RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 46 (1), I Sem. 2008, pp. 81-91.
CL ISSN 0033 - 698X

DISPOGRAFO: UNA NUEVA HERRAMIENTA COMPUTACIONAL PARA EL ANALISIS DE RELACIONES SEMANTICAS EN EL LEXICO DISPONIBLE[1]

DISPOGRAFO. A NEW COMPUTATIONAL TOOL FOR THE ANALYSIS OF SEMANTIC RELATIONS IN LEXICAL AVAILABILITY

MAX S. ECHEVERRIA
mechever@udec.cl
ROBERTO VARGAS
robvargas@gmail.com
PAULA URZUA
paurzua@udec.cl
Universidad de Concepción, Chile
ROBERTO FERREIRA
raf502@york.ac.uk
University of York, Heslington, York, UK

RESUMEN

Se desarrolló un programa computacional destinado a apoyar el análisis psicolingüístico de los términos elicitados mediante encuestas de léxico disponible. Utilizando un algoritmo basado fundamentalmente en las relaciones de secuencia de las palabras disponibles, nuestro programa, DispoGrafo, ingresa los términos elicitados y genera luego automáticamente grafos cuyos nodos representan palabras y cuyas aristas simbolizan las relaciones entre ellas. Los grafos se interpretan como redes semánticas cuya configuración expresa las relaciones semánticas subyacentes en el corpus. El software permite, además, eliminar las conexiones débiles (de menor peso) para así dejar sólo las relaciones más robustas y visualizar de este modo las relaciones más relevantes.

Palabras claves: Disponibilidad léxica, redes semánticas, psicolingüística computacional.

ABSTRACT

A computer program was developed to support psycholinguistic analyses of words elicited in lexical availability tests. By means of an algorithm based on word sequence relations, our program, DispoGrafo, inputs elicited words to automatically generate graphs in which

[1] Trabajo financiado mediante Proyecto Fondecyt 1050598-2005.

nodes represent words and lines the links between them. Graphs are then interpreted as semantic networks displaying the latent semantic ties underlying the data. The software also allows users to remove weak connections so that only those representing stronger relations remain to ensure a better representation of semantics.

Keywords: Lexical availability, semantic networks, computational psycholinguistics.

Recibido: 28-12-2007. Aceptado: 24-03-2008.

INTRODUCCION

LAS coincidencias en los resultados de las distintas sintopías constituyen un fenómeno lingüístico de mayor importancia: No es por azar que las voces más disponibles en cada centro se repitan de una a otra comunidad hispanohablante. La explicación del fenómeno, sin embargo, dista mucho de ser simple. Intervienen en la selección léxica que realizan los sujetos variados factores tales como su memoria operacional, memoria semántica y episódica, saliencia (*salience*) o prominencia psicológica del término, experiencias del individuo, etc. El mejor trabajo sobre esta problemática del 'ser disponible' de un vocablo es la tesis doctoral de Natividad Hernández (Hernández, 2005). Remitimos allí al lector interesado en profundizar el tema.

La tarea de producción lingüística a que se somete a los informantes tiene sin duda un carácter artificial, aunque no mayor que muchos experimentos psicolingüísticos on-line de decisión léxica, por ejemplo, donde el sujeto debe pulsar una tecla del computador para señalar si el término es o no propio de la lengua.

La mayor parte de los estudios de disponibilidad léxica son esencialmente cuantitativos y de carácter sociolingüístico: se calcula el índice de disponibilidad, el promedio de vocablos por centro, el número de palabras diferentes en cada uno de ellos, el índice de cohesión, la convergencia/divergencia entre grupos, la incidencia del sexo, nivel sociocultural y tipo de enseñanza en los resultados obtenidos. Desde el punto de vista cualitativo no escapa a los investigadores la particular relación que se da entre los términos entregados por los sujetos. Diversos autores advierten en sus corpora ciertas agrupaciones categoriales o conjuntos asociativos. Es más: como muy bien lo señala Natividad Hernández (2005: 51), muchos de ellos afirman que los vocablos disponibles se organizan en forma de redes semánticas (propias de un paradigma conexionista) como por ejemplo, López Morales (1998), Urrutia (2003), Galloso (2002), Gómez Devís (2003), pero ninguno de ellos establece con precisión cómo son estas redes, qué propiedades formales presentan o cómo se llega a ellas.

