posibilidad de presentar en pantalla una serie de fichas de desambiguación léxica resulta una de las características más innovadoras de TRADUNED. Para cada sentido de una palabra ambigua, nuestra aplicación crea una ficha que nos informa con detalle de las relaciones sintagmáticas y paradigmáticas que se han podido identificar a partir del texto seleccionado. Cada relación sintagmática o paradigmática establecida entre dos sentidos de dos palabras del aducto forma parte de un procedimiento de puntuación que permite la designación del sentido del término teniendo en cuenta la puntuación más alta.[29] En otras palabras, nuestra

[29] En caso de empate, es decir, que varios sentidos tengan la puntuación más alta, se selecciona aquel sentido que tenga una mayor puntuación en el plano sintagmático. Si persiste el empate, el sistema elegirá entonces aquella definición cuyo número de acepción sea más bajo, decisión que se fundamenta en el siguiente comentario de Higinbotham: "Often, the easiest way to pick the correct sense of a word is

aplicación presenta las diversas puntuaciones establecidas a lo largo del análisis junto a sus correspondientes explicaciones. El propósito de este informe es precisamente describir con detalle cómo se ha obtenido la puntuación de un sentido "ganador".

6. RESULTADOS Y CONCLUSIONES

Hemos aplicado la presente versión de TRADUNED sobre tres textos de entrada en inglés que contienen instrucciones de uso de determinados productos de peluquería y cosmética, como el que se muestra a continuación:

> A non-irritating body lotion to gently moisturize sensitive skin.
>
> DIRECTIONS: Smooth lotion daily over clean, dry skin.
> CAUTION: AVOID CONTACT WITH EYES, AND INFLAMED OR IRRITATED SKIN.

Podemos hablar aquí de un caso de sublengua,[30] que a nivel gramatical se refleja por el uso sistemático de imperativos y una estructura sintagmática nominal muy simple: un núcleo nominal premodificado a veces por un adjetivo (deverbativo o no), otro nombre o una combinación de éstos. Incluso es evidente la presencia de estructuras sintácticas que se alejan a menudo de los cánones establecidos para un uso normalizado de la lengua inglesa. En cuanto al léxico, existe cierta preferencia por los verbos de acción y los lexemas relacionados con el campo semántico del cuerpo humano. El propósito de utilizar este tipo de textos no ha sido precisamente el reducir el número de ambigüedades léxicas. Nuestro objetivo no es la construcción de un componente léxico *ad hoc*,[31] ya que en futuras versiones de TRADUNED pretendemos utilizar textos que pertenezcan a una sublengua diferente, lo que posiblemente repercutirá en el componente gramatical.[32] La dependencia temática nos permite, en cambio, reducir al máximo la complejidad sintáctica, pudiendo así centrar mejor nuestro trabajo en la descripción del proceso de desambiguación léxica a partir de un sistema orientado al lexicón.

---

simply to choose the sense that is most frequent in general text" (1990: 70).

[30] Todos los lingüistas destacan las ventajas que nos ofrece el uso de una sublengua en un sistema de traducción automática (Boitet 1988: 234; Lehrberger y Bourbeau 1988: 52). La sublengua se puede definir básicamente como una variedad especializada de una lengua que se utiliza sobre un tema específico. No obstante, la sistematicidad en el uso de ciertos patrones gramaticales y léxicos es el rasgo que realmente caracteriza a una variedad de la lengua. La dependencia temática que implica la utilización de una sublengua ofrece diversas ventajas, como la reducción del tamaño y la complejidad de los componentes léxico y gramatical: "First, the parsing time is reduced, since sublanguage grammars are always smaller than the grammars of whole languages. Second, the problem of structural and lexical ambiguity is greatly reduced, since many interpretations or analyses which are possible in the standard language are not 'legal' (i.e., they are meaningless) in the sublanguage, and therefore can be ruled out" (Kittredge 1987: 63).

