UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA
Departamento de Ingeniería del Software e Inteligencia Artificial

MODELO COMPUTACIONAL DE LECTURA COGNITIVA PARA LA REPRESENTACIÓN AUTOMÁTICA DE TEXTOS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

José Ignacio Serrano Moreno

Bajo la dirección de la doctora
María Dolores del Castillo Sobrino

Madrid, 2007

- **ISBN: 978-84-669-3162-5**

Universidad Complutense Madrid

Facultad de Informática

isia

Modelo Computacional de Lectura Cognitiva para la Representación Automática de Textos

Tesis Doctoral

JOSÉ IGNACIO SERRANO MORENO

Ingeniero en Informática

Año 2007

Modelo Computacional de Lectura Cognitiva para la Representación Automática de Textos

TESIS DOCTORAL

José Ignacio Serrano Moreno

Ingeniero en Informática

Año 2007

Universidad Complutense de Madrid

Facultad de Informática

Departamento de Ingeniería del Software
e Inteligencia Artificial

Modelo Computacional de Lectura Cognitiva para la Representación Automática de Textos

TESIS DOCTORAL

Programa:

Sistemas Informáticos y Programación

Autor:

José Ignacio Serrano Moreno

Directora:

María Dolores del Castillo Sobrino

Año 2007

A Juana, allá donde estés

Puesto que nadie camina a solas por los laberintos de la ciencia, ya sea en cuerpo o en espíritu, me gustaría dar las gracias al Consejo Superior de Investigaciones Científicas, y más concretamente al equipo directivo del Instituto de Automática Industrial, por su apoyo y financiación. Así mismo, desearía mostrar mi más sentido agradecimiento a mis padres y mis hermanos, por su sempiterna confianza y dedicación, a mis amigos, por su incomprensible aunque alentadora fe ciega y apoyo incondicional, a mis colegas y compañeros, por hacer de lo cotidiano un alivio y por su disponibilidad intemporal, a mi directora, por poner los pilares tanto técnicos como estéticos y éticos de lo que esta tesis y yo mismo somos en este momento, a todo el que me tendió su mano, a todo el que me dio aliento, a todos los que me regalaron consejos…

...y por supuesto, a Almudena, por absolutamente todo y más aún.

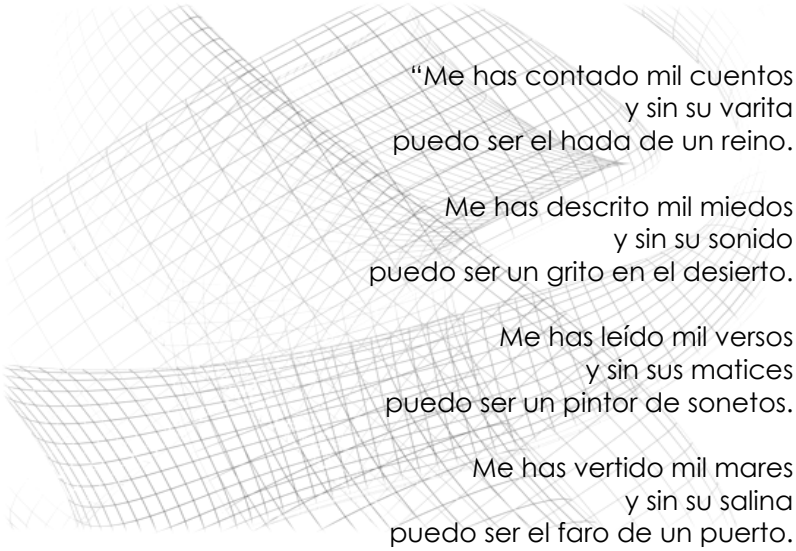
"Cuando el amor hiere cualquiera se hace poeta, aunque nunca antes hubiese sido favorecido por las musas"

Platón (427/428 a.C. – 347 a.C.)

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS



"Me has contado mil cuentos
y sin su varita
puedo ser el hada de un reino.

Me has descrito mil miedos
y sin su sonido
puedo ser un grito en el desierto.

Me has leído mil versos
y sin sus matices
puedo ser un pintor de sonetos.

Me has vertido mil mares
y sin su salina
puedo ser el faro de un puerto.

Me has descrito mil tiempos
y sin su fragancia
puedo ser la bodega de un pueblo.

Y ahora que no me hablas
ahora
en este silencio
voy a hacer un viaje al lugar
de tus sueños."

M. D. C.

Resumen

El modelado del lenguaje natural en los ordenadores conlleva ciertas restricciones debido a la estructura lógica y a las limitaciones de tiempo y espacio de las máquinas, además de la complejidad intrínseca del lenguaje. Uno de los mayores problemas de dicho modelado es la representación de la semántica. Los primeros modelos conexionistas del lenguaje se situaban próximos a la cognición humana pero no eran lo suficientemente generales y eficientes para aplicaciones reales. Estos primeros sistemas de procesamiento de lenguaje natural hacían uso de redes de asociación como formalismo de representación. Debido a las limitaciones de almacenamiento y procesamiento de los ordenadores de aquella época, y al crecimiento de la información textual almacenada electrónicamente, los sistemas de procesamiento del lenguaje adoptaron formalismos matemáticos y estadísticos. Hoy en día, a causa de esa cantidad creciente de información textual los sistemas que son capaces de procesar textos son de extrema utilidad. Hasta hace relativamente poco tiempo, la mayoría de estos sistemas utilizaban la clásica representación de los textos como "bolsa de palabras", un formalismo de tipo vectorial que sólo tiene en cuenta las apariciones de las palabras de manera independiente.

A mediados de los noventa, surgen los hiperespacios de palabras como un formalismo de representación alternativo al de "bolsa de palabras" tradicional. LSA (Análisis de Semántica Latente) fue el precursor de todos ellos, seguido por HAL (Hiperespacio Análogo al Lenguaje), PMI-IR, Indexado Aleatorio, WAS (Espacio de Asociación de Palabras) o ICAN (Construcción Incremental de una Red Asociativa), entre otros. Este tipo de sistemas construyen una representación en forma de matriz del conocimiento semántico lingüístico almacenado en una colección de textos dada. Este hiperespacio tiene en cuenta las relaciones entre las palabras y el contexto sintáctico y semántico en el que aparecen. Sin embargo, estos sistemas también representan los textos como vectores, llevando a cabo operaciones con las filas y las columnas de la matriz correspondientes a las palabras de los documentos. Aunque la representación mediante hiperespacios contiene mucha más información que la representación tradicional, puesto que los valores de los vectores son el resultado de la interacción entre las palabras y el contexto, los textos siguen siendo presentados como un conjunto de números sin estructura. A pesar de ello, los sistemas basados en hiperespacios han aportado una mejora significativa con respecto a los sistemas basados en la representación clásica. De los sistemas anteriormente mencionados, sólo ICAN introduce una representación estructural, almacenando el conocimiento en forma de red contextual asociativa de palabras y no como una matriz. Este modelo, a diferencia del resto de sistemas mencionados, hace posible la actualización del conocimiento sin necesidad de la reconstrucción total del mismo.

A pesar del progreso realizado utilizando los hiperespacios de palabras, los seres humanos continúan realizando tareas de procesamiento de lenguaje natural, como la clasificación de textos o la recuperación de información, de manera mucho más precisa que los ordenadores aunque, por supuesto, más despacio. Es difícil concebir el conocimiento lingüístico representado como una matriz en el cerebro humano, así como que la lectura suponga realizar operaciones matemáticas sobre dicha matriz. La lectura es un proceso secuencial de percepción en el tiempo, durante el cual los mecanismos

mentales construyen imágenes e inferencias que se van reforzando, actualizando o descartando hasta la conclusión de la lectura del texto, momento en el que la imagen mental generada permite a los seres humanos resumir o clasificar el texto, recuperar documentos similares o simplemente expresar opiniones sobre el mismo. Esta es la filosofía que subyace en el sistema presentado en esta tesis. Este sistema, denominado SILC (Sistema de Indexación por Lectura Cognitiva), está ligeramente inspirado en el formalismo que sugiere el sistema ICAN. Lo que se propone en este trabajo de tesis doctoral es un modelo computacional de lectura que construye una representación de la semántica de un texto como resultado de un proceso en el tiempo. Dicha representación posee una estructura que posibilita la descripción de las relaciones entre los conceptos leídos y su nivel de significación en cada momento del proceso de lectura.

Existen otros modelos computacionales de lectura cuyo objetivo es más teórico que aplicado. La mayoría de ellos parten del modelo conexionista de Construcción-Integración y se centran en diferentes fases u objetivos de la lectura. Todos estos sistemas ponen de manifiesto la gran variedad y complejidad de los procesos cognitivos implicados en la lectura. El modelo propuesto en esta tesis, SILC, es un método sencillo que incluye sólo algunos de dichos procesos cognitivos y, aunque trata de ser útil en aplicaciones prácticas, está inspirado en los seres humanos tratando de asemejarse más a su proceder que el resto de sistemas del mismo campo de aplicación.

El modelo que implementa SILC intenta simular, en parte, procesos cognitivos de alto nivel que operan en el tiempo. Primero, el sistema construye una red de asociación conceptual como una memoria lingüística base a partir de una colección de textos que representan el espacio de conocimiento semántico. A continuación, el modelo genera representaciones de los textos de entrada como redes de conceptos con niveles de activación, que recogen el nivel de significación semántica de los mismos. Para ello, el modelo utiliza el conocimiento semántico lingüístico previamente construido realizando inferencias sobre el mismo mediante la propagación por la red de la activación de los conceptos leídos en orden secuencial. La representación generada se usa posteriormente para indexar documentos con el fin de clasificarlos automáticamente. Los métodos de indexación tradicionales representan los textos como resultado de procesos matemáticos. Puesto que los seres humanos superan ampliamente a los ordenadores en tareas de procesamiento de lenguaje natural, el modelo de SILC se inspira en la cognición humana para mejorar su eficacia en dichas tareas. Se han realizado experimentos para comparar el modelo con sujetos humanos, tanto durante la lectura, mediante la predicción o inferencia de conceptos, como al final de la misma, mediante la comparación con resúmenes generados por los sujetos. Los resultados muestran que el sistema es adecuado para modelar de manera aproximada el proceder humano en la lectura y sustentan la hipótesis de partida de SILC: cuanto más se asemeje el sistema a los seres humanos, mejor realizará las tareas prácticas del lenguaje. Los resultados también demuestran que el sistema es adecuado como marco experimental de validación de hipótesis relacionadas con aspectos cognitivos de la lectura. Otros experimentos de aplicación práctica han mostrado que, una vez que los parámetros del modelo han sido optimizados, la representación generada obtiene mejores resultados en clasificación de textos que otras representaciones generadas por los sistemas existentes. Se han definido tres medidas de similitud semántica entre textos a partir de las representaciones generadas por SILC. Los resultados experimentales muestran que la mejor de ellas es más eficaz y eficiente que otras medidas de similitud existentes. Además, la sinergia de dicha medida con el modelo de lectura implementado hace a SILC apropiado para su aplicación a tareas reales de procesamiento de lenguaje natural.

Abstract

Modelling of natural language in computers implies some restrictions due to the logical structure and to the time and space limitations of machines, in addition to the language complexity itself. The computational representation of language semantics is one of the major problems. Connectionist models of language are close to human cognition but they are not general and efficient enough for real applications. The first natural language processing systems made use of association nets or graphs, generally speaking, as representation formalism. Due, in one hand, to the storage and processing limitations of the computers in that time and, in the other hand, to the need of processing a growing amount of textual information electronically stored, the language processing systems adopted mathematical and statistical formalisms. Nowadays, because of this huge amount of digital information stored in natural language, systems that automatically process text are of crucial importance and extremely useful. Until fairly recently, most of the systems used the highly common electronic text representation, "bag of words". No information other than independent occurrences of words is considered in this latter vector-like formalism.

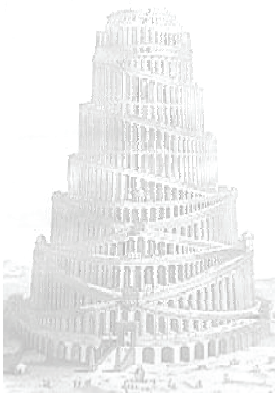
In the mid-nineties, word hyperspaces were proposed as an alternative to the traditional "bag of words" approach. LSA (Latent Semantic Analysis) was the first of these systems, followed by HAL (Hyperspace Analogue to Language), PMI-IR, Random Indexing, WAS (Word Association Space) or ICAN (Incremental Construction of an Associative Network), among others. These kind of systems build a representation, a matrix, of the linguistic knowledge contained in a given text collection. The representation, or hyperspace, takes into account the relationship between words and the syntactic and semantic context where they occur, and this is the main difference with the common "bag of words" representation. However, once the hyperspace has been built, word hyperspace systems represent the text as a vector, and by doing operations with the rows and the columns of the matrix corresponding to the words in the texts. Although the hyperspace representation contains much more information than the traditional representation because the vector values are the result of word and context interaction, texts are still a set of numbers without a structure. However, this approach has been shown to be a real improvement on the classical representation. Only ICAN introduces a structural representation and does not store linguistic knowledge as a matrix but as a net of associated words. This model makes it possible to incrementally add new words without retraining and recalculating the knowledge, which is psychologically more plausible. This approach proposes the representation of linguistic knowledge as a net of concepts associated by context.

In spite of the progress made with word hyperspaces, human beings continue to do text classification and information retrieval tasks much better than machines, although of course more slowly. It is hard to believe that linguistic knowledge is represented as a matrix in the human mind and that reading implies mathematical operations on this matrix. Human reading is a process of sequential perception over time, during which the mind builds mental images and inferences which are reinforced, updated or discarded until the end of the text. At that moment, this mental image allows humans to summarize and classify the text, to retrieve similar texts or simply to talk about the text by expressing opinions. The latter dynamic is the one in which the system presented in this thesis relies. This system, called CRIM (Cognitive Reading Indexing Model), is

inspired by the ICAN connectionist approach, where words and texts do not share the same structure of representation unlike the systems mentioned above. What is proposed in the PhD work is to build text representations as a result of a process over time, with a structure that makes it possible to indirectly describe the salience and relations of words at every instant during the reading process.

Other computational models of reading exist which search for an assessment of a theory of reading rather than for a real data-intensive application. Most of them are based on connectionist networks inspired by the Construction-Integration model and they focus on different stages of reading and targets. These systems just mentioned show that there is a high number of complex cognitive processes underlying reading. The model proposed in this PhD thesis, so called CRIM, is a simple model that takes into account only a few cognitive processes and although it is aimed at a real application, it is inspired by and closer to human procedure than the other systems in the same application field.

The CRIM tries to simulate in part the high-level cognitive processes in human mind over time. First, the system builds a conceptual association net from a collection of texts representing the semantic knowledge space, as a linguistic base memory. Then the model generates a representation of the input text as a net of concepts, and each concept has an activation value referring to its salience in the text. This representation is then used to index documents in order to automatically categorize them by a supervised learning algorithm. Traditional indexing methods represent texts as the result of a process of mathematical operations. Since humans are able to classify texts much better than machines, the model is inspired in human cognition in order to improve language tasks. Some experiments were carried out to compare the model with humans, either during the reading process by concept prediction, or at the end of the reading process by summary comparison. The results showed that the system is suitable to model human reading process and proved the base hypothesis in which CRIM relies: the closer the system is to human being procedures, the better it performances in natural language processing tasks. Results also make the system suitable as an experimental framework to test hypothesis about other cognitive aspects of reading. Other applied experiments show that, once the model parameters have been optimized, the representation obtained is an improvement on traditional indexing techniques. Given this representation, two different similarity measures between texts have been defined. The similarity measures are based in the distance between the single concepts of the texts and also on the difference of activation or significance. The distance between single concepts is calculated in the context defined by the compared texts and not in the global semantic net. The results show that the latter reduction improves both the efficiency and the accuracy of the comparison method, which is an improvement over other existent similarity measures between words. Other results of comparison between texts of different categories indicate that the synergy between the proposed model and similarity measures are very suitable to be applied in text categorization and information retrieval tasks.



CAPÍTULO 1

Introducción

El estudio del lenguaje ha estado siempre en el punto de mira de los investigadores científicos, aunque comenzó a cobrar mayor interés desde la aparición de los primeros ordenadores y alcanzó su auge, todavía en vigor, coincidiendo con los comienzos de la Inteligencia Artificial (IA). La pretensión inicial de los precursores de la IA era lograr que las máquinas suplantaran en cierto modo a los seres humanos en la realización de actividades intelectuales mediante su automatización. Después de conseguir en parte su objetivo para ciertas tareas sencillas, en las que las máquinas operaban de manera autónoma con un rendimiento incluso mejor al de los seres humanos, la comunidad científica se percató de que, pese a su eficacia y su eficiencia, los ordenadores necesitaban comunicarse con los seres humanos en su mismo lenguaje para asemejarse más a ellos y para que éstos, a su vez, les otorgaran cierta confianza en su autonomía. Era pues necesario diseñar y desarrollar el lenguaje natural en los ordenadores, tratando de incorporar a la teoría Computacional toda la teoría de la Lingüística conocida, surgiendo así un nuevo campo de investigación denominado Lingüística Computacional. Esta aparición dió un impulso a la Lingüística pura mediante el replanteamiento de lo que realmente se conocía sobre el lenguaje, motivado principalmente por las restricciones contextuales computacionales. Puesto que los ordenadores proporcionan un soporte de cómputo con unas restricciones de diseño y ciertas limitaciones de espacio y tiempo, fue necesario adaptar los procesos lingüísticos conocidos a la arquitectura de las máquinas. Esta tarea se antojaba fácil a priori pero posteriormente se mostró casi imposible de manera global, requiriendo el diseño de nuevos formalismos de representación y procesos artificiales que limitaban y distorsionaban la teoría lingüística. La conclusión de esta etapa fue que la capacidad de los ordenadores no era en absoluto comparable a la del cerebro humano, donde se procesa el lenguaje. Así pues, a pesar del avance de la tecnología, de las

representaciones y de los algoritmos de procesamiento de lenguaje natural, las máquinas no han conseguido ni siquiera igualar a los seres humanos en muchos de los aspectos del lenguaje, lo que conduce a un intento de emulación de los procesos mentales que tienen lugar en el ser humano. Este intento sugiere, una vez más, nuevas adaptaciones y preguntas a campos como la Psicología, la Psicolingüística y la Neurociencia, entre otros, perfilando junto con la Lingüística Computacional un nuevo campo de investigación englobado dentro de la Ciencia Cognitiva: la Lingüística Cognitiva.

1.1. Propósito

Muchos de los problemas que originó la incursión de la Inteligencia Artificial en el ámbito del lenguaje natural fueron superados con el desarrollo de los ordenadores, que adquirieron una gran capacidad de almacenamiento y de proceso permitiendo así la aplicación de los algoritmos existentes a grandes cantidades de información lingüística en tiempo real, lo que hacía a los sistemas computacionales de procesamiento de lenguaje natural aptos para su utilización en el mundo real. De hecho, su uso continúa muy extendido en la actualidad, empleándose en diversas tareas de manera eficaz y eficiente entre las que se puede citar, como representantes de la década actual, a los buscadores de Internet (por ejemplo).

Sin embargo, los formalismos de representación y algoritmos existentes hasta principios de los años 90 carecían del uso de conocimiento semántico. Eran algoritmos principalmente basados en métodos estadísticos que "contaban" apariciones de palabras de manera independiente, lo que hace que su eficacia disminuya enormemente en tareas algo más complejas que requieren cierta comprensión del discurso lingüístico para ser llevadas a cabo con éxito. Son pues técnicas "ingenuas", que no pueden reaccionar ante situaciones lingüísticas nuevas puesto que no utilizan el conocimiento semántico.

A principios de esa década, aparecen los sistemas de hiperespacios de palabras. Estos nuevos formalismos de representación eran capaces de considerar ciertos aspectos semánticos y relaciones entre palabras pudiendo tratar, además, con grandes cantidades de información textual. Estos sistemas supusieron un gran avance en la aplicación de los métodos de procesamiento de lenguaje, puesto que estaban preparados para hacer frente a la creciente cantidad de información en lenguaje natural almacenada electrónicamente. Sin embargo, los sistemas de hiperespacios de palabras resultan sumamente rígidos, dando lugar a un alto coste computacional de actualización. Teniendo en cuenta que la actualización de la información hoy en día es un punto crítico para el éxito en muchas de las áreas de negocio existentes, y que la información es accesible y puede ser almacenada por cualquier persona y no sólo por expertos, se hace necesaria la creación de nuevos sistemas de representación y tratamiento del lenguaje natural que sean lo más autónomo posible, flexibles en cuanto a adaptación se refiere, capaces de tratar con

grandes cantidades de información y preparados para entender de una manera más general el lenguaje natural humano, permitiendo así su aplicación a tareas diversas.

1.2. Antecedentes

Uno de los principales propósitos de los sistemas de procesamiento de lenguaje natural es la clasificación automática de textos. Dicha tarea consiste en determinar de manera automática, dado un conjunto de posibles categorías temáticas, a cuáles de dichas categorías, y en qué grado, pertenece un texto dado. Esta tarea no sólo es útil para la organización automática de grandes cantidades de documentos sino también para la posterior realización de consultas de los documentos organizados.

El origen de esta tesis tuvo lugar durante la realización de los proyectos SELECTA, "Clasificación Automática de Contenidos en Internet" (MCYT, PROFIT FIT-070000-2001-193), y CAMOFI, "Construcción Automática de Motores de Filtrado Específicos para Contenidos en Internet" (MCYT. PROFIT FIT-150500-2003-373). El objetivo de los proyectos fue el diseño y desarrollo de un sistema de clasificación y filtrado de páginas web. Es decir, el sistema debía asignar automáticamente una o más categorías a las páginas web accedidas y filtrarlas si las categorías identificadas estaban prohibidas por el usuario. La dificultad radicaba en el tipo de información textual con el que se trató. El lenguaje empleado en las páginas web es heterogéneo y puede provenir de diversas fuentes, por lo tanto puede ser gramatical y semánticamente correcto o no. Además, la división de la información en secciones, en muchos casos dispares, complicaba el proceso de análisis. El sistema desarrollado finalmente [Serrano y Del Castillo, 2007] hacía uso de técnicas clásicas de procesamiento de lenguaje natural, medidas de la Teoría de la Información y algoritmos genéticos, y tuvo que ser completado con heurísticos del dominio, extraídos por seres humanos, para obtener una eficiencia aceptable. Además, fue necesario realizar un estudio de las relaciones entre las palabras que aparecían en los textos y las categorías temáticas a las que pertenecían, seleccionando así el conjunto de términos que mejores resultados de clasificación obtenía [Del Castillo y Serrano, 2004]. Este proceso de selección de términos o características, junto con la aparición de la necesidad de reforzar y ampliar los heurísticos propuestos, pusieron de manifiesto la insuficiencia de la representación clásica para abordar tareas de lenguaje natural ligeramente complejas, donde es necesario tener un cierto conocimiento del discurso. A pesar de esta limitación, la optimización de los métodos de selección de términos y la gran cantidad de ejemplos

disponibles con los que se entrena a los algoritmos de clasificación hacen que los sistemas tradicionales de representación sean suficientes para el éxito de su aplicación en ciertas tareas sencillas y concretas.

La experiencia acumulada hasta el momento con información textual tan confusa como la de las páginas web permitió la realización de los proyectos ICONOSOL, "Sistema para la Identificación, Clasificación y Filtrado de Correo Electrónico" (MCYT, PROFIT FIT-360000-2004-88), y ELECFRA, "Sistema para la Detección y Eliminación del Fraude Electrónico Vía Correo Electrónico e Internet" (MITYC, PROFIT FIT-360000-2005-9). De nuevo, el objetivo fue el filtrado de información textual, en este caso contenida en correos electrónicos. El lenguaje empleado en los correos electrónicos es más incompleto y heterogéneo que el empleado en las páginas web, ya que puede ser muy coloquial e incluso algo críptico. Teniendo en cuenta que se trataba de identificar correos *spam* o correos basura no solicitados, en los que se trata de "engañar" de manera activa a los filtros tradicionales, el objetivo se hizo aún más ambicioso. Dado que sólo existían dos categorías temáticas, *spam*, que puede englobar varios temas, y *no spam*, es decir, una categoría ampliamente general, los heurísticos del dominio son difíciles de determinar y además resultan insuficientes. Así pues, una vez más el sistema desarrollado [Del Castillo y Serrano, 2006] incorporó nuevos heurísticos que tenían que ver, en este caso, con la percepción humana de los textos, y no sólo con el dominio. Los heurísticos empleados solucionaron, en gran medida, el intento de "engaño" por parte de los correos no deseados a los sistemas de filtrado que emplean técnicas de procesamiento de lenguaje tradicionales. Sin embargo, si el objetivo del engaño son los propios usarios y no simples algoritmos computacionales, como es el caso de los correos de fraude electrónico o *phishing*, el problema es distinto. Este tipo de correos aparentan provenir de una entidad bancaria o comercial legítima y solicitan a los usuarios datos confidenciales para después utilizarlos en la extracción no consentida de los bienes monetarios de dichos usuarios. Así pues, el lenguaje empleado en estos correos suele ser correcto y aparentemente normal, por lo que es necesario entender completamente la semántica que conllevan para discernir su naturaleza. Desafortunadamente, este es un objetivo muy exigente para cualquier sistema actual de procesamiento de lenguaje natural, ya que incluso los seres humanos pueden ser y son engañados. De esta manera, el sistema desarrollado utiliza técnicas de procesamiento de

lenguaje natural sólo como complemento al filtrado, empleando métodos de otra índole como núcleo de la identificación del posible fraude.

En definitiva, la investigación llevada a cabo condujo a sistemas híbridos o multiestratégicos integradores, donde la heurística resultante del conocimiento o de la percepción humana pusieron de manifiesto las limitaciones de los métodos de procesamiento de lenguaje natural empleados por dichos sistemas.

1.3. Motivación

Aparte de las limitaciones de los métodos actuales de procesamiento de lenguaje natural mencionadas en la sección anterior, otra de las principales motivaciones del trabajo llevado a cabo en esta tesis doctoral es la inquietud personal por conocer más acerca de la dinámica de las tareas cognitivas de alto nivel, es decir, de las funciones mentales asociadas y de las estructuras del cerebro humano que las subyacen, y que emergen como fruto de la interacción con el medio, entre las que se encuentra el lenguaje [Rubia, 2007]. Dada las dificultades y limitaciones actuales de la tecnología de imagen y de los procedimientos diagnósticos sobre los paradigmas de actividad cerebral relacionados con la lectura y la comprensión del lenguaje, el diseño y construcción de un modelo computacional que permita aislar los procesos, tanto temporal como funcionalmente, puede ayudar a la validación de hipótesis y a la generación de ellas, planteándose así nuevas cuestiones, lo que es a su vez el motor de la ciencia por antonomasia. Además, la naturaleza del trabajo es diversa, por lo que su realización sugiere un enfoque multidisciplinar que requiere una cooperación entre distintas áreas del conocimiento. De hecho, la denominación del trabajo ya denota una intención integradora en cada una de las palabras que la componen: "Modelo Computacional de Lectura". Esta colaboración integradora entre distintos campos tiene también un carácter motivador bajo las creencias de que la evolución de la ciencia actual depende en gran medida de dicha cooperación, y de que toda área de conocimiento aporta una visión distinta sobre el tema objeto de interés.

Desde un punto de vista más práctico, existen otros desafíos que han impulsado la realización de este trabajo de tesis doctoral. Como se desprende de los antecedentes, las exigencias requeridas a los sistemas automáticos de procesamiento de lenguaje natural han ido aumentando en complejidad con el tiempo, hasta el punto de hacerse necesaria una comprensión más general del lenguaje. Además, la información textual almacenada electrónicamente crece de manera exponencial cada día, por lo que los sistemas que la analicen y gestionen deben estar capacitados para tratar con grandes cantidades de información y actualizarse de manera eficiente, demostrando una cierta capacidad de adaptación.

En el marco de los proyectos citados en el apartado anterior siempre se optó por uno de los métodos más comunes de representación del lenguaje, el denominado "bolsa de palabras *tf·idf*". Los modelos de hiperespacios de palabras sí extraen parte de la semántica contenida en los textos y, por tanto, pueden obtener mejores resultados en tareas complejas que requieren cierto nivel de comprensión del lenguaje. Sin embargo, su metodología de funcionamiento no les permite una actualización en absoluto eficiente del conocimiento y de la semántica extraída. Así pues, no son adecuados para sistemas que trabajan en tiempo real y que requieren una actualización diaria para funcionar correctamente, como los filtros, por ejemplo. Esta es la razón por la que se descartó su uso en los proyectos citados.

Como se ha comentado, a medida que se configuraron los proyectos fue necesario incorporar heurísticos de diversa naturaleza. Dichos heurísticos simbolizaban una síntesis de la facilidad con la que el ser humano era capaz de discernir e identificar conceptos críticos para la clasificación. Es pues el entendimiento humano el que ayudó al método computacional en cada caso.

Desde los comienzos de la IA se ha intentado que los ordenadores se comporten de manera similar a los seres humanos. En muchos aspectos han logrado incluso superar al *homo sapiens sapiens* [Rubia, 2007], pero desde luego no en lo que a lenguaje se refiere. Es un hecho que los sistemas computacionales tienen más capacidad de cómputo y de espacio, aunque no son capaces de manejar el lenguaje como lo hace la mente humana. Así pues, es interesante estudiar cómo funciona el lenguaje en el ser humano y tratar de llevar esos principios a un ordenador, con las restricciones que conlleve. Desfortunadamente, muchos de los sistemas recientes abogan por lo contrario, es decir, por la creación analítica de un modelo de lenguaje que luego se atribuye al ser humano si obtienen, en comparación, resultados similares a éste.

La motivación de esta tesis posee un carácter claramente antropocéntrico, que se desprende de preguntas como: ¿cuál es el método de comunicación natural más completo y complejo?, ¿qué especie lo desarrolla desde etapa tempranas de su vida?, ¿quién lo domina?... En las respuestas a estas cuestiones se encuentra la justificación. Así, surge la hipótesis de que cuanto más se asemeje un modelo computacional a lo que se conoce sobre el lenguaje en el ser humano mejor realizará las tareas para las que se diseña.

1.4. Metodología

La gestión y procesamiento del lenguaje se divide comúnmente en la literatura en tres etapas o facetas: adquisición, comprensión y generación. La etapa de adquisición consiste en el aprendizaje del lenguaje, desde la distinción de fonemas pasando por la asociación entre conceptos y palabras hasta las relaciones sintácticas y semánticas entre las mismas (gramática). En la etapa de comprensión se utiliza el conocimiento linguístico adquirido para entender el sentido de las expresiones del lenguaje. En la etapa de generación se utiliza de nuevo el conocimiento linguístico, junto con la intenciones, para producir expresiones linguísticas que contengan la información que se desea comunicar.

Los sistemas de procesamiento de lenguaje natural se caracterizan, entre otros factores, por el nivel de desarrollo de cada una de las tres facetas anteriores. El diseño de modelos que contemplen completamente las tres etapas o facetas es actualmente una meta muy ambiciosa. De hecho, la mayoría de sistemas existentes se centran en una sola de las mismas. El sistema propuesto en este trabajo de tesis pretende ser un substrato para las tres facetas, aunque la generación de lenguaje no está incluída explícitamente en las pretensiones de la misma.

Intuitivamente, las tres facetas son distinguibles e identificables por cualquier ser humano. Sin embargo, ¿cómo se traducen a un modelo computacional? Para modelar la etapa de adquisción es necesario definir un modelo de representación del conocimiento linguístisco que involucra las siguientes tareas:

- Determinación del conocimiento linguístico que va a ser adquirido y almacenado.
- Diseño de un formalismo de representación del conocimiento linguístico.
- Diseño de mecanismos de construcción, actualización (ampliación, modificación o y reducción) y acceso a la información representada por el formalismo.
- Determinación de las fuentes de las que se extrae el conocimiento linguístico y los mecanismos para realizar dicha extracción.

Estas tareas se antojan complicadas por diversos motivos, entre los que se encuentra el condicionamiento mutuo entre el formalismo de representación y sus mecanismos de construcción y actualización, entre la determinación del conocimiento linguístico y las fuentes de las que se extrae, y entre todos los aspectos anteriores y la cantidad de información que se pretende manejar de manera eficiente.

La etapa de comprensión conlleva tareas parecidas a las mencioandas anteriormente, aunque la primera de todas ellas es determinar qué es comprender el lenguaje en el ámbito de aplicación del modelo que se está desarrollando. En la mayoría de los casos, esta definición de la comprensión deriva en el diseño de un formalismo de representación que contenga la semántica comprendida. Así pues, se pueden citar las siguientes tareas:

- Determinación de la información que va a representar la semántica comprendida, dependiendo del objetivo del sistema.
- Diseño de un formalismo de representación para almacenar dicha semántica y los correspondientes mecanismos de acceso a la misma.
- Diseño de los mecanismos para construir la representación de la semántica, dadas las expresiones en lenguaje natural que se pretenden comprender y dado el conocimiento linguístico previamente adquirido.

Por último, es necesario idear los métodos que empleen la representación resultante de la comprensión para lograr el objetivo de aplicación final del sistema. Estos objetivos comprenden desde la clasificación de textos hasta la propia generación de lenguaje, con la salvedad de que en esta última se ha de hacer uso del conocimiento linguístico adquirido, además de que la representación semántica que utilice puede haber sido generada por otros medios distintos al de la comprensión.

El modelo computacional de representación del lenguaje presentado en esta tesis empleará una red asociativa como formalismo de representación del conocimiento linguístico, que construirá a partir de colecciones de textos mediante la relación semántica contextual de las palabras que aparecen en los mismos. Dicha relación vendrá dada por la concurrencia de las palabras en el contexto definido. El proceso de comprensión será realizado mediante un modelo del proceso de lectura del ser humano. Dicho modelo construye una representación semántico-conceptual del texto, como resultado de su comprensión, representado de nuevo por una red asociativa de conceptos

con un determinado nivel de activación. Esta representación semántica se utiliza para asignar categorías semánticas a los textos (clasificación), lo que hace posible, por ende, una posible aplicación a la recuperación de los textos de una colección que estén relacionados con una consulta expresada en lenguaje natural (Recuperación de Información - RI). Para la consecución de dichos objetivos de aplicación se emplean métodos de aprendizaje automático y medidas de similitud en grafos o redes asociativas.

1.5. Caracterización de los Sistemas de Procesamiento de Lenguaje Natural

Existen varios criterios mediante los cuales se pueden caracterizar los diferentes sistemas y modelos de procesamiento de lenguaje natural. Uno de ellos, como ya se ha comentado anteriormente, es el nivel de desarrollo de cada una de las tres facetas del lenguaje citadas. A continuación, se describen brevemente otros criterios significativos:

- Plausibilidad psicológica. Es decir, si los mecanismos y métodos que emplea el sistema son susceptibles de ser atribuidos a la mente humana, basándose en aquello que de ésta se conoce.
- Tipo de conocimiento linguístico. Se refiere al nivel de detalle y a la naturaleza de la información que se adquiere y almacena. Comúnmente, el conocimiento almacenado puede ser de naturaleza fonética, sintáctica y semántica.
- Paradigma de representación. Indica la naturaleza del formalismo de representación que se emplea, tanto para el conocimiento lingüístico como para la semántica comprendida. Esta naturaleza puede ser diversa índole, encontrándose entre las más significativas:
    - la conexionista, que emplea redes de asociación y grafos de nodos interconectados entre sí con niveles de activación que se propagan por la red o grafo,
    - la lógica, que utiliza su notación y sus operadores para representar y manejar el conocimiento contenido en el lenguaje,
    - la analítica, que emplea estructuras y métodos mátemáticos analíticos y algebraicos para el mismo fin,
    - y la probabilista, que utiliza modelos de probabilidad y sus funciones para el modelado de la información.

  La naturaleza de la representación determina, en cierto modo, la plausibilidad psicológica. De hecho, tanto el paradigma conexionista, por su

semejanza estructural con el cerebro humano, y la lógica, por su semejanza funcional con los procesos deductivos de la mente, son a priori más plausibles psicológicamente que los paradigmas analítico y probabilista.

Es necesario decir que la mayoría de los sistemas de procesamiento de lenguaje existentes son sistemas híbridos en cuanto a su naturaleza se refiere, es decir, combinan varios paradigmas distintos.

- Dinamismo temporal. Hace referencia a la posibilidad de ampliación y modificación eficiente del conocimiento almacenado.
- Nivel de automatización. Se refiere a la autonomía del sistema, es decir, la no intervención humana, en cada una de las facetas del lenguaje. También puede considerarse como la cantidad de conocimiento proveniente de un experto humano que requiere el sistema.
- La fuente de información. El tipo y cantidad de información que va a ser procesada y de la que se va a extraer el conocimiento linguístico. En muchos casos está relacionado con el nivel de automatización. Por ejemplo, si el sistema requiere que un usuario le proporcione una gramática, entonces la parte de adquisición del lenguaje ya no será completamente autónoma. Además, es necesario determinar si se van a procesar sólo palabras, frases, textos completos, diálogos, etc., aparte de la cantidad de los mismos que se pretende manejar, aspecto que condicionará la elección y el diseño del formalismo de representación.
- El objetivo de aplicación. Es decir, la tarea práctica para la que el sistema está concebido. Algunos sistemas se centran en el mero modelado y su posterior validación en comparación con los seres humanos. Otros sistemas son púramente prácticos y están orientados a una determinada tarea concreta, como la clasificación de textos o la recuperación de información, por ejemplo. Por último, algunos sistemas tienen un propósito más general, y persiguen la aportación de un núcleo linguístico que permita realizar varias tareas relacionadas con el procesamiento del lenguaje natural.

Existen, por supuesto, más criterios que pueden caracterizar a los sistemas de lenguaje natural, pero los aquí mentados son suficientes para situar y diferenciar al sistema objeto de esta tesis en el contexto actual del campo de investigación.

1.6. Objetivos

El objetivo general de la tesis que se presenta es el diseño y desarrollo de un sistema de procesamiento de lenguaje natural denominado SILC (Sistema de Indexación mediante Lectura Cognitiva) que, dado un texto o un fragmento de texto, derive una representación de la semántica del mismo que pueda ser posteriormente empleada, entre otras tareas, para clasificar dicho texto, recuperarlo de una colección, resumirlo, compararlo o generar lenguaje a partir de él. En definitiva, el objetivo del sistema es de carácter general, tratando de proporcionar un núcleo linguístico que permita aplicaciones diversas.

Más concretamente, los objetivos perseguidos son los siguientes:

- El sistema debe llevar a cabo las etapas de adquisición y de comprensión de manera totalmente automática. Aunque la representación resultante de la comprensión sea aplicable a la generación de lenguaje de manera directa, esta faceta será abordada en el futuro.
- El sistema deberá extraer y almacenar una representación del conocimiento linguístico contenido en colecciones compuestas por grandes cantidades de textos, de tal forma que dicho conocimiento sea lo más completo posible.
- Dicho conocimiento será de naturaleza semántica y recogerá las relaciones conceptuales entre las palabras.
- El conocimiento debe ser actualizado, es decir, ampliado, modificado o reducido de manera eficiente, pudiendo incorporar el conocimiento contenido en grandes cantidades de información textual con un bajo coste computacional.
- El modelo debe ser psicológicamente plausible en lo que a la faceta de comprensión se refiere. Es decir, el proceso de generar la representación de la semántica de un texto, a partir del conocimiento linguístico previo y del texto en lenguaje natural, debe ser compatible con las evidencias sobre el proceso de lectura en la mente humana. Esto involucra a la memoria, a las inferencias y al seguimiento del orden de las palabras en el texto durante el proceso de lectura en el tiempo.