Uno de los problemas a que se enfrenta la ciencia cognitiva es el de la representación del conocimiento. Los cognitivistas deben ofrecer teorías que permitan un modelamiento de las representaciones que se utilizarán. Actualmente dos enfoques intentan dar solución al problema de las representaciones: el simbólico y el conexionista. El primero concibe la cognición como cómputo basado en la ma-

nipulación de símbolos. Por su parte el conexionismo, que es un tipo especial de asociacionismo, se basa en las asociaciones entre diferentes tipos de información modeladas en redes neuronales artificiales. Aunque se presentan como paradigmas antagónicos, hay teóricos que intentan una hibridación (Ver por ejemplo Gärdenfors, 2004: 1).

Este es pues el problema del psicolingüista que intenta caracterizar las relaciones que se establecen entre los elementos del léxico disponible: cómo representar las unidades y al mismo tiempo dar cuenta de sus interrelaciones. Una de las respuestas más promisorias es el constructo llamado red semántica.

La idea de representar el significado lingüístico y las relaciones entre conceptos mediante diagramas es tan antigua como Aristóteles. El ejemplo más famoso es el *Arbol de Porfirio* que consiste en una representación gráfica de las *Categorías* de Aristóteles en la forma de un árbol de relaciones (Sowa, 2000: 4 ó Widdows, 2004: 80).

Para nuestros propósitos, elegiremos la noción de *redes semánticas* tal como se ha usado en Inteligencia Artificial. Lehmann (1992) ha definido este constructo en la siguiente forma:

> Una *red semántica* representa el conocimiento como un grafo en forma de red. Una idea, un evento, una situación u objeto tiene casi siempre una estructura compuesta; esto se representa en una red semántica mediante una correspondiente estructura de *nodos* que representan *unidades conceptuales*, y *aristas* direccionadas que representan las *relaciones* entre las unidades (Lehmann, 1992: 2).

Formalmente, nuestras redes serán *grafos* tal como se conciben en la rama de las matemáticas denominada Teoría de Grafos. Un grafo es una colección de puntos u objetos llamados *nodos*. Para entender las relaciones entre estos objetos, ciertos pares de nodos tienen aristas (*links, edges*) que los unen. De este modo un grafo puede considerarse como un típico espacio geométrico (Widdows, 2004: 49).

De acuerdo a la posición de Walter Kintsch, el significado de un concepto/nodo queda definido por su posición en la red cognitiva de la cual participa, es decir, por la fuerza de conexión con los nodos vecinos, tanto los inmediatos como los más lejanos (Kintsch, 1998: 74). El significado se construye en cada instancia de uso, quedando expuesto, por tanto, a una variabilidad importante. Diversos factores contextuales influirán para determinar qué nodos asociados serán activados y, en consecuencia, cuál será el significado del concepto en una instancia particular. Para este autor (Kintsch 1998: 75) hay, sin embargo, una subestructura semántica o red cognitiva básica que subyace al concepto y que permanece en el tiempo, aunque la experiencia y el aprendizaje están continuamente modificándola.

La concepción constructivista conexionista de Kintsch es la que nos servirá de fundamento teórico para interpretar los grafos de léxico disponible que obtendremos de nuestro programa. Veamos en qué forma.

El sujeto enfrentado a la tarea de producir un léxico asociado a un centro de interés específico activará probablemente ciertas estructuras de su memoria semántica ligadas al estímulo dado: profesiones y oficios, problemas del ambiente, etc., y seleccionará así uno o varios nodos/palabras que irá entregando en su encuesta, convirtiéndose cada nodo/palabra en un facilitador para la aparición del próximo nodo/palabra. Debido a esto, cada término al que accede tiene a su vez su propia constelación de nodos, por lo que la secuencia léxica que producirá es difícilmente predecible en su composición y ordenación. (Aquí se habla de la respuesta de un solo sujeto, pero las constelaciones de nodos se producen sólo cuando el grafo está hecho con la respuesta de más de dos sujetos, de lo contrario sólo se obtiene un grafo lineal).