[31] A pesar de que en la mayoría de las aplicaciones de traducción automática los lexicones almacenan un número reducido de sentidos para cada uno de sus lexemas, TRADUNED ofrece la posibilidad de operar con todos los sentidos que recoge el CCED.

[32] Actualmente nuestra aplicación se limita a cubrir verbos primarios del grupo A en términos de la clasificación de Dixon (1991: 88), según la cual estos verbos sólo pueden admitir sintagmas nominales como realizaciones de las funciones de sujeto y objeto directo.

Con el propósito de interpretar los resultados obtenidos en el proceso de desambiguación léxica, hemos diseñado unas tablas a partir de las fichas de desambiguación léxica asociadas a cada uno de los 32 casos de ambigüedad intracategorial que aparecen en los tres textos de entrada.[33] La ambigüedad intercategorial carece de interés ya que nuestra aplicación dispone de un etiquetador gramatical. Al comparar las puntuaciones de los sentidos "ganadores" en los tres textos, llegamos a una serie de conclusiones, las cuales se ilustran perfectamente con el ejemplo del apéndice 2. Podemos medir la asociación que se establece entre los planos sintagmático y paradigmático a través del coeficiente de correlación. Siendo x la variable de las puntuaciones de los sentidos ganadores en el análisis sintagmático e y en el análisis paradigmático, el coeficiente de correlación es el resultado de dividir la covarianza de x e y por el producto de las desviaciones típicas de x e y:

$$\rho = \sigma_{xy} / \sigma_x \sigma_y$$

El resultado de esta fórmula es un número entre -1 y 1, donde la proximidad a 1 indica un fuerte grado de dependencia lineal entre las dos variables. En nuestro caso, el coeficiente de correlación es -0'28, por lo que podemos afirmar que la puntuación del nivel sintagmático de un determinado sentido no permite predecir la puntuación del nivel paradigmático en ese mismo sentido, o viceversa. También podemos señalar que un análisis sintagmático aislado del aducto no es suficiente para resolver todos los casos de ambigüedad intracategorial (p.ej. *body, sensitive, skin*). Observamos igualmente que aquellos sentidos "ganadores" que poseen un alto peso en su plano paradigmático son los sentidos que configuran la línea tópica del discurso: p.ej. *body, skin, smooth* y *dry* son las palabras que mejor definen el tema del primer texto de entrada.

Como ya comentamos en el apartado 3, los ingenieros lingüistas suelen considerar que la división de sentidos que presenta un diccionario convencional es excesivamente sutil para su aplicación en la resolución de la ambigüedad léxica. Por ejemplo, algunos investigadores proponen agrupar las definiciones de un diccionario legible por la máquina según la división de los sentidos de un tesauro.[34] TRADUNED consigue, en cambio, que la máquina utilice el mismo número de sentidos que utilizaría la mente humana en la asignación del significado de una palabra, resolviendo con éxito incluso aquellos casos en que intervienen palabras polisémicas cuyos sentidos son extremadamente parecidos (p.ej. *skin* y *dry* entre otros). Los textos definitorios del CCED contienen toda la información necesaria para llevar a cabo con éxito el proceso de desambiguación léxica de un texto de entrada. Con este propósito, ha sido necesario diseñar una serie de algoritmos que permiten la extracción de toda la información implícita que contienen los textos definitorios para su posterior aplicación en los análisis sintagmático (i.e. las

[33] La tabla del apéndice 2 corresponde al texto de entrada que aparece como ejemplo al principio de este apartado.

[34] Chen y Chang (1998) presentan el algoritmo *TopSense*, el cual utiliza la información del tesauro *Longman Lexicon of Contemporary English* para agrupar automáticamente los sentidos del *Longman Dictionary of Contemporary English*.

preferencias de colocación de los nombres y los adjetivos, los términos *genus* del nombre y el marco predicativo verbal) y paradigmático (i.e. las LDS). Ambos niveles de análisis son complementarios, al no resultar ninguno de ellos primordialmente determinante en la asignación de un sentido a una palabra lexemática ambigua.