- Aportar un marco y herramientas experimentales para evaluar hipótesis teóricas de otros campos dentro del ámbito de la Ciencia Cognitiva, tanto en términos estructurales como en términos funcionales.

Por supuesto, otro de los objetivos importantes es el de la validación del sistema. Así, para validar su aplicación a tareas prácticas concretas se probará en:

- Tareas de clasificación automática de textos.
    - Adecuando la representación de los textos obtenida por el sistema a la representación clásica para la comparación con otros sistemas existentes.
    - Utilizando la propia representación de los textos obtenida por el sistema. Para aplicar esta representación es necesario el diseño y desarrollo de ciertos mecanismos que consisten en:
        - Medidas de similitud semántica entre palabras y entre textos.
        - Caracterización de una categoría temática a partir de las representaciones semánticas de los textos que pertenecen a la misma mediante la construcción de "centroides".

Adicionalmente, se pretende validar el sistema en otro aspecto: su plausibilidad psicológica. Así pues, se compararán las representaciones de la semántica de los textos obtenidas por el sistema con las obtenidas por seres humanos en las siguientes tareas:

- Clasificación de textos, tratando de confirmar la hipótesis de que cuanto más se asemeje el modelo al ser humano mejores resultados de clasificación obtendrá.
- Inferencia de palabras, tratando de validar el modelo del proceso de lectura.

Esta validación de plausibilidad psicológica está también encaminada a probar la validez del sistema en el modelado de seres humanos individuales, de tal forma que se puedan identificar en el modelo disfunciones en sujetos con problemas de lenguaje, además de simular su corrección con diversos estímulos de entrada. Aunque dicho objetivo de aplicación no es una finalidad prioritaria de esta tesis, el diseño del sistema sí debe contemplar esta posible aplicación en el futuro.

A pesar de la búsqueda de la plausibilidad psicológica y de la validación de la semejanza con el ser humano, el sistema desarrollado no pretende ser una teoría sobre el lenguaje en la mente, sino una herramienta computacional y un marco experimental de

procesamiento de lenguaje natural, aplicable a tareas del mundo real y con gran capacidad que, por el hecho de estar más próxima al proceder de los seres humanos que otros sistemas, obtenga mejores resultados en tareas que requieren un entendimiento mayor del lenguaje, tenga un uso más aceptado y confiado por parte de los usuarios, y sirva para validar hipótesis provenientes de otros campos como los citados en la presentación de este capítulo.

1.7. Descripción de los Capítulos

En el siguiente capítulo se presenta una revisión de los formalismos de representación del conocimiento y de los modelos para adquirirlo partiendo de los sistemas clásicos de representación de la semántica, ideados en los primeros años de la IA para modelar el conocimiento encerrado en el lenguaje natural y posteriormente aplicados a otros muchos dominios, y concluyendo con los sistemas alternativos, que se crearon para superar las limitaciones de tiempo y espacio que acarreaban los sistemas clásicos y para afrontar el auge del almacenamiento y distribución de información electrónica. El repaso de dichos sistemas permitirá percibir la evolución de la naturaleza de los mismos, empezando por modelos lógicos y asociativos, pasando por enfoques estadísticos y matemáticos puros y terminando por los modelos híbridos recientes, que mezclan enfoques conexionistas con modelos analíticos y modelos cognitivos.

En el Capítulo 3 se presenta el proceso de lectura en términos funcionales y cognitivos, describiendo las tareas que el ser humano realiza durante la lectura y los subprocesos mentales que las llevan a cabo. El capítulo presenta también diversas teorías existentes, junto con sus evidencias experimentales, sobre varios aspectos y entidades involucradas en la lectura, como son la memoria, las inferencias y los diferentes niveles lingüísticos (fonética, ortografía, sintaxis y semántica), y sobre los mecanismos que integran todos estos aspectos. El capítulo concluye con la definición y caracterización de los modelos computacionales de lectura describiendo sus requisitos, su razón de ser y la manera de evaluarlos, realizando además un recorrido por los más recientes modelos existentes y poniendo de manifiesto sus características más relevantes.

El Capítulo 4 describe en detalle la estructura y funciones del sistema SILC, propuesto en esta tesis. En primer lugar se presentan las estructuras y mecanismos de adquisición del conocimiento semántico lingüístico que sirve de base al posterior proceso de lectura, caracterizados en los mismos términos en los que se caracterizan los sistemas existentes en el Capítulo 2. A continuación se detalla el modelo de lectura, describiendo los algoritmos que se emplean sobre las estructuras de conocimiento previamente creadas y justificando la plausibilidad psicológica de cada uno de ellos. El modelo se caracterizará también en los mismos términos que los modelos de lectura del

Capítulo 3. Finalmente se definen medidas de similitud semántica, entre conceptos en la red de conocimiento creada por el modelo y también entre las representaciones de los textos que produce el modelo, para permitir la aplicación de SILC a tareas de Procesamiento de Lenguaje Natural como la clasificación de textos o la recuperación de información.

En el Capítulo 5 se presenta la evaluación experimental del sistema SILC. En primer lugar se definen las medidas utilizadas para la evaluación. A continuación, se describen los diferentes tipos de procedimientos experimentales y los resultados obtenidos en cada uno de ellos. Cada uno de dichos tipos de experimentación tienen un objetivo distinto: la optimización de los parámetros del modelo y las medidas de similitud para la clasificación de textos, la comparación del modelo con otros sistemas en tareas nuevamente de clasificación de textos y finalmente la similitud del modelo con los seres humanos, tanto en las representaciones textuales producidas como durante el mismo proceso de lectura.

Para finalizar, el Capítulo 6 recoge las conclusiones sobre el trabajo realizado en términos de consecución de objetivos propuestos y de aportaciones a las áreas científicas dentro de la cuales se desarrolla esta tesis doctoral. Se proponen además diversas tareas para la continuación del trabajo a corto plazo y también las líneas de investigación y retos futuros que derivan del mismo.

Los apéndices I y II contienen los ejemplos textuales utilizados en algunas fases de la evaluación experimental, para así permitir la reproducción de dichos experimentos y dar al lector una idea clara del propósito e intención de los mismos.

CAPÍTULO 2

Adquisición y Representación del Conocimiento Lingüístico

El lenguaje es una cualidad común a todos los seres humanos. Incluso las personas con deficiencias innatas en sus sistemas auditivo o fonético son capaces de desarrollar una lengua y comunicarse con sus congéneres. Éste y otros hechos han conducido a que algunos expertos se planteen la teoría del "lenguaje como instinto" en los seres humanos. Sin embargo, existe disparidad de opiniones en cuanto a si el lenguaje es aprendido durante las primeras etapas de la vida o, por el contrario, se nace con el lenguaje total o parcialmente adquirido. Por otra parte, es un hecho perceptible que la complejidad y habilidad en el manejo del lenguaje aumenta a medida que el ser humano crece e interactúa con otras personas. Así pues, esta evolución lingüística en el tiempo es un factor a tener en cuenta por los modelos computacionales de adquisición del lenguaje. Ya sea adquirido o innato, el lenguaje es un producto emergente de estructuras cerebrales específicas en los seres humanos. Cómo está representado el conocimiento lingüístico en el cerebro continúa siendo una incógnita. Sin embargo, los modelos computacionales deben dar forma a esa representación para poder ser diseñados, planteando pues hipótesis aún sin pretenderlo. La elección del formalismo de representación es el punto de partida de cualquier modelo, teniendo en cuenta la capacidad de expresión, la accesibilidad, la flexibilidad, el ámbito de aplicación y la plausibilidad psicológica de la representación elegida entre muchos otros factores. La representación final condicionará el diseño de los procesos de adquisición del lenguaje, así como éstos condicionarán a su vez la representación del conocimiento que se puede obtener a partir de ellos. El modelado del conocimiento lingüístico y su adquisición es pues una tarea difícil que requiere un compromiso entre lo que se conoce sobre la mente humana, las posibilidades y limitaciones de los ordenadores y la ambición de los objetivos perseguidos por los modelos.

2.1. Teoría del Aprendizaje del Lenguaje

Según Weisler [Stillings et al., 1998], el conocimiento lingüístico es inconsciente, es decir, es adquirido sin percatarse de que efectivamente se está aprendiendo y desarrollando una facultad. Es necesario hacer una distinción entre aprender el lenguaje y aprender a leerlo o escribirlo. Cualquier ser humano es capaz de aprender a comunicarse con otras personas aún siendo analfabeto. Aprender a leer o escribir implica el establecimiento en la mente de una correspondencia entre el lenguaje ya conocido y una colección de símbolos gráficos y fonéticos. A pesar de la inconsciencia de la adquisición del lenguaje, éste no es un proceso absolutamente pasivo puesto que los hijos son capaces de aprender aspectos lingüísticos que los padres no poseían. Los adultos del entorno influyen en el aprendizaje del lenguaje de los niños de una manera muy indirecta. Aunque la fonética y la agrupación de fonemas depende del desarrollo cerebral y auditivo (algunas personas dicen ser capaces de recordar mensajes prenatales que memorizaron en el útero sin entender su significado), la gramática se infiere de datos lingüísticos de manera activa, aunque inconscientemente.

Dado el desconocimiento del funcionamiento del cerebro humano en materia de lenguaje, unido a la falta de técnicas de exploración cerebral funcionales fiables y no invasivas [Démonet y Thierry, 2001], la mayoría de los estudios sobre el aprendizaje del lenguaje se realizan mediante la observación de lo que los seres humanos escuchan y de lo que dicen. Este procedimiento de entrada-salida es pues muy adecuado para ser modelado en los ordenadores.

2.1.1. Adquisición de la fonética

Como se ha comentado anteriormente, el lenguaje se empieza a aprender en etapas tempranas del desarrollo del ser humano. Lo primero que los seres humanos aprenden son fonemas, que identifican como tales y distinguen de otros sonidos de manera innata. Pero, ¿cómo distinguen las palabras? Al principio toman las sílabas acentuadas como distintivo y conforme su oído y capacidad intelectual se va desarrollando incorporan el resto de sílabas no acentuadas. Los sonidos del habla son físicamente diferentes a cualquier otro tipo de sonidos [Pinker, 1994; 2005], lo que hace posible que el oído

humano los distinga desde etapas tempranas de su existencia. La sonoridad del habla es muy rica en armónicos con varias componentes en distintas frecuencias, desde 100, 200, 300, etc., hasta 4000 Herzios para la voz masculina y desde 200, 400, 600, etc. para la voz femenina.

Al mismo tiempo que se identifican los fonemas de las palabras se aprende también a generarlos. Este aprendizaje se realiza por prueba y error, de ahí los balbuceos de los bebés. Dichos balbuceos son intentos de imitar los sonidos escuchados previamente. El aire, desde que sale de los pulmones, ha de pasar por las cuerdas vocales, la laringe, la cavidad bucal, la lengua y los dientes para producir el sonido final. Cada variación en la posición de los órganos anteriores produce un sonido distinto, definiendo sus diferentes características (nasal, no nasal, sonoro, sordo, labial, fricativo, etc.). Por esta razón, es necesario practicar desde el principio para aprender a sincronizar un sistema tan complejo.

Al igual que las palabras, los fonemas también obedecen a una serie de reglas específicas de cada lengua para combinarse y formar palabras [Pinker, 1991; 2005]. De hecho, los humanos son capaces de inferir si una palabra no conocida pertenece o no pertenece a su lengua gracias a las reglas fonéticas. Sin embargo, existen combinaciones de fonemas que no se dan en ninguna lengua debido a la dificultad que supone su pronunciación para el sistema articulatorio vocal. Así pues, tanto el desarrollo como la estructura del aparato auditivo y del aparato fonador son los principales condicionantes del aprendizaje de la fonética del lenguaje.

Existen varios trabajos científicos, principalmente en el campo del reconocimiento del habla, que describen modelos de la adquisición, identificación y generación de la fonética del lenguaje. La mayoría de dichos modelos pertenecen al paradigma conexionista, es decir, están representados por redes de nodos interconectados por las que fluye información. Como ejemplo representativo se puede citar el trabajo de Daniel Ellis [Ellis et al., 2001]: en primer lugar emplea una red neuronal para discriminar fonemas en la señal acústica. Los fonemas detectados son después introducidos en un Modelo de Markov Oculto (*HMM – Hidden Markov Model*) que es el encargado de agrupar los fonemas según las reglas fonéticas. El aprendizaje en las dos etapas se realiza a partir de fragmentos de lenguaje hablado, es decir, de lo que se podría considerar como experiencia previa, obteniendo estadísticas del uso de las combinaciones de fonemas.

2.1.2. Adquisición de la sintaxis

Todos los lenguajes existentes, así como los muertos o en desuso de los que se tiene constancia, poseen un conjunto de reglas para combinar las palabras de manera ordenada formando oraciones con contenido semántico. Es lo que se conoce como gramática. Este punto es el de más controversia en cuanto a lo innato del lenguaje. Ya Von Humboldt hacia 1830 observó y postuló que el lenguaje hacía uso infinito de medios finitos. Así pues, una gramática de una lengua es un sistema combinatorio discreto y finito que puede generar un número infinito de combinaciones de palabras, demasiado elevado como para tenerlo almacenado en el cerebro de manera exhaustiva. Entonces, dado que el tamaño de las reglas gramaticales sí es limitado, ¿son éstas aprendidas o, por el contrario, se nace con ellas en el cerebro? Si la gramática fuese innata debería estar almacenada en los genes. Es más, dado que a priori cualquier humano puede aprender cualquier lengua, los genes deberían contener todas las gramáticas de todos los posibles lenguajes, lo que supone demasiada información. Sin embargo, el lingüista Derek Bickerton reunió pruebas que apuntan a la existencia de una gramática intrínseca al cerebro humano [Bickerton, 1981; 1990]. Bickerton afirmó que un conjunto de niños separados de sus padres desde su nacimiento, que sólo escuchaban un lenguaje inconexo, entrecortado y casi agramatical, mezcla de varias lenguas importadas de los adultos con los que mantenían contacto, dotaron por sí solos a dicho lenguaje de una gramática compleja y expresiva a lo largo de su aprendizaje [Mosterín, 2006].

De esta manera se llega a Noam Chomsky, que propone la existencia una gramática universal innata común a todas las lenguas [Chomsky, 1986; 2006]. Dicha gramática consiste en una especie de armazón o molde configurable. El aprendizaje de una lengua consiste, según esta hipótesis, en el refinamiento y especificación de la gramática universal. Esta teoría implica la existencia de una raíz común a todas las lenguas y de un árbol filogenético de las mismas de manera análoga al de la evolución de las especies. Hasta el momento, lo que se ha podido demostrar es que todas las lenguas conocidas poseen una serie de parámetros compartidos que las definen y distinguen sintácticamente del resto [Pinker, 1991; 1994]. Así, las lenguas pueden ser aislantes, flexionales o aglutinantes dependiendo de la posibilidad de variación de las unidades léxicas, pueden presentar un orden fijo de palabras o no, pueden ser acusativas o

ergativas según el tratamiento de los verbos transitivos e intransitivos, pueden ser de tipo "SVO" (Sujeto Verbo Objeto), "SOV" o "VSO", etc.

De esta manera, la teoría más aceptada (y más intuitiva) dice que se nace con los mecanismos para aprender la sintaxis, es decir, el modelo general y los procesos para refinarlo, más que con las propias gramáticas. Es lo que Piaget denomina "*modus operandi*" [Piaget, 1970].

Los niños suelen cometer errores en la pronunciación y en el uso del significado de las primeras palabras que aprenden. Estos errores suelen estar provocados por una imitación imperfecta o por la intención de expresar pensamientos complejos con un lenguaje extremadamente simple (el que conocen en ese momento), ya que su pensamiento conceptual se desarrolla más rápido que su capacidad lingüística. Así pues, los niños asocian las palabras a objetos o hechos que perciben, y la asociación de esos objetos y hechos es la primera fuente de aprendizaje de la sintaxis. Según se asocien los conceptos que se perciben en el mundo real así se asociarán las palabras que los identifican. Según Braine [Braine, 1976], los niños organizan las palabras en dos categorías léxicas: pivotes, palabras frecuentes y de fácil uso que juegan el papel de referentes, y abiertas, palabras menos usuales que se unen a los pivotes, con la restricción de que no puede aparecer una palabra pivote suelta (es lo que se llama "gramática pivote"). Bowerman [Bowerman, 1973] postula, sin embargo, que la primera gramática en los seres humanos se describe en términos de agente-acción, más que en términos léxicos. Con el desarrollo intelectual, las relaciones entre las palabras se establecerán además en términos de modificador-modificado.

Como ejemplo del desarrollo de aprendizaje sintáctico, Brown y Fraser [Brown y Fraser, 1963; Brown, 1973] realizaron un experimento con varios niños de 28 meses de edad, infieriendo de sus resultados que todos empleaban la siguiente gramática en sus producciones lingüísticas:

$$O \rightarrow (SN)\ SV$$
$$O \rightarrow SN\ (SV)$$
$$SN \rightarrow (DET)\ N$$
$$SN \rightarrow N\ N$$
$$SV \rightarrow (V)\ SN$$

O – Oración; SV – Sintagma Verbal; SN – Sintagma Nominal;
N – Nombre; V- Verbo; DET – Determinante

Aparte de las componentes innatas del lenguaje existen otros dos factores teóricos que influyen en el aprendizaje de la sintaxis: el estímulo externo y el entorno. Las características relacionadas con los estímulos que inducen el aprendizaje son la cantidad de información lingüística proporcionada, la corrección de los errores o estimulación negativa y la distinción entre información lingüística y no lingüística. En cuanto al entorno se refiere, la simpleza del lenguaje que se recibe y el refuerzo o redundancia son los aspectos más relacionados con el aprendizaje gramatical. Newport y Gleitman [Gleitman y Newport, 1995] dividieron los elementos del lenguaje natural en dos categorías: los conceptos globales o núcleo y los conceptos específicos o periferia. Según los autores, las variaciones de los aspectos del entorno sólo afectan a los conceptos específicos. El núcleo del lenguaje se aprende siempre sea cual sea el entorno.

El proceso de descomponer una oración en sus partes y subpartes constituyentes de acuerdo a una gramática es lo que se conoce como análisis sintáctico. Las gramáticas son conjuntos formalizados de reglas que indican cómo se deben agrupar las palabras para formar fragmentos sintácticamente correctos de acuerdo a una lengua, y cómo se deben agrupar dichos fragmentos para formar oraciones de la misma. Así pues, el análisis sintáctico proporciona la estructura total de una oración cuando ésta pertenece al lenguaje en cuestión. La obtención de gramáticas que contemplen la totalidad de las estructuras del lenguaje natural real es muy costosa, ya que han de ser gramáticas muy amplias cuya construcción requiere una enorme cantidad de conocimiento experto. Incluso disponiendo de este conocimiento y de recursos suficientes se obtendrían gramáticas con carencias. Como observó Edward Sapir, "todas las gramáticas tienen fugas", [Sapir, 1921].

Además del coste, tanto en trabajo como en conocimiento, la definición de gramáticas que tratan de abarcar todas las estructuras del lenguaje presenta más problemas a tener en cuenta [Briscoe, 1994]. Uno de los principales es el de la ambigüedad sintáctica, donde dada una oración existen varios análisis sintácticos distintos. La ambigüedad es un fenómeno que crece con el tamaño de la gramática, cuantas más estructuras recoja más ambigüedad se introduce. Otro problema añadido es el de la subgeneración, es decir, que la gramática no puede dar cobertura a todo el lenguaje puesto que éste evoluciona y cambia con el tiempo, el entorno, etc.

El análisis sintáctico automático realizado por ordenador es uno de los objetivos últimos de la investigación en Procesamiento de Lenguaje Natural. Dados los problemas mencionados anteriormente, el análisis completo es muy costoso y computacionalmente intratable para la totalidad del lenguaje natural real. Así, existen algoritmos y técnicas que son capaces de realizarlo para lenguajes restringidos [Allen, 1995]. Sin embargo, para tratar textos no restringidos de manera automática es necesario imponer las restricciones en los propios algoritmos y no en los lenguajes. Dichas restricciones se han centrado en la completitud del análisis. Para muchas de las tareas del Procesamiento de Lenguaje Natural no es necesario realizar un análisis sintáctico completo. La identificación de ciertas partes de las oraciones, sin necesidad de descomponer esas partes o identificarlas en mayor profundidad, es en muchas ocasiones suficientemente útil para que el proceso que utilice dicho análisis tenga éxito. Por ejemplo, el etiquetado léxico comúnmente conocido como *POS (Part-Of-Speech) Tagging*, que asocia a cada palabra de la oración con su etiqueta léxica (verbo, nombre, adjetivo, etc.), es ya un tipo de análisis útil y suficiente del que parten muchas aplicaciones.

Como paradigma de técnicas computacionales de aprendizaje de la sintaxis que permiten aprender gramáticas de la experiencia se encuentran los algoritmos de Inferencia Gramatical. La Inferencia Gramatical se puede definir como el proceso de aprender u obtener patrones estructurales del lenguaje y representarlos mediante una gramática, dado un conjunto de ejemplos sintácticamente correctos y otro conjunto de ejemplos sintácticamente incorrectos.

Existen varios algoritmos de Inferencia Gramatical [Fu y Booth, 1975], [Rulot, 1992], [Pla, 2000], que se pueden utilizar para obtener modelos del lenguaje parciales. Entre las técnicas de inferencia empleadas se pueden citar las siguientes:

- Inferencia de gramáticas de lenguajes k-reversibles [Angluin, 1982]. Infiere gramáticas para lenguajes regulares empleando conjuntos de ejemplos positivos del modelo de lenguaje que se quiere representar. Va agrupando estados y, por lo tanto, generalizando el lenguaje en un determinado orden para luego refinar las agrupaciones con datos estadísticos.
- Inferencia gramatical de lenguajes regulares [Oncina, 1991]. Al igual que los anteriores, emplea una secuencia de agrupaciones de estados pero la optimiza usando los ejemplos negativos.

- Inferencia de gramáticas de lenguajes *k*-testables [García y Vidal, 1990]. Se elige un valor para *k* y se construye un autómata de tal manera que todos los ejemplos que compartan sus *k*-1 últimos símbolos apunten al mismo estado.
- Inferencia gramatical basada en generadores mórficos [García y Vidal, 1987]. Destaca por identificar un tipo de lenguajes muy concreto y porque permite incorporar el conocimiento a priori del que se disponga.
- Inferencia *Inside-Outside* [Lari y Young, 1991]. Infiere gramáticas no contextuales. Utiliza un método estadístico para estimar la probabilidad de cada regla de la gramática no contextual, utilizando los datos obtenidos de un conjunto de ejemplos de entrenamiento positivos.
- Inferencia basada en corrección de errores [Rulot y Vidal, 1987]. En esta técnica se va construyendo un autómata progresivamente a partir de los ejemplos positivos. El autómata se va transformando al corregir los errores que comete el analizador ante un nuevo ejemplo positivo.

Todos estos algoritmos han sido aplicados en diferentes campos [Dupont, 2002] para modelar, no sólo lenguaje natural, sino cualquier lenguaje que pueda ser descrito por una gramática. Algunos de dichos campos son el reconocimiento del habla, la recuperación de información, la representación de música, analizadores para compiladores de código fuente, la criptografía, etc.

Aunque todas las técnicas mencionadas construyen una gramática progresivamente a partir de ejemplos del lenguaje, como lo hacen los humanos, la mayoría de ellas utilizan las categorías léxicas de las palabras, lo que requiere un etiquetado previo de los ejemplos. Sin embargo, como se comentó anteriormente, en las etapas tempranas de su vida un ser humano no conoce el concepto de categoría léxica, siendo el contexto y el entorno de las palabras las características que permiten asociarlas. A pesar de esto, los modelos de inferencia gramatical han demostrado ser muy útiles en tareas de Procesamiento de Lenguaje Natural, obteniendo resultados que proporcionan la fiabilidad suficiente para utilizarlos de manera automática sin intervención humana.

2.1.3. Semántica

En realidad, como se mencionó anteriormente, la semántica se adquiere antes que el lenguaje, ya que el desarrollo conceptual se produce más rápido que el desarrollo lingüístico. Así, la semántica se corresponde con el mundo real que se percibe. La adquisición de la semántica es pues la asociación del mundo con los elementos lingüísticos aprendidos, ya sean fonéticos o gráficos, y el establecimiento de la correspondencia entre las relaciones observadas en los conceptos del mundo real y las relaciones existentes en los elementos lingüísticos asociados a dichos conceptos.

Esta correspondencia es la que permite el entendimiento y uso de la "composicionalidad", es decir, la noción de que el significado de una expresión lingüística es función de los significados de sus constituyentes [Stillings et al., 1998]. Según este punto de vista de la semántica, el aprendizaje de la "composicionalidad" es un aspecto crucial para el entendimiento del lenguaje, ya que éste último se puede definir como un conjunto de símbolos con significado que se combinan para formar expresiones que se refieren al mundo conocido.

Otro punto de vista es el de la semántica condicionada a la verdad. Según esta visión, conocer el significado de una expresión lingüística es conocer las condiciones bajo las que es cierta o, dicho de otra manera, conocer cómo debería ser el mundo para que fuese cierta. Es necesario para ello formalizar el lenguaje natural en el lenguaje de la lógica, que no posee todo el poder expresivo del lenguaje natural.

Frege hace corresponder a cada término lingüístico dos aspectos semánticos: el sentido y la referencia [Dummet, 1993]. El sentido de una palabra es la cualidad que define su significado dentro de un contexto al tomar su referente. De la misma forma, el sentido de una oración es la manera en la que describe el mundo, de tal forma que si es precisa hace cierta a la oración y si no lo es la hace falsa. Si se observan las expresiones "lucero matutino" y "lucero vespertino", se puede comprobar que el referente es el mismo (un mismo planeta que refleja la luz del sol), pero no el sentido. Así pues, además de hacer referencia a conceptos del mundo las palabras también poseen un sentido.

Entre los modelos computacionales del procesamiento semántico se pueden distinguir dos clásicos de la semántica condicionada a la verdad:

- Teórico de Evidencia: Formaliza las premisas de lenguaje natural en lenguaje lógico y se siguen axiomas y reglas de deducción para llegar a su entendimiento. Lo que importa es la estructura lógica de la frase y no el significado de sus componentes.

- Teórico de Modelo: También es necesario formalizar en lenguaje lógico. Deriva las condiciones que hacen cierta la expresión que se quiere entender. Si existe otra expresión que es cierta dada las condiciones derivadas, entonces se da por cierta la expresión inicial y, por tanto, por comprendida.

Según Johnson-Laird [Johnson-Laird, 1983] el Teórico de Modelo es más plausible psicológicamente que el Teórico de Evidencia, aunque existan evidencias de que el cerebro humano es capaz de realizar, y de hecho realiza, operaciones lógicas de deducción pero posiblemente no durante el entendimiento del lenguaje.

Existen además otros modelos computacionales, en este caso conexionistas, del procesamiento semántico. Smolensky [Smolensky, 1988] define un conjunto de atributos semánticos que caracterizan el significado de cualquier palabra. Estos atributos pueden ser compartidos por varias palabras, lo que las une formando una red de asociación. Los patrones de activación en dicha red son tomados como la semántica de las expresiones lingüísticas. Shastri [Shastri, 1992] emplea redes semánticas y también obtiene el significado de patrones de activación que se propagan por las redes. Los modelos conexionistas tienen la desventaja de no admitir representaciones de expresiones simbólicas estructuradas sobre las que aplicar la lógica y la deducción. Sin embargo, al contrario que los modelos basados en la lógica, los modelos conexionistas recogen además la semántica funcional, es decir, la relativa a las acciones y procesos.

2.1.4. Ejemplo de modelo computacional completo del aprendizaje del lenguaje

Como se ha visto anteriormente, muchos modelos del aprendizaje del lenguaje se centran en aspectos concretos del mismo y, aunque utilizan ejemplos de lenguaje provenientes de la experiencia, no tienen en cuenta la temporalidad del aprendizaje en los seres humanos.

No obstante, existen modelos que tratan de recoger una teoría completa del aprendizaje del lenguaje tratando de reflejar el ritmo del desarrollo de las capacidades

lingüísticas más comúnmente observadas a determinadas edades. Este es el caso del trabajo de Mallory Selfridge [Selfridge, 1986]. El modelo propuesto por Selfridge trata de explicar y reflejar los siguientes hechos observados en el aprendizaje del lenguaje a lo largo del crecimiento humano:

- La comprensión del lenguaje precede a la generación.
- El vocabulario aprendido, más concretamente la tasa de aprendizaje, primero crece y después disminuye.
- La longitud de las palabras que se es capaz de pronunciar crece con la edad.
- Las palabras irregulares se regularizan.
- Los sonidos con semántica confusa se malinterpretan.
- Las oraciones pasivas reversibles presentan problemas de compresión.

Selfridge establece, además, las preguntas que según ella debería tratar de contestar cualquier teoría de aprendizaje del lenguaje:

1. ¿Cuál es el problema central del aprendizaje de lenguaje?
2. ¿Qué capacidades cognitivas existen ya antes del lenguaje?
3. ¿Qué se aprende?
4. ¿Cuáles son los mecanismos del aprendizaje?
5. ¿Cuál es la naturaleza de la experiencia de los niños con el lenguaje?
6. ¿Cómo se infiere el significado del lenguaje a partir de fuentes no lingüísticas?

Selfridge propone en su modelo una respuesta para cada una de las preguntas propuestas:

1. El problema central del aprendizaje del lenguaje es explicar el aprendizaje de las habilidades de comprensión y generación.
2. Las capacidades cognitivas previas son: conocimiento del mundo, mecanismos básicos de comprensión, y generación y mecanismos de inferencia y aprendizaje.
3. Lo que se aprende son los significados de las palabras y la sintaxis.

4. Los significados de las palabras se aprenden mediante mecanismos de aprendizaje de conceptos. La sintaxis se aprende por acumulación de características sintácticas que subyacen bajo el significado de una palabra.
5. El lenguaje se asimila en contextos en los que el significado buscado con los sonidos, que no se entienden de manera completa, puede ser inferido.
6. La inferencia del significado de unas palabras no entendidas completamente se realiza mediante un mecanismo de búsqueda que opera en conceptos que se desprenden de las palabras entendidas y del concepto.

En general, el modelo que propone Selfridge se puede esquematizar como muestra la Figura 2.1.

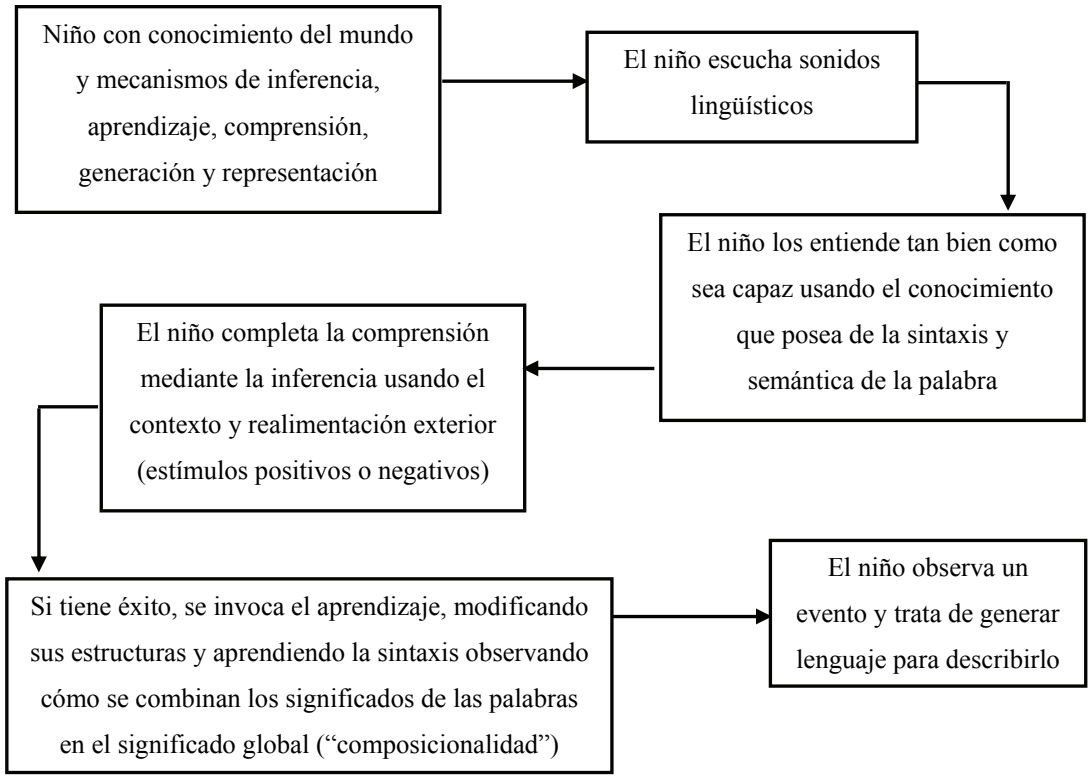


Figura 2.1. Modelo del aprendizaje del lenguaje en seres humanos, según Selfridge.

La implementación del modelo se realiza utilizando la representación *slot-filler* de la Dependencia Conceptual (CD) de Schank [Schank, 1982]. Así, la semántica tanto del lenguaje como del conocimiento del mundo se representa mediante conceptos caracterizados con *slots* que tienen unos requisitos semánticos para ser rellenados. La sintaxis se representa mediante el orden sucesor-antecesor entre *slots*. El mecanismo de comprensión se realiza rellenando *slots* de todas las formas posibles y asignando una puntuación a cada uno. La más puntuada será la interpretación final. Para inferir se

realiza una búsqueda de los conceptos que más *slots-fillers* tengan en común con la parte entendida en cada momento. Finalmente, el aprendizaje de la semántica se realiza asignando a las palabras nuevas el significado del contexto en el que aparece por primera vez. Si la palabra ya se conocía, se refinan sus estructuras de acuerdo al nuevo significado.

Como se puede apreciar, el modelo de Selfridge propone una teoría del aprendizaje basada en observaciones psicolingüísticas y cognitivas, posteriormente implementada en un ordenador. Es pues también un ejemplo de interdisciplinaridad. La adquisición del lenguaje es objeto de estudio de muchas áreas como la Ligüística, la Psicología, la Inteligencia Artificial, la Filosofía o La Neurociencia entre otras. La cooperación y aportación de expertos de todos los campos es necesaria para la realización e implementación de modelos lo más cercanos posibles al ser humano.

Puesto que estos modelos van a ser aplicados en un ordenador, es necesario buscar una representación que sea adecuada para ser tratable en la máquina (como por ejemplo la anteriormente comentada), tanto en tiempo como en espacio, lo que siempre supone un sesgo en la teoría que se pretende reflejar. Es necesario, pues, hacer hincapié en perder la menor cantidad de información teórica posible. Puesto que este sesgo es principalmente debido a la naturaleza computacional de los modelos, la Inteligencia Artificial y la Lingüística Computacional son las áreas que más han estado contribuyendo al avance de este aspecto del modelado del lenguaje desde hace más de dos décadas.

2.2. Representación del Conocimiento Lingüístico

A pesar de la controversia sobre si el lenguaje emerge o no de manera innata de estructuras cerebrales específicas en los seres humanos, está ampliamente aceptado el hecho de que el ser humano adquiere con el paso del tiempo aspectos y elementos lingüísticos que no poseía previamente. Uno de los ejemplos más claros es el vocabulario, que va creciendo en tamaño a un ritmo variable según la edad. Dicho vocabulario y la relación semántica entre las palabras que lo forman en diferentes contextos es el objeto principal de la representación. En definitiva, se trata de crear una representación del conocimiento en general.

En lo que respecta a lenguaje, además del conocimiento semántico general se tiene también el conocimiento sintáctico, que básicamente consiste en la gramática del lenguaje, y el conocimiento fonológico, que reside en la correspondencia entre sonidos y símbolos gráficos o sonidos y conceptos. Para estos y otros tipos de conocimiento la investigación en áreas como la Inteligencia Artificial (IA), la Lingüística Computacional (LC) y la Ciencia Cognitiva (CC) entre otras, ha estado proponiendo durante las últimas tres décadas diferentes tipos de representación de tal forma que pudiese ser almacenada y procesada en un ordenador.

2.2.1. Sistemas clásicos de representación del conocimiento

La mayoría de las propuestas tradicionales de representación del conocimiento surgieron en la etapa de énfasis en el conocimiento que sufrió la IA entre los años 70 y los años 80, motivadas por el entusiasmo por los sistemas basados en conocimiento (SBC) y los sistemas expertos (SE). Todos estos sistemas de representación consideran que el conocimiento se adquiere a partir de la experiencia, por lo que todos describen una manera de abordar el aprendizaje del mismo. Así mismo, tienen en cuenta la capacidad del ser humano de emplear el conocimiento que se posee para entender y generar más conocimiento nuevo, por lo que aportan un método de inferencia a partir del conocimiento existente. Puesto que el conocimiento general es muy amplio, los sistemas de representación definen también la naturaleza y ámbito del conocimiento que pretenden representar.

2.2.1.1. Lógica

La lógica es un formalismo de representación con gran riqueza expresiva y con métodos de inferencia derivados directamente de los procesos mentales deductivos. Como se ha comentado anteriormente, la lógica es el sistema de representación requerido si se considera la visión del conocimiento como semántica condicionada a la verdad.

Se pueden distinguir dos tipos de lógica que son los más ampliamente empleados para la representación del conocimiento:

- Lógica proposicional: El conocimiento se representa en forma de variables lógicas, también llamadas proposiciones. Dichas proposiciones se pueden combinar mediante operadores formando expresiones compuestas. Cada variable representa un hecho expresado en lenguaje natural y los operadores conectan dichas variables relacionando su semántica. Los operadores básicos, también llamados conectivas, son: el condicional ("→"), la negación ("¬"), la Y lógica y la O lógica. Así, por ejemplo, si *"El perro come anchoas"* es la variable *p* y *"El gato come chuletón"* es la variable *q* entonces $p \rightarrow q$ expresará *"SI el perro come anchoas ENTONCES el gato come chuletón"*, y $\neg p$ denotará que *"El perro NO come anchoas"*. Dados un conjunto de premisas, ya sean proposiciones o expresiones compuestas, se puede derivar otra expresión mediante métodos deducción automática, aprendiendo o reformulando así nuevo conocimiento.
- Lógica de predicados: Tiene más poder expresivo que la lógica de proposicional, ya que ésta última es un subconjunto de la lógica de predicados. En la lógica de predicados el conocimiento se expresa mediante relaciones, llamadas predicados, entre individuos y objetos, llamados argumentos. Las variables hacen referencia a los individuos y objetos y los operadores combinan las relaciones. De esta manera, *"La anchoa Pepita es salada"* se expresará con la relación *Es_Anchoa(Pepita) Y Salado(Pepita)*, y *"SI algo es salado ENTONCES es una anchoa"* se representará como $Salado(x) \rightarrow Es\_Anchoa(x)$. La introducción de los cuantificadores universal y existencial permite representar oraciones del tipo *"Todas las anchoas son saladas"* o *"Al menos una de las anchoas es salada"*, por lo que aumenta notablemente la capacidad expresiva de la representación con respecto a la

lógica de proposiciones. La generación y validación de conocimiento nuevo se puede realizar mediante métodos de deducción automática. La lógica de predicados puede ser ampliada con predicados de orden superior, es decir, predicados que también pueden ser cuantificados al igual que las variables pudiendo representar frases del tipo *"Todas las anchoas multicolor tienen al menos un color en común"* de la forma:

EXISTE *C(*PARATODO *x (Anchoa(x) Y Multicolor(x) →C(x))*

Es posible, además, añadir la identidad a la lógica de predicados, lo que permite representar frases del estilo *"Como mucho hay una anchoa violeta"*, utilizando la identidad para expresar que si hay dos anchoas violetas entonces la primera es la misma que la segunda, es decir, son la misma anchoa.