Es evidente que esta descripción no pretende explicitar el proceso psicológico real del acceso al lexicón mental del individuo. Se han propuesto diversos modelos de acceso léxico, pero aún no hay consenso respecto al mismo. En su disertación doctoral, Hernández (2005) realiza una extensa discusión del tema.

Sin embargo, existe un modelo de acceso al léxico que podría explicar en parte los procesos que ocurrirían en los sujetos mientras se enfrentan a una prueba de producción de léxico disponible. Nos referimos al modelo de Levelt (1999, 2001; Levelt, Roelofs y Meyer, 1999), principalmente porque está hecho para explicar la producción de palabras aisladas (como es el caso de la disponibilidad léxica) y porque se basa en la dicotomía ontogénica entre el sistema conceptual y el sistema articulatorio. Recordemos que el input o estímulo (centro de interés) recibido por los sujetos es semántico y Levelt sitúa a la base conceptual en el primer nivel de la serie; después se produce una selección léxica por adecuación al contexto para, posteriormente, elegir la forma de la palabra y producirla.

Ahora bien, se sabe que las respuestas de los sujetos podrían variar según las características extralingüísticas que éstos poseen. No obstante, existe una tendencia en los diferentes tipos de sujetos a producir ciertas palabras unidas a otras, independientemente de su propio decurso al responder la encuesta, lo que nos lleva a pensar en la existencia de un cierto patrón de ordenamiento léxico.

Resulta natural a estas alturas preguntarnos cómo será posible establecer las relaciones semánticas entre los vocablos disponibles si cada sujeto sigue su propio decurso al responder la encuesta. La respuesta está en el algoritmo en que se basa el programa y que será descrito en la sección siguiente.

1. ESTRUCTURA Y FUNCIONES DEL PROGRAMA DISPOGRAFO

El objetivo central del programa informático que hemos desarrollado es representar las relaciones semánticas que se establecen entre los vocablos disponibles, mediante grafos generados automáticamente de acuerdo a un algoritmo definido.

Las entradas del programa están constituidas por los protocolos elicitados de los sujetos mediante encuestas de disponibilidad léxica en diversos centros temáticos.

En los grafos resultantes, los nodos son vocablos disponibles y las aristas representan las relaciones semánticas entre ellos.

Las *vecindades* (neighbors) y *agrupaciones* (clusters) definidas en los grafos expresan valores semánticos tanto de unidades (vocablos) como de conjuntos (categorías). En forma inversamente proporcional, la longitud de las aristas expresa la fuerza de la relación entre nodos. Esta propiedad puede aparecer cuantificada como un *peso* asignado a la arista.

Para entender mejor la estructura de los grafos generados por nuestro programa, se explicará en detalle el algoritmo utilizado en la construcción del grafo.

Los datos de entrada se encuentran dispuestos de la siguiente manera: por cada individuo tenemos una sucesión de palabras que corresponden a las respuestas entregadas para cada centro de interés. Si se quiere crear el grafo de un centro de interés C1, se buscan dichas sucesiones para todos los individuos. Hay que destacar que se pueden filtrar los individuos sobre la base de ciertos criterios tales como sexo, nivel sociocultural u otros.

Una vez filtrados los datos de entrada por centro de interés y características de los individuos, se procede a leer las palabras del primer individuo. Al leer la primera palabra se crea un nodo con dicha palabra. Luego con cada palabra se va creando un nuevo nodo, y una arista que una el nuevo nodo con el anterior. Esta arista tendrá un peso 1. Por lo anterior, siempre que leamos las palabras del primer individuo se creará un grafo con forma de línea recta y con todas sus aristas con peso 1, debido a que las palabras no se pueden repetir.

Al pasar al segundo individuo, se toma la primera palabra P1 y a continuación se pueden seguir dos caminos:

1. Si esta palabra ya se había ingresado al grafo, no se cambia nada, pero se debe recordar P1 para el siguiente paso.
2. Si la palabra no se había ingresado, se agrega un nodo con P1 y se almacena en memoria.