Otra de las principales aportaciones de TRADUNED se materializa en forma de fichas de desambiguación léxica. Nuestra aplicación no sólo determina el sentido más apropiado de una palabra lexemática ambigua, sino que además describe el proceso de razonamiento por el cual la aplicación ha tomado esa decisión. De esta manera, demostramos que nuestro sistema se basa en una sólida fundamentación lingüística durante el proceso de desambiguación.

OBRAS CITADAS

Albadalejo Mayordomo, Tomás y Antonio García Berrio 1983: "La Lingüística del Texto". *Introducción a la Lingüística.* Ed. Alicia Yllera et alii. Madrid: Alhambra. 217-60.

Alshawi, Hiyan, Bran Boguraev y Ted Briscoe 1985: "Towards a Dictionary Support Environment for Real Time Parsing". *Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics*. Génova: Association for Computational Linguistics. 171-78.

Barnbrook, Geoff 1993: "The Automatic Analysis of Dictionaries: Parsing Cobuild Explanations". *Text and Technology: In Honour of John Sinclair.* Ed. Mona Baker et alii. Philadelphia-Amsterdam: John Benjamins. 313-32.

Belinchón, Mercedes, José Manuel Igoa y Ángel Rivière 1992: *Psicología del Lenguaje. Investigación y Teoría.* Madrid: Trotta.

Boitet, Christian 1988: "Current Projects at GETA on or about Machine Translation". *Actas del II Congreso Mundial Vasco. Congreso de Inteligencia Artificial.* Vitoria-Gasteiz: Servicio Central de Publicaciones del Gobierno Vasco. 229-252.

Chen, Jen Nan y Jason S. Chang 1998: "Topical Clustering of MRD Senses Based on Information Retrieval Techniques". *Computational Linguistics* 24.1: 61-95.

Chomsky, Noam 1981: *Lectures on Government and Binding.* Dordrecht: Foris.

Collins COBUILD English Language Dictionary 1995 (1987). Londres: Collins.

Dahlgren, Kathleen, Joyce McDowell y Edward P. Stabler 1989: "Knowledge Representation for Commonsense Reasoning with Text". *Computational Linguistics* 15.3: 149-70.

Dik, Simon C. 1978: *Functional Grammar.* Dordrecht: Foris.

——— 1989: *The Theory of Functional Grammar: The Structure of the Clause.* Dordrecht: Foris.

——— 1990: "FG*C*M*NLU: Functional Grammar Computational Model of the Natural Language User". *Functional Grammar and the Computer*. Eds. John H. Connolly y Simon C. Dik. Dordrecht: Foris. 1-28.

——— 1997: *The Theory of Functional Grammar: The Structure of the Clause.* Berlín-Nueva York: Mouton de Gruyter.

Dixon, Robert M. W. 1991: *A New Approach to English Grammar on Semantic Principles*. Oxford: Clarendon Press.

Escobedo Rodríguez, Antonio 1994: *Estudios de Lexicología y Lexicografía.* Almería: Universidad de Almería.

Estévez, Adelina y Manuel De Vega 1999: "Procesamiento de las Palabras Ambiguas en Contextos Narrativos". *Cognitiva* 11: 67-90.

Fass, Dan 1993: "Lexical Semantic Constraints". *Semantics and the Lexicon.* Ed. James Pustejovsky. Dordrecht: Kluwer Academic Publishers. 263-89.

Fellbaum, Christiane 1998: "A Semantic Network of English: The Mother of All Wordnets". *Computers and the Humanities* 32.2-3: 209-20.

Flickinger, Daniel P. 1992: "Natural Language Processing: Lexical Organization". *International Encyclopedia of Linguistics*. Ed. William Bright. Nueva York: Oxford University Press. 66-67.