Aparte de la lógica proposicional y la lógica de predicados existen variaciones de las mismas que surgieron para cubrir ciertas limitaciones que éstas últimas planteaban. Una de esas limitaciones es el concepto de verdad. En las lógicas anteriores las expresiones son falsas o verdaderas. La lógica modal aparece para permitir grados o modalidades de verdad o falsedad introduciendo cualificadores de posibilidad, imposibilidad, necesidad o contingencia.

Otra limitación importante de la lógica proposicional y de predicados es la imposibilidad de reflejar la imprecisión del lenguaje. De esta forma, la lógica difusa surge permitendo tratar con variables que pertenecen en cierto grado a algún o algunos conjuntos de entidades. Por ejemplo, en *"Las anchoas inteligentes"* el término *"inteligente"* es impreciso. En lógica difusa una anchoa con un cociente intelectual de 210 podría tener un grado de pertenencia de 0,99 al conjunto *Inteligentes* (un valor de 0 significa que no pertenece en absoluto al conjunto y un valor de 1 que pertenece totalmente al conjunto), mientras que una anchoa con un cociente intelectual de 80 tendría un grado de pertenencia de 0,1 a dicho conjunto, por ejemplo. El grado se determina mediante una función de pertenencia que es necesario definir para cada conjunto impreciso con el que se trate.

La última de las limitaciones se refiere a los métodos de razonamiento y deducción automática. En la lógica de predicados dichos métodos son muy rígidos, requiriendo la completitud, consistencia y tratabilidad de todas las expresiones empleadas. Sin embargo, las premisas que se creían válidas en un principio pueden no serlo ante nuevos

hechos observados con el paso del tiempo o ante la percepción de excepciones. Para dotar de flexibilidad a los métodos de inferencia aparecieron la lógica no monótona [Doyle y McDermott, 1980], donde se crean reglas condicionadas a la existencia de excepciones, y la lógica por defecto [Reiter, 1980], en la que las reglas de inferencia dependen de las condiciones en las que se aplican.

2.2.1.2. Reglas de producción

Dadas unas condiciones observables en cada momento, el ser humano actúa a partir de la información que tiene del entorno y su conocimiento general del mundo. Según postularon Newell y Simon a principios de los años 70, este comportamiento inteligente se puede describir mediante reglas que, en conjunto, conforman un modelo psicológico del conocimiento funcional humano. Sin embargo, el origen de las reglas surge en el ámbito del lenguaje, como descripción de la gramática, capaz no sólo de validar la corrección de las expresiones lingüísticas sino también de generarlas, de ahí el término "producción".

Las reglas se componen esencialmente de dos partes: el antecedente y el consecuente. El antecedente contiene el conjunto de condiciones que deben darse para que la regla sea aplicable. El consecuente contiene los hechos que se deducen de las condiciones del antecedente. La representación clásica de las reglas tiene la siguiente forma:

$$Antecedente \rightarrow Consecuente$$

Los elementos que forman parte de una regla pueden ser de diversos tipos:

- Hipótesis: Cláusulas que tienen asociado el valor verdadero o falso.
- Datos: Valores numéricos o simbólicos para atributos o cualidades de entidades.
- Relaciones: de comparación o pertenencia entre atributos o atributos y sus valores.
- Cláusulas: Conjunción, disyunción o negación de hipótesis y relaciones.

Así, por ejemplo, se puede representar la expresión "Si hay una anchoa solitaria y aparece un temible tiburón en aguas profundas entonces que se esconda en un arrecife" como:

SI *Anchoa_Solitaria=CIERTO* ***Y*** *Tiburón_Temible=CIERTO* ***Y*** *profundidad > 40 → Esconder_Anchoa*

De esta forma se representa el conocimiento funcional o modal. Si se emplean variables también se puede representar conocimiento declarativo. Por ejemplo, *"Todos los tiburones temibles comen anchoas"* se representará como:

SI *x ES tiburón* ***Y*** *x ES temible → x COME anchoas*

La inferencia con este tipo de representación se realiza mediante correspondencia de patrones. Se seleccionan las hipótesis y datos que cumplen el antecedente de una regla (que satisfacen las variables) y se aplica su consecuente. Si ese consecuente satisface el antecedente de otra regla entonces se puede generar una regla nueva con el primer antecedente y el último consecuente obtenido. De esta forma se genera más conocimiento a partir del que ya se tiene. Si se tiene más de una regla para aplicar (conjunto conflicto) es necesario establecer un criterio de selección, que puede estar orientado hacia la aplicación sucesiva de más reglas o hacia la consecución de antecedentes que cumplan una conclusión objetivo. Además, el mecanismo de selección de reglas ha de tener en cuenta el contenido resultante del orden de aplicación de las reglas, la eficiencia o tiempo que se tarda en obtener un resultado y la autonomía y capacidad explicativa del proceso de razonamiento.

Una de las principales ventajas de las reglas de producción es que son fácilmente interpretables por el ser humano y son suficientemente descriptivas con los procesos de inferencia que emplean, de tal forma que se explican de manera inteligible. Además, son muy eficientes, en cuanto a tiempo de cómputo se refiere, en los procesos de razonamiento e inferencia. Sin embargo, en comparación con la lógica de predicados su expresividad y capacidad de inferencia son mucho menores, no pudiendo expresar el conocimiento que encierran los cuantificadores existenciales. Pese a ello, las reglas de producción sí son capaces de tratar en cierta manera con la incertidumbre, al contrario que la lógica clásica.

2.2.1.3. Marcos

Los marcos fueron concebidos por Marvin Minsky [Minsky, 1975] como forma de representación del conocimiento para la comprensión del lenguaje y para la visión artificial. Es por esto una representación más próxima a la cognición humana que la lógica o las reglas, ya que posee estructura, aunque tal diferencia no se da en los métodos de inferencia y razonamiento.

Un marco es una estructura de datos que representa a un conjunto estereotipo de condiciones del entorno, es decir, a una situación típica. La información que contienen los marcos es heterogénea, indicando desde cómo usar el marco hasta la manera de proceder cuando se dejan de cumplir las condiciones del mismo, pasando por su descripción y las expectativas que se pueden tener en cada estado del entorno [Minsky, 1975].

Toda la información de los marcos está contenida en campos o *"slots"*. Los campos se rellenan mediante valores del dominio de cada campo o *"fillers"*. La estructura de los marcos es recursiva, por lo que un campo se puede rellenar a su vez con otro marco. Puesto que los campos describen a los marcos y éstos últimos representan una situación estereotípica, si se observa un valor de un campo y se identifica el marco en el que se encuentra, se pueden predecir otros elementos a observar en dicha situación, que corresponderán a otros campos del marco identificado. Además de esta información, es posible representar entidades desde distintos puntos de vista, creando distintos marcos para cada visión y actuando como campos de otro marco más general correspondiente a la entidad original. Estas relaciones entre marcos producen una especie de red de asociación que contiene caminos de unos a otros, lo que permite obtener explicaciones de los razonamientos y de las inferencias realizadas.

En la Figura 2.2 se puede ver un ejemplo sencillo de representación con marcos. En el ejemplo de la Figura 2.2, se tienen dos marcos: ANCHOA y COMIDA. Sin embargo se tienen dos instancias del marco ANCHOA y una del marco COMIDA. El marco define la "armadura" que describe a la entidad o situación. Una vez rellenos los campos lo que se tiene son casos particulares de esa entidad o situación, llamados ejemplares.

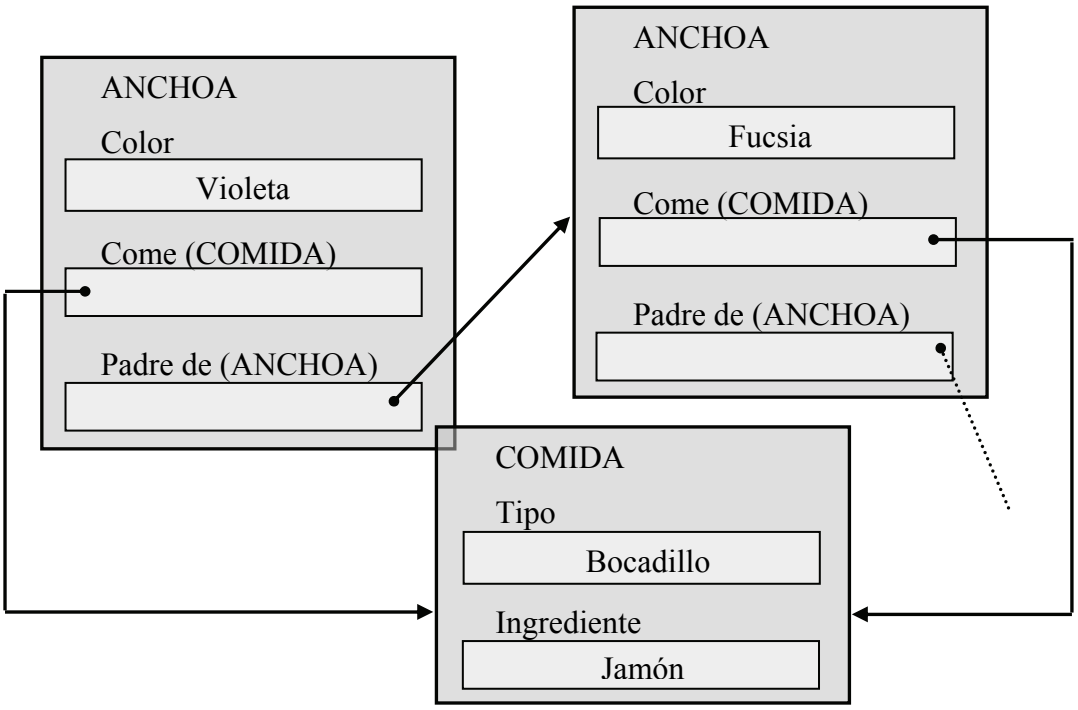


Figura 2.2. Ejemplo de representación con marcos.

Como se comentó anteriormente, la inferencia en los marcos se realiza mediante mecanismos de herencia o seguimiento de caminos en la jerarquía de los marcos y sus instancias. Así, por ejemplo, en la Figura 2.2 se puede deducir quién es el nieto de la anchoa violeta siguiendo los campos "Padre de" en la jerarquía. De esta manera, también se podría inferir que una anchoa padre y su hijo comen lo mismo.

Para dotar de más poder expresivo a la representación se introducen además una serie de aspectos referentes a los campos de los marcos:

- Valor por defecto: Es el valor que se asigna inicialmente a los campos de nuevas instancias.
- Dominio de conjuntos: Un campo puede tomar varios valores del dominio a la vez.
- Reglas de asignación: Restricciones a los valores que pueden tomar los campos.
- Confianza: Certeza que se tiene sobre el valor con el que se rellena un campo. Es análogo al grado de pertenencia en lógica difusa.

Todos estos aspectos sirven además para salvar limitaciones que presentan los marcos, comunes también a los sistemas de representación de la lógica.

2.2.1.4. Guiones

Los guiones o *"scripts"* fueron propuestos por Schank [Schank, 1982]. La noción y estructura de guión es muy similar a la de los marcos, aunque los guiones provienen de un sistema de representación de naturaleza completamente distinta como son las redes conceptuales, que serán tratadas más adelante.

La diferencia con los marcos radica en la naturaleza y las relaciones de los campos contenidos. Un guión también está formado por campos, pero todos ellos se refieren a sucesos o acciones. El guión representa una secuencia de acciones en el tiempo por lo que el orden de los campos es determinante.

El conocimiento se representa utilizando los grafos de dependencia conceptual que se verán en la siguiente sección. Los guiones establecen relaciones de causalidad entre el conocimiento almacenado en los grafos. La introducción de variables permite que los guiones se puedan instanciar, al igual que los marcos, permitiendo distintos roles. Lo elementos de un guión son:

- Pasajes: la secuencia de sucesos causalmente relacionados.
- Roles y objetos: Individuos o entidades que pueden aparecer en el guión y que instancian las variables del mismo.
- Desencadenantes: Información para activar el guión en el momento en el que se dan las condiciones para ello.

Así pues, un pasaje típico representado por un guión podría ser el siguiente: "La anchoa va al restaurante y espera que le asignen mesa. La anchoa se sienta y la sepia le atiende. La anchoa elige y la sepia le sirve. La anchoa come. Al terminar paga y se va. Si no tiene dinero la anchoa se quedará fregando los corales".

La inferencia en los guiones consiste en asignar las variables de los mismos. Dado un pasaje en lenguaje natural, es necesario seleccionar el guión que más se ajusta al mismo. Una vez seleccionado, se asignan sus roles y entidades. A partir de este momento, toda la información que contenga el guión ejemplarizado será considerada como conocimiento inferido que el pasaje original no contenía de manera explícita. La realización de inferencias o el hallazgo de correspondencias con información ya

conocida, dado un texto en lenguaje natural, pueden considerarse en cierta manera como la comprensión de dicho texto. De hecho, los guiones se crearon con el objetivo de la comprensión de lenguaje natural.

A pesar de la independencia del idioma y del control de la coherencia al realizar inferencias, los guiones presentan ciertas limitaciones. El principal obstáculo es la falta de flexibilidad. La secuencia de sucesos es fija y predefinida de antemano, por lo que no se adapta a posibles cambios o situaciones excepcionales. Además, esta rigidez se extiende también a la posibilidad de compartir información entre distintos guiones, al contrario que los marcos, lo que limita enormemente su contexto de aplicación. Pese a ser adecuados para la comprensión del lenguaje natural en cualquier idioma, los guiones, así como los sistemas de representación presentados hasta ahora, no son capaces de captar y tratar con toda la semántica que encierra el lenguaje, principalmente la relativa a intenciones, motivaciones, emociones o sentimientos. Para ello es necesario tener un conocimiento del mundo enorme y ser capaz de almacenarlo en un ordenador usando las estructuras de representación elegidas de manera que los procesos de razonamiento sean factibles en términos de tiempo de cómputo.

2.2.1.5. Redes asociativas

Una red es un conjunto de nodos unidos por enlaces que indican un tipo de asociación entre los nodos que conectan. Por este motivo los modelos con forma de red reciben el nombre de conexionistas. Los nodos representan conceptos, entidades o proposiciones. Las asociaciones pueden ser de diferentes tipos, dependiendo del conocimiento que se quiera representar. Típicamente, se distinguen tres tipos de redes asociativas: las redes semánticas, las redes de clasificación y las redes causales.

2.2.1.5.1. Redes semánticas

El objeto principal de las redes semánticas es la representación del conocimiento para la comprensión del lenguaje. Entre los principales tipos de redes se encuentran los grafos relacionales y las redes proposicionales.

De los grafos relacionales el sistema más representativo es el de Memoria Semántica de Quillian [Quillian, 1968], producto de su tesis doctoral, en la que pretendía abordar el almacenamiento de la semántica de las palabras para la comprensión del lenguaje. Cada concepto o palabra se representa mediante una red distinta de manera análoga a

los marcos. Los nodos de la red representan términos de la definición del concepto. Existen seis diferentes tipos de asociaciones entre los nodos:

- Subclase: Representa una relación de inclusión del nodo destino en el nodo origen.
- Modificación: El nodo origen modifica el alcance del nodo destino.
- Disyunción: Representa diferentes opciones para un mismo significado.
- Conjunción: Denota la presencia en el significado de todos los nodos implicados en la asociación.
- Propiedad: Realiza un enlace entre tres nodos, siendo el nodo origen la relación y los dos nodos destino el sujeto y el objeto de la misma.
- Referencia al tipo: Denota la identidad de un nodo con otro grafo referente a otro concepto, análogamente a los campos de los marcos que se rellenaban con otro marco.

En la Figura 2.3 se puede ver un ejemplo sencillo de representación de la memoria semántica del significado de "ANCHOA".

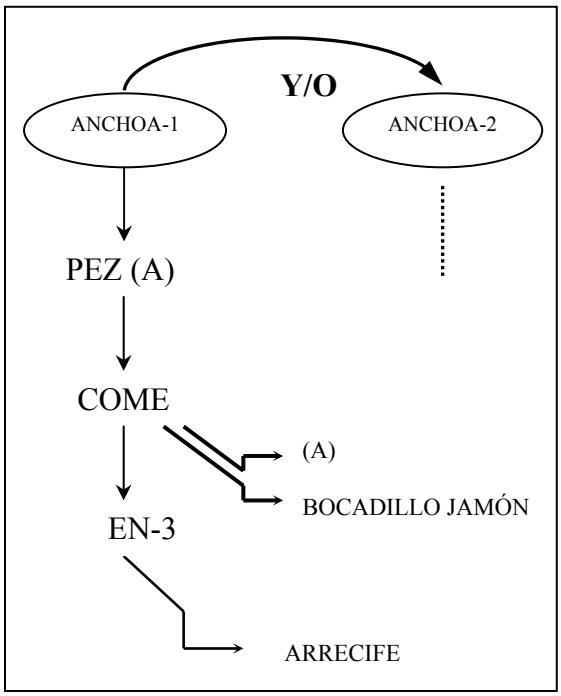


Figura 2.3. Ejemplo de representación con el sistema de memoria semántica.

El razonamiento en esta representación está determinado por la comparación del significado de dos palabras. Para ello, se parte del primer nodo de cada uno de los

grafos correspondientes a las dos palabras. Desde ese nodo se recorren los grafos buscando intersecciones entre ambos. Cada subgrafo intersección es la parte semántica que ambas palabras tienen en común.

Las intersecciones obtenidas en el proceso de comparación son interpretables para el ser humano. Sin embargo, puesto que las etiquetas de la representación y las definiciones provienen de una cierta lengua concreta, el sistema de representación está limitado a dicha lengua.

Otro caso típico de grafo relacional es el sistema SCHOLAR de Carbonell [Carbonell, 1970]. Es uno de los primeros sistemas de pregunta/respuesta. Dada una base de conocimiento representada en forma de grafo relacional, el sistema es capaz de generar y responder preguntas. Es aún más similar a los marcos que la memoria semántica de Quillian, ya que propone la definición de cada nodo como la clase a la que pertenece y una lista de propiedades que pueden ser a su vez otros nodos. Sin embargo, introduce la diferenciación entre nodos conceptuales y nodos ejemplo o instancias.

Con el objetivo de la comprensión del lenguaje y tratando de solucionar el problema de la dependencia con el idioma se presentan los Grafos de Dependencia Contextual de Schank [Schank, 1982]. Para salvar la limitación del idioma Schank pensó en la representación de conceptos a un nivel abstracto, en oposición a la representación de palabras. El objetivo de este sistema de representación es más ambicioso que el de los dos sistemas anteriores. En este caso, lo que se persigue es la comprensión de oraciones completas. Así pues, se basa en la idea de representar cualquier oración mediante un número de primitivas semánticas, independientes de cualquier idioma. Dichas primitivas se clasifican de la siguiente manera:

- 6 categorías conceptuales: objeto físico, acción, atributo de un objeto físico, atributo de una acción, tiempo y espacio.
- 16 reglas sintácticas.
- Acciones primitivas: Transferir físicamente (PTRANS), transferir una relación abstracta (ATRANS), transferir mentalmente (MTRANS), empujar (PROPEL), mover un miembro de un animal (MOVE), coger (GRASP), ingerir (INGEST), etc.

Un ejemplo de la representación de *"La anchoa come un bocadillo de jamón"* se presenta en la Figura 2.4.

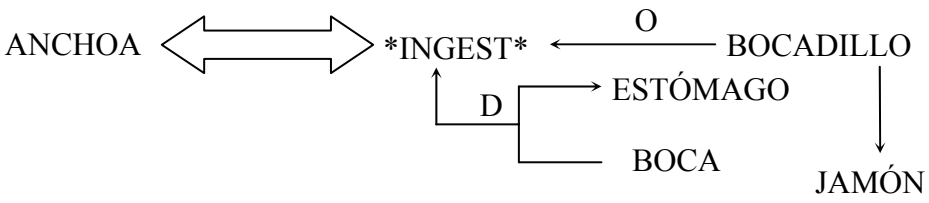


Figura 2.4. Ejemplo de representación con grafos de dependencia contextual.

En la Figura 2.4, la doble flecha representa relación entre sujeto y verbo, la "O" denota relación de objeto de la acción, la "D" representa dirección y la flecha simple indica posesión o composición.

Las inferencias que se pueden realizar en este sistema son las condiciones que están implícitas en las frases, las causas y la intencionalidad. En *"La anchoa come un bocadillo de jamón"* se puede inferir que la anchoa tiene boca, estómago y probablemente dientes, por ejemplo.

Las ventajas de este sistema de representación son aportadas principalmente por el uso de primitivas. Estas proporcionan independencia de los idiomas y, al ser limitadas, homogeneidad en la representación así como eficiencia en la obtención de inferencias. Además, permiten dividir las oraciones en fragmentos más simples facilitando la aplicación directa del principio de "composicionalidad" del que se habló al comienzo de este capítulo. Así mismo, las mismas primitivas plantean una serie de inconvenientes como la validez de su supuesta universalidad, su definición para dar la cobertura deseada y el incremento exponencial de complejidad y tamaño de los grafos a medida que aumenta la longitud de las oraciones.

A pesar de la capacidad expresiva e intuitiva de las redes semánticas presentadas hasta el momento, éstas continúan presentando limitaciones a la hora de tratar la comprensión del lenguaje natural, siendo incapaces de igualar a la lógica de predicados. Además, estos sistemas hacen difícil la representación de las relaciones existentes entre distintas proposiciones. Para superar esta dificultad surgen las redes proposicionales. Las redes proposicionales son un tipo de red semántica en la que los nodos pueden representar entidades de cualquier nivel, desde palabras y frases hasta textos completos.

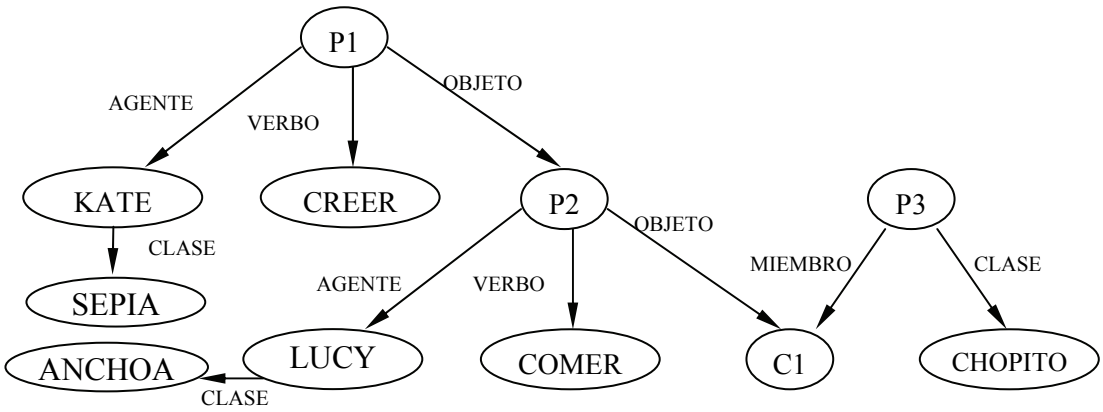


Figura 2.5. Ejemplo de representación con redes de Shapiro.

Las dos implementaciones clásicas de las redes proposicionales son las redes de Shapiro y los grafos de Sowa. Como su nombre indica, las redes de Shapiro fueron propuestas por Stuart Shapiro [Shapiro, 1979]. La flexibilidad introducida en esta representación hizo que el poder expresivo de las redes proposicionales igualara al de la lógica de primer orden. La flexibilidad consiste en la versatilidad de los nodos y también de las relaciones, representadas mediante una etiqueta en lenguaje natural. Por ejemplo, *"La sepia Kate cree que la anchoa Lucy se come los chopitos"* se representa mediante la red de Shapiro correspondiente que aparece en la Figura 2.5.

La representación mediante grafos de Sowa [Sowa, 1991] introduce el elemento contexto, que determina qué nivel de profundidad con respecto a la raíz tiene cada nodo. En la Figura 2.6 se muestra el ejemplo anterior representado con un grafo de Sowa.

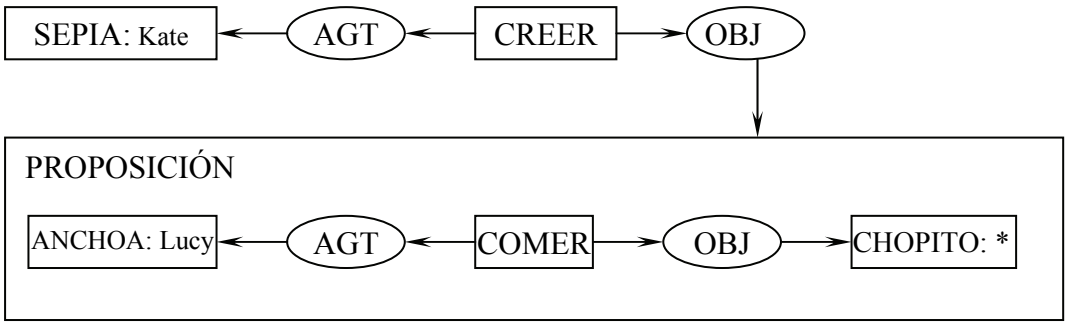


Figura 2.6. Ejemplo de representación con grafos de Sowa.

Tanto en las redes de Shapiro como en los grafos de Sowa se distingue entre conceptos e instancias de ese concepto, es decir, permiten la representación del cuantificador existencial. En el ejemplo anterior, *"el chopito"* es un individuo concreto de la clase de todos los chopitos. En la red de Shapiro se crea un nodo de la clase

"CHOPITO" y se enlaza con el individuo concreto mediante una relación "MIEMBRO". En el grafo de Sowa se representa escribiendo dentro del nodo la clase y a continuación un "*", haciendo referencia a un individuo cualquiera de dicha clase puesto que se desconoce su identidad, al contrario que la sepia Kate y la anchoa Lucy.

Para realizar inferencia en las redes proposicionales se utilizan tres mecanismos básicos:

- La especificación. Es decir, la introducción de elementos para hacer a la representación menos general mediante, por ejemplo, la ejemplificación de un nodo clase.
- La generalización. Es la operación inversa a la anterior. Por ejemplo, se puede generalizar una instancia convirtiendo su nodo correspondiente en un nodo que represente a toda su clase.
- La unión entre dos redes o grafos y su simplificación posterior eliminando la información redundante.

Mediante la combinación de estos tres mecanismos se puede igualar a la capacidad expresiva de la lógica de primer orden y a algunos aspectos de orden superior.

2.2.1.5.2. Redes de clasificación

Las redes de clasificación representan, como su nombre indica, información sobre las categorías a las que pertenecen las entidades contenidas en ellas. La mayoría de los sistemas de representación anteriores contienen información sobre categorías semánticas. Las relaciones y atributos del tipo ES_UN son un ejemplo claro de ello. Sin embargo, las redes de clasificación se constituyen como modelo individual a partir de la introducción de la herencia. Una red de clasificación es un grafo dirigido conexo sin ciclos que parte siempre de un concepto superior, generalmente denotado por la letra T, que es más general que cualquier otro. Los arcos entre conceptos representan relaciones de pertenencia. Así, un arco del nodo A al nodo B significa que B pertenece a A, o que A es más general que B.

El razonamiento o inferencia que se realiza en este modelo de representación es el de la herencia de propiedades. Existen dos tipos de herencia posibles: la herencia estricta y la herencia por defecto. En la herencia estricta los conceptos relacionados con otro más general tienen exactamente las mismas propiedades que dicho concepto. En la herencia

por defecto ocurre lo mismo, siempre y cuando no se indique lo contrario. De esta forma, si un pez come bocadillos de jamón se puede inferir que una anchoa también los come, por ejemplo.

2.2.1.5.3. Redes causales

Las redes causales reciben su nombre del tipo de relaciones que representan. En una red causal, la asociación entre dos nodos indica que el nodo origen influye en el nodo destino y más concretamente que el nodo origen es la causa del nodo destino. Puesto que el objetivo originario de estas redes no era la comprensión del lenguaje natural sino los problemas de diagnóstico, los nodos representan variables que generalmente describen los parámetros del problema. Los nodos se pueden agrupar por niveles permitiendo también relaciones causales entre dichos niveles. CASNET [Weiss et al., 1978] es un ejemplo representativo aplicado al diagnóstico médico del glaucoma.

Hasta el momento, todos los sistemas de representación eran simbólicos, es decir, no empleaban números para cuantificar aspectos de las relaciones o de los conceptos. En las redes causales es donde se introduce dicha cuantificación por primera vez, dando lugar a las redes bayesianas [Jensen, 2001].

Una red bayesiana es un grafo dirigido sin ciclos donde sus arcos tienen asociados un número. Dicho número corresponde a la probabilidad de que se produzca el nodo destino dado el nodo origen, calculándose dicha probabilidad mediante el Teorema de Bayes, de ahí su denominación. Estas redes asumen además que un nodo sólo depende de los nodos que están asociados a él y es independiente del resto. La figura 2.7 muestra un ejemplo sencillo de red bayesiana.

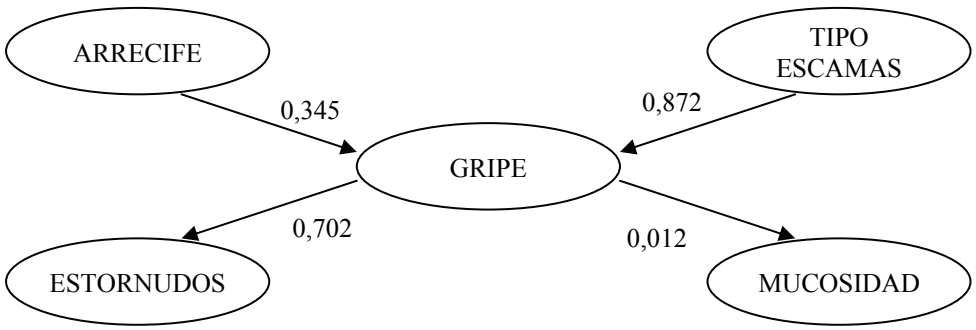


Figura 2.7. Ejemplo de representación con redes bayesianas.

En la figura 2.7 se representa la relación que existe entre el arrecife donde viven las anchoas y el tipo de escamas que poseen, y la presencia o ausencia de gripe, y cómo se manifiesta dicha gripe a través de estornudos y de la presencia de mucosidad en las anchoas.

En este caso, las inferencias que se pueden realizar en dichas redes consisten en el cálculo de las probabilidades de que se produzcan ciertos eventos dadas combinaciones del resto de nodos, como por ejemplo la probabilidad de que una anchoa grazne dado el arrecife donde vive. También se puede realizar razonamiento abductivo, como encontrar el nodo más probable dado otro nodo. El razonamiento deductivo es inmediato, consistiendo en el seguimiento de caminos en la red. Además, la confirmación de ciertos nodos puede desechar otros nodos como causas de uno dado. A este último tipo de razonamiento se le ha llamado "intercausal" [Mira y Delgado, 1995].

Las redes causales, a pesar de permitir diversos tipos de razonamiento y de aportar explicaciones inteligibles de los mismos, tienen un ámbito de aplicación muy definido, puesto que fueron diseñadas *ad hoc* para el diagnóstico. En otro tipo de aplicaciones su uso no es posible o bien resulta muy complicado y artificioso.

Las redes bayesianas se benefician de las ventajas de las redes causales y de la potencia del cálculo de probabilidades. Sin embargo, ese cálculo requiere mucha información numérica de la que es difícil disponer a priori. Además, en determinadas circunstancias el proceso de cálculo puede ser muy ineficiente en términos de tiempo, lo que conduce al uso de simplificaciones y, por tanto, a la pérdida de información y capacidad de representación.

Muchos de los sistemas clásicos de representación descritos fueron ideados para tratar la representación del lenguaje natural, como ya se ha comentado. El resto de sistemas clásicos creados con otros propósitos también han sido empleados alguna vez para tareas relacionadas con el lenguaje. Los sistemas de reglas son particularmente adecuados para representar conocimiento sintáctico y fonológico y, más concretamente, las gramáticas que lo describen [El-Iman, 2005], aunque también se han empleado ciertos tipos de redes bayesianas [Narayanan y Jurafsky, 2003], [Livescu et al., 2002].

Todos los sistemas clásicos descritos parten de conocimiento experto que, como mínimo, prepara la representación para recibir cierto tipo de información. Es decir, estos

sistemas no "aprenden" de la experiencia de manera autónoma como lo hacen los humanos, por lo que estarán limitados por ese conocimiento inicial del dominio que siempre resulta insuficiente para tratar con problemas de lenguaje natural de manera global, dada su extensión y complejidad. Para superar estas limitaciones surgen, desde comienzos de los años 90 hasta la actualidad, nuevos sistemas de representación capaces de almacenar y razonar o inferir sobre cantidades enormes de conocimiento lingüístico que adquieren y extraen de ejemplos reales de lenguaje natural.

2.2.2. Sistemas de representación masiva de conocimiento lingüístico

Los sistemas de representación masiva del conocimiento lingüístico persiguen dos objetivos principales: el tratamiento de gran cantidad de información y la aplicación a problemas de lenguaje natural reales. La forma de almacenamiento de la información y sus mecanismos de tratamiento son pues cruciales para el éxito de los sistemas. Una de las principales características de estos sistemas es el escaso conocimiento a priori que requieren. De hecho, lo único que precisan son ciertos presupuestos en los que se basan para extraer la información lingüística de la mayor cantidad posible de textos en lenguaje natural. La cantidad y naturaleza de dichos textos pueden determinar el nivel lingüístico de un ser humano de acuerdo a una cierta edad y nivel cultural, por lo que estos sistemas pueden modelar el aspecto temporal de la adquisición del lenguaje.

Uno de los primeros sistemas de representación que permite el tratamiento masivo de lenguaje es el denominado comúnmente como "bolsa de palabras" [Sebastiani, 2002]. Es el modelo más extendido hasta el momento. En dicho modelo, los textos se consideran como vectores de tamaño igual al tamaño del vocabulario, es decir, al número total de palabras que aparecen en los textos de entrada después de preprocesarlos. Si una palabra no aparece en un texto, el vector contendrá en la posición correspondiente a la palabra un valor igual a 0. En caso contrario, el valor contenido en dicha posición será igual a la frecuencia global de la palabra multiplicado por la inversa de la frecuencia de los textos donde aparece dicha palabra. Esta expresión matemática es la que da el nombre a la representación *TF·IDF* (*Term Frequency · Inverse Document Frequency*). El apelativo de "bolsa de palabras" se deriva de la consideración de independencia entre las palabras en un mismo texto, tratándolas como entidades independientes y ponderando su aportación semántica de manera global mediante un

valor estadístico estático. A pesar de que el empleo de esta representación como entrada para los algoritmos de descubrimiento de conocimiento permitió el tratamiento masivo de textos con unos resultados lo suficientemente aceptables para su utilización en aplicaciones reales, lo cierto es que no recoge ni aprovecha en absoluto la información lingüística encerrada en los textos y, por tanto, no permite la realización de inferencias y razonamientos en lenguaje natural.

2.2.2.1. "Concurrencia" léxica

Cubriendo la carencia principal de la representación clásica de "bolsa de palabras", el supuesto principal en el que se basan la mayoría de los sistemas de representación masiva es la hipótesis de Miller y Charles [Miller y Charles, 1991] de "coocurrencia" o concurrencia léxica: dos palabras que aparecen en un mismo contexto, es decir, que "coocurren", están semánticamente relacionadas. De esta forma el conocimiento lingüístico se construye asociando las palabras que aparecen en el mismo contexto y asociando las palabras a su vez al contexto en el que aparecen. Así pues, la sintaxis y semántica del lenguaje queda determinada por relaciones contextuales entre las palabras extraídas de ejemplos de uso en textos en lenguaje natural. Son sistemas que representan la semántica léxica y utilizan una noción matemática de "composicionalidad", es decir, la semántica de una oración en función de la semántica de las palabras que contiene, en contraste con la semántica estructural que propone la integración de la semántica léxica guiada por la sintaxis.

Aparte de las relaciones sintácticas obvias, las relaciones semánticas comúnmente consideradas hasta finales de los años 90 son las llamadas relaciones semánticas clásicas. Estas consisten en la hiponimia, hiperonimia, toponimia, meronimia, antonimia y sinonimia. Existen diferentes ontologías que almacenan estas relaciones y que son empleadas por los sistemas de representación del lenguaje para realizar inferencias y razonamientos. La más representativa de estas ontologías o taxonomías léxicas es WordNet [FellBaum, 1998]. Estas relaciones requieren que las palabras tengan la misma categoría sintáctica.

Sin embargo, existen otros tipos de relaciones semánticas en los textos. Los resultados de un estudio realizado por Morris [Morris y Harris, 2004], donde se pedía a una serie de sujetos que agruparan las palabras que aparecían en un texto en diferentes categorías, según criterios personales y subjetivos, desvelan que la mayoría de los

sujetos coinciden en la identificación de trece categorías distintas de palabras, entre las que se encuentran relaciones sintácticas y algunas de las relaciones clásicas, aunque la mayoría de ellas relacionan las palabras de otras maneras. Entre estas relaciones no clásicas se pueden mencionar relaciones estereotípicas, de definición, positividad o negatividad y contextuales. Según Burgess, la naturaleza de estas últimas asociaciones que implica la concurrencia de dos palabras en un mismo contexto puede ser de tres tipos [Burgess, 1998]:

- Semántica: Sus significados son similares o bien pertenecen a una misma categoría semántica, como por ejemplo "anchoa-mar" o "música-arte".
- Asociación: Sus significados individuales no son similares, pero aparecen con frecuencia en el mismo contexto para matizar otro significado. Ejemplos: "pan-molde" o "pez-martillo".
- Semántica-asociación: Se dan las dos anteriores al mismo tiempo, como en "tío-abuelo" o "negro-azulado".

Estudios realizados por Lund [Lund y Burgess, 1995; 1996] revelan que los seres humanos parecen almacenar dichas asociaciones en su mente de alguna manera, puesto que responden ante las mismas mucho más rápido que ante palabras no relacionadas de alguna de las tres formas anteriores. Así pues, la manera de definir el contexto y la asociación de las palabras dentro del mismo condiciona qué tipo de relaciones va a almacenar el sistema de representación.

La concurrencia léxica ha abierto una nueva línea de investigación en los modelos del lenguaje, puesto que supone una manera simple de adquirir sintaxis y semántica descriptiva de manera automática y masiva a partir de fragmentos de lenguaje natural real, es decir, no diseñado *ad hoc*, y sin requerir conocimiento experto.

2.2.2.2. Características generales

Según Lemaire [Lemaire y Denhière, 2004], los sistemas de representación masiva del conocimiento lingüístico se caracterizan por seis aspectos básicos:

- Tipo de entrada. A partir de qué naturaleza y de qué cantidad de información va a adquirir conocimiento el sistema. Pueden ser colecciones de textos, normas de asociación preestablecidas por un experto o combinaciones de ambas.

- Formalismo de representación. Puede ser cualquiera de los sistemas clásicos que se ha repasado anteriormente pero, teniendo en cuenta que se trata con grandes cantidades de información, lo más común es que empleen distribuciones de probabilidad, matrices o ciertos tipos de redes asociativas.
- Dinamismo temporal. Indica si es posible o no añadir conocimiento lingüístico una vez construido uno inicial y la dificultad y problemas que puede conllevar esa actualización.
- Definición del contexto. Tamaño del contexto en el que se va a considerar que las palabras concurren y determinación y cuantificación de las asociaciones encontradas.
- Concurrencias de orden superior. Indica si el sistema captura y hace uso o no de asociaciones indirectas y de qué orden. Si A está asociado a B y B a C, entonces una asociación indirecta de orden 2 se daría entre A y C.
- "composicionalidad". La manera en que el significado de un fragmento de lenguaje es inferido a partir de los significados de las palabras que lo forman.