Luego se toma la segunda palabra P2, y se pueden seguir dos caminos:

1. Si la palabra P2 ya existía, se subdivide el algoritmo en dos pasos:
   1.1. Si existe una arista entre P1 y P2, se aumenta el peso de dicha arista
   1.2. Si no existe una conexión entre P1 y P2, se crea dicha arista.

2. Si la palabra P2 no existe, se crea el nodo con P2 y se genera una arista entre P1 y P2.

Este algoritmo se repite hasta que no queden individuos en el archivo de entrada[2].

El resultado final es un grafo de variados niveles de complejidad. A modo de ilustración, véase el siguiente ejemplo:

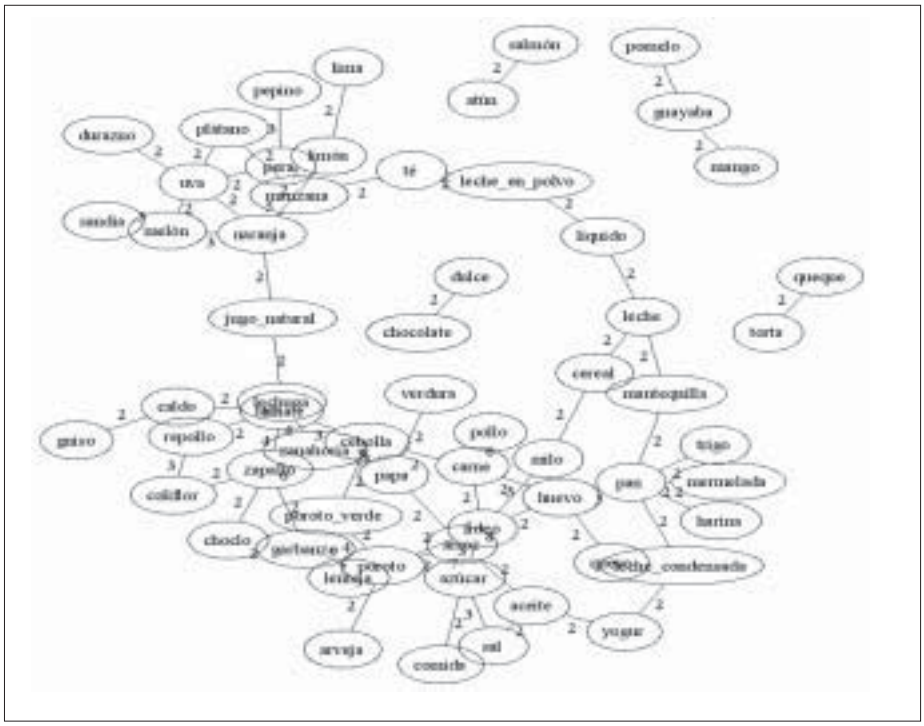

Figura 1. Grafo construido a partir de las respuestas de 23 sujetos en el centro *Alimentos*.

Además de la construcción de un grafo total, el software tiene otras funcionalidades para el posterior análisis de éste. A continuación se detallarán las principales funciones de DispoGrafo, las que serán numeradas para que el lector pueda ubicarlas dentro de la Figura 2.

a) Obtención de un subgrafo: Un subgrafo es un subconjunto de nodos y aristas de un grafo.
b) Funciones de *poda* de aristas y nodos. Para realizar análisis y para la visualización de grafos muy grandes es necesario podar el grafo para limpiar los datos de menor peso y dejar sólo las estructuras más relevantes.

[2] Para mayor información técnica véase la tesis de grado de Vargas (2006).

c) Longitud de las aristas y distancia entre nodos. Una función de *escala* permite dimensionar mejor cada grafo.
d) Generación de un grafo visual: Los grafos se almacenan en una estructura propia del software, diseñada pensando en la optimización de los algoritmos de análisis. Debido a esto, es necesario aplicar un algoritmo que convierta los grafos en otra estructura, la cual podrá ser impresa en un archivo de imagen.
e) Obtención de Cliques: Se diseñó un algoritmo que obtenga conjunto de nodos completamente conectados, o sea, que cada nodo esté conectado mediante una arista con el resto.
f) Obtención de Curvatura: El software posee un algoritmo que es capaz de encontrar la *curvatura* de cada uno de los nodos, esto es el grado de interrelación que presentan entre sí los nodos adyacentes al nodo principal. La noción geométrica de *curvatura* es compleja y puede verse en Widdows (2004:123). La importancia de esta noción para el análisis de los grafos de disponibilidad es aún una tarea en estudio.