Foley, William A. y Robert D. van Valin 1980: "Role and Reference Grammar". *Syntax and Semantics 13: Current Approaches to Syntax.* Eds. E. Moravcsik y J. Wirth. Nueva York: Academic Press. 329-51.

Fontenelle, Thierry 1992: "Automatic Extraction of Lexical-Semantic Relations from Dictionary Definitions". *EURALEX '90: Fourth International Congress on Lexicography*. Barcelona: Biblograf. 89-103.

Garrido Medina, Joaquín 1988: *Lógica y Lingüística.* Madrid: Síntesis.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum e Ivan A. Sag 1985: *Generalised Phrase Structure Grammar*. Oxford: Basil Blackwell.

Halliday, Michael y Ruqaiya Hasan 1976: *Cohesion in English.* Londres: Longman.

Handke, Jürgen 1995: *The Structure of the Lexicon: Human versus Machine.* Berlín-Nueva York: Mouton de Gruyter.

Hickman, Paul y John Langdon 1996: The Use of *Visual Basic* as a CALL Development Tool: The Production and Refinement of *GramEx, GramDef* and *Ça Sonne Français* for the TELL Consortium. *Proceedings EUROCALL '95. Technology Enhanced Language Learning: Focus on Integration.* Ed. Ana Gimeno. Valencia: Universidad Politécnica de Valencia. 185-194.

Higinbotham, Dan Walter 1990: *Semantic Cooccurrence Networks and the Automatic Resolution of Lexical Ambiguity in Machine Translation.* Tesis doctoral inédita. Austin: University of Texas.

Hoey, Michael 1991: *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Johnson, Roderick 1987: Translation. *Linguistic Theory and Computer Applications.* Eds. Peter Whitelock et alii. Londres: Academic Press. 257-85.

Kaplan, Ronald M. y Joan Bresnan 1982: "Lexical-Functional Grammar: A Formal System for Grammatical Representation". *The Mental Representation of Grammatical Relations*. Ed. Joan Bresnan. Cambridge (Massachusetts): MIT Press. 173-281.

Kilgarriff, Adam 1997: "I Don't Believe in Word Senses". *Computers and the Humanities* 31.2: 91-113.

Kittredge, Richard I. 1987: "The Significance of Sublanguage for Automatic Translation". *Machine Translation: Theoretical and Methodological Issues*. Ed. Sergei Nirenburg. Cambridge: Cambridge University Press. 59-67.

Laffling, John 1991: *Towards High-Precision Machine Translation: Based on Contrastive Textology*. Berlín-Nueva York: Foris.

Lázaro Carreter, Fernando 1984: *Diccionario de Términos Filológicos*. Madrid: Gredos.

Leech, Geoffrey 1974: *Semantics*. Harmondsworth: Penguin.

Lehrberger, John y Laurent Bourbeau 1988: *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. Philadelphia-Amsterdam: John Benjamins.

Lesk, Michael E. 1986: "Automated Sense Disambiguation Using Machine-Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone". *Proceedings of the 1986 ACM SIGDOC Conference*: 24-26.

Little, Patrick 1990: "METAL – Machine Translation in Practice". *Translating and the Computer 11. Preparing for the Next Decade*. Ed. Catriona Picken. Londres: Aslib. 94-107.

Mairal, Ricardo 1999: "El Componente Lexicón en la Gramática Funcional". *Nuevas Perspectivas en Gramática Funcional*. Eds. Christopher Butler *et al.* Barcelona: Ariel. 41-98.

Meijs, Willem 1989: "Spreading the Word: Knowledge-Activation in a Functional Perspective". *Functional Grammar and the Computer*. Eds. John H. Connolly y Simon C. Dik. Dordrecht: Foris. 201-15.

——— 1992: "The Expanding Lexical Universe: Extracting Taxonomies from Machine-Readable Dictionaries". *EURALEX '90: Fourth International Congress on Lexicography*. Barcelona: Biblograf. 119-28.