A estos seis aspectos propuestos por Lemaire se pueden sumar también los siguientes:

- el tipo de asociaciones contextuales que recoge de entre las tres citadas anteriormente,
- la aplicación para la que está diseñado el modelo,
- qué conocimiento a priori se incluye o en qué interviene el experto,
- si recoge o no conocimiento sintáctico,
- si permite la generación de lenguaje y
- su plausibilidad psicológica.

La plausibilidad psicológica indica si el sistema de representación y sus mecanismos de adquisición son posibles en la mente de los humanos, basándose en estudios psicológicos y en métodos introspectivos. Este último aspecto es muy controvertido, ya que algunos de los sistemas existentes son presentados como una teoría cognitiva del lenguaje como proceso mental a pesar de que, aunque eficientes y eficaces en su ámbito

de aplicación, son de una naturaleza estrictamente matemática y, por tanto, dudosos de ser atribuidos a la mente humana.

Aunque todos los aspectos citados son importantes a la hora de describir un sistema de representación del conocimiento lingüístico se utilizará el formalismo de representación como criterio de agrupación de los sistemas existentes, puesto que es el más visual. De una u otra forma, los sistemas descritos a continuación tratan de dar respuesta a las preguntas fundamentales, propuestas por Dewey [Dewey, 1916; 1938], de la psicolingüística y de la lingüística cognitiva: ¿cómo están relacionados los significados entre sí? y ¿cómo están representados y estructurados los significados de manera individual?

2.2.2.3. Espacios de memoria de alta dimensión

Los espacios de memoria de alta dimensión se caracterizan por representar la semántica léxica como espacios matemáticos de *n* dimensiones, siendo *n* un número del orden de las decenas de millar en adelante. El formalismo empleado es el de matriz. Habitualmente, las filas de las matrices contienen la información sobre las palabras y las columnas contienen la información relativa a los contextos. Las dimensiones del espacio y, por tanto, de la matriz pueden venir dadas como resultado de procesar la entrada del sistema o pueden ser fijadas por un experto de antemano.

El ejemplo más representativo y el precursor de todos los sistemas de representación masiva es el modelo de Análisis de Semántica Latente o LSA (*Latent Semantic Analysis*) [Landauer y Dumais, 1996], [Landauer et al., 1997]. Este sistema fue ideado por Dumais y sus colaboradores hacia 1990, aunque alcanzó su reconocimiento y auge a partir de 1996. En LSA, el conocimiento semántico se obtiene de colecciones de textos compiladas con el objetivo de representar cierto nivel cultural. Este conocimiento se almacena en una matriz con tantas filas como palabras distintas se encuentren en los textos de entrada y tantas columnas como contextos se definan. Estos contextos no son completamente definidos por un experto pero sí delimitados de antemano. Los contextos referidos son grupos de párrafos, llamados pasajes, existentes en los textos. Así, el conocimiento queda representado por una matriz de palabras por pasajes, como se muestra en la Figura 2.8.

De esta forma, el sistema representa a las palabras como vectores de longitud fija correspondientes a las filas de la matriz. La distancia semántica entre dos palabras

vendrá dada por la distancia Euclídea de sus vectores. Los fragmentos de texto están representados también por vectores que son el resultado de realizar la media ponderada de los vectores correspondientes a las palabras que aparecen en ellos, implementando así la "composicionalidad" de manera sencilla y eficiente. El sistema representa pues el significado de las palabras y el de las oraciones y textos de la misma manera, con vectores.

	Pasaje 1	**…**	**Pasaje j**	**…**	**Pasaje n**
Palabra 1	v_{11}	…	v_{1j}	…	v_{1n}
…	…	…	…	…	…
Palabra i	v_{i1}	…	v_{ij}	…	v_{in}
…	…	…	…	…	…
Palabra k	v_{k1}	**…**	v_{kj}	**…**	v_{kn}

Figura 2.8. Formalismo de representación de la semántica léxica en el modelo de Análisis de Semántica Latente (LSA).

Para obtener la matriz final que corresponderá al conocimiento lingüístico latente, primero se crea una matriz como la de la Figura 2.8, donde v_{ij} es la frecuencia con la que la palabra *i* aparece en el pasaje *j*. A continuación, cada valor de la matriz es ponderado mediante una función que tiene en cuenta la relevancia de la palabra en el pasaje y la información que la palabra aporta al dominio del discurso de manera global. A dicha matriz se le aplica una descomposición en valores singulares SVD (*Singular Value Decomposition*) [Golub y Van Loan, 1996]. La descomposición en valores singulares es un método matemático de análisis de factores en el que, dada una matriz *M*, la descompone en un producto de otras tres matrices $S \cdot V \cdot D$, con la particularidad de que la matriz central *V* del producto es una matriz que sólo contiene valores distintos de cero en su diagonal. Dicho valores son los valores singulares. Si se realiza el producto de las tres matrices se obtiene la matriz original. Sin embargo, el sistema LSA realiza una reducción intermedia de dimensiones. Para ello, selecciona las filas y las columnas de las tres matrices que se corresponden con las filas y columnas de los *m* mayores valores de la diagonal de la matriz central *V*. Una vez eliminadas las columnas y filas restantes se realiza el producto de matrices, obteniendo una matriz *M'* distinta de la original. Es lo que llaman espacio de semántica latente de *m* dimensiones. Esta última matriz *M'* captura mejor que la matriz original *M* las relaciones semánticas entre las

palabras, mostrando una correlación alta entre palabras semánticamente similares que no aparecían en el mismo contexto en los textos de entrada y una correlación baja entre palabras que, a pesar de aparecer en el mismo contexto, tienen significados muy alejados el uno del otro. De esta forma, el proceso de reducción de dimensiones da lugar a relaciones indirectas de órdenes superiores entre las palabras.

La elección del número de dimensiones *m* es crucial para determinar la eficacia del modelo y se establece de manera empírica. Estudios realizados por los autores concluyen que el número de dimensiones ha de oscilar entre 50 y 400 dependiendo del ámbito de aplicación. Fuera de estos límites el modelo no es capaz de reflejar el conocimiento semántico de manera apropiada.

La descomposición en valores singulares y la reducción de dimensiones son procesos que requieren una gran cantidad de tiempo en términos computacionales, teniendo en cuenta además que los problemas que se abordan pueden implicar cientos de miles de palabras y, por lo tanto, el mismo número de filas en las matrices. Así pues, dado que la introducción de nuevo conocimiento en el espacio semántico requiere la realización entera del proceso de construcción y reducción de dimensiones a partir de los textos de entrada junto con la información nueva, la actualización temporal del modelo resulta muy costosa.

En cuanto al conocimiento sintáctico se refiere, los autores señalan que se halla presente de manera implícita en el modelo a pesar de que el proceso adquisición del conocimiento no lo recoge de manera explícita. La prueba de dicha existencia es experimental. Dado un conjunto de palabras pertenecientes a distintas categorías léxicas, principalmente nombres, verbos, adverbios y artículos, sus correspondientes vectores de la matriz de conocimiento son escalados al espacio de dos dimensiones y representados gráficamente. En dicha representación se observa que las palabras quedan agrupadas en el plano por categorías léxicas, siendo linealmente separables.

El modelo ha sido probado con éxito en diversos ámbitos de aplicación. Dumais aplicó el modelo en la recuperación de información [Dumais, 1994], obteniendo un 16% de mejora con respecto a sistemas de representación similares. El resto de aplicaciones implican una comparación del modelo con los humanos: juicios de sinonimia, relación de pares de conceptos [Landauer et al., 1998], identificación la semántica de palabras polisémicas en el contexto [Landauer y Dumais, 1997], medida de la calidad de respuestas a preguntas de exámenes [Landauer et al., 1998] y cuantificación del

conocimiento que puede aportar un texto. Los resultados de todos estos trabajos indican que el modelo se comporta de manera muy similar a los humanos en todas las tareas mencionados. Las diferencias que aparecen son justificadas por los autores basándose en la idea de que los humanos se comportan de manera "rara" en algunas ocasiones y eso es lo que genera la diferencia de comportamiento. De estos resultados concluyen que la mente humana debería funcionar de manera similar al modelo LSA puesto que de otra manera no se podrían dar los resultados obtenidos. Así pues, presentan el modelo como una teoría de la semántica de realidad psicológica.

Otro modelo similar a LSA pero de metodología más sencilla e intuitiva es el de Hiperespacio Analógico para el Lenguaje, HAL (*Hyperspace Analogue to Language*), ideado por Curt Burgess hacia 1998 [Burgess, 1998]. En el caso de HAL, el conocimiento semántico también se adquiere de la concurrencia de las palabras en un mismo contexto. Sin embargo, existen tres diferencias significativas:

- El conocimiento que se almacena son las relaciones entre las palabras, no entre palabras y pasajes.
- El contexto es una ventana de un tamaño fijo de palabras que se desplaza a lo largo del texto.
- Se cuantifica la concurrencia de las palabras en un mismo contexto mediante la distancia que las separa en dicho contexto.

De esta forma, el sistema es más autónomo que LSA puesto que el experto no ha de definir contextos, sólo la forma de cuantificar las distancias. Sin embargo, un contexto de tamaño fijo que se desplace a lo largo del texto es cuestionable desde el punto de vista semántico, puesto que la ventana contextual puede abarcar en algún momento a dos párrafos distintos dando lugar a que se relacionen las palabras de ambos. El modelo HAL establece el tamaño de la ventana contextual en 10 palabras, que es el tamaño que mejor resultado produce según un estudio experimental de Lemaire [Lemaire y Denhière, 2004].

El sistema HAL almacena, además de palabras, la semántica de cifras, emoticonos y otros símbolos. Para recoger un conocimiento lingüístico amplio y preciso se requieren una gran cantidad de textos de entrada. Esto supone que la matriz de representación puede alcanzar dimensiones enormes, del orden de los cientos de miles de palabras,

números y símbolos. Para favorecer la eficiencia, el sistema HAL sólo emplea las 70,000 palabras (y cifras o símbolos) más frecuentes en los textos de entrada. En ocasiones realizan una selección sobre esas 70,000 palabras calculando la varianza de filas y columnas y seleccionando las *k* palabras con mayor varianza, generalmente con *k* entre 100 y 200.

	Palabra 1	**…**	**Palabra j**	**…**	**Palabra n**
Palabra 1	v_{11}	…	v_{1j}	…	v_{1n}
…	…	…	…	…	…
Palabra i	v_{i1}	…	v_{ij}	…	v_{in}
…	…	…	…	…	…
Palabra n	v_{n1}	**…**	v_{nj}	**…**	v_{nn}

Figura 2.9. Formalismo de representación de la semántica léxica en el modelo de Hiperespacio Analógico para el Lenguaje (HAL).

La Figura 2.9 muestra el formalismo de representación empleado por HAL. En cada posición v_{ij} se almacena la suma de las distancias que existen entre la palabra *i* y la palabra *j* en cada contexto en el que aparece *i* antes que *j*. Puesto que no existe ningún tipo de tratamiento posterior, el sistema HAL sólo trata con relaciones de concurrencia directas o de primer orden. Así, el significado de una palabra viene dado por el grado de relación que tiene con las demás palabras y por el grado de relación que las demás palabras tienen a su vez con ella, como se muestra en la Figura 2.10. De esta manera, la representación de la semántica de una palabra queda representada mediante su correspondiente vector fila concatenado con su correspondiente vector columna.

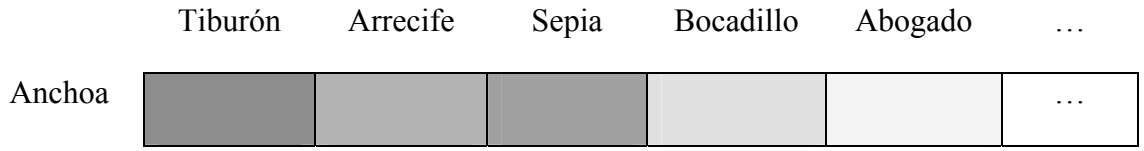


Figura 2.10. Ejemplo de representación de la semántica de la palabra "anchoa" en el sistema HAL.

Este tipo de representación permite recoger la semántica tanto de conceptos básicos como de conceptos abstractos. Para medir la similitud entre palabras simplemente se calcula la distancia Euclídea entre los vectores correspondientes a las palabras. Al igual

que el sistema LSA, el significado de una secuencia de palabras en el sistema HAL se representa como la media de los vectores de dichas palabras.

Para cuantificar y cualificar el conocimiento contenido en la representación, HAL ha sido aplicado a diversas tareas comparándolo además con los seres humanos. En primer lugar se realizan experimentos de categorización semántica, donde dado un conjunto amplio de palabras, se calcula en el modelo la distancia de cada palabra a las demás y se agrupan por dicha distancia, obteniéndose grupos que efectivamente corresponden a categorías semánticas diferenciadas. En segundo lugar se realizan pruebas para determinar el "vecindario" semántico: dado un conjunto de palabras que están relacionadas con otra palabra objetivo en un determinado contexto, se ha de encontrar dicha palabra objetivo. En estas pruebas los humanos resultaron ser más precisos que el modelo, encontrando éste la palabra objetivo en un 20% de las ocasiones y palabras muy similares en el resto de los casos. En cuanto a conocimiento semántico se refiere, los últimos experimentos se refieren a la reacción semántica, donde dada una palabra, se ha de responder con la primera palabra que surja en la mente de los individuos. El modelo HAL responde con la palabra más similar a la palabra dada. Los resultados demuestran que el modelo se comporta de manera similar a los humanos en el caso de relaciones semánticas, sin embargo no es capaz de capturar las relaciones de asociación y semánticas-asociación.

El conocimiento sintáctico implícito se pone de manifiesto con pruebas similares a las realizadas con el modelo LSA. Los vectores de un conjunto de palabras de categorías sintácticas distintas se escalan a dos dimensiones y se representan en el plano, observando que se agrupan por categorías sintácticas. El mismo experimento se realiza con palabras de la misma categoría sintáctica, concretamente determinantes, observando que en la representación de dos dimensiones se agrupan en artículos, demostrativos y posesivos.

Al contrario que LSA, los creadores de HAL no defienden su plausibilidad psicológica, presentándolo como una mera herramienta para la compresión automática del lenguaje natural.

El modelo Espacios de Asociación de Palabras, WAS (*Word Association Spaces*) [Steyvers et al., 2004] es un sistema de representación que emplea también la

descomposición en valores singulares de manera análoga a LSA. WAS utiliza una matriz como formalismo de representación que recoge relaciones entre cada par de palabras como el sistema HAL. Sin embargo, la mayor diferencia radica en la adquisición de dichas relaciones. En este caso, las asociaciones entre las palabras son obtenidas directamente de sujetos humanos mediante una encuesta, llamadas normas de asociación libres. Así pues, el valor v_{ij} de la matriz M corresponde a la proporción de sujetos encuestados que respondieron la palabra j cuando se les presentó la palabra i. A esta matriz se le aplican métodos de escalado multidimensional mediante descomposición de valores singulares, posicionando las palabras en un espacio multidimensional de tal forma que aquellas con patrones de asociación similares están en regiones del espacio próximas. Este es el espacio denominado Espacio de Asociación de Palabras. El número de dimensiones del escalado depende de la tarea a realizar, pero el número óptimo se estima experimentalmente entre 200 y 500 dimensiones.

Previamente al proceso de escalado se crean otras dos matrices simétricas a la obtenida con las normas de asociación libres. En la primera matriz S^1 se establece cada posición como: $a_{ij} = v_{ij} + v_{ji}$. En la segunda matriz S^2, los valores r_{ij} son calculados como: $r_{ij} = a_{ij} + \sum_k a_{ik} \cdot a_{kj}$. A continuación se aplica la reducción de dimensiones a las dos matrices.

Una vez más, la semántica de las palabras queda representada como el vector fila correspondiente. Para que se pueda medir la distancia semántica entre palabras como la distancia Euclídea entre sus vectores, se crea un espacio de métricas. Con ese fin se genera una nueva matriz D que contiene en cada posición d_{ij} el producto de los de los valores del camino más corto en alguna de las matrices S^1 o S^2, entre la palabra i y la palabra j. A continuación se aplica un escalado multidimensional a la matriz D. Así, la distancia entre dos palabras se calcula mediante la distancia Euclídea entre sus correspondientes vectores en la matriz D.

La mayoría de las pruebas realizadas para determinar la calidad de la representación WAS se refieren a la predicción de los efectos de similitud semántica en la memoria, comparando los resultados con los obtenidos por los humanos y por el sistema LSA. Los experimentos realizados consisten en la asignación, a palabras objetivo, de un grado de pertenencia a ciertas categorías semánticas. También se realiza la recuperación de palabras similares a una dada a partir de una lista y la inferencia de palabras objetivo, dadas las listas de palabras que más se parecen a ellas. Las conclusiones que se

desprenden de estos experimentos es que tanto el beneficio de la aplicación de la descomposición en valores singulares como el número óptimo de dimensiones para el escalado y la conveniencia del uso de la matriz S^1 o de la matriz S^2 para obtener la matriz *D* dependen de la tarea que se realice. El sistema WAS supera en todas las pruebas al sistema LSA, aunque es necesario decir que las normas de asociación libres forman parte de la predicción de los fenómenos de la memoria episódica, razón por la cual WAS funciona mejor. Los autores finalmente concluyen que ningún sistema de representación masivo es mejor que otro de manera general sino que depende de la tarea a la que se aplique.

Los sistemas descritos hasta el momento están compuestos por información en un solo nivel, el léxico, aunque contengan conocimiento de nivel superior de manera latente o implícita. El sistema de Espacio Semántico Unitario de Características [Vigliocco et al., 2004], FUSS (*Featural and Unitary Semantic Space*) es de los pocos modelos que considera expresamente dos niveles de conocimiento, el léxico-semántico y el conceptual, previamente identificados en diversos trabajos de neurociencia y lingüística [Damásio y Damásio, 1992], y los integra de forma cuantitativa. Presentan su modelo como una hipótesis teórica delimitada por evidencias neuropsicológicas y neuroanatómicas, por lo que persigue la plausibilidad psicológica en todos sus aspectos. Dicha hipótesis se formula bajo los siguientes supuestos:

- Las representaciones de las palabras mediante características son componentes esenciales de la estructura conceptual subyacente.
- Estas representaciones están enmarcadas en representaciones léxico-semánticas cuya organización es dictada por las propiedades de las características empleadas.
- El mismo tipo de representación puede ser empleado para describir palabras en diferentes dominios de contenido.

El nivel conceptual está formado por estructuras que contienen características específicas de modalidad, es decir, las características del nivel conceptual se refieren al modo mediante el que los humanos adquieren el conocimiento de la experiencia. Estudios que emplean técnicas de exploración cerebrales PET y fMRI [Démonet y Thierry, 2001] demuestran que multitud de objetos y eventos no se relacionan por su

categoría semántica, sino por su manera de usarlos o llevarlos a cabo. Así pues, las palabras estarán representadas en un espacio de características. Dichas características no están definidas por un experto sino que se obtienen de sujetos humanos: se pide a 20 personas que usen suficientes características para definir cada palabra perteneciente a una lista de 456 términos, de manera análoga al trabajo de McRae [McRae et al., 1997]. De esas 456 palabras, 230 son nombres y 216 verbos. De los nombres, 169 se refieren a objetos y 71 a acciones. Todos los verbos hacen referencia a acciones. Los objetos pertenecen a 7 categorías semánticas distintas (animales, frutas, ropa, etc.) y las acciones a otras 15 categorías semánticas (comunicación, cocina, acciones corporales, cambio de estado, etc.). El formalismo de representación es pues una matriz de palabras por características como muestra la Figura 2.11. El valor de la posición v_{ij} corresponde al número de personas que han generado la característica *j* para describir a la palabra *i*. A continuación se realiza una selección de características, eliminando aquellas usadas por menos de seis personas.

	Ctca. 1	**…**	**Ctca. j**	**…**	**Ctca. k**
Palabra 1	v_{11}	…	v_{1j}	…	v_{1k}
…	…	…	…	…	…
Palabra i	v_{i1}	…	v_{ij}	…	v_{ik}
…	…	…	…	…	…
Palabra n	v_{n1}	**…**	v_{nj}	**…**	v_{nk}

Figura 2.11. Formalismo de representación de la semántica conceptual en el modelo de Espacio Semántico Unitario de Características (FUSS).

Para validar su hipótesis, los autores clasifican las características obtenidas de los sujetos en varias categorías como "visual", "funcional" y "movimiento", entre otras, y las ordenan según el número de descripciones de palabras en el que hayan sido empleadas. Así, por ejemplo, para el conjunto de características pertenecientes a la categoría "visual", la mayoría de las palabras descritas por dichas características son las pertenecientes a la categoría "emisión de luz". De esta forma demuestran la relación entre la categoría semántica de las palabras y la forma en la que se aprenden al percibirlas.

El nivel léxico-semántico se deriva del nivel anterior. Los autores de FUSS asumen la existencia de este nivel por tres razones:

- Usar el lenguaje requiere asociar representaciones conceptuales con información lingüística.
- Provee de una hipótesis explícita para explicar la variabilidad entre concepto y su palabra correspondiente en distintas lenguas.
- Proporciona un nexo entre el conocimiento conceptual y la información sintáctica.

El formalismo empleado para la representación léxico-semántica es el de Mapas Auto-Organizativos SOM (*Self-Organizing Maps*) [Kohonen, 1995]. Definen 100 mapas de 20x40 unidades y los entrenan con los vectores fila correspondientes a los conceptos del nivel conceptual presentados cada vez en distinto orden. Una vez entrenados, y dado un vector de entrada, la unidad más activa una vez estabilizado el mapa es la representación léxico-semántica del concepto asociado al vector de entrada. La similitud entre dos conceptos viene dada por la media de las distancias de las representaciones léxico-semánticas de dichos conceptos en todos los mapas. Puesto que esta medida de similitud no es simétrica y la recopilación de características de sujetos humanos tiende a minimizar las relaciones, la medida no tiene por qué reflejar asociaciones de palabras. Sin embargo, la agrupación de palabras mediante esta medida permite diferenciar las categorías semánticas de los objetos y relacionarlos con las acciones en las que están implicados.

Siguiendo la línea de extraer características para representar la semántica de las palabras se encuentra el sistema de Similitud No Latente NLS (*Non-Latent Similarity*) [Cai et al., 2004]. En este modelo, al contrario que en FUSS, las características empleadas se extraen de los textos de entrada y no de sujetos humanos. Dichas características son de naturaleza sintáctica y se componen de una palabra y una relación que denota el rol sintáctico de dicha palabra. La información extraída es del tipo:

Palabra 1	*Relación*	*Palabra 2*	
Vivir	V:sujeto:N	Gente	
Comer	V:cd:N	Jamón	…

Características son, por ejemplo, (**Vivir**, **V**:sujeto:N) o (**Jamón**, V:cd:**N**). De los textos de entrada se extraen cerca de 400,000 características distintas de este tipo a las

que se asigna un peso, mediante una modificación del algoritmo de Lin [Lin, 1998] para denotar su importancia. Para cada palabra de los textos de entrada se construye su vector de características correspondiente y, a continuación, se construye una matriz que representará al espacio de conocimiento y que contendrá en cada posición v_{ij} la similitud entre la palabra *i* y la palabra *j*. Dicha similitud corresponde al coseno entre los vectores de cada palabra. De esta forma se tiene la matriz de distancias denominada FOM (*First-Order Similarity Matrix*). Un fragmento de texto estará representado por un vector de las palabras que contiene. La similitud entre dos textos es una función que implica la multiplicación de ambos vectores de palabras por la matriz FOM.

Dado que los autores piensan que palabras similares tienen en común muchas de las palabras que más se parecen a ambas a su vez, información que se recoge en las relaciones indirectas, se construye una matriz de orden superior denominada SOM (*Second-Order Similarity Matrix*). Dicha matriz se deriva de FOM multiplicando ésta por sí misma y por una matriz diagonal formada por las recíprocas de las normas de los vectores columna de FOM.

Los experimentos realizados están encaminados a medir la independencia del contexto del conocimiento representado, por provenir las características de la sintaxis. Para ello tratan de identificar asociaciones de verbos con verbos, nombres con nombres y modificadores con modificadores, y las comparan con las obtenidas por el sistema LSA. Puesto que los verbos y modificadores dependen en gran medida del contexto, NLS obtiene resultados iguales o ligeramente peores que LSA en este tipo de asociaciones. Sin embargo, en las asociaciones de nombres los resultados son significativamente mejores que los obtenidos por LSA, cubriendo así la dependencia exclusiva del contexto. Además, NLS contiene información sintáctica, ya que la adquiere de manera explícita de los textos de entrada.

Hasta ahora, los sistemas descritos representan palabras en un solo contexto. El sistema de Agrupaciones de Sentido (*Sense Clusters*) [Kulkarni y Pedersen, 2005] extrae características léxicas de un conjunto de textos de entrada que conforman contextos de distinto tamaño para representar fragmentos de texto. Inicialmente, los autores definen unas palabras objetivo dependientes del dominio de aplicación de la tarea a realizar. A continuación extraen los diferentes tipos de características: las

palabras, los bigramas (o grupos de dos palabras consecutivas) y los grupos de palabras que contengan alguna de las palabras objetivo.

Con estas características crean una matriz de documentos por característica como la de Figura 2.12a. Es lo que los autores llaman representación de primer orden. Para capturar relaciones indirectas de orden superior entre las palabras crean otra matriz de características por características, mostrada en la Figura 2.12b, que llaman representación de segundo orden.

	Palabra 1	**…Palabra k**	**Bigrama 1**	**… Brigrama h**	**…**
Texto 1	v_{11}	…v_{1k}	v_{1k1}	…v_{1kh}	…
…	…	…	…	…	…
Texto i	v_{i1}	…v_{ik}	v_{ik1}	…v_{ikh}	…
…	…	…	**…**	…	…
Texto n	v_{n1}	…v_{nk}	v_{nk1}	…v_{nkh}	…

a)

	Palabra 1 **Bigrama 1**	**…**	**Palabra m** **Bigrama n**
Palabra 1 - Brigrama 1	v_{11}	…	v_{1mn}
…	…	…	…
Palabra m - Brigrama n	v_{nm1}	…	v_{mnmn}

b)

Figura 2.12. Formalismo de representación a) de primer orden y b) de segundo orden de la semántica léxica en el modelo de Agrupaciones de Sentido.

A ambas matrices se les aplica una reducción de dimensiones mediante descomposición en valores singulares, análogo al procedimiento del sistema LSA. En la representación de primer orden, la matriz contiene los valores de frecuencia de las características en los textos, quedando éstos últimos representados como su vector fila correspondiente. En la representación de segundo orden, la matriz contiene el número de veces que las palabras aparecen en el mismo documento y los textos se representarán por el vector promedio de los vectores fila correspondientes a las palabras que contienen. Utilizando ambas matrices, los textos y las palabras se someten a un

algoritmo de agrupación, concretamente el algoritmo conocido como *Repeated Bisections* [Zhao y Karypis, 2003]. Así, se obtienen grupos de sentido semántico etiquetados por las características más discriminantes.

El sistema ha sido utilizado para discriminar nombres propios y para clasificar correos electrónicos. Los resultados muestran que la representación de primer orden funciona mejor cuando sólo existen dos categorías temáticas distintas en el problema de clasificación. Cuando hay más categorías los dos niveles de representación son comparables.

Cualquiera de los sistemas anteriores resulta difícilmente adaptable a la inclusión de nuevo conocimiento. Con el objetivo principal de lograr adaptabilidad y flexibilidad se diseñó el sistema de Indexado Aleatorio RI (*Random Indexing*) [Sahlgren, 2005]. Su autor postula que el significado es dinámico, varía con el tiempo y es relativo al contexto. La ambigüedad, levedad y no completitud del lenguaje son propiedades esenciales que, contrariamente al pensamiento de los defensores de la lógica, es necesario tener en cuenta por ser aspectos de prosperidad en la comunicación y no de deterioro o malfuncionamiento. Según el autor, todas estas propiedades se recogen en la representación con vectores:

- Los vectores no significan nada por sí solos. Se miden relaciones entre ellos, luego son relativistas. Este es el aspecto esencial en el que se basa el sistema de Indexado Aleatorio.
- La semántica se extrae de manera automática de texto no estructurado, sin conocimiento previo. Esto hace que sean adaptativos y dinámicos a la hora de incluir nuevos textos y conocimiento lingüístico.
- No requieren el concepto de significado y semántica. Sólo requieren la hipótesis de Miller y Charles de la "coocurrencia léxica".

Sin embargo, los vectores presentan ciertas limitaciones también presentes en los sistemas descritos hasta el momento. La principal desventaja es que necesitan una reducción de dimensiones para ser tratables y esa reducción implica estacionalidad, perdiendo el dinamismo y la flexibilidad, además de introducir un alto coste computacional.

El modelo de Indexado Aleatorio propone representar el significado de las palabras como vectores de un número de dimensiones fijo suficientemente pequeño para que sea tratable y eficiente. Los valores de dicho vector son aleatorios en el rango {0,1}, generando un vector distinto para cada palabra en los textos de entrada. Cada vez que dos palabras coocurran el vector de una se sumará al de la otra y viceversa. De esta manera, una vez tratados todos los textos de entrada, una palabra está representada por un vector resultado de la suma de todos los vectores de las palabras con la que ha aparecido en el mismo contexto. De igual forma, el significado de un fragmento de texto está representado por el vector resultado de la suma de todos los vectores de las palabras que lo forman. Así pues, la introducción de nuevo conocimiento sólo supone la generación aleatoria de un vector y una serie de sumas.

El sistema fue comparado con LSA en la búsqueda de palabras relacionadas a una dada. Los resultados de ambos sistemas son equiparables pero el modelo RI es, obviamente, mucho más eficiente y flexible. El autor expone, sin embargo, una serie de desventajas que presenta su modelo causadas principalmente por la naturaleza de los vectores, y que no permiten situar al Indexado Aleatorio como un modelo teórico del lenguaje:

- Los vectores pueden reflejar sinonimia pero no polisemia. Una palabra polisémica siempre se representa con el mismo vector.
- Un vector no aporta información alguna por sí solo por lo que no se puede validar el modelo mediante la mera observación de los mismos.

El autor concluye con una serie de recomendaciones para entender el conocimiento semántico adquirido y representado con los vectores:

- Es necesario investigar la influencia de los textos de entrenamiento en el modelo.
- Se requiere un mejor conocimiento de cómo el significado reside en el lenguaje y de cómo emerge en los textos.
- Se requiere también un mejor conocimiento de la dinámica del espacio de vectores y de sus parámetros
- Son necesarias mejores métricas de evaluación para decidir la validez de las relaciones capturadas con las estadísticas de concurrencia.

También en busca de una solución para el problema de la reducción de dimensiones, Koldo y O'Leary propusieron una modificación del sistema LSA [Kolda y O'Leary, 1998]. Dicha modificación consiste en sustituir el método de descomposición en valores singulares por el método de descomposición semi-discreta SDD (*Semi-Discrete Decomposition*). El método fue originalmente creado para la compresión de imágenes [O'Leary y Peleg, 1983] pero después fue aplicado a la reducción de dimensiones, obteniendo valores pertenecientes al conjunto {-1,0,1} en la matriz final. Los resultados de los experimentos en recuperación de información muestran que la precisión del modelo es comparable a la del sistema LSA original. Sin embargo, el método SDD consume la mitad de tiempo y de espacio que el método de descomposición en valores singulares. Además, eliminar conocimiento en la matriz obtenida con SDD simplemente se consigue mediante la eliminación de las columnas o filas correspondientes, mientras que LSA requiere realizar de nuevo todos los cálculos. Así, el método SDD facilita uno de los aspectos relativos a la actualización del conocimiento, dotando al modelo de cierta flexibilidad, aunque la eliminación de conocimiento linguístico no sea un proceso muy común en el ser humano.

Uno de los problemas que plantean la mayoría de los modelos anteriores es que, puesto que la representación de los textos es una combinación de los vectores de las palabras que los forman, dicha representación contiene todos los posibles significados de las palabras en la semántica de los textos, lo que no es psicológicamente plausible. Con esta motivación principal se crea el modelo de Abstracción de Frases Dependiente del Contexto CDSA (*Context-Dependent Sentence Abstraction*) [Ventura et al., 2004]. En este modelo las palabras se representan mediante vectores de sólo "vecinos", es decir, de palabras que coocurren directamente con ellas en el mismo contexto. El conocimiento queda representado en una matriz de palabra por palabra donde cada componente v_{ij} contiene el peso o la importancia que la palabra *j* tiene como vecino de la palabra *i*. Si *i* y *j* no son vecinos, entonces la matriz contiene un valor de 0 en la posición correspondiente. Así, el modelo supone que dos palabras similares compartirán gran parte del vecindario. Cuantos más vecinos en común tengan dos palabras más similitud semántica tendrán. La similitud se calcula, por lo tanto, utilizando los pesos de los vecinos en común.

Los experimentos realizados comparan el modelo CDSA con el modelo LSA y con sujetos humanos en la tarea de cálculo de similitud entre frases de distinta longitud. Cuando se trata de frases de longitud igual a uno o dos el modelo LSA se asemeja un poco más a los humanos que el modelo CDSA. Sin embargo, a medida que crece la longitud de las frases comparadas también crece la semejanza del modelo CDSA con las respuestas de los sujetos, superando ampliamente al modelo LSA, incluso cuando las frases comparadas son de distinta longitud. Así, CDSA es capaz de especializar el significado de las palabras según el contexto en el que se encuentren, precisando así la semántica y, por tanto, aliviando la ambigüedad.

2.2.2.4. Modelos probabilistas

Aunque los modelos de espacio de memoria de la sección anterior están basados en la concurrencia léxica y por tanto en la estadística, su concepto de base no es el de modelos de probabilidad, aunque la empleen. Por el contrario, los modelos probabilistas parten de distribuciones de probabilidad para representar el conocimiento encerrado en el lenguaje. En general, estos modelos están diseñados para aplicaciones específicas, típicamente para clasificación de textos y recuperación de información, puesto que el conocimiento que son capaces de recoger es limitado.

El modelo de Análisis de Semántica Latente Probabilista PLSA (*Probabilistic Latent Semantic Analysis*) [Hofmann, 1999; 2001] propone un modelo de probabilidad, dados un documento y una palabra, de dos maneras distintas, como muestra la Figura 2.13. En el caso a) el modelo es asimétrico y en el caso b) simétrico, siendo *d* el documento, *w* la palabra, *z* la categoría o tópico y *P* la función de probabilidad.

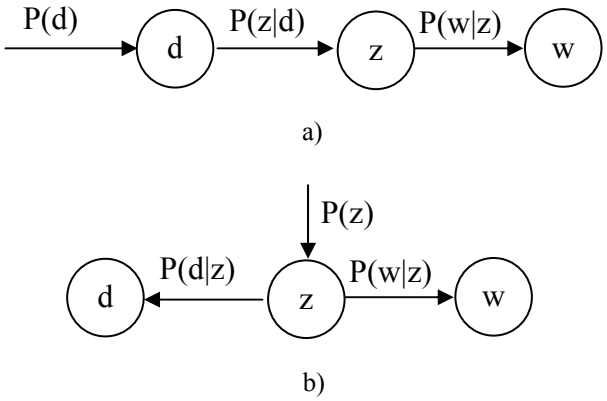


Figura 2.13. Modelos probabilistas NLSA a) asimétrico y b) simétrico.

Como en LSA, el experto ha de definir clases, tópicos o pasajes en los que incluir los textos y palabras. Así pues, los modelos de la figura responden a las expresiones siguientes:

a) $P(d,w) = \sum_z P(w|z) \cdot P(z|d)$

b) $P(d,w) = \sum_z P(z) \cdot P(d|z) \cdot P(w|z)$

Estos modelos son posteriormente ajustados mediante un algoritmo EM (*Expectation Maximization*) [McLachlan y Krishnan, 1997], obteniendo así tres matrices: *U*, que contiene las probabilidades de los documentos dadas las categorías, *V*, que contiene las probabilidades de las palabras dadas las categorías y la matriz diagonal $\sum$, que contiene las probabilidades de las categorías. El modelo queda definido finalmente como $P = U \cdot V \cdot \sum$.

Aunque se pueda establecer cierta analogía con la descomposición en valores singulares de LSA, la matriz *P* presenta ciertas diferencias ventajosas:

- *P* es una distribución de probabilidad bien definida y los factores tienen un claro significado probabilista, contrariamente a LSA.
- Las direcciones de los vectores en el espacio LSA no tienen interpretación. En PLSA son distribuciones multinomiales de palabras.
- La elección del número de dimensiones del espacio tiene un trasfondo teórico en PLSA. En LSA se hace de manera experimental.

Experimentos realizados en recuperación de información demuestran que PLSA obtiene una precisión entorno al 10% mejor que LSA en varios corpus de textos.

El modelo de Localización de Dirichlet Latente LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003] proporciona, según sus autores, una semántica completa probabilística generativa para documentos. Los documentos se modelan mediante una variable aleatoria oculta de Dirichlet. Al igual que en PLSA, asume un conjunto de categorías predefinidas. Cada categoría se representa como una distribución multinomial sobre el conjunto de palabras del vocabulario. El modelo queda descrito por la siguiente expresión:

$$P(d) = \int_\theta [\prod_n \sum_z P(w_n|z_n;\beta) \cdot P(z_n|\theta)] \cdot P(\theta;\alpha) \ \delta\theta$$

siendo $P(\theta;\alpha)$ una distribución de Dirichlet, $P(z_n|\theta)$ una multinomial que indica el grado en que el tema z_n es tratado en un documento y β una matriz de clases por palabras del vocabulario. De esta manera, la probabilidad de un documento depende de las probabilidades de que sus palabras denoten ciertas categorías dentro de una distribución de Dirichlet. Para aprender e inferir en el modelo usan un algoritmo EM análogo al modelo PLSA descrito anteriormente.

El modelo de Mezcla de Unigramas [Nigam et al., 2000] es un modelo sencillo y muy similar a los dos sistemas anteriores. Está descrito por la siguiente expresión:

$$P(d) = \sum_z \left(\prod_n P(w_n|z) \right) \cdot P(z)$$

Como se aprecia, la probabilidad de un documento depende de las probabilidades de que sus palabras pertenezcan a las categorías.

Experimentos en clasificación de textos y en recuperación de información muestran la superioridad del modelo LDA con respecto a PLSA y al modelo de Mezcla de Unigramas. Además, LDA recoge la posibilidad de que un documento contenga más de una categoría temática, al contrario que la Mezcla de Unigramas, y no está condicionado por los ejemplos de entrenamiento, como es el caso de PLSA.

Como los sistemas anteriores, el conocimiento almacenado por el modelo PMI-IR (*Pointwise Mutual Information Information Retrieval*) [Turney, 2001] también está condicionado por la tarea de aplicación. En este caso, PMI-IR se creó con el propósito de reconocer sinónimos. Así, dada una palabra y una lista de sinónimos, el modelo recoge el grado de dependencia estadística entre la palabra y cada una de las opciones de sinonimia mediante la siguiente expresión:

$$PMI\ (opción_i) = P(palabra \wedge opción) / P(opción_i)$$

siendo P la función de probabilidad y '$\wedge$' el operador de concurrencia. Los modelos probabilistas suelen fallar ante palabras inusuales puesto que no tienen información suficiente para que las probabilidades calculadas sean fiables. Para solventar este problema, el sistema PMI-IR calcula las probabilidades a partir de documentos recolectados utilizando un buscador de Internet (de ahí la parte IR de su denominación),

realizando consultas que contienen pares de palabra-opción de sinonimia. Emplea cuatro tipos de contexto de concurrencia distintos:

1. El contexto es un documento completo. La probabilidad de una opción es la proporción de documentos recolectados en los que aparece la opción pero no la palabra objeto.
2. El contexto es una ventana de 10 palabras.
3. El contexto es una ventana de 10 palabras, pero no se considera concurrencia si en el contexto aparece la palabra "no". Así se evita relacionar la palabra con antónimos.
4. Igual que el contexto 3 pero considerando además el resto de palabras del contexto (relaciones indirectas).