Otra funcionalidad que no es propia del software en sí, sino de la interfaz que se creó para el sistema operativo Windows, es la posibilidad de crear filtros propios (g). Esto es de gran utilidad ya que no todos los datos de entrada se encuentran catalogados con los mismos parámetros.

Para mayor claridad respecto a cómo se comporta el programa, entregamos a continuación una pantalla de la interfaz del mismo.

Figura 2. Interfaz del Generador de Grafos.

2. ANALISIS DE UN GRAFO

Sólo a modo de ilustración analizaremos el grafo que entregáramos más arriba (Figura 1).

Este es un grafo pequeño basado solamente en un conjunto de 23 sujetos. Para comodidad del lector copiaremos aquí el grafo mencionado.

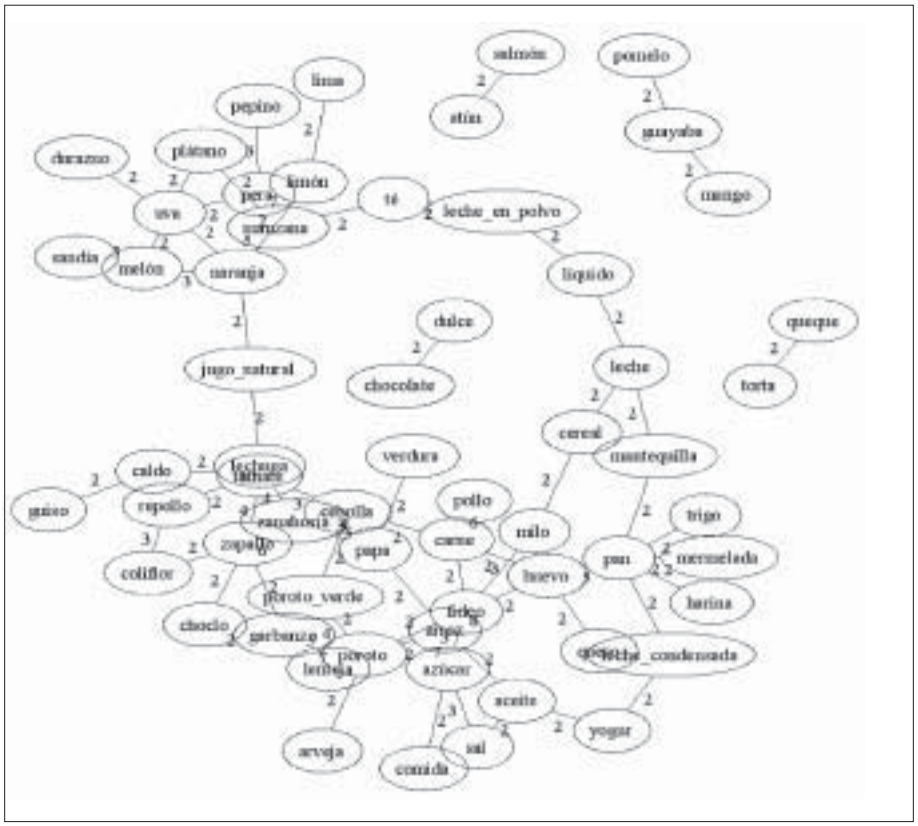

Figura 3. Centro de interés *Alimentos* (23 sujetos, poda 1).

El grafo original ha sido podado de las aristas con peso 1 y eliminados los nodos huérfanos (sin aristas). La estructura general muestra categorías fácilmente reconocibles tales como ***frutas*** (superior izquierda), ***frutas tropicales*** (superior derecha): *pomelo, guayaba, mango*. Como subconjuntos aparte aparecen *salmón, atún* (superior central), y *queque, torta* (medio derecha). Alrededor de *lechuga* y *tomate* se agrupan las ***hortalizas*** (inferior izquierdo).