Meijs, Willem y Piek Vossen 1992: "In So Many Words: Knowledge as a Lexical Phenomenon". *Lexical Semantics and Knowledge Representation: First SIGLEX Workshop*. Eds. James Pustejovsky y Sabine Bergler. Berlin: Springer-Verlag. 137-53.

Miller, George A. 1995: "WordNet: A Lexical Database for English". *Communications of the ACM* 38.11: 39-41.

Moon, Rosamund 1987: The Analysis of Meaning. *Looking up. An account of the COBUILD Project*. Ed. John Sinclair. Glasgow: Collins. 86-103.

Nirenburg, Sergei y Lori Levin 1992: "Syntax-Driven and Ontology-Driven Lexical Semantics". *Lexical Semantics and Knowledge Representation: First SIGLEX*

Workshop. Eds. James Pustejovsky y Sabine Bergler. Berlin: Springer-Verlag. 5-20.

Onyshkevych, Boyan A. y Sergei Nirenburg 1992: "Lexicon, Ontology, and Text Meaning". *Lexical Semantics and Knowledge Representation: First SIGLEX Workshop*. Eds. James Pustejovsky y Sabine Bergler. Berlin: Springer-Verlag. 289-303.

Pustejovsky, James 1995: *The Generative Lexicon*. Cambridge (Massachusetts): MIT Press.

Quillian, M. Ross 1968: "Semantic Memory". *Semantic Information Processing*. Ed. Marvin Minsky. Cambridge (Massachusetts): MIT Press. 216-70.

Raskin, Victor 1987: "Linguistics and Natural Language Processing". *Machine Translation: Theoretical and Methodological Issues*. Ed. Sergei Nirenburg. Cambridge: Cambridge University Press. 42-58.

Ritchie, Graeme 1987: "The Lexicon". *Linguistic Theory and Computer Applications*. Ed. Peter Whitelock et alii. Londres: Academic Press. 225-56.

Russell, Stuart y Peter Norvig 1996: *Inteligencia Artificial: Un Enfoque Moderno*. Méjico: Prentice Hall.

Schubert, Lenhart K. 1992: "Natural Language Processing: Semantics and Knowledge Representation". *International Encyclopedia of Linguistics*. Ed. William Bright. Nueva York: Oxford University Press. 69-72.

Sinclair, John 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tucker, Allen B. 1987: "Current Strategies in Machine Translation Research and Development". *Machine Translation: Theoretical and Methodological Issues*. Ed. Sergei Nirenburg. Cambridge: Cambridge University Press. 22-41.

Vossen, Piek 1989: "The Structure of Lexical Knowledge as Envisaged in the LINKS-Project". *Functional Grammar and the Computer*. Eds. John H. Connolly y Simon C. Dik. Dordrecht: Foris. 177-99.

——— 1994: "The End of the Chain: Where Does Decomposition of Lexical Knowledge Lead Us Eventually?" *Function and Expression in Functional Grammar*. Ed. Elisabeth Engberg-Pedersen et alii. Berlín-Nueva York: Mouton de Gruyter. 11-39.