El modelo fue probado en experimentos con el objetivo de encontrar la palabra más parecida a otra dada de entre una lista de opciones. El contexto que mejores resultados obtuvo fue el de tipo 3, superando al modelo LSA en un 10% de acierto. A pesar de que realizar búsquedas en Internet para obtener probabilidades resulta muy costoso en tiempo, lo es significativamente más el proceso de reducción de dimensiones empleado por LSA.

El modelo Sintagmático-Paradigmático SP (*Syntagmatic Paradigmatic*) [Dennis y Harrington, 2001] está basado en la idea de que el procesamiento de oraciones comprende la recuperación de fragmentos de oraciones de la memoria y la correspondencia de estos fragmentos con la oración a interpretar. Así pues, representa las probabilidades de sustituir una palabra por otra en un contexto dado. Dichas probabilidades son inducidas mediante un algoritmo EM, que implica comparar cada fragmento de frase con el resto. Para reducir tiempo de procesamiento, el modelo crea clases de equivalencia de los fragmentos. Un fragmento es considerado como una secuencia de palabras delimitada por palabras altamente frecuentes o por final o principio de frase. Por ejemplo, en la oración "*La anchoa comía un bocadillo de jamón de un primo sardina de su pueblo*" se distinguen los siguientes fragmentos:

1. LA anchoa comía UN
2. UN bocadillo DE
3. DE jamón DE
4. UN primo sardina DE
5. DE su pueblo

Dado que los elementos de una misma clase han de seguir el mismo patrón, el modelo crearía las clases de equivalencia {2,4}, {1}, {3} y {5}. La probabilidad de sustitución de un palabra por otra se calcula, en otras dos clases de ejemplo a continuación, de la siguiente manera:

	Fragmento	**Palabras en común en la misma posición**	**Probabilidad de recuperación**
Clase 1 Representante: UN bocadillo DE LA	UN arrecife DE LA	3	0,33
	UN chopito DE LA	3	0,33
	UN jamón DE LA	3	0,33
Clase 2 Representante: EN EL bocadillo	EN EL arrecife	2	0,5
	EN EL mar	2	0,5

La probabilidad de reemplazar “bocadillo” y “arrecife” es pues:

$$P(<\text{bocadillo,arrecife}>) = (0,5+0,33)/2 = 0,415$$

Los experimentos realizados demuestran, como es de suponer, que el modelo encierra conocimiento sintáctico, ya que un alto porcentaje de las palabras más similares a otra dada, o más reemplazables en este caso, tienen la misma categoría sintáctica, superando al modelo LSA. Si se observa la categoría semántica de dichas palabras el porcentaje de palabras más reemplazables con la misma categoría es comparable al obtenido mediante LSA. El modelo es capaz de adquirir relaciones tanto sintagmáticas o semánticas (“correr-rápido”) como paradigmáticas o de asociación (“correr-nadar”), de ahí su nombre.

2.2.2.5. Modelos conexionistas

A pesar de sus diferencias en cuanto a formalismo de representación y naturaleza, la mayoría de los sistemas anteriores comparten una característica común: representan la semántica mediante vectores. Los vectores presentan la ventaja de facilitar una

implementación directa y simple de la "composicionalidad", además de proporcionar una forma sencilla de medir la similitud semántica. Sin embargo, dicha medida de similitud es simétrica, lo que no es coherente con los hallazgos psicolingüísticos de Tversky [Tversky, 1977]. Además, para inferir relaciones y sus grados es necesario comparar todas las palabras entre sí. Los modelos conexionistas no introducen estas desventajas y además conservan los beneficios de los vectores.

En los modelos conexionistas el formalismo de representación consiste en una red asociativa de nodos interconectados, donde el conocimiento se expresa mediante patrones de conectividad y activación. Este formalismo permite la inferencia y la representación explícita de relaciones de cualquier orden sin la realización de ningún cálculo. Además, las redes permiten la actualización del conocimiento en tiempo real, ya sea por eliminación, inserción o modificación, sin necesidad de rehacer las estructuras, lo que resulta eficiente a la vez que psicológicamente plausible. Las redes hacen posible también el almacenamiento de conocimiento heterogéneo, pudiendo contener, por ejemplo, información semántica y sintáctica al mismo tiempo de manera explícita.

Si se trata de modelos conexionistas de conocimiento linguístico es necesario comenzar nombrando a WordNet [Miller et al., 1990]. WordNet fue ideado en la universidad de Princeton a cominezos de los años 90, inspirándose en las teorías psicolinguísticas de la memoria léxica del ser humano, y aún se continúa manteniendo y ampliando, contando con más de 150,000 términos en la actualidad. Contiene en cada nodo de la red lo que sus autores denominan "synsets" o conjuntos de palabras sinónimas etiquetadas con su categoría léxica correspondiente (nombre, verbo, adjetivo o adverbio). Las aristas que unen los nodos denotan relaciones de diversos tipos, como por ejemplo, hiperonimia e hiponimia (más general que o más específico que), meronimia y holonimia (parte de o contiene a) y antonimia, entre otras. La principal diferencia con el resto de sistemas de representación es que está íntegramente construida "a mano" por expertos lingüistas, que se encargan de situar las nuevas palabras en los "synsets" correspondientes y de establecer las relaciones entre los mismos. WordNet ha sido ampliamente empleada en diversas tareas de procesamiento de lenguaje natural debido, principalmente, a la variedad de medidas de similitud semántica entre palabras que permite su estructura [Budanitsky y Hirst, 2001].

Aunque algunos de los modelos de representación masiva recogen información sintáctica ninguno de ellos trata de adquirirla expresamente. No es el caso de la red del estudio de Ferrer [Ferrer et al, 2004]. Ferrer emplea el formalismo de gramática de dependencias. Se basa en el hecho, comprobado experimentalmente, de que el 87% de las relaciones sintácticas se dan a distancia menor o igual que dos, aunque en esta distancia el 56% de las relaciones establecidas son erróneas o vacías. El conocimiento se almacena en un grafo donde los nodos son palabras y las aristas son relaciones dirigidas desde el modificador a su correspondiente núcleo. Estas relaciones se extraen a partir de colecciones de texto anotadas léxica y sintácticamente.

Una vez construido el grafo de conocimiento sintáctico, los autores estudian ciertas propiedades del mismo citadas a continuación:

- Estructura de Mundo Pequeño [Watts y Strogatz, 1998]: Sea *d* la longitud de camino mínimo media entre cada par de nodos de un grafo y sea *C* el coeficiente de agrupamiento medio, que indica la probabilidad media de que dos nodos vecinos de otro nodo sean vecinos entre sí, se cumple la propiedad de mundo pequeño si *d* es muy similar a la *d* de un grafo aleatorio y *C* es mucho mayor que el *C* de un grafo aleatorio. La propiedad de mundo pequeño implica que la probabilidad de tener un nodo con grado *k*, es decir, conectado con *k* nodos, es una potencia de *k*. Además, en una red con la propiedad de Mundo Pequeño la distancia mínima media entre cualquier par de palabras está por debajo de tres aristas.
- Heterogeneidad: es la probabilidad, según una distribución de Poisson, de que la probabilidad de que un nodo tenga cierto grado sea potencia de ese grado.
- Organización jerárquica: Una red será jerárquica si *C*, en función de un grado *k*, es una potencia de *k*.
- Centralización de caminos: El número medio de caminos mínimos que pasan por cada nodo.
- Asertividad: Una red es asertiva si nodos muy conectados están conectados a su vez a otros nodos muy conectados. Es disertiva en caso contrario.

Ferrer realiza el estudio para tres colecciones de textos en distintos idiomas: Alemán, Rumano y Checo. Los resultados concluyen que todos los grafos cumplen la propiedad de Mundo Pequeño, son heterogéneos, demuestran una organización jerárquica

razonable y los tres son disertivos, presentando los grafos Alemán y Rumano una centralización de caminos similar.

Un modelo computacional del conocimiento semántico del lenguaje debe imitar las representaciones semánticas humanas, es decir, qué se representa, pero también la manera de construir esas representaciones, es decir, cómo se representa. Los sistemas de memoria de gran dimensión y los sistemas probabilistas están más centrados en el contenido de la representación, utilizando técnicas matemáticas costosas alejadas del proceder humano intuitivo. Así pues, Lemaire y Denhière proponen el modelo ICAN (*Incremental Construction of an Associative Network*) [Lemaire y Denhière, 2004], en el que el conocimiento está representado en una red o grafo cuyos nodos representan palabras y cuyas aristas representan relaciones de concurrencia y orden entre las palabras que unen, como muestra la Figura 2.14.

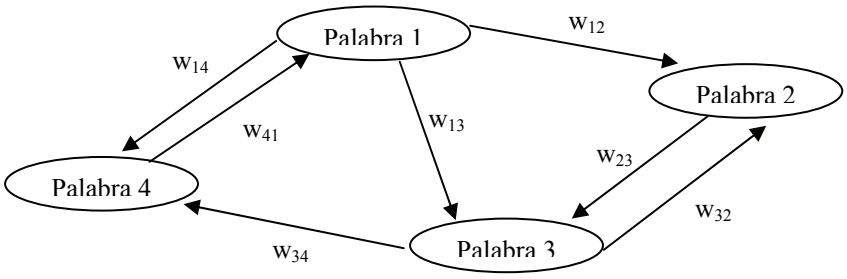


Figura 2.14. Modelo ICAN de representación del conocimiento semántico.

Un estudio realizado por los autores muestra las siguientes conclusiones sobre las relaciones contextuales:

- la concurrencia de dos palabras tiende a aumentar su similitud,
- la ocurrencia de una palabra y la ausencia de otra tiende a disminuir la similitud entre ambas,
- las concurrencias de segundo y tercer orden tienden a incrementar levemente la similitud.

Siguiendo estas bases se construye la red a partir de un conjunto de textos de entrada. Se define el contexto como una ventana deslizante de tamaño fijo. Las palabras en el mismo contexto se asocian, actualizando los pesos de manera distinta dependiendo de qué caso de entre los tres anteriores se dé: si se dan el primero y el tercero, el peso de la

asociación aumenta en función del peso ya existente, siendo el aumento mayor en caso de concurrencia directa que en el caso de concurrencia de orden superior. Para la no concurrencia los pesos de las relaciones entre las palabras del contexto y las que no aparecen en el mismo disminuye en un porcentaje fijo. La similitud entre palabras queda determinada por el producto de los pesos de los enlaces del camino más corto entre las mismas. Esta similitud no es simétrica, así como no lo es en el razonamiento humano.

Un fragmento de texto está representado por la subred formada por los nodos correspondientes a las palabras del fragmento y las asociaciones entre las mismas. Para evitar considerar todos los significados de las palabras en la semántica de un fragmento de texto proponen una poda de la subred resultante. En este caso, la representación de las palabras y de los textos no es la misma, al contrario que en los sistemas anteriores. Además, los autores postulan que no es cognitivamente razonable pasar de manera directa y sin esfuerzo de la semántica de las palabras a la semántica de los textos. A pesar de esta afirmación, la representación como subred que proponen es una manera directa y poco costosa de llevar a cabo la "composicionalidad", salvo por el proceso de poda.

En los experimentos realizados se compara el modelo con LSA y con sujetos humanos. Dada una palabra, se presentan otras seis palabras relacionadas con ella y los modelos deben seleccionar la más similar de las seis. Esta prueba se realiza con una lista de 200 palabras. Los resultados muestran un 0,5 de coeficiente de correlación con los resultados humanos, frente a un 0,39 de LSA. Si el modelo ICAN se construye teniendo en cuenta sólo concurrencias directas, la correlación cae a 0,39. Si además de las concurrencias directas se tienen en cuenta las no concurrencias, la correlación que se obtiene es de 0,48. Los autores concluyen de estos resultados que la frecuencia de concurrencia tiende a sobrestimar la similitud semántica.

Ferrer, además de las redes de dependencia sintáctica, estudió también las propiedades de las redes de dependencia contextual análogas a ICAN. En este caso, tomó un contexto de ventana con un tamaño igual a tres, descartando las relaciones entre palabras más distantes en los textos al considerar que casi todas las relaciones se dan a distancia menor o igual que dos. Rememorando los conceptos de grado de un nodo como número de conexiones que salen de él, y de longitud media de camino mínimo, Ferrer concluyó que las redes construidas con relaciones de concurrencia

cumplen la propiedad de mundo pequeño. Concluye también que para alcanzar cualquier palabra desde otra dada se requieren caminos de menos de tres aristas en media. Esto explica en cierta forma la velocidad de producción hablada del lenguaje.

Ferrer también introduce el concepto de palabras "núcleo". El cerebro humano es capaz de almacenar entre 10,000 y 100,000 palabras. Sin embargo, sólo una pequeña parte de ellas es usada con frecuencia. Estas palabras son las denominadas "núcleo" y con ellas se puede expresar casi todo lo que se quiera decir. Por ejemplo, en la lengua inglesa se puede expresar con menos de 1,000 palabras todo lo que se podría expresar con un léxico mucho más complejo a costa de una alta redundancia (Romaine, 1992). Así pues, cuanto más frecuente es una palabra más disponible está para ser usada y es más fácil de comprender [Aaronson y Scarborough, 1977] o, dicho en términos más analíticos, cuanto mayor es el grado de una palabra en la red mayor es su disponibilidad. A este hecho se le denomina "efecto frecuencia" [Akmajian et al., 1995]. La constatación de este efecto corrobora la hipótesis de la existencia de un léxico común básico para la comunicación. El resto del léxico es dependiente de la edad, educación, localización, contexto social, etc. Según lo expuesto anteriormente, las palabras "núcleo" están a distancia menor que tres del resto de palabras por lo que se consigue enriquecer el núcleo con poco esfuerzo durante la producción del lenguaje.

Artículos, preposiciones, conjunciones y el resto de partículas lingüísticas son palabras pertenecientes al núcleo y poseen grados muy elevados en la red. Sin embargo, están vacías de semántica, por lo que se pueden suprimir sin afectar a la comprensión, como se demuestra en los telegramas [Akmajian et al., 1995]. Sin embargo, muchos trastornos del lenguaje están relacionados con la ausencia de palabras con alto grado pero, en este caso, con contenido semántico.

2.2.2.6. Consideraciones sobre los sistemas basados en "coocurrencia" léxica

Algunos de los sistemas de representación masiva del conocimiento lingüístico descritos se presentan como nuevas teorías de realidad lingüística. Estas afirmaciones son las que motivaron la crítica que Charles Perfetti realiza a los sistemas basados en concurrencia léxica [Perfetti, 1998]. Perfetti distingue dos procesos en los sistemas de representación lingüística:

- El análisis del texto mismo: Una teoría psicolingüística propondría cuáles son las unidades objetivo del análisis en base a hipótesis o presunciones cognitivas y posteriormente realizaría la validación experimental. Sin embargo, sistemas como LSA o HAL primero realizan cálculos y, basándose en los resultados, tratan de relacionar esos cálculos con alguna hipótesis de realidad psicológica en un intento por revelar estructuras internas del lenguaje que pudieran tener una realidad lingüística.
- La aplicación de métodos cuantitativos al tipo de análisis anterior.

Así pues, Perfetti cuestiona la validez, como teoría de la semántica, de sistemas como LSA o HAL argumentando que sus conclusiones son accidentales y no sistemáticas.

Todos los sistemas de representación masiva descritos tienen en común la explicación de los procesos del discurso del lenguaje mediante datos de concurrencia léxica. En lo que a realidad psicológica se refiere, Perfetti critica ciertos aspectos de la concurrencia léxica:

- El tamaño y contenido del corpus de textos de entrada. Es un problema práctico que limita el éxito de los sistemas. Algunos de los fallos para simular el comportamiento humano están causados por limitaciones prácticas. Si no se pueden superar, entonces el modelo las hereda y por tanto pierde realidad psicológica.
- Cuantificación de la concurrencia. La manera en que se cuenta el número de veces que las palabras concurren. En principio, un sistema basado solamente en concurrencia presenta limitaciones para capturar el conocimiento lingüístico. La concurrencia es necesaria para el aprendizaje del lenguaje pero otras habilidades no se adquieren de la concurrencia, como por ejemplo la sintaxis. LSA, HAL y otros sistemas capturan algunos efectos contextuales de la sintaxis pero no la sintaxis en sí misma. El contexto en el que se mide la concurrencia es también un factor que aleja a los sistemas del rol teórico. El contexto como ventana, sin limitación de frases, es absolutamente inconsistente con los resultados de investigación sobre la memoria y su dependencia con los límites de las frases en la lectura [Goldman et al., 1980].

Perfetti también habla de dos tipos de fallos que cometen los sistemas de representación masiva:

- Fallos de base. No se comprueba si el modelo es coherente con lo que mejor conocemos sobre la cognición humana. Tampoco se comprueba si los datos modelados difieren con los que puede manejar un ser humano. Un modelo puede escapar a ciertos fallos de base y funcionar, como por ejemplo LSA con la sintaxis. LSA no captura la sintaxis y, sin embargo, se comporta bien en las aplicaciones. Esto puede ser debido a que las creencias sobre el papel de la sintaxis no sean ciertas, a que la sintaxis sí sea relevante pero esté implementada *ad hoc*, o bien a que los datos no reflejen el papel real de la sintaxis, como es el caso de LSA que asume frases sintácticamente bien formadas.
- Fallos teóricos. Los que se refieren a fundamentos del sistema que son intrínsecamente diferentes a la cognición humana. Por ejemplo, los experimentos que comparan LSA con sujetos humanos presentan resultados que muestran una fuerte similitud en los aciertos. Sin embargo, LSA no yerra de manera similar a los humanos, por lo que no se puede considerar como una teoría de realidad psicológica.

Los sistemas basados en concurrencia léxica son buenas herramientas que han posibilitado la aplicación real de sistemas de procesamiento de lenguaje natural que emplean información semántica. Además, son capaces de capturar la coherencia del discurso, considerando la dependencia existente entre el estilo de escritura y el aprendizaje y comprensión del lenguaje. Sin embargo, no se pueden posicionar como teorías de la cognición semántica humana ya que tampoco necesitan hacerlo para ser útiles.

CAPÍTULO 3

El Proceso Mental de la Lectura.

Modelos Computacionales

La naturaleza de la cualidad innata del lenguaje en el ser humano sigue siendo hoy una cuestión de debate abierto. Sin embargo, la comunidad científica parece estar de acuerdo en que el lenguaje escrito, junto con la habilidad para generarlo y comprenderlo, son cualidades que deben ser adquiridas y desarrolladas durante el proceso de maduración intelectual del ser humano. De hecho, toda persona sin ningún tipo de problema vocal, auditivo o relativo al cerebro desarrolla un lenguaje de manera involuntaria e inconsciente, mientras que la lectura y escritura del mismo no se adquieren a no ser que se realice un esfuerzo de aprendizaje de las mismas. La comunicación mediante el lenguaje escrito carece de ciertos tipos de información importantes con respecto al lenguaje hablado, como son la entonación y los gestos. Dicha carencia se suple con otros aspectos del lenguaje escrito, como el discurso y la sintaxis, entre otros. Sin embargo, dichos aspectos junto con el proceso necesario de transformación de información visual en información lingüística demandan un uso más intenso que en el lenguaje hablado de estructuras y procesos mentales como la memoria o las inferencias. Por esta razón, la comprensión del lenguaje escrito depende en gran medida de la capacidad y el manejo de las estructuras y procesos mencionados. Así pues, es importante que un modelo computacional de lectura contemple al menos dichos aspectos, tanto su diseño como la interacción entre los mismos. Además, existen otros factores no lingüísticos que condicionan en gran medida la comprensión del lenguaje escrito mediante la lectura. Dichos factores son inherentes al ser humano y en muchas ocasiones incontrolables por el mismo. Las intenciones, las emociones y la imaginación, entre otros, también forman parte del proceso de lectura y pueden ser pues objeto del modelado.

3.1. El Proceso de Lectura

La lectura es un proceso mental mediante el cual se percibe, comprende e interpreta una secuencia de símbolos dispuestos de acuerdo a una gramática que describe un determinado lenguaje.

A diferencia del lenguaje hablado, la comprensión del lenguaje escrito es un proceso que comienza con un proceso de percepción visual. A grandes rasgos, el proceso de lectura se puede dividir en tres fases [Perfetti, 1999]:

1. Percepción visual.

2. Conversión de la representación visual en representación lingüística.

3. Procesamiento de la representación lingüística.

Por supuesto, para cada una de las fases anteriores existe al menos un proceso que la lleva a cabo. Cada una de las fases está relacionada con diversos aspectos y niveles del lenguaje, como son la fonología (autoinducida), grafología, ortografía, sintaxis, semántica y discurso. En términos de unidades básicas correspondientes a cada uno de los aspectos anteriores, se puede decir que los procesos que llevan a cabo las fases mencionadas perciben y procesan sonidos o fonemas, letras, palabras, oraciones, conceptos y modelos de situación, respectivamente. La comprensión mediante la lectura depende pues de la identificación de las palabras y de los procesos que transforman esas palabras en mensajes con sentido utilizando la sintaxis y la semántica.

3.1.1. Tareas del proceso de lectura

Las tres fases enunciadas anteriormente se dividen en tareas, que involucran a los diferentes niveles lingüísticos y procesos que operan sobre ellos. Así pues, las principales tareas de la lectura son [Perfetti, 1999]:

1. Identificación de palabras.
2. Asignación del significado.
3. Comprensión de oraciones a partir del significado de las palabras.
4. Comprensión de textos a partir del significado de sus oraciones.

Como se ha mencionado, cada tarea actúa sobre determinados niveles del lenguaje. Existen diversas teorías sobre la manera y el orden al que obedecen los procesos que realizan las tareas y sobre el modo de interacción entre los mismos.

Para el proceso de identificación de las palabras, por ejemplo, los modelos cognitivos tradicionales abogan por la representación de un lexicón en la memoria, es decir, una especie de representación mental de la correspondencia entre la forma y significado de las palabras, de tal manera que una palabra se lee con éxito si se activan una o más formas del lexicón. Sin embargo, otros modelos teóricos conexionistas [Plaut et al., 1996] definen el éxito de la lectura de una palabra como la consecución de un patrón de activación estable en una red de palabras que representan el vocabulario disponible. Éstos últimos defienden la identificación de palabras como un proceso emergente y son más compatibles con modelos neurocognitivos de lectura.

Existe también otro aspecto controvertido en el proceso de identificación de las palabras relacionado con la intervención o no de la fonética en dicho proceso y el lugar que ocupa en la secuencia definida por el mismo. A pesar de que existen estudios a favor de la intervención fonética que indican que las áreas cerebrales correspondientes a la percepción fonética se activan prácticamente al mismo tiempo que las correspondientes a la percepción visual de las palabras [Stone et al., 1997], [Lukatela y Turvey, 1998], no se ha demostrado de manera concluyente dicha influencia de los fonemas en el proceso de identificación de las palabras, puesto que es difícil aislar de manera experimental los aspectos fonético, ortográfico y semántico de las mismas. Por ello, existe una corriente que niega dicha intervención, postulando la existencia de un camino directo entre los grafemas de las palabras y su significado en el vocabulario. Dentro de la corriente contraria, es decir, la que defiende que la fonética participa activamente en el proceso de identificación de las palabras, existen diversas teorías. Una corriente de dichas teorías defiende el proceso de identificación como una secuencia de subprocesos en el tiempo pudiendo situar, por tanto, la percepción fonética como un evento previo a la identificación, o bien entre la percepción visual y la búsqueda de correspondencia en el lexicón [Perfetti y Tan, 1998], o bien posterior a la identificación. Otra de las corrientes pertenecientes a las teorías que apoyan la participación de la fonética afirma que la identificación de las palabras es un proceso de negociado distribuido entre sus constituyentes [Van Orden y Goldinger, 1994], [Plaut et al., 1996] y, por tanto, no sitúan temporalmente a la fonología en el proceso de identificación. Esta

distinción en el modo de interacción se hace también extensible a otros procesos de la lectura, como se verá más adelante. Como contrapunto a estas dos corrientes se encuentra una tercera, el modelo de Ruta Dual [Colheart et al., 1993]. En dicho modelo, la identificación de las palabras depende de ellas mismas. Existen palabras que se identifican de manera directa desde sus grafemas, existen palabras en las que es necesario recuperar su fonética a partir de su ortografía para acceder al lexicón y existen palabras que siguen los dos caminos en paralelo, teniendo éxito el más rápido de los mismos. Estudios realizados por Colheart indican que las palabras muy frecuentes se identifican por el camino directo, mientras que para palabras poco comunes tiene éxito en ciertas ocasiones el camino que implica a la fonética.

Una vez identificada una palabra percibida, la siguiente tarea llevada a cabo en la lectura es la de la activación y selección del significado apropiado de la misma en el contexto en el que aparece. Nuevamente, existen varias teorías. En el modelo de Acceso Selectivo [Glucksberg et al., 1986] la activación y selección se realiza siempre teniendo en cuenta la consistencia con el contexto. Sin embargo, en el modelo de Acceso Múltiple [Kintsch y Mross, 1985] se activan todos los significados de la palabra al mismo tiempo. A continuación, un proceso de selección restringido por el contexto elige el más consistente con el mismo. Finalmente, el modelo de Acceso Múltiple Ordenado [Neil et al., 1988] se presenta como un híbrido de los dos anteriores, en el que sólo los significados más frecuentes se activan seleccionándose el primero, en orden de frecuencia, que no quebrante la coherencia contextual.

La siguiente tarea, consistente en comprender oraciones a partir del significado de las palabras que las componen, involucra a un aspecto lingüístico no considerado hasta el momento: la sintaxis. Como el caso de la fonética en la identificación de palabras, el tipo de influencia de la sintaxis en la tarea de la comprensión de oraciones también ha creado una polémica que ha derivado en diversas teorías. De nuevo, la principal cuestión es la naturaleza secuencial o concurrente de los procesos. Así, se plantean teorías que siguen la tesis modular de Fodor [Fodor, 1983], que presenta la comprensión del lenguaje como la ejecución de procesos aislados en cadena. Entre dichas teorías se encuentra la del Principio de Agregación Mínima [Frazier y Rayner, 1982], donde una palabra leída se añade a la estructura sintáctica de la oración construida hasta el momento de la forma más sencilla posible, de tal forma que se comprueba a continuación si dicha agregación sintáctica es correcta mediante la semántica de la

misma. Otras teorías, sin embargo, defienden que la semántica de una palabra leída es la que influye en la asignación sintáctica de la palabra. Como contrapartida a la tesis modular de Fodor aparece la tesis interactiva [MacDonald et al., 1994], que plantea la tarea de comprensión de oraciones como un proceso distribuido con realimentación recíproca entre los distintos niveles de información lingüística. Esta teoría interactiva defiende, además, la existencia de interacciones de otras fuentes externas de carácter no lingüístico en el proceso de compresión, como puede ser el conocimiento general del individuo [Altmann y Steedman, 1988].

La última tarea que realiza el proceso de lectura es la comprensión de textos, más allá de frases aisladas. Por supuesto, esta tarea requiere el entendimiento de las oraciones individuales pero también la integración del significado de todas ellas. En este caso, se necesitan representaciones mentales de alto nivel que capturen el significado global de un texto. Durante esta tarea se manejan diferentes niveles de información mutuamente excluyentes, que juntos conforman el resultado mental del proceso de lectura en cada momento. Así pues, la lectura de un texto genera una representación con los siguientes componentes o niveles [Ericsson y Kintsch, 1995], [Perfetti, 1999]:

- *Estructura lingüística superficial*. Está compuesto por la traza de las palabras del texto que han sido leídas en cada momento. Dicha traza se compone de las palabras mismas y su información sintáctica y semántica interpretada en el contexto de las oraciones en las que aparecen.
- *Texto base*. Es una representación conceptual coherente de la estructura de un texto leído. Se compone de microproposiciones, ya sean derivadas directamente del texto o inferidas, y macroproposiciones, que son el resultado de procesos de selección y generalización aplicados sobre las microproposiciones.
- *Modelo de situación*. Integra información textual y conocimiento general del mundo. La información que presenta no es exclusivamente proposicional (o lingüística) sino que también puede almacenar información espacio-temporal, proporcionando al individuo una base para la inferencia, la elaboración y una posible actuación.

Estos tres componentes son los que permitirán llevar a cabo los diversos objetivos de la lectura de un texto, como son la respuesta a preguntas, su resumen, su verificación, etc. [Kintsch, 1988]. El nivel de detalle de cada uno de los componentes generados para

un texto depende de la forma y el contenido de dicho texto, siendo más completo el texto base en ciertas ocasiones y el modelo de situación en otras.

Durante esta tarea también participan procesos que realizan inferencias. Por una parte, se realizan inferencias en el texto base para mantener la coherencia referencial. Por otra, son las inferencias las que permiten crear el modelo de situación a partir del texto base.

Nótese que al igual que en la tarea de comprensión de oraciones también se tiene en cuenta la influencia del conocimiento general del mundo. Sin embargo, en el caso de la comprensión de textos la creencia sobre la participación del conocimiento general en los procesos que la llevan a cabo es unánime y no ha lugar a debate. Nótese también que son necesarios procesos que construyan y traten los tres componentes de representación mental citados, por lo que también son partícipes de la tarea de comprensión de textos la interpretación de las oraciones en el discurso y la generación de imágenes espacio-temporales.

3.1.2. Componentes y procesos generales en la lectura

Como resumen de todo lo expuesto en la sección anterior, Perfetti [Perfetti, 1999] propone un esquema genérico que refleja todas los posibles interacciones entre los procesos implicados en las tareas que se llevan a cabo durante la lectura y los componentes y niveles lingüísticos, independientemente de las diferentes teorías o modelos para cada caso. En la Figura 3.1 se recoge dicho esquema.

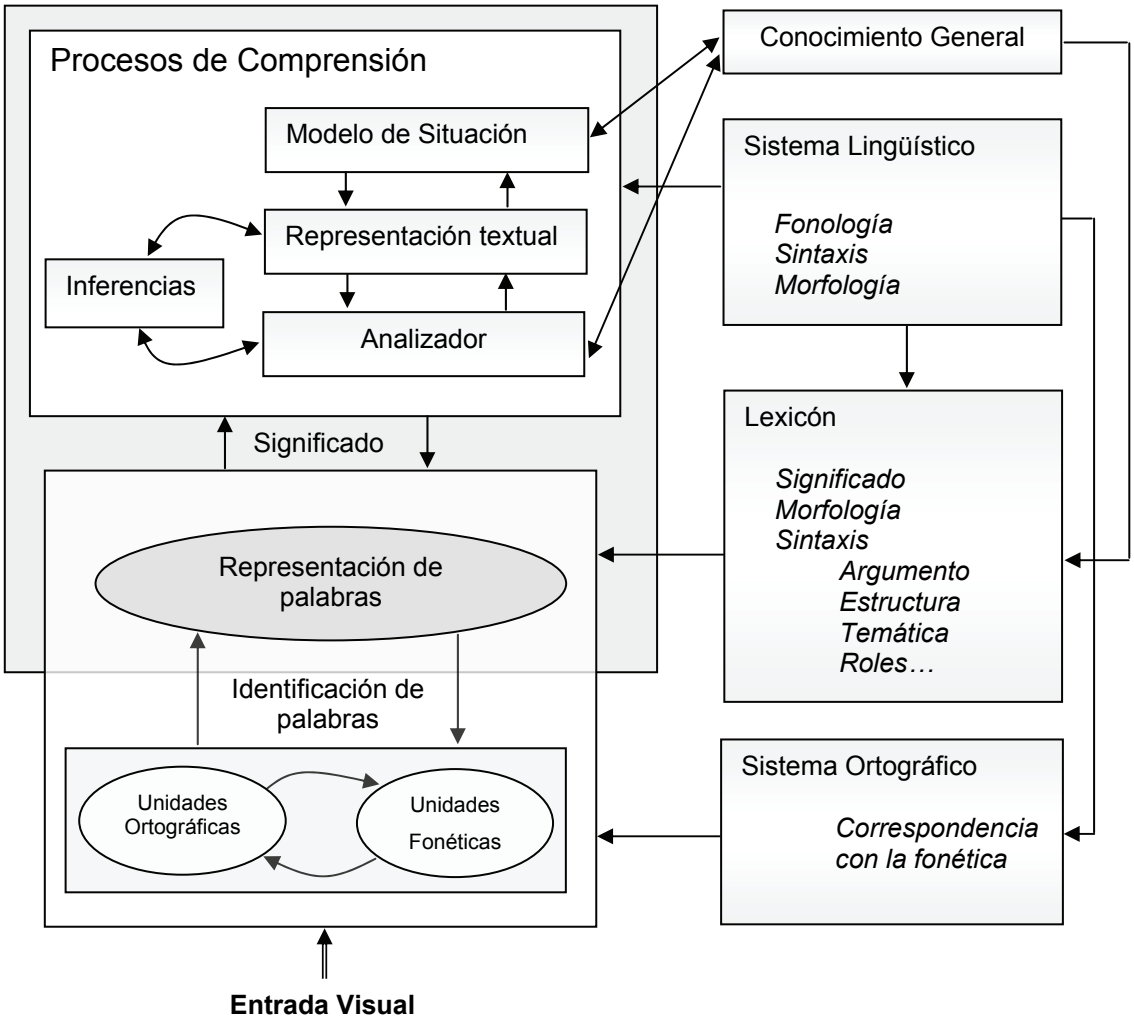


Figura 3.1. Posibles interacciones entre los procesos y componentes generales de la lectura.

3.1.3. Modelos de flujo de información en el proceso de lectura

El debate acerca de la naturaleza secuencial de la interacción entre las diversas fuentes de información en cada una de las tareas descritas, es también extensible a la naturaleza de la interacción global entre los niveles de información lingüísticos durante el proceso general de lectura. Así pues, existen tres modelos de flujo generales clásicos, los dos primeros abogando por la seriación de los componentes y el tercero por la interacción distribuida entre los mismos [Zakaluk, 1998].

El primero de ellos es el modelo conocido como "Top-Down" (Arriba-Abajo) [Goodman, 1970], cuyo esquema se presenta en la Figura 3.2. En dicho modelo, la semántica y la sintaxis dirigen el flujo de información en la manera que muestra la figura. Según este esquema, el lector procede de la siguiente manera:

1. Percibe las letras.
2. Realiza predicciones hipotéticas (inferencias) sobre la identificación de la palabra basada en el conocimiento a priori de la temática del texto y el sentido de la oración.
3. Lee la palabra para confirmar la hipótesis.
4. Construye el significado.
5. Asimila el nuevo conocimiento.

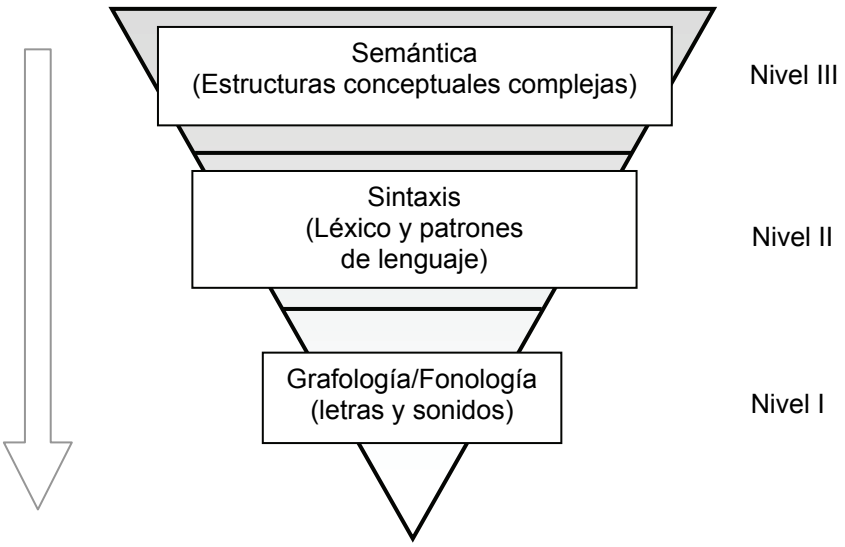


Figura 3.2. Esquema del modelo de lectura Arriba-Abajo ("Top-Down").

Las críticas de los detractores de este modelo "Top-Down" están apoyadas por distintas evidencias. La primera de ellas es resultado de estudios experimentales que demuestran que los lectores rápidos son tales cuando las oraciones son muy comunes y cortas, pero no lo son tanto en oraciones más complejas donde es necesario fijar la atención de la percepción visual para identificar las palabras. Además, el propio contexto semántico también determina la facilidad de predicción de una palabra. Si éste es limitado es fácil realizar una hipótesis sobre el significado que se va a leer a continuación. En el caso contrario, es necesario leer la palabra antes de poder inferir un significado. Otro argumento en contra es el hecho de que realizar hipótesis a todos los niveles y luego verificarlas, utilizando niveles de información inferiores, es un proceso costoso que contradice la velocidad de lectura y procesamiento de los seres humanos por la cual, paradójicamente, aboga el modelo. Para finalizar, la mayoría del material de lectura que emplea el ser humano no es suficientemente redundante ni predictivo para que el método hipótesis-comprobación opere exclusivamente de manera precisa.

Como contraposición se encuentra el modelo "Bottom-Up" (Abajo-Arriba) [Bobrow y Norman, 1975], que propone el flujo de información inverso al del modelo anterior y cuyo esquema se presenta en la Figura 3.3.

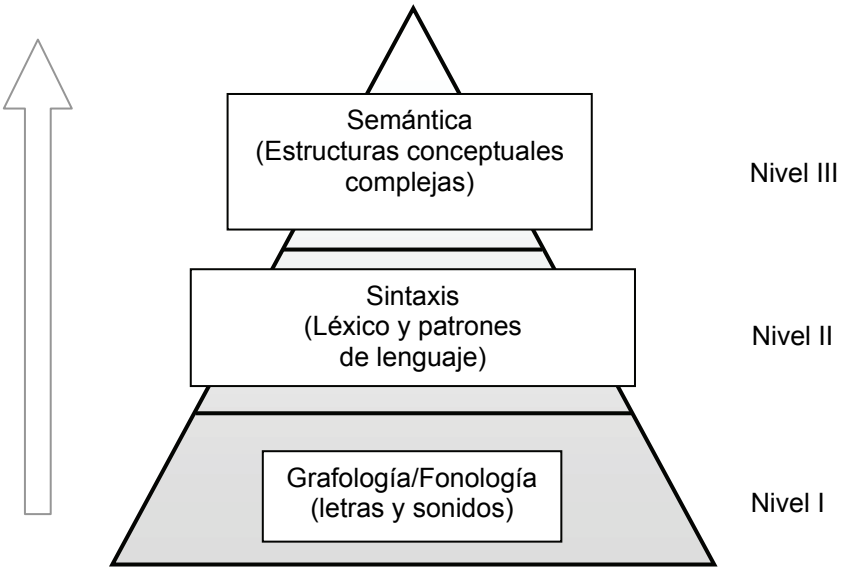


Figura 3.3. Esquema del modelo de lectura Abajo-Arriba ("Bottom-Up").

Según el modelo "Bottom-Up", el lector procede de la siguiente manera:

1. Percibe las letras y las transforma en representaciones fonéticas.

2. Transforma las letras y representaciones fonéticas en palabras.
3. Asigna a las palabras un significado.
4. Combina los significados de las palabras formando frases con sentido parcial.
5. Asocia las frases en oraciones con un sentido pleno.
6. Asimila la información leída.