La complejidad nodal del grafo de la Figura 3 en su parte inferior nos hace necesario hacer zoom sobre las agrupaciones más importantes, lo que se logra mediante poda de las aristas con peso 2. El resultado será el grafo de la Figura 4 a continuación.

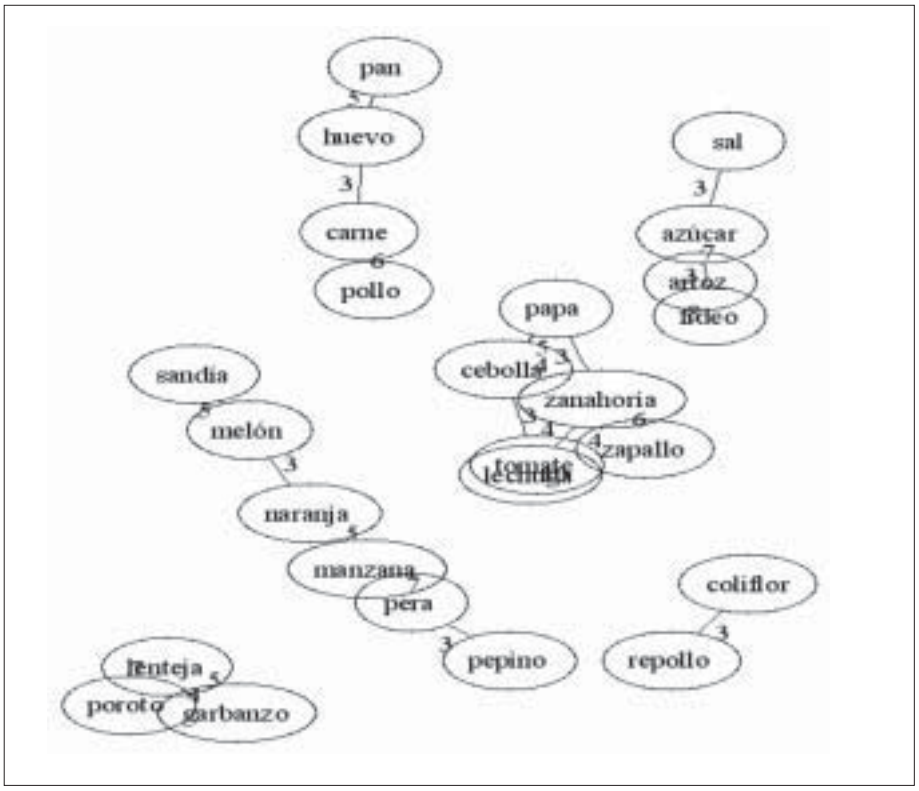

Figura 4. Centro de interés *Alimentos* (23 sujetos, poda 2).

Ahora se puede visualizar mejor la categoría ***hortalizas*** (en posición central) con *papa, cebolla, zanahoria, zapallo, tomate y lechuga*, con núcleo en estas dos últimas. A la izquierda abajo aparecen claramente las ***legumbres***, y a la derecha es interesante observar el par autónomo de *coliflor* y *repollo*.

Es posible que algunas asociaciones presentes en los grafos comentados sean extrañas o "dudosas" para el lector, pero debe tenerse en cuenta que este no es un grafo representativo del centro en estudio ya que sólo se han ingresado los datos de 23 sujetos. Cuando se trabaja con 200 ó 400 sujetos, los grafos tienden a presentar estructuras más decantadas y representativas.

3. CONCLUSIONES

El estudio de la disponibilidad léxica que era hasta ahora fundamentalmente cuantitativo, podrá en el futuro entrar al área cualitativa de las relaciones semánticas presentes en los datos recogidos, gracias a DispoGrafo, una aplicación informática descrita en este trabajo.