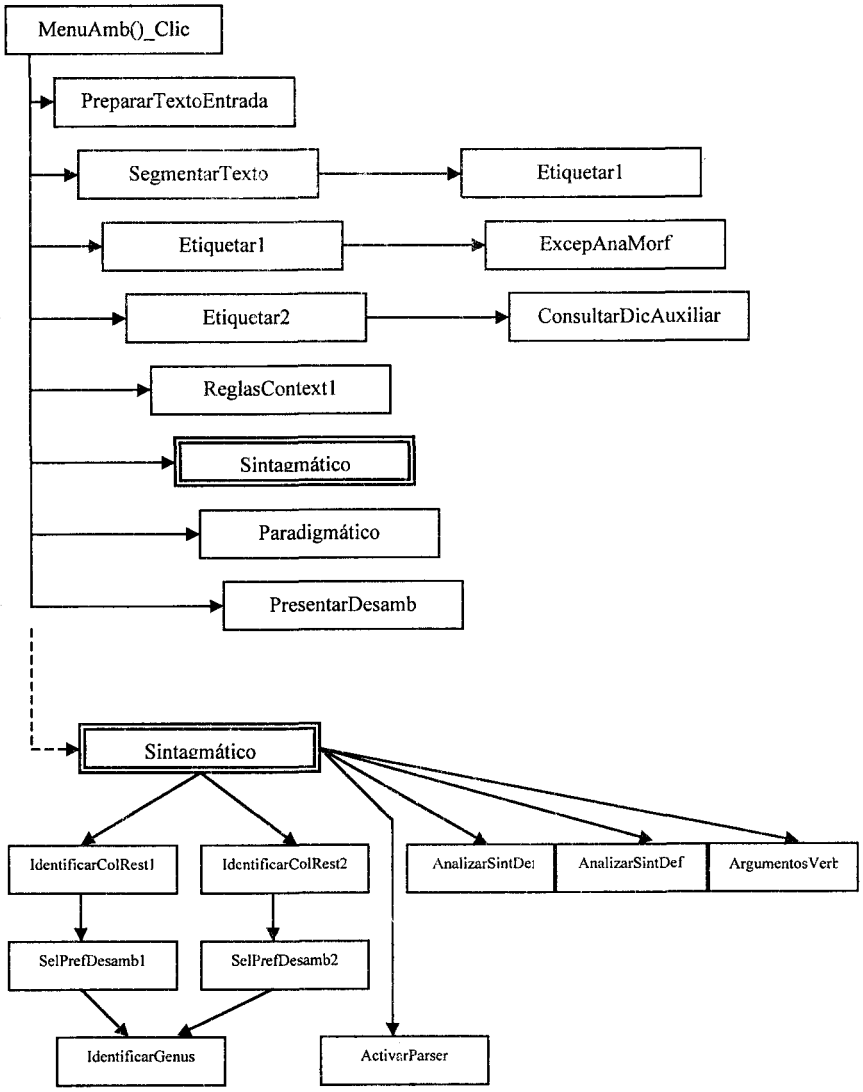
——— 1998: "Introduction to EuroWordNet". *Computers and the Humanities* 32.2-3: 73-89.

Wilks, Yorick A. 1975: "Preference Semantics". *Formal Semantics of Natural Language*. Ed. E. L. Keenan. Cambridge: Cambridge University Press. 329-48.

Wilks, Yorick A., Dan Fass, Cheng-ming Guo, James E. McDonald, Tony Plate y Brian M. Slator 1993: "Providing Machine Tractable Dictionary Tools". *Semantics and the Lexicon*. Ed. James Pustejovsky. Dordrecht: Kluwer Academic Publishers. 341-401.

Wilks, Yorick A., Brian M. Slator y Louise M. Guthrie, eds. 1996: *Electric Words. Dictionaries, Computers and Meanings*. Cambridge (Massachusetts): MIT Press.

APÉNDICE 1: LA ARQUITECTURA DE TRADUNED PARA EL PROCESO DE DESAMBIGUACIÓN LÉXICA



APÉNDICE 2: LOS RESULTADOS DE LA DESAMBIGUACIÓN LÉXICA EN UN TEXTO DE ENTRADA[35]

TEXTO 1 [White Pond Lily Sensitive Skin Body Lotion]

A non-irritating body lotion to gently moisturize sensitive skin.

DIRECTIONS: Smooth lotion daily over clean, dry skin.

CAUTION: AVOID CONTACT WITH EYES, AND INFLAMED OR IRRITATED SKIN.