Existen también evidencias que contradicen este modelo. Existen estudios psicológicos que demuestran que los seres humanos son capaces de agrupar más letras en una oración con sentido que en una cadena de palabras no relacionadas. Dichos estudios demuestran también que muchas cadenas de letras son identificadas con la palabra correspondiente incluso cuando alguna de las letras es errónea o está ausente. Además, los estudios también evidencian que las letras se perciben más acertadamente cuando se presentan formando parte de palabras que cuando se presentan por separado. Un último argumento en contra del modelo "Bottom-Up" es la prueba de que los procesos mentales procesan e identifican más rápido pares de palabras semánticamente relacionadas que pares de palabras que poseen significados dispares.

La alternativa a los modelos anteriores es el denominado modelo "Interactivo" [Rumelhart, 1977]. En dicho modelo, el flujo de información entre los distintos niveles lingüísticos no es secuencial. Por el contrario, es una interacción de todos ellos en cualquier instante del proceso de lectura, donde cada uno recoge la información necesaria de cualquier otro en los momentos que lo requiera.

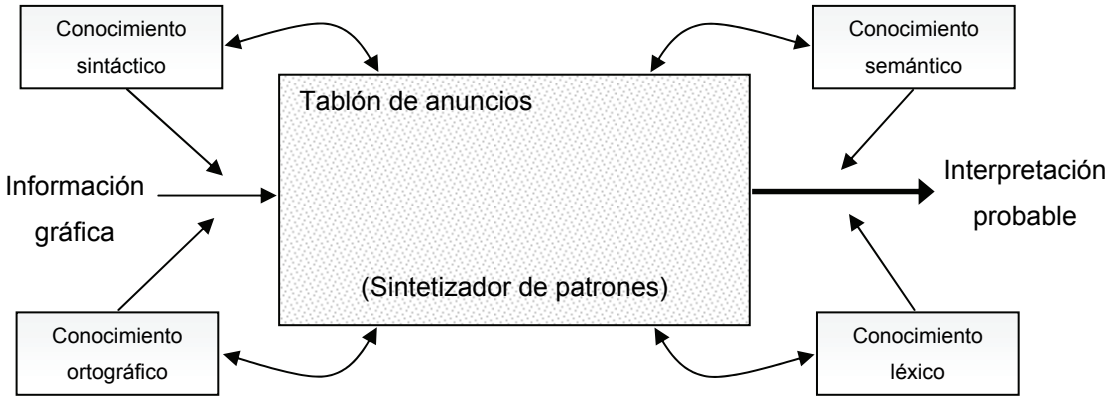


Figura 3.4. Esquema del modelo Interactivo de lectura.

En la Figura 3.4 se presenta una analogía computacional con un tablón de anuncios que Rumelhart propone para explicar el modelo, donde cada nivel deposita en el tablón su información quedando ésta disponible para cualquier componente que la requiera.

El modelo Interactivo es el más aceptado de los tres descritos, ya que es más general que los anteriores y da cabida a todas las evidencias experimentales halladas.

3.1.4. La memoria en el proceso de lectura

A pesar de las diferentes visiones sobre el proceso de lectura y de los debates sobre la interacción de los componentes que toman parte en la misma, es obvio que cualquier tipo de información que manejen los procesos mentales debe estar almacenado en algún lugar. Si se habla del cerebro humano, el almacén de información por antonomasia es la memoria. Así pues, todos los niveles de información lingüística así como el conocimiento general el mundo y las representaciones de los textos leídos en cada momento se almacenan en la misma.

3.1.4.1. Teorías clásicas de la memoria

La teoría clásica de la memoria [Cowan, 1988], distingue dos tipos básicos de almacenamiento: la llamada memoria a corto plazo y la memoria a largo plazo [Atkinson y Shiffrin, 1968]. El almacenamiento en la memoria a corto plazo es temporal y la información que allí se sitúa se vuelve rápidamente inaccesible si se desvía la atención hacia otra tarea. Además, la capacidad de dicha memoria es muy limitada. Por el contrario, la memoria a largo plazo es mucho más amplia y estable. El almacenamiento en la memoria a largo plazo es principalmente asociativo y relaciona elementos entre sí y elementos con la situación o contexto actual. Por el contrario, el tiempo necesario para almacenar una unidad de información en la memoria a largo plazo es bastante mayor que el necesario para almacenarla en la memoria a corto plazo [Simon, 1973].

En teorías clásicas como la de Atkinson [Atkinson y Shiffrin, 1968], la información sólo se almacena en la memoria a largo plazo si ha sido almacenada previamente en la de corto plazo, y no siempre con éxito. Según estas teorías, el almacenamiento en la memoria a largo plazo es una función probabilística del tiempo que la información estuvo previamente en la memoria a corto plazo o incluso del número de veces que

dicha información se almacenó en ella [Anderson, 1983]. Para estas teorías, la información sólo puede ser recuperada de la memoria a largo plazo. La memoria a corto plazo actúa como una antesala limitada donde se acumulan temporalmente los elementos mínimos para poder crear una unidad de almacenamiento para la memoria a largo plazo.

Las teorías de la memoria de trabajo consideran que sólo una parte de la memoria puede recuperarse en cada instante. Dicha parte consta de los elementos activos debido al contexto o a peticiones expresas. Dicha memoria es también volátil y con una capacidad limitada. Por ello, algunas de las teorías de la memoria de trabajo sí contemplan la recuperación de información desde la propia memoria a corto plazo, aunque la escasa capacidad de la misma, unas siete unidades [Miller, 1956], hizo que fueran descartadas por la comunidad científica, ya que se requiere una capacidad mucho mayor para llevar a cabo con éxito tareas cognitivas complejas. En el modelo de Newell y Simon [Newell y Simon, 1972], por ejemplo, se considera a la memoria de trabajo como un componente independiente con su propia capacidad, y la recuperación de la memoria a largo plazo se realiza haciendo un emparejamiento de las unidades almacenadas en ésta con las unidades activas en la memoria de trabajo. El modelo ACT* [Anderson, 1983], sin embargo, propone que la memoria de trabajo es una parte activa de la propia memoria a largo plazo, por lo que hereda su capacidad. Con el fin de adaptarse a la naturaleza volátil y escasa de la memoria de trabajo, cualidades globalmente aceptadas, el modelo ACT* las justifica mediante límites en los niveles de activación y en el tiempo que se mantiene la misma. Para finalizar, Ericsson y Kintsch [Ericsson y Kintsch, 1995] mantienen que existe una memoria de trabajo a largo plazo. A pesar de que el tiempo de acceso a dicha memoria se opone a la velocidad de procesamiento de la mente humana en tareas en tiempo real, los autores justifican su existencia dada la alta capacidad de almacenamiento e inferencia que requieren ciertas tareas cognitivas complejas. Según Ericsson y Kintsch, este tipo de memoria se desarrolla para dichas actividades en las que los seres humanos necesitan ser expertos y que requieren un aprendizaje intenso para ser realizadas con éxito, dada su complejidad cognitiva. Este aprendizaje y entrenamiento hace que se generen consultas específicas predefinidas y que se asocien directamente con elementos de la memoria a largo plazo, además de anticipar otras consultas futuras, disminuyendo notablemente la velocidad de acceso a la memoria a largo plazo y deshaciendo así la contradicción con las evidencias

experimentales sobre la memoria de trabajo. Entre las actividades complejas a las que los autores se refieren se encuentran los juegos como el ajedrez o los naipes, la interpretación de instrumentos musicales, el diagnóstico médico y, por supuesto, la lectura del lenguaje natural escrito.

3.1.4.2. La memoria durante el proceso de lectura

Independientemente del tipo de memoria en la que se almacene, la información que se genera y se recupera durante el proceso de lectura se guarda en distintos tipos de contenedores, asociados a distintas tareas. Desde que se lee una oración hasta que se almacena el resultado de su comprensión en la memoria a largo plazo se producen distintas representaciones intermedias que se almacenan en distintos contenedores. Todos estos contenedores han sido estudiados en detalle de manera experimental [Potter, 1983], [Baddeley, 1986], y se presentan como:

1. Icono retinotrópico. En este contenedor se almacena la información sobre una palabra proveniente de los fotorreceptores y otros mecanismos neuronales. Diversos estudios indican que este contenedor no influye en el proceso normal de lectura.
2. Memoria visual espaciotópica. Recoge la representación del texto escrito desde la retina como una estructura estable localizada en el espacio.
3. Memoria visual reatópica. Recoge la información resultado de agrupar segmentos de texto.
4. Memoria conceptual. Contiene los conceptos a los que hacen referencia las palabras y objetos aislados.
5. Lazo articulatorio. Contiene la información fonética de las palabras que se leen.
6. Lazo espacio-visual. Contiene el mismo tipo de información que el lazo articulatorio pero en el dominio espacio-visual. Se corresponde con la memoria fotográfica del texto.
7. Memoria de trabajo. Almacena la información resultado de todos los procesos cognitivos llevados a cabo. Al mismo tiempo, sirve esa información a dichos procesos. La información registrada en la memoria de trabajo es compleja y está estructurada en diferentes niveles de representación (recuérdese los tres

componentes o niveles de la representación de un texto: superficial, base y modelo de situación)

Así, los fotorreceptores registran información visual en la retina, seguida de una serie de transformaciones hasta que llega al córtex, donde se derivan representaciones de más alto nivel de la información temporalmente almacenada en todos los contendores que componen la memoria de trabajo. El acceso a algunos de los contendores está restringido a ciertos procesos y su existencia no puede ser percibida de manera introspectiva sino sólo mediante estudios experimentales. En cambio, la información de otros contenedores puede llegar a la memoria de trabajo y hacerse consciente para el individuo, retroalimentando así a los procesos que han contribuido a generarla.

En términos más generales, Glanzer [Glanzer et al., 1984] atribuye a la memoria a corto plazo la función de almacenar la estructura superficial correspondiente a las oraciones en el texto que se lee, sin ningún tipo de semántica asociada. La estructura superficial de una oración se almacena hasta el final de la misma y se pierde rápidamente. Según Ericsson y Kintsch, la memoria de trabajo a corto plazo, sin embargo, parece almacenar las consultas para recuperar estructuras necesarias de la memoria a largo plazo [Ericsson y Kintsch, 1995]. Los mismos autores atribuyen a la memoria de trabajo a largo plazo, puesto que son los que defienden su existencia en tareas complejas, el almacenamiento del texto base y el modelo de situación correspondiente al texto leído en cada momento. Dicho almacenamiento sólo se produce después de la comprensión e identificación del significado de lo que se lee. De esta forma, el texto base se almacena cada vez que se lee una oración. Dicha información presente en la memoria de trabajo a largo plazo no está activa y disponible de manera directa, sino bajo peticiones que suelen venir impuestas por el propio texto, unas veces para mantener la coherencia y otras para reforzar conceptos previamente leídos. Finalmente, el modelo de situación es la información más duradera en la memoria a largo plazo, permitiendo la posterior respuesta a preguntas sobre el texto que requieran incluso la integración de información presente en distintas partes del mismo.

3.1.4.3. Influencia de la memoria sobre la capacidad de comprensión del lenguaje escrito

Es un hecho que existen grandes diferencias de habilidad lectora entre los seres humanos. Suponiendo que los sujetos no poseen ninguna deficiencia estructural ni

funcional, dichas diferencias se han tratado de justificar mediante diferencias en la memoria. Así, por ejemplo, diversos estudios han descartado la diferencia de capacidad de la memoria a corto plazo, ya que el aumento de dicha capacidad durante la maduración intelectual del ser humano no es proporcional al aumento de la capacidad de comprensión durante la misma etapa [Dempster, 1981]. Además, otros estudios demuestran que individuos que presentan una gran diferencia en su habilidad lectora poseen una capacidad de memoria a corto plazo muy similar [Farnham-Diggory y Gregg, 1979].

Con respecto a la memoria de trabajo, y más concretamente a la capacidad para mantener información activa, los estudios sí presentan diferencias significativas [Engle et al., 1992]. Es más, la diferencia no sólo afecta a la capacidad de mantener la activación sino también a la de perderla en el caso que no sea necesaria, es decir, cuando corresponde a información irrelevante, como se produce comúnmente en los personas de edad avanzada [Hasher y Zacks, 1988].

A pesar de las diferencias halladas en la memoria de trabajo, la diferencia más significativa se presenta en la memoria a largo plazo. Dicha diferencia no se define en términos de capacidad, sino de estado de la misma. Así pues, cuanto más conocimiento se posee sobre lo que se lee mejor se comprende. Estudios experimentales demuestran que lectores hábiles pero con poco conocimiento sobre la temática de los textos que leen tiene menos capacidad de comprensión que lectores con poca habilidad de lectura pero expertos en dicha temática [Walker, 1987]. Sin embargo, a igualdad de nivel conocimiento los lectores más hábiles comprenden mejor que los no habituales. Este hecho corrobora la hipótesis de que la lectura es una habilidad adquirida y desarrollada, y apoya la teoría de la memoria de trabajo a largo plazo de Ericsson y Kintsch, en la que los lectores hábiles tienen más capacidad para crear y almacenar consultas que recuperan la información activa relevante en cada momento.

3.1.5. Las inferencias en el proceso de lectura

La inferencia es un proceso mediante el cual se reformula o se accede a cierto conocimiento a partir de otro dado. Dicho proceso está presente en diversas tareas de la lectura y se da a todo los niveles, desde el fonético hasta el semántico [Perfetti, 1999]. Por ejemplo, en la identificación de palabras, según los distintos modelos, se infiere el

significado de la palabra identificada. Así mismo, se realizan inferencias para predecir la siguiente palabra. Las inferencias, junto con el conocimiento general del mundo, permiten también transformar la representación del texto base en la representación del modelo de situación. Diversos estudios denotan la existencia de diferentes tipos de inferencias. Entre los principales se encuentran:

- Inferencias para mantener la coherencia referencial.
- Inferencias para mantener la coherencia causal.
- Inferencias predictivas. Anticipan eventos. Se desconoce el motivo que las desencadena.

Existen distintas opiniones sobre si los lectores realizan inferencias de una manera selectiva o si las realizan de manera automática. Existen dos teorías formales al respecto:

- Hipótesis Minimalista [McKoon y Ratcliff, 1992]. Las inferencias para mantener la coherencia se realizan de manera automática y el resto sólo bajo demanda.
- Hipótesis Construccionista [Graesser et al., 1994]. Se generan todos los tipos de inferencia al mismo tiempo, sean necesarias o no.

Los tipos y cantidad de inferencias generadas en cada caso distinguen no sólo a los lectores hábiles de los inexpertos, sino también el modo de lectura que se practica, es decir, si se lee de manera activa volcando toda la atención en la lectura o por el contrario se lee de manera pasiva o superficial.

3.2. Modelos Computacionales de Lectura

Un modelo computacional de lectura es un conjunto de formalismos de representación y métodos para procesarlos, materializados mediante estructuras de datos y algoritmos computacionales respectivamente, que simulan los procesos cognitivos mentales que tienen lugar durante la comprensión de un texto en lenguaje natural escrito.

Según Ashwin Ram y Keneth Moorman [Ram y Moorman, 1999], un modelo computacional de lectura debe cumplir los siguientes requisitos:

1. Ha de ser descrito en términos funcionales, computacionales y de representación:
   - Funcional. Descripción del proceso, y todos los subprocesos, en términos de sus entradas y sus salidas.
   - Computacional. Descripción del proceso de transformación de las entradas en las salidas mediante un conjunto de estructuras de datos y algoritmos programables en un ordenador.
   - De representación. Descripción del formalismo de representación del conocimiento general y de los mecanismos que lo procesan, así como del uso que se hace de éste para la generación de la salida.
2. Las funciones, procesos y representaciones deben ser psicológicamente plausibles, es decir, deben estar apoyadas y justificadas por resultados experimentales en el ámbito de la psicología o, en su defecto, por argumentos neurológicos, filosóficos o antropológicos.

La exigencia de la plausibilidad psicológica es de crucial importancia para impedir que se ideen modelos con el objetivo de superar el test de Turing [Turing, 1950]. El objetivo principal de los modelos computacionales de lectura es el de tratar de comprender el lenguaje escrito y no sólo el de aparentar que lo comprenden, puesto que dicha apariencia puede no ser indicativa del nivel de comprensión, aunque esta cuestión continúa siendo objeto de debate filosófico.

3.2.1. Utilidad de los modelos computacionales de lectura

La utilidad última de un modelo computacional de lectura es la ayuda a la comprensión del proceso de lectura en el ser humano [Ram y Moorman, 1999]. Según este fin, la utilidad de un modelo se basa pues en los siguientes aspectos:

- Aportación de una base sobre la que realizar una evaluación empírica.
- Aportación de una herramienta computacional que ayude a probar o refutar hipótesis teóricas.
- Adecuación de los datos de salida para facilitar la comparación con datos psicológicos provenientes de seres humanos.
- Flexibilidad para la incorporación de asunciones teóricas.
- Aportación de modularidad que permita aislar y evaluar los procesos por separado, de tal forma que se pueda generalizar una teoría si los resultados la apoyan.

3.2.2. Aspectos a tratar por un modelo computacional de lectura

La principal tarea que debe tener en cuenta un modelo computacional de lectura es el del tratamiento de palabras y frases, es decir, de los procesos que agrupan los significados individuales de las palabras en el significado de las oraciones y el de éstas en el discurso global. Adicionalmente, un modelo de lectura puede describir también los siguientes aspectos:

- Inferencias. Los procesos que determinan significados implícitos a partir del contexto definido por los propios textos y por el conocimiento del mundo.
- Novedad. Los procesos que tratan palabras desconocidas hasta el momento o su uso metafórico en nuevos contextos.
- Control del proceso. Modelado de la atención que se dedica a la lectura. En este aspecto intervienen también conceptos como el interés y las expectativas sobre lo que se lee.

3.2.3. Formalismos de representación en los modelos computacionales de lectura

La representación del conocimiento es esencial en el modelado cognitivo y en los sistemas de Inteligencia Artificial en general. Aunque el lenguaje es una forma de representación en sí misma, es necesario idear formalismos que representen a su vez al lenguaje para resolver en dicha representación todos los aspectos planteados. Dado que un modelo de lectura debe incluir los procesos de inferencia, es deseable un formalismo de representación canónica que optimice dichos procesos. Los modelos computacionales de lectura actuales plantean una teoría de la representación más pragmática y funcional que los modelos teóricos clásicos de la semántica, donde la representación captura la verdad si ésta se corresponde con el mundo real. Dicha visión pragmática contempla tres aspectos básicos de los formalismos de representación:

- Forma. Sintaxis de la representación y mecanismos de construcción y actualización de las estructuras, junto con los requerimientos computacionales.
- Contenido. Dimensiones y dominio que se emplean para representar el mundo.
- Organización. Asociaciones dentro del conocimiento y accesibilidad del mismo.

Domeshek [Domeshek et al., 1999] propone que el conocimiento que se representa en un modelo computacional puede ser de tres tipos: Físico, Mental o Social. Sobre el diseño del contenido para los modelos computacionales de lectura, el mismo Domeshek y sus colegas definen cuatro aspectos a tener en cuenta:

- Diseño de una ontología básica que agrupe los términos descriptivos en categorías.
- Definición de los símbolos principales y la asignación de sus significados a las categorías.
- Definición de métodos para combinar símbolos de distintas categorías.
- Diseño de mecanismos de inferencia a partir de patrones de símbolos.

Así mismo, estos autores proponen también diversas propiedades del contenido de un formalismo de representación, deseables para un modelo de lectura. El contenido de la representación debería ser pues:

- No ambiguo.
- Canónico. Sólo un significado por cada estructura.
- Coherente. Significados similares deben dar lugar a representaciones similares.
- "Composicional". Los significados de estructuras compuestas son una función sencilla de los significados de sus partes.
- Elaborado. El detalle del conocimiento se adapta a cada tarea en cuestión.
- Integrado. Con la capacidad de razonar con variables en múltiples niveles.
- Completo. Debe proveer maneras para expresar todo el conocimiento requerido por la tarea.
- Útil. Restringido exclusivamente a la tarea objetivo.
- Reutilizable. Para otras tareas, por ejemplo.
- Compacto. Simplicidad tanto en complejidad como en extensión.
- Eficiente.

En cuanto a la organización del formalismo de representación, los autores anteriormente citados distinguen entre dos tipos distintos:

- Organización orientada a objetos. La representación se compone de objetos y asociaciones entre los mismos. Las inferencias se realizan a través de dichas asociaciones.
- Organización basada en plantilla. Ejemplificación de plantillas previamente concebidas. Las inferencias se dan a través de los componentes de las propias plantillas.

Los autores proponen también dos propiedades deseables para la organización de la representación en el modelado de la lectura. La primera es que permita la construcción eficiente de significados plausibles y la segunda es que permita la distinción entre éstos y los significados no plausibles. Por último, Domeshek destaca la importancia de los

índices que dan acceso al conocimiento, y propone dos aspectos a tener en cuenta en su diseño: la especificación del contenido de los índices y el diseño de estructuras y mecanismos para poder realizar un encaje parcial eficiente.

En cualquier modelo computacional el formalismo de representación es crucial para el éxito del mismo. La justificación de tal importancia es esencialmente intuitiva: es difícil aprender y comprender si se desconoce lo que se quiere aprender y comprender.

3.2.4. Modelos computacionales de lectura

Los primeros modelos computacionales de lenguaje, descritos en el Capítulo 2, adolecen de falta de flexibilidad. Incluso siendo capaces de generar significados correctos y coherentes con el contexto, dichos modelos están restringidos a un número muy reducido de posibles interpretaciones y a un dominio muy particular del lenguaje. Otros sistemas, por el contrario, son muy flexibles pero poco precisos y por lo tanto ineficientes.

3.2.4.1. Modelos de Construcción-Integración

Con el objetivo de alcanzar un equilibrio entre flexibilidad y precisión, Walter Kintsch propuso un modelo que es el referente de muchos de los modelos computacionales de lectura actuales. Dicho modelo se conoce como el modelo de Construcción-Integración [Kintsch, 1988]. Es, en parte, un modelo conexionista donde la representación del conocimiento viene expresada como una red de nodos con asociaciones entre los mismos, tanto positivas como negativas. Cada nodo está compuesto por una cabecera y un número variable de campos. Los campos guardan distinta relación con la cabecera, haciendo referencia a atributos, componentes, propiedades, funciones, etc. De manera general, el modelo propuesto por Kintsch genera un gran número de inferencias entre las que siempre se encuentran las interpretaciones o significados correctos. Esta generación de inferencias es muy flexible, puesto que no se ha de adaptar a ningún contexto. A continuación, el modelo realiza una integración de las inferencias generadas por el contexto, siendo este proceso de integración el precio a pagar por la flexibilidad. Así, el modelo consta de dos fases que le dan nombre:

- La primera es la fase de construcción. En primer lugar se crean los conceptos y proposiciones correspondientes a la entrada textual mediante un analizador. A continuación, se elaboran estos elementos seleccionando los vecinos más simples y cercanos a ellos en la red de conocimiento, es decir, los que tengan una probabilidad de asociación mayor. A partir de éstos se infieren proposiciones adicionales utilizando un sistema de producción y, por último, se asigna un peso a todos los elementos inferidos a través de la herencia del peso de los elementos fuente de las inferencias.
- La segunda fase es la llamada de Integración. En esta fase se comienza con un vector que contiene las activaciones iniciales de todos los elementos generados en la fase de construcción. Dicho vector se multiplica reiteradamente por la matriz de pesos, normalizando los valores a cada iteración. El proceso se detiene cuando el vector de activaciones se estabiliza, es decir, no varía con la multiplicación. Si transcurrido un número determinado de iteraciones la representación no se estabiliza, entonces se añaden nuevos elementos y se procede a integrar una vez más. Una vez que la activación de los elementos está estable, los nodos con mayor activación son los que componen la representación final del texto de entrada.

De esta forma, el modelo construye una representación de texto base a partir de la entrada textual y del conocimiento general, integrando a continuación dicho texto base en un contexto coherente. El modelo se define así como un híbrido, ya que la fase de construcción se realiza mediante un sistema de producción y la fase de integración se lleva a cabo mediante una propagación de la activación en una red conexionista.

Como modelo computacional de lectura, Kintsch justifica la plausibilidad psicológica del mismo mediante evidencias experimentales. La primera de ellas se basa en el tiempo de respuesta del ser humano en reconocer la siguiente palabra del texto tras leer la palabra inmediatamente anterior [Fischler y Goodman, 1978]. Dicho tiempo es muy reducido (aproximadamente 50 ms) debido a que la percepción visual reduce rápidamente los posibles candidatos, con lo que se justifica la activación de sólo los vecinos más cercanos y simples a la palabra leída. La segunda evidencia apoya la fase de integración propuesta por el modelo. Los estudios experimentales correspondientes demuestran que el tema central de una oración no se percibe hasta más de un segundo después de haberla leído, descartándose los conceptos no asociados a los 200 ms y los asociados, pero no por el contexto, a los 500 ms después de haberlos leído [Till et al.,

1988]. Este es el proceso que simula la estabilización de la activación en el modelo, donde los nodos más relevantes van ganando activación con respecto a los nodos menos relacionados, alcanzando dicha diferencia de activación una magnitud significativa al final del proceso.

Langston utiliza también un enfoque conexionista para el modelado de la lectura [Langston et al., 1999], concretamente una modificación del modelo de Construcción-Integración de Kintsch. Utilizan esta variante para modelar medidas de la capacidad lectora desde dos puntos de vista: "on-line" y "off-line". Las medidas "on-line" cuantifican la capacidad lectora durante el procesado del texto. Las medidas "off-line", por el contrario, asignan valores numéricos a la capacidad de comprensión después de haber finalizado la lectura del texto. Las medidas "on-line" que proponen se definen como el número de ciclos para que se estabilice la red en la fase de integración, el nivel de activación de la memoria y el peso medio de las asociaciones. Comparan estas medidas con otras medidas utilizadas para evaluar "on-line" a sujetos humanos, como el tiempo de lectura por sílaba o el grado de relación que una frase guarda con frases leídas previamente. Los resultados experimentales muestran que las medidas propuestas en el modelo tienen una alta correlación con las medidas tomadas de los sujetos humanos. Por otro lado, las medidas "off-line" que los autores proponen son el nivel final de activación y el peso medio final de las asociaciones. En este caso no realizan comparación alguna con otras medidas de comprensión "off-line" comúnmente tomadas de sujetos humanos, como son la calidad de resúmenes elaborados o de las respuestas a preguntas sobre el texto leído.

Una de las limitaciones del modelo de Construcción-Integración es que sólo admite conocimiento en forma de proposiciones, y requiere para ello analizadores y sistemas de producción que siguen sin aportar flexibilidad. Además, el sistema de Kintsch no contempla el modelado ni la posible inclusión de otros aspectos que posteriormente fueron reconocidos como integrantes inequívocos del proceso de lectura. Dichos aspectos han sido el objeto de modelos computacionales de lectura posteriores. De hecho, el énfasis que dichos modelos vuelcan en cada uno de estos aspectos establece un criterio de diferenciación entre los mismos.

3.4.2.2. Modelos de interacción entre fuentes lingüísticas

El modelo propuesto por Mahesh [Mahesh et al., 1999] se centra en el equilibrio entre la sintaxis y la semántica en la lectura y comprensión de oraciones aisladas. Adoptan la teoría interactiva de la lectura, en la que las fuentes de información están siempre disponibles y participan en el proceso bajo petición expresa. El mecanismo de peticiones es, precisamente, el objeto principal de este modelo llamado COMPERE. Los autores describen este mecanismo como un modo de arbitraje que integra todos los niveles de información en los momentos precisos. Representan el conocimiento sintáctico y semántico de manera independiente pero homogénea, es decir, utilizando el mismo formalismo de representación para todos ellos con el propósito de favorecer la integración. Dicho formalismo consiste en un conjunto de reglas y roles agrupados en categorías e indexados mediante precondiciones. El módulo de análisis sintáctico decide cuándo concatenar una unidad sintáctica a un nodo padre de la estructura sintáctica actual de la oración, proponiendo posibles concatenaciones, seleccionado un conjunto apropiado de ellas y llevándolas a cabo. El módulo de análisis semántico realiza las mismas acciones pero concatenando roles en lugar de unidades sintácticas. El arbitraje consiste en asignar un valor de preferencia a cada una de las posibles concatenaciones en un momento dado, ya sean de carácter semántico o sintáctico. De todas esas concatenaciones se escoge aquella que esté propuesta por las dos fuentes de información citadas y que tenga el valor de preferencia más alto. De esta forma se integran los dos tipos de conocimiento. Contemplan también la posibilidad de errores de comprensión, reflejados mediante la no completitud de la estructura construida al final de la oración o mediante la imposibilidad de encontrar concatenaciones comunes a ambas fuentes. El modo de arbitraje propuesto aporta escalabilidad al sistema, puesto que la incorporación de cualquier otra fuente de información es directa.

Debido al dinamismo del lenguaje, es decir, a su constante evolución en el tiempo y en las comunidades que lo emplean, un modelo de lectura debería poder contemplar la comprensión de términos desconocidos o nuevos usos de términos conocidos. Con este objetivo, Peterson y Billman proponen un modelo computacional denominado modelo de Correspondencia Semántica [Peterson y Billman, 1999]. En dicho modelo la interacción de los niveles lingüísticos es secuencial y se ajusta a un flujo de información "Abajo-Arriba". Así, cada oración leída se analiza sintácticamente y se genera un árbol sintáctico como resultado de dicho análisis. Dicho árbol es la entrada de un analizador

semántico que genera interpretaciones del árbol sintáctico de entrada. Dichas interpretaciones son nuevamente la entrada de un módulo de compresión conceptual que produce como salida el significado final de la oración leída. El conocimiento se representa mediante conceptos de dos tipos: objetos y personas. Cada concepto tiene asociada una serie de atributos, como eventos, estados, lugares y causas, entre otros. El nombre del modelo se deriva del proceso de transformación de la estructura sintáctica en interpretación semántica. En dicho proceso se hace corresponder, mediante reglas, cada elemento del árbol sintáctico (rol sintáctico y léxico) con un concepto, como muestra la Figura 3.5.

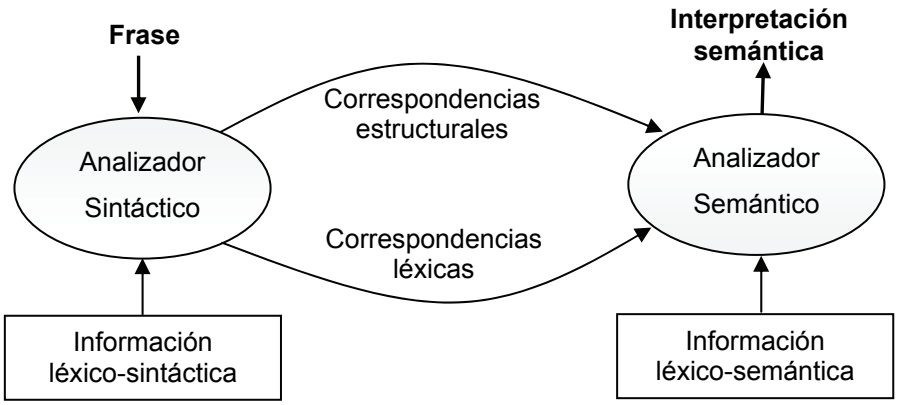


Figura 3.5. Diagrama de flujo de información en el modelo de Correspondencia Semántica.

El léxico y la sintaxis restringen pues el significado de las oraciones. Puesto que el árbol semántico es coherente con el sintáctico, el modelo permite predecir pues el significado de nuevos términos asignándoles conceptos compatibles con la estructura. De esta forma, aunque el sistema está restringido a oraciones que describen estados espaciales o de posesión, es capaz de producir interpretaciones coherentes del texto leído en presencia términos desconocidos. Ésta es una cualidad que no poseen la mayoría de los sistemas de procesamiento de lenguaje natural existentes.

3.4.2.3. Modelos de influencia del texto en el lector

Otro aspecto relevante a tener en cuenta es la influencia del estilo y formato del texto en la comprensión del lenguaje escrito. Meyer propone un modelo representado por conjuntos de variables que recoja las características individuales tanto de los textos como de los lectores [Meyer y Poon, 2001]. Meyer destaca la existencia de estrategias de lectura por parte de los lectores y pone de manifiesto cómo el tipo y formato de los

textos hace variar dichas estrategias. Además, hace énfasis en la integración de las intenciones tanto del escritor como del autor. El descubrimiento del lector de la intención del autor influye en la estrategia de lectura a seguir. Así pues define los siguientes tipos de variables:

- Variables de estrategia:
  - Plan.
  - Relectura.
  - Inducción de preguntas.
- Variables de intención del autor. Normalmente denotadas por señales lingüísticas:
  - Informar.
  - Entretener.
  - Persuadir.
  - Proveer una experiencia.
- Variables de tarea:
  - Modo de presentación del texto. En papel, en formato electrónico, por oraciones, por párrafos, tipo de letra, maquetación, etc.
  - Requerimientos de la tarea.
  - Tipo de tarea.
  - Forma de evaluación del lector en la tarea.
- Variables del texto:
  - Estructura
  - Tema central.
  - Señales lingüísticas de intención.
  - Género.
  - Cohesión.
  - Niveles de la estructura.
  - Detalles.
  - Estilo sintáctico.
- Variables del lector:
  - Habilidad verbal.
  - Nivel académico.

- Edad.
- Cultura general.
- Perspectiva.
- Valores.
- Interés.
- Capacidad de la memoria.
- Estilo de lectura.
- Hábito de lectura.

A pesar de que algunos expertos piensan que el texto no tiene significado por sí mismo, sino que lo forma el lector de manera individual, está constatado que la forma e intenciones con las que el autor escribe influyen en la interpretación de los textos. Meyer concluye que la planificación de la escritura de un texto, con el fin de inducir la utilización de estrategias en la lectura, ayuda a compensar muchas de las limitaciones que el lector pueda poseer. Así pues, un modelo que identifique y evalúe esas limitaciones puede ayudar a superar las carencias de sujetos con problemas de lectura mediante cambios en el estilo y la forma de los textos que se les presentan.

Una de las variables que define Meyer para las estrategias de escritura es la inducción de preguntas al lector. El planteamiento introspectivo de preguntas durante la lectura es central para la comprensión del texto que se lee. Éste es el aspecto principal del modelo computacional propuesto por Ram, denominado AQUA [Ram, 1999]. Según Ram, las preguntas que emergen en la mente de un sujeto durante la lectura denotan el grado de comprensión. Además, los objetivos o intenciones del lector se pueden ver como preguntas sobre el dominio de interés. Así, el modelo AQUA se presenta como un proceso de generación de preguntas y respuesta a las mismas durante la lectura. El conocimiento, es decir, las preguntas y respuestas, se almacena en forma de marcos y se organiza en una taxonomía definida *ad hoc* por el autor, que describe qué preguntas llevan a otras, cómo se pueden responder y qué consecuencias tiene su respuesta. Así, el modelo lee una porción de texto, centrando o no la atención (centrar la atención implica no explorar todas las posibilidades a la hora de recuperar preguntas). A continuación determina la relevancia del fragmento leído basándose en preguntas anteriores. Posteriormente, recupera preguntas relevantes de la memoria y las emplea para centrar la atención. Seguidamente, trata de responder a las preguntas recuperadas con el fragmento de texto leído. Inmediatamente después genera preguntas con relación al

fragmento de texto leído y las almacena en memoria, para proceder a leer el siguiente fragmento de texto. Es pues un modelo "Abajo-Arriba" ya que el texto sugiere preguntas, aunque también contempla el enfoque "Arriba-Abajo" puesto que las preguntas en memoria condicionan la atención con la que se lee un fragmento de texto, como muestra la Figura 3.6.

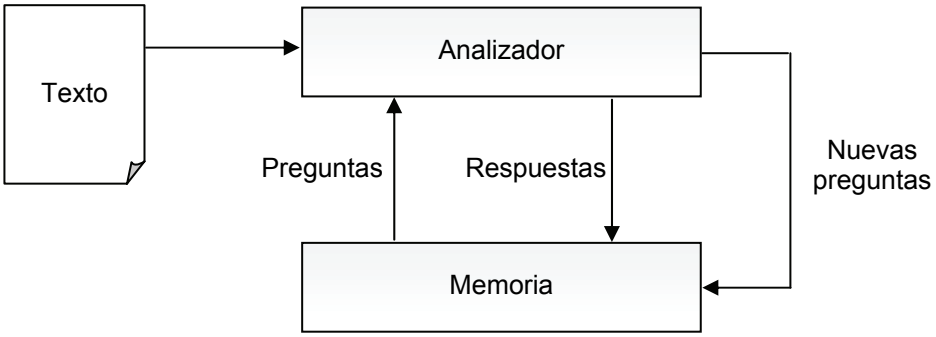


Figura 3.6. Diagrama de flujo de información en el modelo AQUA.

3.4.2.4. Modelos de influencia del lector en la lectura del texto

Otro aspecto importante en la lectura es la memoria episódica. Lange y Wharton defienden que la comprensión de un pasaje depende en gran medida de la capacidad de recordar episodios similares que fueron entendidos previamente en circunstancias parecidas [Lange y Wharton, 1993]. Comúnmente, el estudio de la comprensión de textos y el de la recuperación de información de la memoria episódica han seguido caminos independientes. El modelo propuesto por los autores citados, denominado REMIND, trata de integrar ambos aspectos de la cognición. Evidencias experimentales muestran que la base para el recuerdo es la estructura superficial del texto, más que el texto base o el modelo de situación. El conocimiento del modelo REMIND está representado por una red de nodos interconectados. Dichos nodos son marcos con diversos roles. Las interpretaciones de los textos así como los episodios en memoria se representan como conjuntos de nodos de dicha red. Así, el modelo recibe una representación textual como entrada y construye una interpretación de la misma, identificando los nodos y roles que aparecen en la misma. Dichos nodos se activan y propagan dicha activación por la red. La propagación de la activación hace que se recuperen episodios similares a la interpretación elaborada. En definitiva, el modelo interpreta un texto de entrada y recupera de la memoria interpretaciones similares previas. Posteriormente, las inferencias a partir de la información recordada se utilizan

para refinar la interpretación del texto de entrada y así comprenderlo de manera completa. De esta manera, los autores proponen un modelo computacional psicológicamente plausible de la interacción entre la comprensión del lenguaje y recuperación de la memoria episódica.

El modelo propuesto por Gerrig contempla un aspecto más abstracto que los citados hasta el momento: la implicación del lector en los textos que lee, concretamente en las historias [Rapp y Gerrig, en prensa]. Gerrig explica cómo se generan oraciones en la mente producidas por la intensidad con la que el lector está inmerso en el proceso de lectura. Estas oraciones se denominan respuestas participativas y se dan con mayor frecuencia cuando la historia sugiere un final negativo. Las respuestas participativas no son inferencias, aunque se derivan de las mismas o son propiciadas por ellas. Gerrig defiende la integración de las respuestas participativas en los modelos computacionales de lectura. Para dicha integración, es necesario contemplar cuestiones tales como qué aspectos del texto hacen que emerjan las respuestas, qué respuestas emergen de manera automática y qué papel juega la memoria a largo plazo en la generación de las mismas. Gerrig también aporta las claves para incorporar las respuestas participativas a la representación o interpretación de los textos. La primera de las claves se basa en una definición de perspectiva, clasificada en dos tipos: perspectiva externa, la que el lector tiene del mundo fuera del texto, y perspectiva de diferencia, la que el lector percibe con respecto a la perspectiva que tienen los personajes de las historias a las que se refieren los textos. La segunda de las claves está basada en el concepto de expectativa, clasificada análogamente en los dos tipos en los que se clasifica la perspectiva. El autor concluye que la diferencia entre los tipos de perspectiva y expectativa es la que genera "suspense" y, por lo tanto, induce a la producción de respuestas participativas.