DispoGrafo permitirá diversas observaciones cualitativas que hasta el momento no habían sido posibles de forma automática, sin edición o intervención de los datos básicos. Algunas de las más importantes son la *categorización*, a través de las agrupaciones nodales claramente visibles en el grafo; la configuración de *núcleos* y *vecinos* inmediatos; la *fuerza de la relación* establecida por el peso de las aristas, etc. No menos importantes serán los análisis de la incidencia de las variables sexo, nivel sociocultural, ubicación geográfica u otras, en la configuración semántica de los mismos centros de interés. Nuestras incursiones iniciales en el tema nos indican que ciertos centros son mucho más sensibles que otros a diferenciar las configuraciones semánticas según las variables intervinientes. DispoGrafo permitirá indagar sobre la naturaleza de las relaciones entre los nodos ya que ésta puede ser semántica, fonológica o morfológica. Si se toman determinadas categorías y se estudian por separado se pueden determinar más específicamente las relaciones entre las palabras. Por otra parte, se pueden tomar determinadas palabras representativas de algunas categorías en un grupo de individuos y compararlas con las de otro grupo para establecer el nivel de consolidación que tienen estos términos en diferentes grupos de individuos.

Esperamos continuar nuestras investigaciones en esta y otras direcciones.

4. REFERENCIAS

Ayora, C. 2006. *Disponibilidad léxica en Ceuta: Aspectos sociolingüísticos.* Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

Galloso, V. 2002. *El léxico de los estudiantes preuniversitarios en el distrito universitario de Salamanca (Avila, Salamanca y Zamora).* Salamanca: Ediciones Universidad de Salamanca.

Gärdenfors, P. 2004. *Conceptual Spaces: The Geometry of Thought*. Cambridge: MIT Press.

Gómez Devís, M. B. 2003. La disponibilidad léxica de los estudiantes preuniversitarios valencianos: reflexión metodológica, análisis sociolingüístico y aplicaciones. Tesis doctoral. Valencia, España: Universidad de Valencia.

Hernández, N. 2005. Hacia una teoría cognitiva integrada de la disponibilidad léxica: El léxico disponible de los estudiantes castellano-manchegos. Tesis doctoral. Salamanca, España: Universidad de Salamanca.

Kintsch, W. 1998. *Comprehension. A paradigm for cognition*. Cambridge: Cambridge University Press.

Lehmann, F. 1992. "Semantic networks", en F. Lehmann (Ed.) *Semantic Networks in Artificial Intelligence*. Oxford: Pergamon Press.

Levelt, W. J. L. 1999. "Models of word production", en *Trends in cognitive science* 3 (6), 223-232.

Levelt, W. J. L. 2001. "Spoken word production: a theory of lexical access", en *PNAS Proceedings of the National Academy of Sciences* 98 (23), 13464-13471.

Levelt, W. J. L., Roelofs, A. & Meyer, A.S. 1999. "A theory of lexical access in speech production", en *Behavioural and Brain Sciences* 22, 1-75.

López Morales, H. 1984. *La enseñanza de la lengua materna. Lingüística para maestros de español*. Madrid: Playor.

López Morales, H. 1998. "Los estudios de disponibilidad léxica: pasado y presente", en *Boletín de Filología de la Universidad de Chile* 35, 245-259.

Samper, J.A., Hernández, C.E. & Bellón, J.J. 2003. "El proyecto de estudio de la disponibilidad léxica en español", en R. Avila, J.A. Samper, H. Ueda *et alii* (Eds.) *Pautas y pistas en el análisis del léxico hispano(americano)*.

Samper, J. A. & Samper, M. 2006. "Aportaciones recientes de los estudios de disponibilidad léxica", en. *LynX, Panorámica de Estudios Lingüísticos* 5, 5-95.

Sowa, J. F. 2000. *Knowledge Representation. Logical, Philosophical and Computational Foundations*. Pacific Grove, CA.: Brooks Cole.

Urrutia, M. 2003. "Redes semánticas en línea: una tarea de acceso léxico a partir de un estudio experimental", en *RLA Revista de Lingüística Teórica y Aplicada* 41, 119-141.

Valencia, A. & Echeverría, M. S. 1999. *Disponibilidad léxica en estudiantes chilenos*. Santiago de Chile y Concepción: Universidad de Chile y Universidad de Concepción.

Vargas, Roberto. 2006. Software de generación automática de grafos. Tesis de Ingeniería Informática. Concepción, Chile: Universidad de Concepción.

Widdows, D. 2004. *Geometry and Meaning*. Stanford: CSLI Publications.