	S	P	T
BODY			
1 Your body is all your physical parts, including your head, arms and legs.	6 6	16 15	22 21
2 You can also refer to the main part of your body, excluding your arms, head, and legs, as your body.	3	-	3
3 You can refer to a person's dead body as a body.	6	-	6
4 A body is an organized group of people who deal with something officially.	3	-	3
5 A body of people is a group of people who are together or who are connected in some way.	3	-	3
6 The body of something such as a building or a document is the main part of it or the largest part of it.	3	-	3
7 The body of a car or aeroplane is the main part of it, not including its engine, wheels, or wings.	3 3	- 1	3 4
8 A body of water is a large area of water, such as a lake or a sea.	6	-	6
9 A large body of information is a large amount of it.			
10 If you say that an alcoholic drink has body, you mean that it has a full and strong flavour.			
SKIN			
1 Your skin is the natural covering of your body.	6	18	24
2 An animal skin is skin which has been removed from a dead animal. Skins are used to make things such as coats and rugs.	6 3	- 3	6 6
3 The skin of a fruit or vegetable is its outer layer or covering.	6	1	7
4 If a skin forms on the surface of a liquid, a thin, fairly solid layer forms on it.			

[35] Esta tabla sólo presenta la puntuación obtenida en los planos sintagmático y paradigmático y el total correspondiente para todos los sentidos que presenta el CCED en cada uno de los casos de ambigüedad léxica intracategorial que aparecen en el texto 1.

SENSITIVE			
1 If you are sensitive to other people's needs, problems or feelings, you show understanding and awareness of them.	3	-	3
2 If you are sensitive about something, you are easily worried and offended when people talk about it.	3	-	3
3 A sensitive subject or issue needs to be dealt with carefully because it is likely to cause disagreement or make people angry or upset.	6	-	6
4 Sensitive documents or reports contain information that needs to be kept secret and dealt with carefully.	6	1	7
5 Something that is sensitive to a physical force, substance, or treatment is easily affected by it and often harmed by it.	6	6	12
6 A sensitive piece of scientific equipment is capable of measuring or recording very small changes.	6	2	8
DIRECTION			
1 A direction is the general line that someone or something is moving or pointing in.	6	2	8
2 A direction is the general way in which something develops or progresses.	6	-	6
3 Directions are instructions that tell you what to do, how to do something, or how to get somewhere.	9	-	9
4 The direction of a film, play, or television programme is the work that the director does while it is being made.	3	-	3
SMOOTH			
8 If you smooth something, you move your hands over its surface to make it smooth and flat.	9	14	23
9 If you smooth something somewhere, you use your hands to spread it there.	12	15	27
CLEAN			
1 Something that is clean is free from dirt or unwanted marks.	9	2	11
2 You say that people or animals are clean when they keep themselves or their surroundings clean.	6	-	6
3 A clean fuel or chemical process does not create many harmful or polluting substances.	6	-	6
6 If you describe something such as a book, joke, or lifestyle as clean, you think that they are not sexually immoral or offensive.	6	-	6
7 If someone has a clean reputation or record, they have never done anything illegal or wrong.	6	1	7
8 A clean game or fight is carried out fairly, according to the rules.	6	-	6
9 If you describe a flavour, smell, or colour as clean, you like it because it is light and fresh.	6	-	6
10 A clean sheet of paper has no writing or drawing on it.	6	-	6