3.4.2.5. Modelos de aspectos cognitivos complejos

Rapaport y Shapiro proponen un modelo de lectura que represente y trate la ficción [Rapaport y Shapiro, 1999]. A partir de cuatro teorías sobre la ficción, los autores definen qué es un objeto ficticio, cuáles son sus propiedades, cómo se representan en los procesos mentales y cómo se distinguen de los objetos reales, todo ello en términos computacionales. De esta manera, idean varias estructuras y mecanismos computacionales que se pueden resumir en:

- Un operador de historia. Establece el tipo de espacio o mundo en el que se desarrolla la historia y permite distinguir las creencias propias del modelo de los predicados leídos.
- Un único modo de predicado. Las entidades, ya sean ficticias o no, se representan de igual forma en los procesos mentales.
- Un único tipo de propiedad para los objetos.

Con estos elementos construyen redes asociativas capaces de representar el conocimiento propio y los predicados de las historias leídas.

El concepto de novedad es el aspecto central de ISAAC, el modelo computacional ideado por Moorman y el mismo Ram [Moorman y Ram, 1999]. En este caso, el concepto de novedad es tratado a más alto nivel, en el ámbito de las historias, e introducen el término creatividad para referirse a la compresión de dicha novedad. Dado el ámbito específico de las historias, la creatividad debe ser útil y no disparatada. Los autores distinguen además cuatro tipos de novedad: novedad "absoluta", cuando un concepto es desconocido, novedad "ejemplificada", cuando un concepto es desconocido pero realiza una función conocida, novedad "evolutiva", cuando un concepto desconocido realiza una función conocida de manera más precisa y novedad "revolucionaria", cuando un concepto desconocido realiza una función conocida de una manera completamente distinta. El modelo representa el conocimiento como marcos con atributos (primarios y secundarios), roles y funciones. Los conceptos están organizados en un matriz, donde las filas denotan tipos de conceptos y las columnas el dominio de los mismos, como se muestra en el ejemplo de la Figura 3.7.

	Físico	**Mental**	**Emocional**	**Social**	**Temporal**
Acciones	*andar*	…	…	…	…
Objetos	*roca*	*idea*	…	…	…
Agentes	…	…	…	*jefe*	…
Estados	…	…	…	…	*temprano*

Figura 3.7. Ejemplo de representación del conocimiento conceptual en el modelo ISAAC.

La comprensión de un concepto desconocido supone la transacción de dicho concepto de una posición a otra en la matriz. Las transacciones están descritas mediante un conjunto de reglas y su complejidad, en términos de celdas implicadas en el

desplazamiento, modela la dificultad de comprensión del nuevo concepto. El modelo incluye además mecanismos de comprensión, mediante inducción y abducción, mecanismos de análisis de oraciones, tanto léxicos como sintácticos, y un modelo de memoria. De manera general, y dado el dominio para el que está diseñado el modelo, ISAAC descompone el proceso de lectura en tres supertareas:

- Control del proceso. Integra el resto de tareas. Específicamente controla el foco de atención, gestiona el tiempo y lleva a cabo la aceptación temporal de conceptos que violan el conocimiento del mundo, disparando el mecanismo de comprensión creativa.
- Comprensión del escenario. Identifica y describe eventos, agentes y acciones, así como sus intenciones y creencias.
- Comprensión de la estructura de la historia. Identificación tanto de personajes como del género y del tiempo.

El modelo ha sido evaluado comparando las respuestas generadas por el modelo y por sujetos humanos a preguntas planteadas sobre varias historias de ficción previamente leídas por ambos, obteniendo un alto grado de correlación.

La resolución de problemas y el aprendizaje son otros dos aspectos que Cox propone como integrantes del proceso de lectura [Cox y Ram, 1999]. Para el autor, tanto el aprendizaje como la comprensión de texto escrito consisten en tres fases: identificación, generación y verificación. El modelo computacional de lectura propuesto se centra en la fase de generación, e incluye, además de un subsistema de comprensión, un módulo de aprendizaje como componente activo en la lectura. El modelo, denominado Meta-AQUA [Cox y Ram, 1999], es una modificación del modelo AQUA descrito anteriormente. Los conceptos están representados por marcos y se distribuyen entre conocimiento general del mundo y modelo del texto. El modelo también recoge el interés del lector. Así, los autores definen dos tipos de interés:

- un concepto es interesante si es anómalo, es decir, si contradice el conocimiento previamente adquirido sobre el mundo.
- un concepto es intrínsecamente interesante si se encuentra entre los objetivos del lector o ha sido aprendido recientemente.

Así, además de reaccionar ante lo contradictorio o desconocido, el modelo contempla los objetivos del lector. El módulo de comprensión del sistema se centra en la generación de explicaciones ante anomalías. Dicha tarea se lleva a cabo en cuatro fases: elabora la anomalía, genera preguntas que representen las diferencias entre la anomalía y el modelo del texto, construye la explicación, consistente en las respuestas a las preguntas planteadas, y finalmente verifica la posible explicación. Una vez verificada la explicación, ésta se aprende para subsanar la carencia que había provocado la anomalía. El módulo de aprendizaje utiliza un enfoque multiestratégico, componiéndose de un conjunto de algoritmos de aprendizaje que combina para llevar a cabo su tarea. Dicho módulo opera de la siguiente manera: detecta el fallo o anomalía, busca la causa, crea los objetivos de aprendizaje, elige los algoritmos de aprendizaje que va a aplicar y los ejecuta. La Figura 3.8 muestra un esquema de la dinámica del modelo.

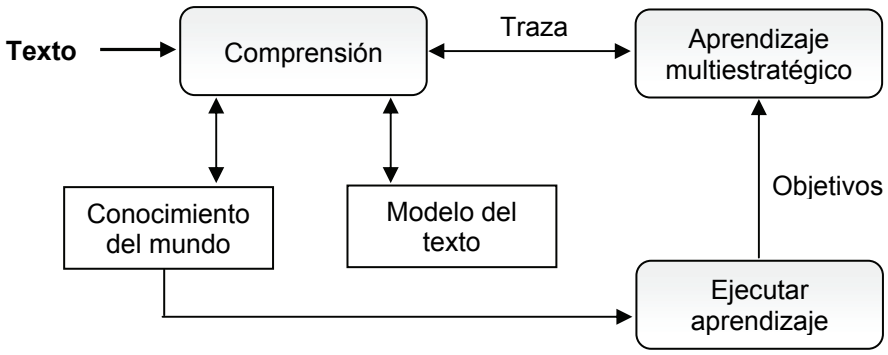


Figura 3.8. Esquema general del modelo Meta-AQUA.

Los autores evaluaron el modelo sometiéndole a preguntas sobre los textos leídos y evaluando sus respuestas. Los resultados mejoraron notablemente cuando el modelo utilizó el módulo de aprendizaje junto con el de comprensión.

3.4.2.6. Modelos de identificación y extracción del conocimiento

Muchos de los modelos descritos anteriormente representan el conocimiento conceptual mediante marcos. Dichos marcos han de ser descritos y creados manualmente, lo que es costoso y limita el dominio de aplicación del modelo. Por ello, Riloff defiende la Extracción de Información (EI), campo perteneciente al Procesamiento del Lenguaje Natural que trata de identificar de manera automática entidades junto con sus roles y propiedades que aparecen en los texto leídos, como

componente del proceso de lectura [Riloff, 1999]. Riloff propone un sistema, denominado AutoSlog TS, que es capaz de identificar conceptos y generar sus marcos correspondientes de manera automática a partir de textos escritos en lenguaje natural. Tan sólo es necesario aportar al sistema dos conjuntos de textos: uno formado por textos relevantes para el dominio objetivo y el otro compuesto de textos distantes de dicho dominio. Tras un análisis sintáctico, el sistema genera todos los posibles marcos para los conceptos identificados basándose en su rol sintáctico, junto con un índice de relevancia con respecto al dominio. De esta forma, el sistema modela la capacidad de los seres humanos para identificar y ubicar conceptos en un determinado campo semántico.

Los sistemas descritos anteriormente están centrados en aspectos puntuales y psicológicamente plausibles del proceso de lectura. La razón de dicha especialización es la complejidad del proceso mismo de lectura. Aunque pueda parecer lo contrario, la capacidad de aislar los diferentes componentes o tareas de la lectura es de gran utilidad. Dado que dicho aislamiento de subprocesos entraña una dificultad muy alta en los experimentos realizados con sujetos humanos, la especialización de los modelos permite el estudio individual de las tareas y componentes implicados en la lectura, si bien es cierto que limita la aplicación de dichos modelos a otros aspectos de la cognición en la comprensión del lenguaje natural escrito.

3.4.3. Evaluación de los modelos computacionales de lectura

Es difícil cuantificar y evaluar la bondad de los modelos computacionales de lectura dado que tratan con aspectos de diferente naturaleza y complejidad y que el referente es el ser humano. Así pues, Fletcher propone cuatro criterios de evaluación que califican al modelo desde diferentes puntos de vista [Fletcher, 1999]:

- Suficiencia computacional de las teorías que implementan. Tipo de entrada que requiere el modelo y qué variaciones de la misma permiten las simplificaciones en las que se basa.
- Provisión de explicaciones no conocidas de las teorías. Es decir, los modelos pueden permitir el aislamiento y evaluación individual de los procesos, discriminación que resulta imposible mediante experimentación psicológica.

Además, esta capacidad de aislamiento permite estudiar interacciones individuales entre diversos procesos, nuevamente imposibles de evaluar mediante otro tipo de experimentación. Un modelo computacional puede pues aportar un marco de evaluación experimental que permite obtener información inaccesible hasta el momento.

- Provisión de medidas de similitud con el comportamiento humano. El modelo puede aportar ciertas medidas de evaluación tanto "on-line" y "off-line", como se comentó en la sección anterior, y su correspondencia con las medidas comúnmente tomadas de los seres humanos para evaluar su capacidad de lectura.
- Utilidad para resolver problemas reales. Los modelos pueden ser directamente aplicados al filtrado y clasificación de contenido textual, a la realización automática de resúmenes y a diversas aplicaciones de Procesamiento de Lenguaje Natural. Tanto el requisito de la definición manual de sus dominios como la especialización de los modelos en aspectos concretos son impedimentos para la aplicación en tareas reales dado su alto coste y sus limitaciones funcionales, respectivamente. Como resultado de la búsqueda de la utilidad por parte de los modelos cognitivos se produce al mismo tiempo un acercamiento hacia la cognición por parte de los sistemas computacionales puros ya útiles por naturaleza, en un intento por mejorar su eficacia en las tareas para las que se han diseñado.

CAPÍTULO 4

SILC, Sistema para la Indexación de Textos mediante un Modelo Computacional de Lectura Cognitiva

Un modelo computacional de lectura trata de describir, mediante estructuras de datos y algoritmos que las manejan, el proceder de los seres humanos durante el proceso de lectura de textos. Si dichas estructuras y algoritmos están sujetas desde el principio a la plausibilidad psicológica, es decir, se ciñen a lo que se conoce sobre la mente humana, los propios procesos de diseño e implementación plantean a su vez hipótesis estructurales y funcionales sobre el proceso de lectura humano. Así, no sólo se han de tener en cuenta las limitaciones computacionales de espacio y tiempo sino también las restricciones psicológicas que estas limitaciones suponen.

4.1. Descripción General

El sistema SILC (Sistema de Indexación por Lectura Cognitiva) propuesto en este trabajo de tesis está basado en un modelo computacional de lectura que se utiliza para indexar o representar la semántica de los textos una vez leídos por el sistema. Esta representación producida pretende contener la semántica conceptual del lenguaje contenido en los textos. Además, la representación trata de reflejar la diferente significación o importancia de los conceptos y sus relaciones semánticas. Es decir, tanto el contenido como la organización resultante de la aplicación del modelo de lectura han de permitir un uso eficaz y sencillo de la representación en diferentes aplicaciones o tratamientos posteriores, como la realización automática de resúmenes, la recuperación de información, la clasificación en categorías temáticas o los sistemas de pregunta y respuesta, entre muchas otras.

El proceso para llegar a esa representación semántico-conceptual queda pues en manos del modelo computacional de lectura en sí, cuyo diseño está condicionado por la plausibilidad psicológica de los elementos que lo componen y siempre bajo la hipótesis de que cuanto más se asemeje el modelo al ser humano mejores resultados obtendrá en sus aplicaciones.

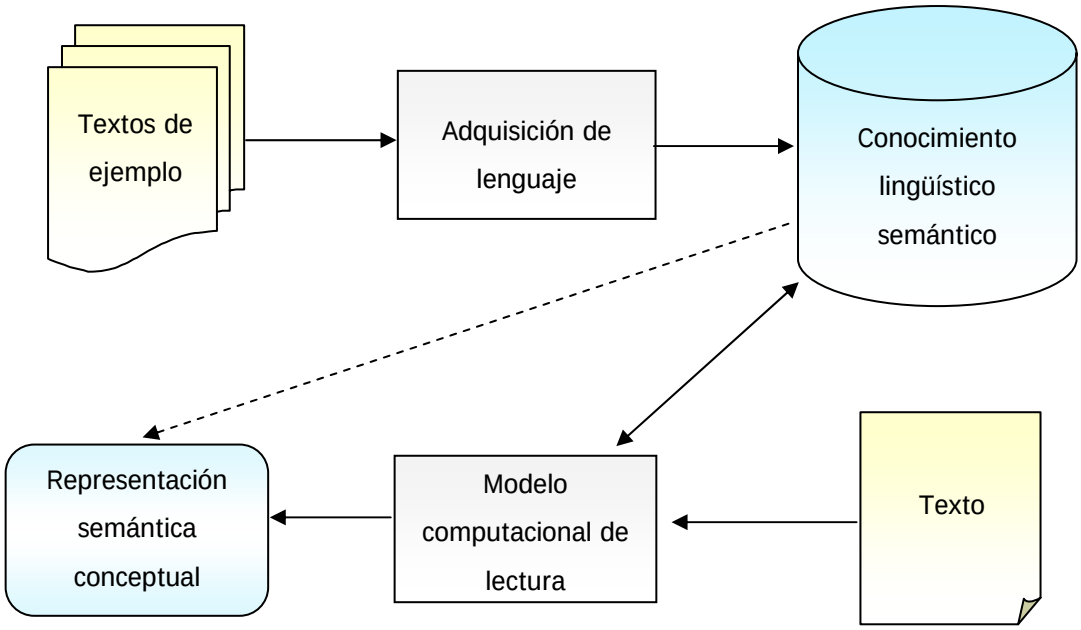


Figura 4.1. Esquema general del sistema SILC.

Como ya se comentó en el primer capítulo, el modelo planteado en esta tesis sólo se ocupa de las etapas de adquisición del lenguaje y comprensión del mismo, centrándose en esta última, aunque también pretende proporcionar una vía y punto de partida para la generación de lenguaje natural. El modelo de lectura opera sobre un conocimiento lingüístico adquirido previamente, basando en él las representaciones producidas como resultado de la lectura secuencial de un texto dado, como se muestra en la Figura 4.1. Así pues, se describen a continuación las estructuras empleadas para almacenar el conocimiento lingüístico adquirido y los algoritmos para crearlas a partir de textos de entrenamiento o ejemplo. Posteriormente, se define la representación semántica conceptual de los textos que producirá el sistema, los algoritmos que utilizarán el conocimiento lingüístico almacenado para generarla y los parámetros que los caracterizan, modelando así el proceso de lectura propiamente dicho. Finalmente, se describe el proceso que permite establecer la comparación entre las representaciones semánticas producidas por el modelo y los algoritmos que los implementan. Dichos procesos han sido concebidos para la utilización eficiente de las representaciones semántico-conceptuales en la clasificación de textos o en la recuperación de información, entre otras áreas de aplicación.

4.2. Preprocesamiento

El preprocesamiento de los textos [Sebastiani, 2002], [Feldman y Sanger, 2006] es una etapa común en la gran mayoría de los sistemas de procesamiento de lenguaje natural que tienen un objetivo práctico. Consiste en realizar ciertas transformaciones a todos los textos de entrada al sistema. Dichas transformaciones suelen tener la eficiencia como objetivo. Puesto que la mayor parte de los lenguajes naturales constan de vocabularios muy extensos y están descritos por gramáticas complejas, es necesario hacer simplificaciones y reducciones de los mismos para que puedan ser almacenados y tratados por un ordenador en un periodo razonable de tiempo.

Obviamente, las simplificaciones y reducciones casi siempre conllevan una pérdida de información y, en ciertos casos, la aparición o el aumento de ambigüedad. Así pues, la elección de las medidas y tipos de simplificación a aplicar se ha de realizar teniendo en cuenta los problemas derivados anteriormente mencionados y la manera en que éstos afectan al objetivo del sistema.

Uno de los métodos de preprocesamiento aplicados por el sistema SILC presentado en esta tesis es el de la aplicación de una lista de corte [Sebastiani, 2002]. Es una etapa que requiere la definición de una lista compuesta por palabras muy frecuentes y vacías de significado a priori. De esta forma, se reduce el vocabulario que ha de manejar el sistema con una pérdida de información relativamente pequeña. Algunos tipos de palabras que se suelen incluir en las listas de cortes son las preposiciones, conjunciones, interjecciones, artículos, etc. Dado que el sistema SILC pretende producir representaciones conceptuales de los textos y que los citados tipos de palabras no poseen una semántica descriptiva propia, la aplicación de la lista de corte no afecta significativamente al objetivo. Aunque muchas de las palabras contenidas en la lista aportan información semántica relacional, SILC recurre a otras fuentes presentes en los textos para establecer en la representación semántica conceptual las relaciones entre conceptos. Nótese que la lista de corte es dependiente del idioma, por lo que es necesario la construcción de una diferente para cada idioma.

Otro de los métodos comunes de preprocesamiento, empleado también por el sistema SILC, es el de la reducción de las palabras a su lexema raíz [Hull, 1996]. De este modo,

se agrupan palabras de la misma familia con un mismo significado bajo un único concepto representado por la raíz léxica común de las mismas. De esta forma, se pretende reducir de nuevo el tamaño del vocabulario con una pérdida mínima de información semántica. Así pues, todas las formas y tiempos de todo verbo son consideradas como una única palabra y concepto, y se eliminan el género y el número de todos los términos agrupándose a su vez en uno sólo. De este modo, cada aparición de "amo", "amaba" y "amaré", por ejemplo, se considera como la aparición del concepto "am", y los términos "amigo", "amiga", "amigos" y "amigas" se reúnen bajo el concepto "amig".

Dado que el sistema SILC produce representaciones conceptuales de alto nivel que reflejan el sentido y semántica general de los textos, la pérdida de información que supone la reducción de las palabras a su lexema raíz no tiene efectos significativos en las representaciones resultantes, si bien es cierto que las limitaciones de los algoritmos existentes que implementan la reducción introducen otro tipo de problema: la ambigüedad. Algunas palabras ortográficamente similares con significados dispares, como por ejemplo "costa" y "coste", pueden ser reducidas a la misma raíz debido a la dinámica ingenua de los algoritmos actuales. Sin embargo, este efecto no influye significativamente al sistema SILC, puesto que define y representa la semántica en términos de relaciones con otros conceptos y no de manera individualizada. Así, si la raíz "cost" está relacionada con conceptos como "mar" y "puerto" se asigna a dicha raíz el significado de "costa", mientras que si se relaciona con conceptos como "dinero" o "compra" se considera el significado de "coste". El algoritmo de reducción empleado por el sistema SILC es el algoritmo de Porter [Porter, 1980]. Es uno de los algoritmos más populares y eficaces, existiendo una gran variedad de versiones para diferentes idiomas (fue inicialmente diseñado para la lengua inglesa) [Frakes, 1992]. Concretamente, SILC contempla el filtrado por lista de corte y la reducción de raíz para el español, el inglés, el francés, el alemán, el italiano, el portugués, el finés, el danés, el noruego y el sueco.

Otros métodos comunes de preprocesamiento son el filtrado de palabras "raras" y la reducción de características [Liu y Motoda, 1998]. En el primero de ellos se eliminan palabras con muy poca frecuencia en los textos empleados para entrenar al sistema, al considerar que no van a contribuir en absoluto a la eficacia del mismo. En el segundo, sólo se conservan un porcentaje de las palabras que mayor significación semántica

poseen según ciertas medidas de la teoría de la información o según validación experimental [Yang y Pedersen, 1997], [Guyon y Elisseff, 2003], [Del Castillo y Serrano, 2004]. Ambos son métodos que dependen en gran medida de la aplicación práctica a la que esté orientado el sistema. El sistema SILC, dada su generalidad e independencia de la aplicación, no utiliza pues ninguno de los dos métodos citados anteriormente.

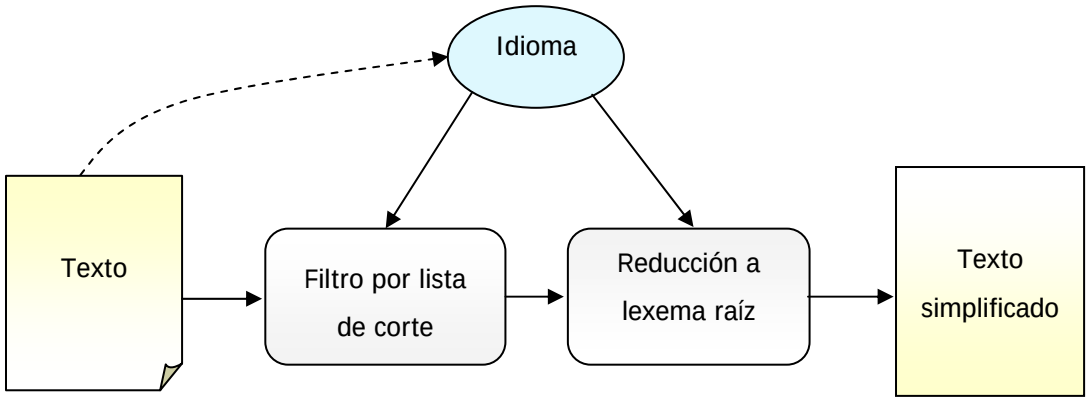


Figura 4.2. Etapas del preprocesamiento de los textos de entrada al sistema SILC.

De esta forma, cualquier texto en lenguaje natural que vaya a utilizar el sistema, bien para construir el conocimiento lingüístico o bien para ser leído y representado, sufrirá una etapa de preprocesamiento como la descrita en la Figura 4.2.

4.3. Adquisición del Conocimiento Lingüístico

La etapa de adquisición del conocimiento lingüístico es previa a la aplicación del modelo de lectura y representación de textos por parte del sistema. En esta etapa se tratan de obtener las relaciones semánticas entre conceptos adquiridas a través de la experiencia en la lectura de textos. En cierto modo, trata de simular el aprendizaje de la semántica del lenguaje aunque sin la pretensión de modelar ningún aspecto humano del mismo. El diseño del método de adquisición está encaminado a la obtención de la información necesaria para la correcta aplicación posterior del modelo de lectura. Dicho método está también orientado a la adquisición de manera eficiente de todo ese conocimiento a partir de un gran número textos de entrenamiento de modo que permita su aplicación práctica en ámbitos reales. El conocimiento que se adquiere durante esta etapa consta del vocabulario del que dispondrá el sistema posteriormente y las relaciones semánticas entre las palabras que lo forman. Aunque el proceso de adquisición diseñado para el sistema no guarde analogía alguna con el proceso llevado a cabo en el ser humano, el resultado del mismo, es decir, el mencionado vocabulario y sus relaciones sí representan una componente psicológicamente plausible como es la memoria lingüística a largo plazo, comentada en el capítulo anterior. En realidad, el método de adquisición definido sí que modela un aspecto presente en muchos procesos mentales: la influencia de la experiencia previa en el conocimiento adquirido actual. La cantidad y el orden en el que se presentan los textos al módulo de adquisición de SILC influyen notablemente en la cantidad y forma del conocimiento adquirido por el mismo, por lo que son parámetros a tener en cuenta.

4.3.1. Representación del conocimiento lingüístico semántico

El formalismo de representación elegido para describir el conocimiento semántico adquirido es claramente de carácter conexionista. Esta elección está condicionada recíprocamente por el método de adquisición y, principalmente, por la dinámica del modelo de lectura. Una representación conexionista consta de nodos que simbolizan entidades e interconexiones entre los mismos por donde fluye información de uno a otro, haciendo que dichos nodos cobren un determinado nivel de activación en función de la información entrante. Es una analogía con la fisiología del cerebro humano a nivel neuronal.

En el caso del sistema SILC, el formalismo anterior se instancia de manera similar a la representación propuesta por el sistema ICAN [Lemaire y Denhière, 2004], descrito en el capítulo 2. Así, el conocimiento lingüístico adquirido queda representado en SILC por una red conexionista con las siguientes características:

- Cada nodo representa un concepto y se etiqueta mediante su lexema raíz correspondiente.
- Las conexiones entre nodos indican relación semántica entre los mismos y son unilaterales, es decir, relacionan un concepto *C1* origen con otro destino *C2*. También se da siempre la unión contraria de *C2* a *C1*, pero se presenta como una conexión distinta e independiente de su inversa. Éste es también un aspecto psicológicamente plausible, puesto que para un ser humano la relación entre conceptos no es simétrica [Tversky, 1977], [Ortony et al., 1985]. Ésta es además una de las principales diferencias con la representación propuesta por el sistema ICAN.
- La ausencia de unión no denota la independencia semántica de los conceptos conectados mediante la misma, sólo el desconocimiento de la relación.
- Dicha relación semántica se define en términos de aparición simultánea en un mismo contexto de los conceptos relacionados.
- Las uniones poseen un valor asociado entre 0 y 1, que denota la significación o importancia de la relación.
- Los pesos de las uniones que salen de un mismo concepto suman siempre 1. Es decir, los pesos son relativos a los conceptos implicados en las uniones. Este aspecto supone también una diferencia con respecto al sistema ICAN.

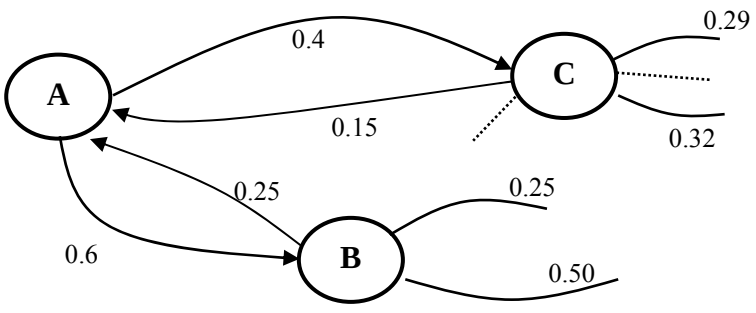


Figura 4.3. Representación del conocimiento lingüístico semántico en SILC.

De esta manera, el conocimiento lingüístico previo utilizado por el modelo de lectura se representa como una red de conceptos interconectados mediante relaciones semánticas contextuales cuya importancia se pondera mediante un valor numérico, como se presenta en la Figura 4.3, donde *A*, *B* y *C* representan conceptos que aparecen en los textos de entrenamiento ya preprocesados. La noción de significado de un concepto queda determinada por sus relaciones con los conceptos asociados, de manera similar a los sistemas HAL [Lemaire y Denhière, 2004] y CDSA [Ventura et al., 2004].

4.3.2. Extracción y construcción de la red conceptual de conocimiento semántico-lingüístico

Puesto que los textos que trata el sistema han sido preprocesados previamente, todos los términos que aparecen en los mismos se incorporan al conocimiento semántico. Sin embargo, también se han de extraer y almacenar las relaciones entre los mismos. Para ello es necesario determinar cómo están semánticamente relacionados dos conceptos y cómo extraer dicha información del lenguaje escrito sin ningún otro tipo de conocimiento previo.

Al igual que los sistemas LSA, HAL o ICAN, entre otros, el sistema SILC emplea la concurrencia léxica de dos palabras como indicador de relación semántica entre los conceptos que aparecen en el mismo contexto, basándose en la hipótesis de Miller y Charles [Miller y Charles, 1991]. De este modo, SILC procesa todos los textos de entrenamiento de manera secuencial, siguiendo el orden natural de las oraciones en cada uno e incorporando al vocabulario todos los términos que aparecen en los mismos así como las relaciones entre los conceptos que poseen un contexto común.

La mayor diferencia del sistema SILC con respecto a otros sistemas basados en concurrencia léxica es la definición de contexto. Para los sistemas previos existentes, el contexto es una ventana, es decir, una secuencia de palabras consecutivas de tamaño invariable. A continuación, dicha ventana se desplaza una palabra hacia el sentido natural de la lengua en cuestión (de izquierda a derecha en español, por ejemplo). Así, dada la frase “En un lugar de la Mancha…”, si se define una ventana de tamaño igual a 3, los términos “en”, “un” y “lugar” estarían relacionados semánticamente entre sí, y por tanto se incorporarían dichas relaciones al conocimiento. A continuación, la ventana se desplazaría una palabra hacia la derecha, con lo que se incluiría el término “de” en el

vocabulario y las relaciones entre "un", "lugar" y "de" en el conocimiento, y se procedería de la misma manera hasta el final del texto.

Como demuestran algunos experimentos realizados con el sistema LSA y HAL, esta forma de asociar los conceptos también recoge, de manera implícita e involuntaria, cierto conocimiento sintáctico. Sin embargo, plantea los problemas de la definición del tamaño óptimo de la ventana y de la captura de relaciones semánticas entre conceptos que aparecen en oraciones o párrafos diferentes, lo que podría introducir ruido en el conocimiento. Para el primer problema, los autores del sistema ICAN experimentaron con el tamaño de la ventana en problemas de clasificación de textos, determinando que 13 palabras es el tamaño óptimo de la ventana, es decir, que los mejores resultados en la aplicación a la clasificación de textos, utilizando diferentes métodos, se alcanzan procesando los textos de la manera descrita con una ventana contextual de 13 palabras. A pesar de los resultados, y puesto que sólo experimentaron con un conjunto de textos, no queda demostrado que dicho tamaño sea óptimo para cualquier tipo y cantidad.

Tratando de superar los problemas descritos, SILC utiliza un contexto de tamaño variable. Para SILC, el contexto es la oración, relacionando los conceptos que aparecen en las mismas. De esta forma, se busca expresamente la captura de información sintáctica y se evita relacionar conceptos que podrían no tener relación dada su posición en el discurso. Además, no es necesario determinar un tamaño óptimo dependiendo del tipo o la cantidad de textos puesto que el tamaño viene implícito en los mismos. Así, por ejemplo, en el fragmento de texto "El ayer es hoy. El pasado es mañana. Pero yo soy yo", se incorporarían al vocabulario todos los términos que aparecen y las relaciones entre los conceptos "el", "ayer", "es" y "hoy", entre los conceptos "el", "pasado", "es" y "mañana" y entre los conceptos "pero", "yo", "soy" y "yo" (nótese que ciertos términos del ejemplo no serían considerados después de aplicar el preprocesamiento al texto).

Como se ha descrito en la sección anterior, la representación del conocimiento semántico del sistema SILC no sólo recoge las relaciones entre conceptos sino que también las cuantifica. Dicha cuantificación se ve reflejada en los pesos de las asociaciones o aristas de la red. Análogamente, el sistema ICAN, entre otros, también cuantifica las relaciones. ICAN distingue tres tipos de asociaciones: directas, indirectas y ausencia de asociación, empleando una función distinta para calcular la significación o importancia de cada una de ellas. Dichas funciones refuerzan o debilitan las

asociaciones existentes en porcentajes dependientes de ciertos parámetros. Las expresiones de las mismas no son justificadas por los autores en modo alguno. Otros sistemas tienen también en cuenta la distancia a la que aparecen las palabras en el contexto para la cuantificación de las asociaciones. Así, la primera y última palabras dentro del contexto son las menos relacionadas por ser las más alejadas la una de la otra. Intuitivamente, este criterio no parece ser muy coherente puesto que relaciona de manera directa la distancia espacial (dentro del texto) con la distancia semántica. De hecho, los autores de los trabajos que emplean esta cuantificación no la justifican ni experimentalmente ni en términos de plausibilidad psicológica.

La ponderación del sistema SILC no emplea el concepto de distancia espacial dentro del contexto ni distingue expresamente entre tipos de asociaciones, aunque dicha distinción sí se tenga en cuenta en el modelo de lectura. La cuantificación está basada en la idea de que cuantas más veces aparecen en un mismo contexto dos conceptos más fuertemente relacionados están. Así pues, los pesos o significación de las asociaciones son la proporción del número de veces que los conceptos asociados concurren, con respecto al concepto origen de la asociación. Es decir, el peso de la asociación entre un concepto *A* y un concepto *B* es la proporción de apariciones de *A* y *B* en el mismo contexto con respecto al número total de apariciones de *A*, y viceversa para la asociación inversa entre *B* y *A*. De esta manera, todos los pesos de las asociaciones salientes de cada concepto suman 1. Además, las relaciones semánticas entre dos conceptos no son simétricas, es decir, la relación de *A* con *B* no tiene la misma significación que la relación de *B* con *A*, lo que también parece suceder en la mente humana según ciertos experimentos psicolingüísticos [Tversky, 1977], [Nosofsky, 1991].

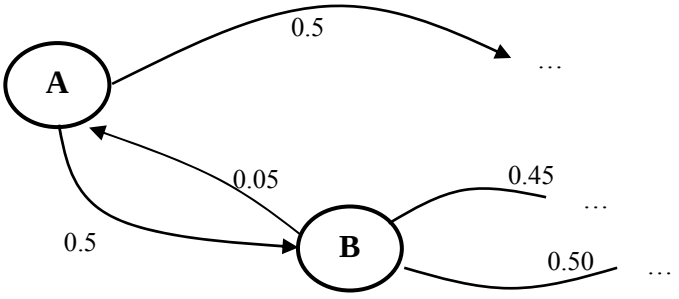


Figura 4.4. Ponderación de las asociaciones en SILC.

La Figura 4.4 muestra un ejemplo de la ponderación de las asociaciones. Si *A* aparece 10 veces en los textos de entrada y *B* aparece 100 veces, y los dos conceptos aparecen en la misma oración 5 veces, entonces el peso de la relación entre *A* y *B* es igual a 0.5 (50%), mientras que el peso de la relación entre *B* y *A* toma un valor de 0.05 (5%).

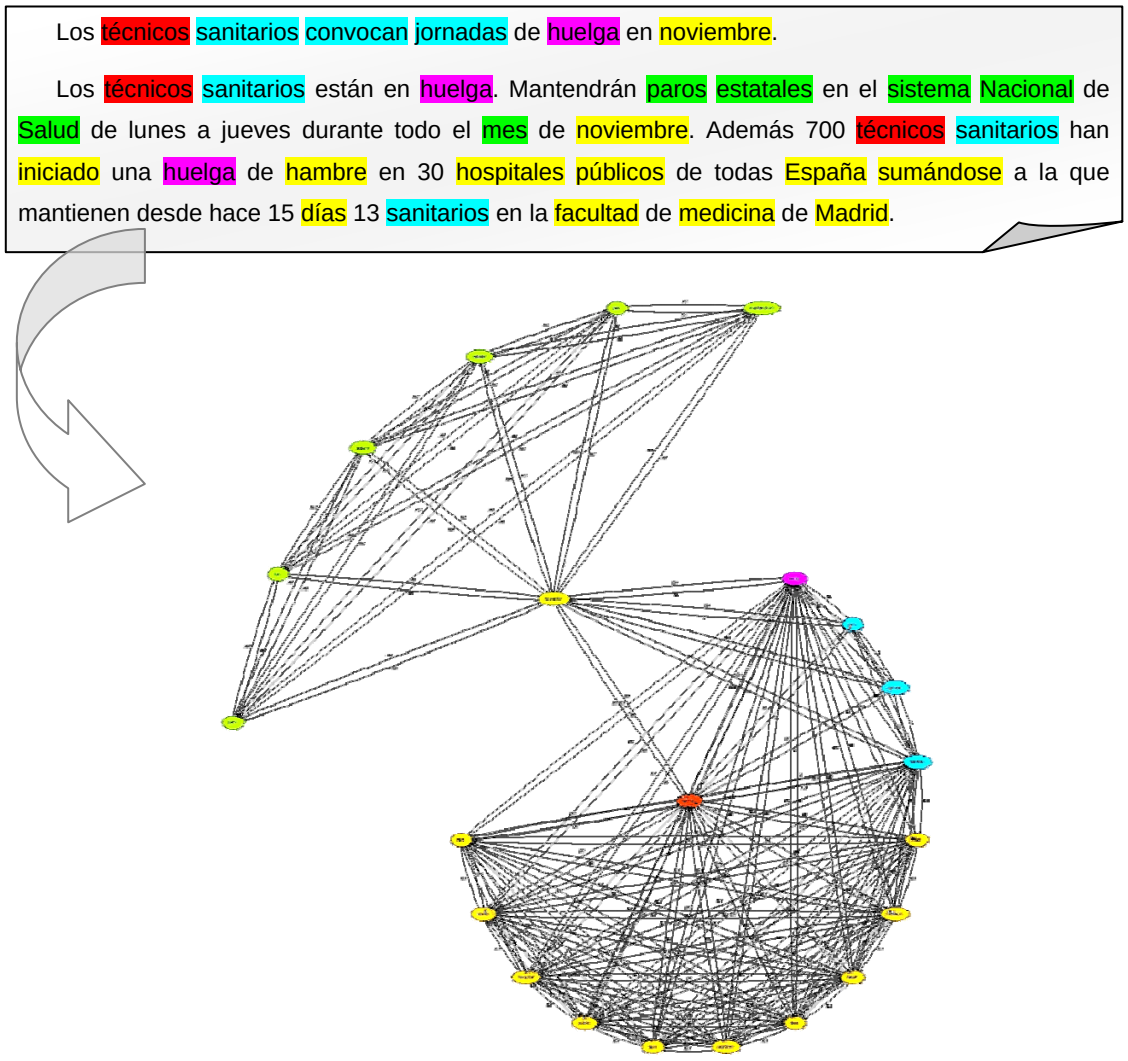


Figura 4.5. Representación del conocimiento adquirido a partir de un texto de ejemplo.

De esta forma, se crea un vocabulario y se establecen relaciones semánticas entre los componentes del mismo. Esta información se extrae de un conjunto de textos escritos que pretenden simbolizar la experiencia previa del sistema con el lenguaje. Así pues, debido al método de adquisición del conocimiento éste se ve afectado por los textos de

entrada, tanto por su cantidad como por su forma, de igual manera que la experiencia lectora de un ser humano influye en su conocimiento actual del lenguaje. El método de adquisición permite igualmente la actualización (nuevos pesos) y ampliación (nuevos conceptos y relaciones) eficiente del conocimiento actual a partir de nuevos fragmentos de lenguaje, una vez más como sucede en el cerebro humano. Todo el conocimiento adquirido se almacena de manera homogénea en una red neuronal de conceptos con conexiones ponderadas simulando, lo que se podría denominar, la memoria a largo plazo del sistema [Ericsson y Kintsch, 1995].

La Figura 4.5 muestra la representación del conocimiento resultante adquirido a partir de un texto de ejemplo. En dicha figura, el concepto central en color amarillo es el de "noviembre". Los conceptos del arco de la parte superior en color verde son los correspondientes a la tercera oración del texto. El concepto central de la circunferencia inferior en color rojo es el correspondiente a "técnicos". El concepto "huelga" se muestra en color rosa y los conceptos contiguos de color azul son los correspondientes a la primera y segunda oración del texto. El resto de conceptos en color amarillo son los que conforman la última oración del texto. El texto ha sido preprocesado previamente, por lo que las cifras y otros términos no aparecen en el conocimiento adquirido. Como se puede apreciar, cada asociación es bidireccional y está ponderada con un valor numérico.