DRY			
1 If something is dry, there is no water or moisture on it or in it.	9 9	- 15	9 24
4 If you say that your skin or hair is dry, you mean that it is less moist, oily, or soft than average or than normal.	6	-	6
5 If the weather or a period of time is dry, there is no rain or there is much less rain than average.	6	-	6
6 A dry place or climate is one that gets very little rainfall.	6	-	6
8 If a river, lake, or well is dry, it is empty of water, usually because of hot weather and lack of rain.	6	-	6
9 If an oil well is dry, it is no longer producing any oil.	6	1	7
10 If you are dry, you are thirsty and need to drink something; an informal use.	6	2	8
11 If your mouth or throat is dry, it has little or no saliva in it, and so feels very unpleasant, perhaps because you are tense or ill.	6 6	1 8	7 14
12 A dry cough is one that does not produce any phlegm.	6	2	8
13 If someone has dry eyes, there are no tears in their eyes; often used with negatives or in contexts where you are expressing surprise that they are not crying.	6	-	6
14 If a country, state, or city is dry, it has laws or rules which forbid anyone to drink, sell, or buy alcoholic drink; an informal use.	6 6	- -	6 6
16 Dry humour is very amusing, but in a subtle and clever way.	6	5	11
17 If you describe a voice as dry, you mean that it is cold or dull, and does not express any emotions; mainly used in written English.	6	-	6
18 If you describe something such as a book, play, or activity as dry, you mean that it is dull and uninteresting.			
19 Dry bread or toast is plain and not covered with butter or jam.			
20 Dry sherry or wine does not have a sweet taste.			
CONTACT			
1 Contact involves meeting or communicating with someone, especially regularly.	6	2	8
7 When people or things are in contact, they are touching each other.	6 6	4 1	10 7
8 Radio contact is communication by means of radio.	3	1	4
9 A contact is someone you know in an organization or profession who helps you or gives you information.			

EYE			
1 Your eyes are the parts of your body with which you see.	9	18	27
3 You use eye when you are talking about a person's ability to judge things or about the way in which they are considering or dealing with things.	9 9	- -	9 9
4 An electric eye or infra-red eye is a device which can recognize the presence of people or objects by detecting the light or heat coming from them.	3	-	3
5 People sometimes talk about the eye of the camera when they are talking about something being filmed or photographed, or the way something appears in a photograph or film.	6 9	2 3	8 12
6 An eye on a potato is one of the dark spots from which new stems grow.	6	-	6
7 An eye is a small metal loop which a hook fits into, as a fastening on a piece of clothing.	3	-	3
8 The eye of a needle is the small hole at one end which the thread passes through.			
9 The eye of a storm, tornado, or hurricane is the centre of it.			

TELLING WOMEN'S LIVES: VISION AS HISTORICAL REVISION IN THE WORK OF MICHÈLE ROBERTS

Sonia Villegas López

Universidad de Huelva

Within the recent trend in postmodern fiction which disavows the *grand récit* of History and promotes individual stories by means of autobiography, in *Impossible Saints* (1997) the feminist author Michèle Roberts rewrites an important body of texts belonging to the Christian tradition: hagiographies, or the lives of women saints, in her case. In this novel, the former inspired texts appear merely as accounts of "personal history", since they go through a process of revision that results in the demystification of both the patriarchal discourse, and the female prototypes imposed by the gender construction of femininity, and fostered by Christianity. In *Impossible Saints*, the constant fictionalization of Josephine's experiences, similar in many respects to Teresa of Ávila's, as well as the rewriting of many female saints's life stories, illustrate, in the first place, the end of history as we know it, and secondly, the political end of autobiography, in so far as it provides the only space for female representation in the official discourse of Christianity, and eludes a unique interpretation in favor of plurality and heterogeneity.

In *Impossible Saints* (1997), the British writer of French origin Michèle Roberts sets out on a journey towards the origins of women's (auto)biography, and thus, of women's history. In this novel, the author alludes to the historical figure of Teresa of Ávila through Sister Josephine, a character of her invention, whose life story becomes the object of analysis and reflection for her niece Isabel. Josephine's private and public experiences become fictionalized by the skilled pen of an intuitive she-narrator who is believed to know most of Josephine's life first hand, but who does not claim to be telling the truth at all times. The central, though elusive and fragmented, narrative of Teresa-Josephine, is interspersed with other minor portraits of famous women mystics and saints of the Christian tradition, which differ in various degrees from the conventional account of their lives. It is my contention in this paper that Michèle Roberts makes use of genres and subgenres like (auto)biography and hagiography in order to rewrite history from a feminist perspective. In fact, precisely by writing other(s') lives the author manages to revise

ATLANTIS
Vol. XXIII Núm. 1 (Junio 2001): 173-88.
ISSN 0210-6124

INDICE