4.4. Modelo Computacional de Lectura

El modelo computacional de lectura pretende obtener una representación mental de la semántica de un texto a partir de las palabras que lo componen y su organización dentro del mismo. Así pues, la transición de secuencia de palabras al modelo conceptual no es directa como en otros sistemas de representación, sino que obedece a un proceso secuencial en el tiempo donde las palabras son leídas en su orden natural y donde se tiene en cada instante un modelo de la semántica del fragmento leído hasta el momento. Las representaciones conceptuales parciales del texto se almacenan en la memoria de trabajo a largo plazo, según la teoría de Kintsch [Ericsson y Kintsch, 1995]. Dichas representaciones provisionales comparten el mismo formalismo que la representación final y son subredes del conocimiento semántico adquirido, con la peculiaridad de que los nodos conceptuales tienen un grado de activación.

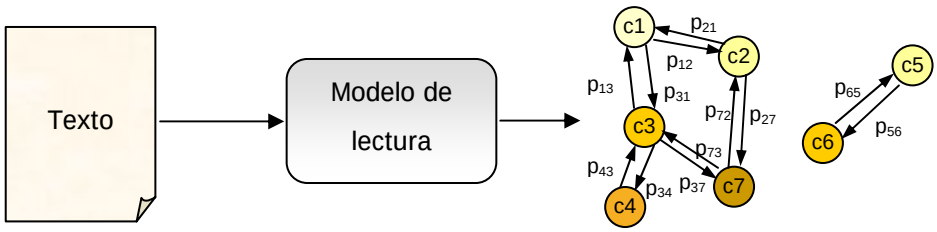


Figura 4.6. Representación conceptual resultante del proceso de lectura.

La Figura 4.6 muestra el modelo de lectura como caja negra y el formalismo de representación conceptual de los textos, representando el grado de activación de los conceptos mediante colores (más claro denota más activo).

La dinámica del modelo es la que sigue: se obtiene la primera palabra del texto y se busca en el conocimiento semántico. Si el concepto correspondiente ya era conocido, es decir, formaba parte del léxico adquirido, entonces se recupera y se almacena en la memoria de trabajo con un cierto grado de activación. En caso contrario, se ignora. Si dicho concepto ya estaba presente en la memoria de trabajo se incrementa la activación que posee en ese momento. A continuación se realizan inferencias en el conocimiento semántico a partir del concepto identificado. Los conceptos que se infieran a partir del

concepto leído serán también añadidos a la estructura de la memoria de trabajo con un nivel de activación producto del proceso de inferencia. Si los conceptos inferidos ya estaban en la memoria de trabajo se incrementa su nivel de activación. Se añadirán también a la estructura de la memoria de trabajo todas las asociaciones existentes en el conocimiento adquirido de la memoria a largo plazo entre los conceptos activos en la misma. Cada cierto intervalo de tiempo se disminuirá el nivel de la activación de todos los conceptos de la memoria de trabajo. Si la activación de alguno de los conceptos decae por debajo de un umbral, entonces el concepto se elimina de la estructura, así como sus asociaciones con los demás conceptos.

De esta forma, si un concepto no se trata a lo largo del texto, bien expresamente o bien a través de conceptos relacionados, se acaba olvidando. Por el contrario, si un concepto se trata de manera explícita o a través de conceptos relacionados a lo largo de todo el texto tendrá mucha significación en la representación semántica final. Se define pues un proceso en el que entran en juego la percepción, la recuperación de la memoria, las inferencias y el olvido a lo largo de la lectura ordenada de un texto en el tiempo. En la Figura 4.7 se muestra un esquema de la dinámica general de dicho proceso.

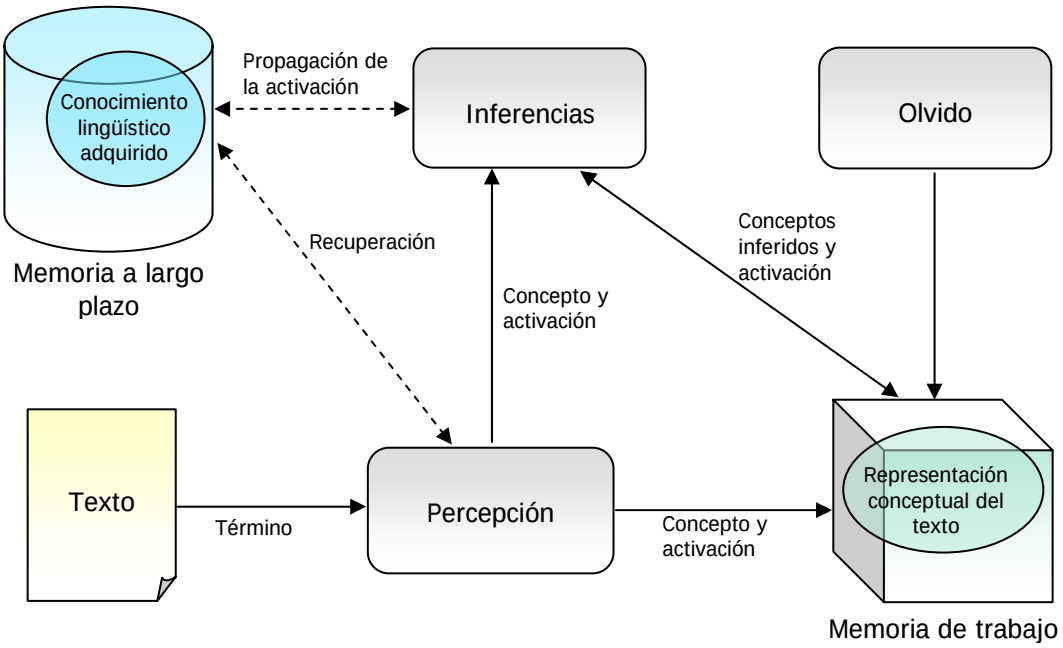


Figura 4.7. Esquema general del modelo computacional de lectura.

4.4.1. Percepción del texto

En la etapa de percepción se procesan en orden las palabras del texto de entrada, comprobando si ya se conocen, es decir, si existen en el conocimiento semántico previo, y llevándolas a la memoria de trabajo con un determinado nivel de activación. Aunque en el sistema SILC no se ha implementado ningún mecanismo de ponderación de la percepción, otorgando siempre un nivel de activación inicial igual a 1, su diseño posibilita de forma inmediata la incorporación de dicho mecanismo. Así, se puede incrementar la plausibilidad psicológica del modelo cuantificando la significación de los términos leídos según su formato, posición en el texto o incluso según expectativas o intereses predefinidos.

4.4.2. Inferencia

A continuación se produce un proceso de inferencia a partir del concepto percibido. Dicho proceso de inferencia se simula mediante la propagación de la activación que el concepto tiene en la memoria de trabajo hacia los conceptos que tiene asociados en el conocimiento lingüístico semántico de la memoria a largo plazo. Así, un concepto muy activo hace que los conceptos relacionados con él también lo estén. Durante el proceso de propagación, todo concepto afectado por la misma se recupera y almacena en la memoria de trabajo, así como las asociaciones existentes del concepto recién incorporado con todos los elementos de la misma. Si un concepto al que se propaga activación ya está en la memoria de trabajo, entonces su activación se incrementa en una cantidad igual a la activación que le ha sido transferida. Para que la ponderación de las asociaciones del conocimiento semántico lingüístico tenga influencia en las inferencias, la propagación de la activación se realiza multiplicando la activación del concepto origen por el peso de la asociación por donde se propaga. De este modo, se infieren con más significación los conceptos más asociados al concepto origen. Con el propósito de considerar las asociaciones indirectas o de orden superior entre conceptos, la propagación se realiza de manera recursiva. De esta manera, una vez sumada la activación entrante a la que un concepto posee (recuérdese que pudiera no estar activado) éste la propaga de igual forma a sus conceptos asociados. Así, por ejemplo, dado el conocimiento de la Figura 4.5 en la memoria a largo plazo, si SILC encuentra en

un texto el término "noviembre" lo incorpora a la memoria de trabajo con un nivel de activación igual a 1. A continuación, este concepto propaga su activación a los conceptos asociados como "huelga" o "técnico" y se incorporan estos dos conceptos a la memoria de trabajo con una activación igual a 1 multiplicado por el correspondiente peso de las asociaciones. Se incorporan también todas las asociaciones entre estos tres conceptos. Los conceptos así inferidos propagan su activación actual a sus conceptos asociados repitiendo el mismo proceso. Cuando este proceso de inferencia termina, el sistema lee la siguiente palabra del texto y procede análogamente hasta el final del mismo.

Como se ha comentado en capítulos anteriores, existe una influencia mutua entre el formalismo de representación y los métodos que lo procesan. En este caso, la existencia de ciclos en la red de conocimiento lingüístico semántico y la naturaleza recursiva del proceso de inferencia hacen necesario el planteamiento de determinadas consideraciones. La primera de ellas es evitar que la propagación de la activación caiga en un bucle infinito realimentado entre conceptos asociados. Para ello, no se permite la propagación de la activación a un concepto al que ya le haya sido transferida previamente mediante otra asociación. De esta restricción se deriva la noción de orden en la propagación y puesto que la implementación del proceso es secuencial, la segunda consideración tiene que ver con la definición de este orden, es decir, a qué conceptos asociados se transfiere en primer lugar. El sistema SILC plantea dos opciones de implementación análogas a los algoritmos de recorrido de árboles [Wood, 1993]. La primera de ellas consiste en una propagación por niveles, es decir, transferir primero a todos los conceptos asociados directamente con el concepto origen en orden de ponderación de las asociaciones. La segunda es una propagación en profundidad, consistente en transferir la activación al concepto más asociado y que éste la transfiera a su vez a su concepto más asociado de manera recursiva. La Figura 4.8 muestra un ejemplo esquemático de las dos opciones citadas (Figura 4.8a y Figura 4.8b respectivamente). Ambas tienen una correspondencia psicológicamente plausible con las teorías de activación del significado relativas al modelo de acceso múltiple y al modelo de acceso múltiple selectivo, respectivamente, comentadas en el capítulo anterior.

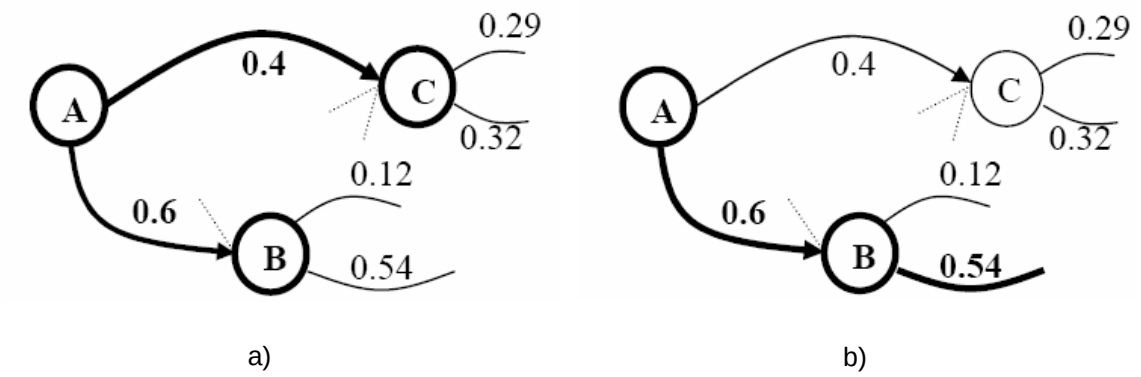


Figura 4.8. Tipos de inferencia en SILC: a) por niveles y b) en profundidad.

A pesar de que la activación propagada va disminuyendo a medida que recorre diferentes niveles de conceptos, puesto que los pesos de las asociaciones son números positivos menores que 1, es necesario establecer un límite o criterio de parada que indique cuándo un concepto deja de propagar la activación que le es transferida. Este criterio se puede definir en términos de niveles de propagación o en términos de umbrales de activación. En el primer caso, se indica el número máximo de conceptos que puede atravesar la activación propagada desde el nodo origen, controlando de esta manera el orden o grado máximo de asociación que se desee incluir. En el segundo caso, se indica el valor de activación mínimo que puede ser propagado, simulando en cierta forma la actividad eléctrica cerebral. Así, si la activación no se propaga al estar por debajo del umbral y, por lo tanto, no se infiere el concepto asociado, se debe a que éste está semánticamente lejos del concepto origen o a que su relación con él es muy débil con respecto a otros conceptos. Estos dos parámetros sirven, además, para caracterizar la habilidad lectora de individuos humanos. En la Figura 4.9 se muestra un ejemplo de la influencia del primer parámetro, el nivel máximo de propagación o asociación. En dicha figura se pueden apreciar los conceptos recuperados en la memoria de trabajo y su activación (representada por la intensidad del color) después de inferir a partir del concepto "lunes" para valores del parámetro nivel de propagación desde 1 hasta 3, que permiten sólo asociaciones directas con el concepto (Figura 4.9a), asociaciones de hasta de segundo orden (Figura 4.9b) y asociaciones de hasta de tercer orden (Figura 4.9c) en el proceso de propagación, respectivamente. El conocimiento lingüístico semántico adquirido es el de la Figura 4.5. El concepto "lunes" es uno de los nodos de color verde en dicha figura.

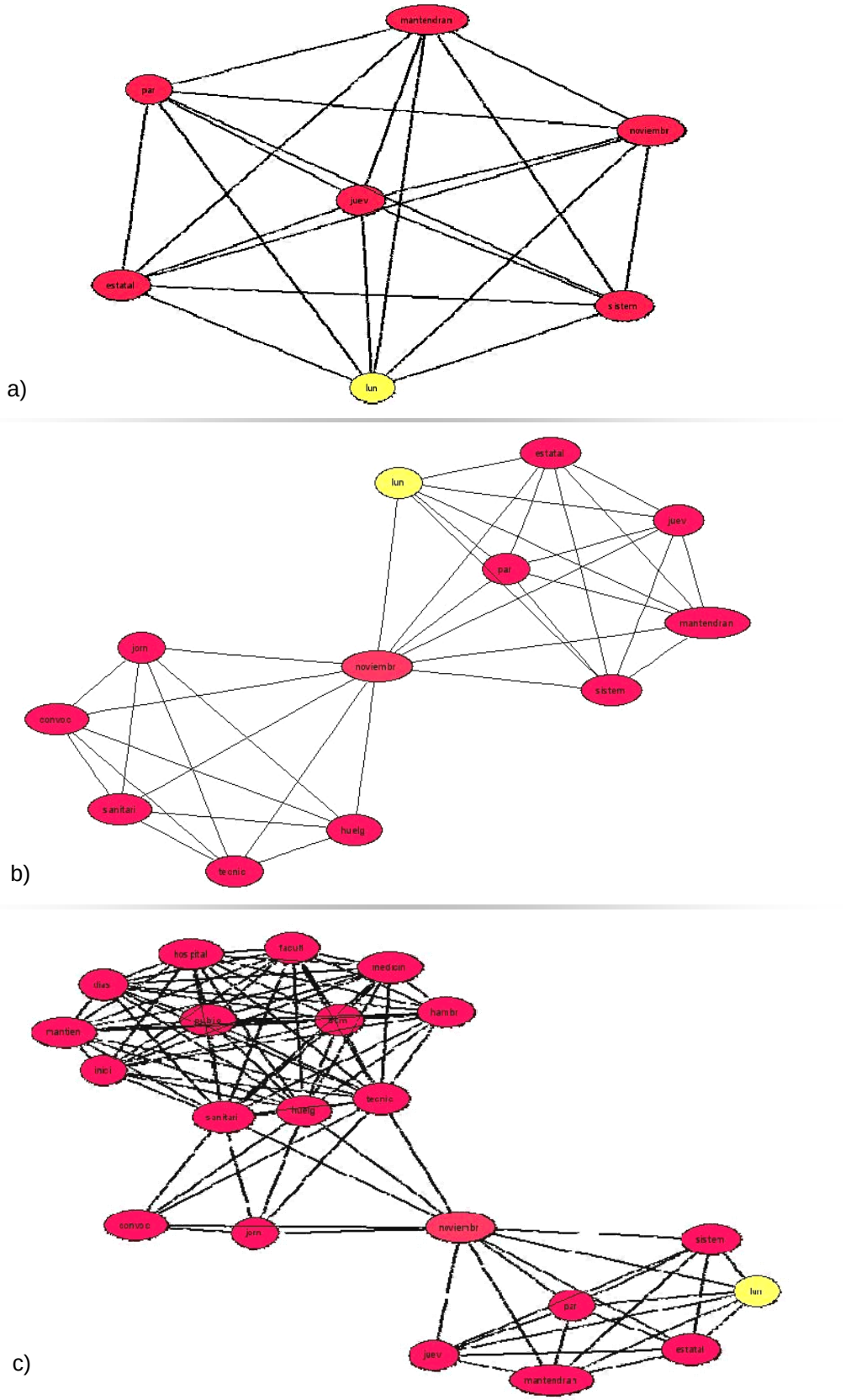


Figura 4.9. Conceptos inferidos utilizando niveles de propagación máximos de a) 1, b) 2 y c) 3.

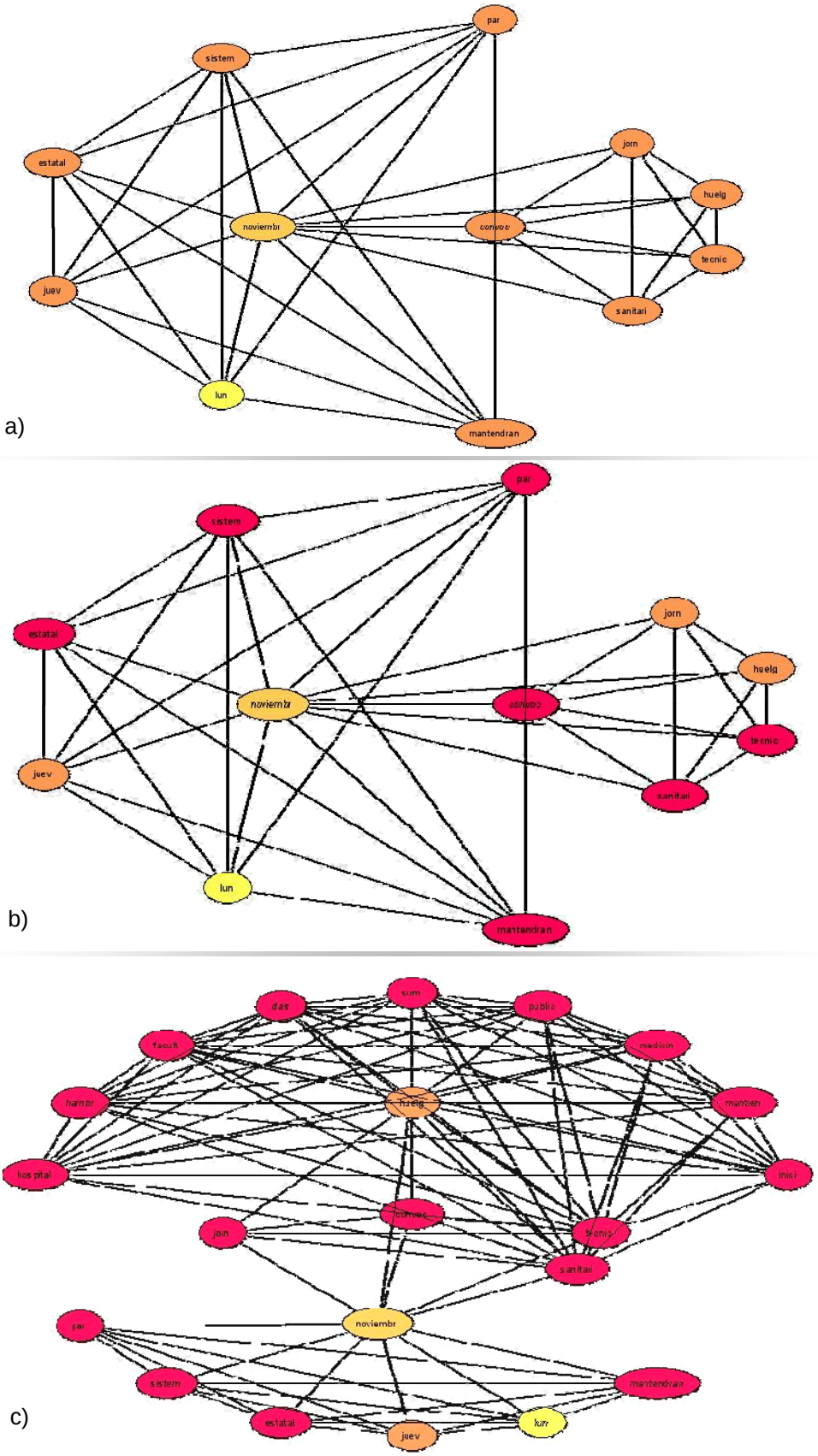


Figura 4.10. Conceptos inferidos a partir de otro concepto leído utilizando umbrales de propagación mínimos de a) 0.01, b) 0.001 y c) 0.0001.

La Figura 4.10 muestra, análogamente, los conceptos inferidos a partir del concepto "lunes" para distintos valores del segundo de los parámetros mencionados, el valor mínimo de activación, por debajo del cual la activación ya no es propagada. El conocimiento semántico del que se dispone es el mismo descrito anteriormente y el nivel de propagación no está limitado. Los valores del umbral son 0.01 (Figura 4.10a), 0.001 (Figura 4.10b) y 0.0001 (Figura 4.10c). Como se puede apreciar, la variación de la activación mínima no sólo afecta al número de conceptos inferidos sino también al valor de activación con el que se infieren.

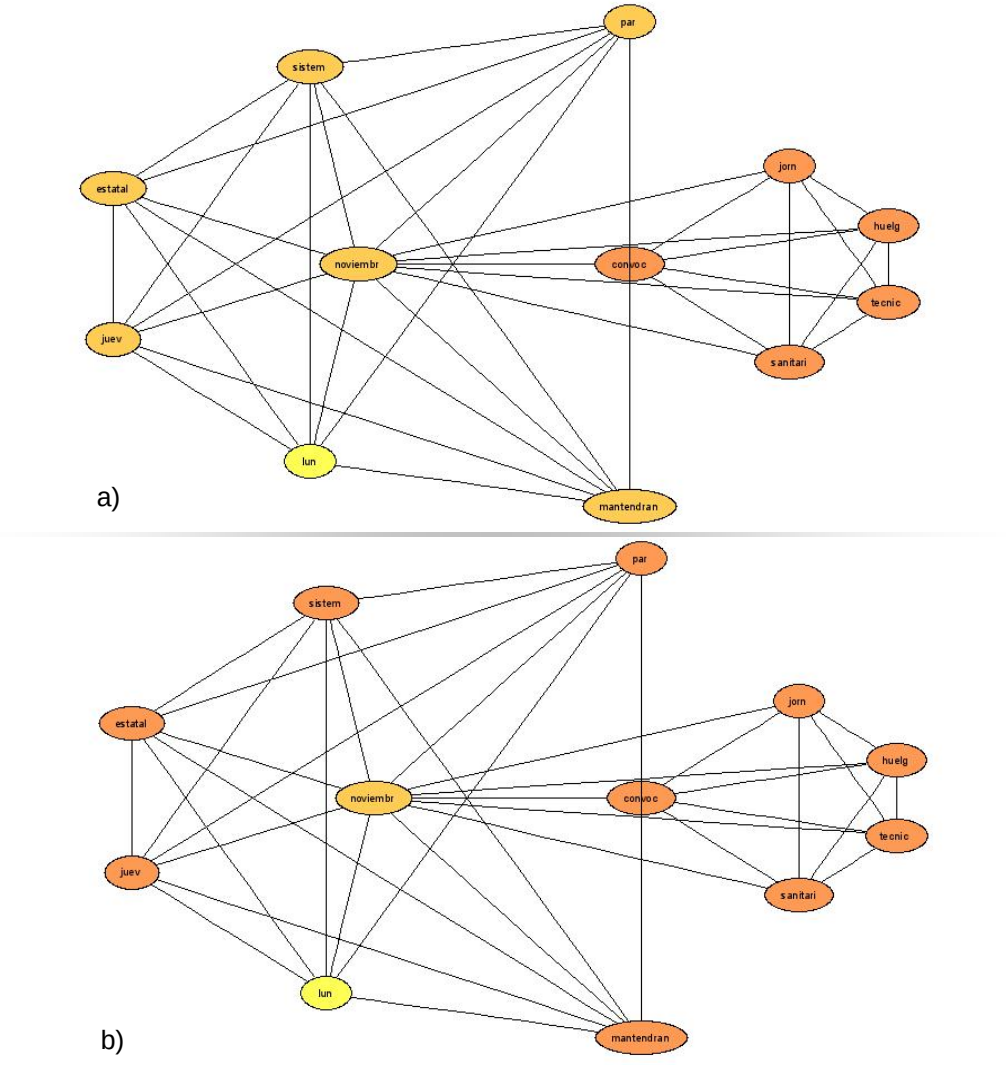


Figura 4.11. Conceptos inferidos a partir de otro concepto leído utilizando propagación a) por niveles y b) en profundidad.

Esta diferencia en el valor de activación inferido se da también entre los dos tipos de propagación descritos anteriormente, como se aprecia en la Figura 4.11, donde se

muestran las inferencias al leer el concepto "lunes", como en los ejemplos anteriores, empleando la propagación por niveles (Figura 4.11a) y la propagación en profundidad (Figura 4.11b).

El proceso de inferencia mediante la propagación de la activación hace que la representación final de los textos pueda contener conceptos que no aparecen explícitamente en los mismos, es decir, que no estaban representados por una palabra en los textos leídos. Esta es una de las principales diferencias con el resto de sistemas de representación de textos existentes. Además, en el sistema SILC la representación de la semántica de un texto es diferente a la de la semántica de las palabras y sus conceptos asociados. Esta distinción también se da en los seres humanos [Lemaire y Denhière, 2004]. Sólo los sistemas ICAN y CDSA, además de SILC, proponen esta jerarquía de representación.

4.4.3. Olvido

La mente humana no es capaz de recordar cada palabra de un texto leído a no ser, por supuesto, que se realice un esfuerzo premeditado para memorizarlo debido a sus limitaciones de memoria. Así pues, el sistema SILC también modela este aspecto persiguiendo una vez más la plausibilidad psicológica y también la eficiencia, ya que el proceso de inferencia descrito aumenta significativamente el tamaño de la representación del texto en la memoria de trabajo a cada palabra leída. Para evitar este crecimiento, se define el factor de olvido. El olvido hace que todos los conceptos en la memoria de trabajo pierdan un porcentaje de su activación cada cierto periodo de tiempo, denominado factor de olvido, de tal forma que los conceptos cuya activación decaiga por debajo de un umbral son eliminados de la memoria de trabajo y, por tanto, olvidados. El proceso de olvido hace que los conceptos no tratados directa o indirectamente durante un fragmento relativamente extenso del texto se olviden y pierdan importancia en la representación semántica final del mismo, ya que su activación sólo disminuye y no se incrementa durante dicho fragmento.

Puesto que el sistema no mide el tiempo real, es necesario establecer los términos en los que se define el intervalo a cuyo paso se produce una disminución de la activación de los conceptos en la memoria de trabajo. Dado que el modelo procesa las palabras en orden secuencial, éstas pueden ser consideradas como unidades de tiempo. Así, el

intervalo de olvido queda definido por un número de palabras consecutivas. Dicho número puede ser fijo o variable. En el caso de un valor fijo *k*, la representación actual pierde un porcentaje de su activación cada *k* palabras. Este valor puede ser considerado como el tamaño de la memoria a corto plazo [Miller, 1956]. El intervalo variable viene dado una vez más por el discurso del texto en sí, y más concretamente por las oraciones del mismo, al igual que en el contexto para la adquisición del conocimiento previo. Es decir, los conceptos de la representación actual en la memoria pierden activación a cada oración leída. La definición de este intervalo variable es compatible con las evidencias sobre el procesamiento de oraciones completas por parte de los seres humanos y su influencia sobre el modelo de situación del texto (representación semántica) en cada momento [Goldman et al., 1980].

Otras cuestiones del diseño de este aspecto son la determinación del factor de olvido y la del umbral que determina cuándo un concepto se olvida, es decir, cuándo se elimina de la representación en memoria por carecer de importancia con respecto a la semántica global del texto. Para mantener la coherencia del formalismo de representación conexionista, el umbral de olvido se establece en el mismo valor que el umbral de mínimo de propagación, ya que ambos cuantifican la misma magnitud, una activación mínima requerida.

4.4.4. Dinámica del modelo de lectura

Para ilustrar el resultado de la combinación de todos los procesos que componen el modelo de lectura se presenta a continuación un ejemplo detallado del mismo. El conocimiento semántico lingüístico utilizado fue adquirido a partir de 300 textos obtenidos del servicio de noticias de Google[1] más 500 textos de cultura general recopilados por el escritor Jorge Orellana[2], obteniendo una red de más de 12,000 conceptos empleando la oración como ámbito de contexto. Las inferencias fueron realizadas mediante la propagación en profundidad, con un nivel máximo de propagación de 3 y un umbral mínimo de propagación de 0.05. El intervalo de olvido viene determinado por las oraciones y el umbral de olvido es el mismo que el umbral

[1] http://news.google.es

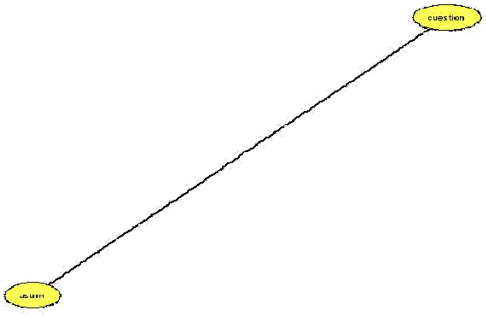
[2] http://j.orellana.free.fr/

mínimo de propagación, 0.05. El factor de olvido tiene un valor de 10, es decir, los conceptos pierden un 10% de su activación cada vez que se aplica el proceso de olvido. Con el modelo así definido, se muestra a continuación el estado de la memoria de trabajo, es decir, la representación semántica del texto junto con la activación de los conceptos, después de leer cada palabra no eliminada en la etapa de preprocesamiento del siguiente fragmento de texto: *"No es cuestión de asumir más riesgos. Nadal se retira de la competición de tenis. Su médico le aconsejó que no jugase."*

"No es **cuestión** *de asumir más riesgos. Nadal se retira de la competición de tenis. Su médico le aconsejó que no jugase."*

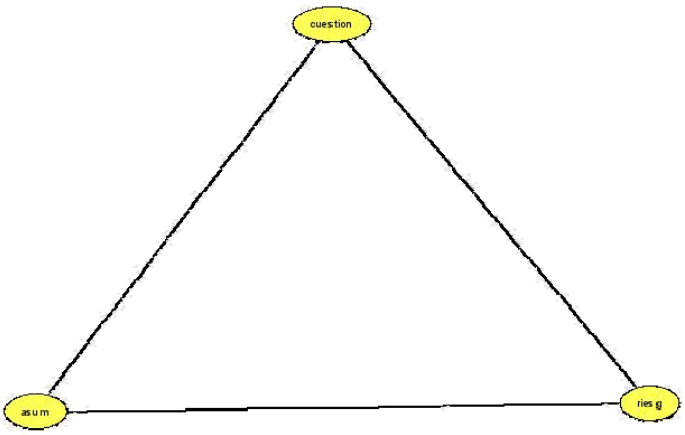
cuestion=1

"No es cuestión de **asumir** *más riesgos. Nadal se retira de la competición de tenis. Su médico le aconsejó que no jugase."*



asum=1
cuestion=1

"No es cuestión de asumir más **riesgos**. *Nadal se retira de la competición de tenis. Su médico le aconsejó que no jugase."*

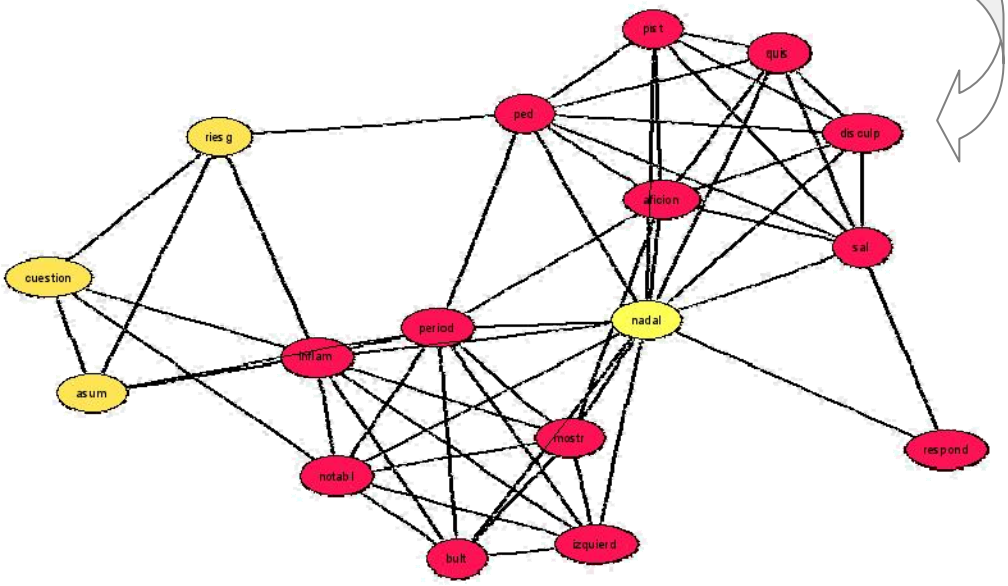


asum=1
cuestion=1
riesg=1

Olvido

asum=0.9
cuestion=0.9
riesg=0.9

"No es cuestión de asumir más riesgos. **Nadal** *se retira de la competición de tenis. Su médico le aconsejó que no jugase."*



nadal=1
respond=0.077
notabl=0.077
bult=0.077

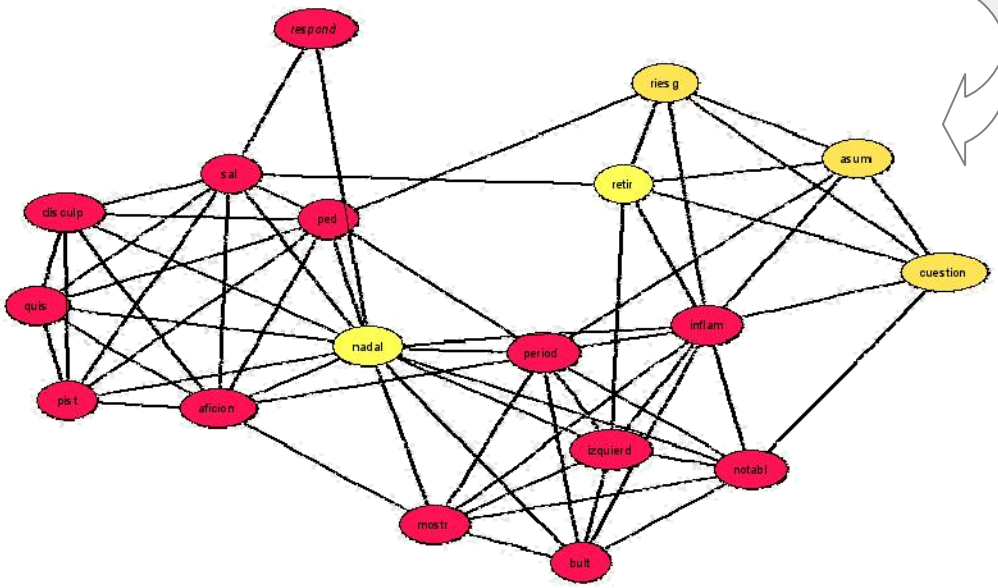
riesg=0.9
quis=0.077
mostr=0.077
aficion=0.077

cuestion=0.9
pist=0.077
izquierd=0.077

asum=0.9
period=0.077
inflam=0.077

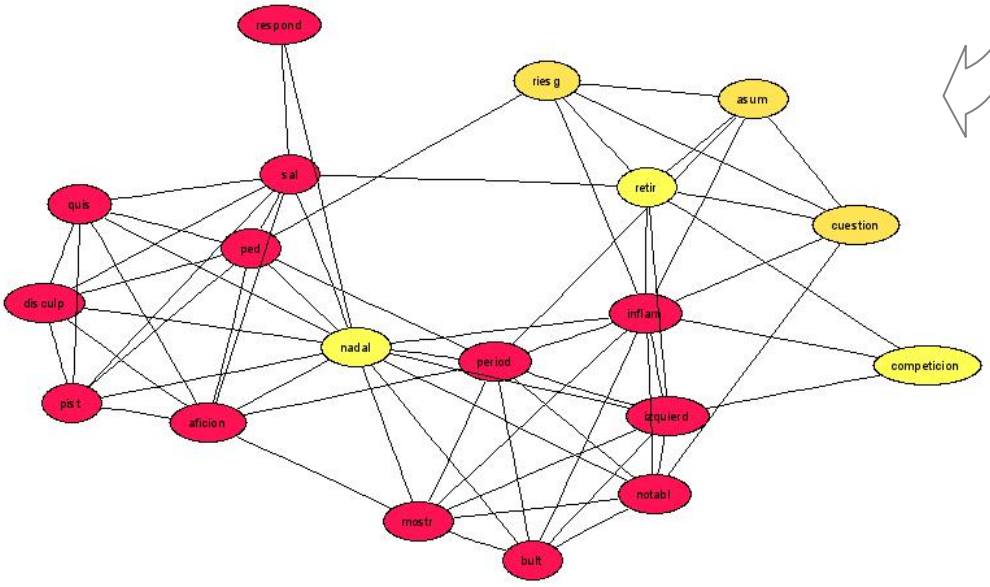
sal=0.077
ped=0.077
disculp=0.077

"No es cuestión de asumir más riesgos. Nadal se **retira** *de la competición de tenis. Su médico le aconsejó que no jugase."*



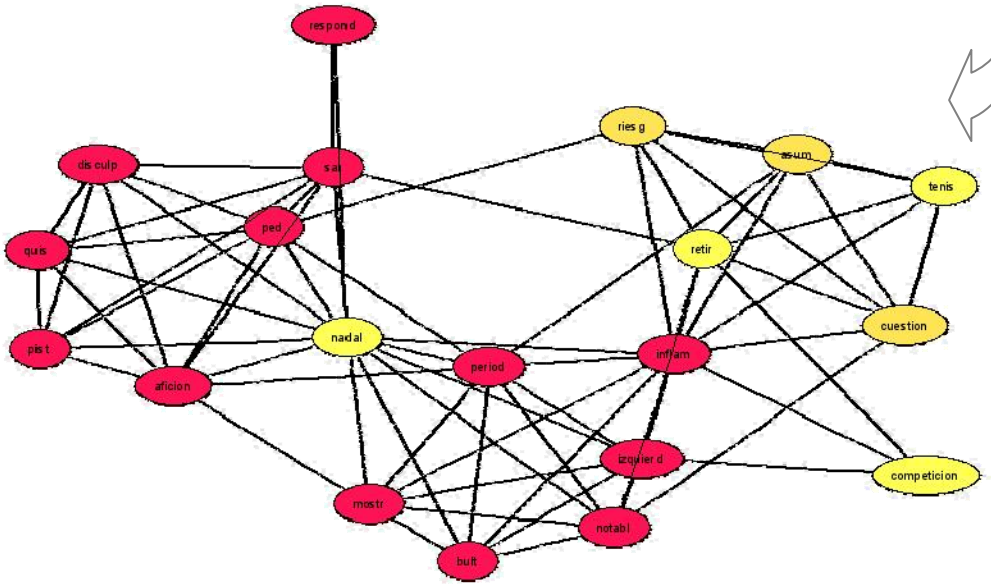
retir=1 nadal=1 riesg=0.9 cuestion=0.9 asum=0.9
sal=0.077 respond=0.077 quis=0.077 pist=0.077 period=0.077
ped=0.077 notabl=0.077 mostr=0.077 izquierd=0.077 inflam=0.077
disculp=0.077 bult=0.077 aficion=0.077

"No es cuestión de asumir más riesgos. Nadal se retira de la **competición** *de tenis. Su médico le aconsejó que no jugase."*



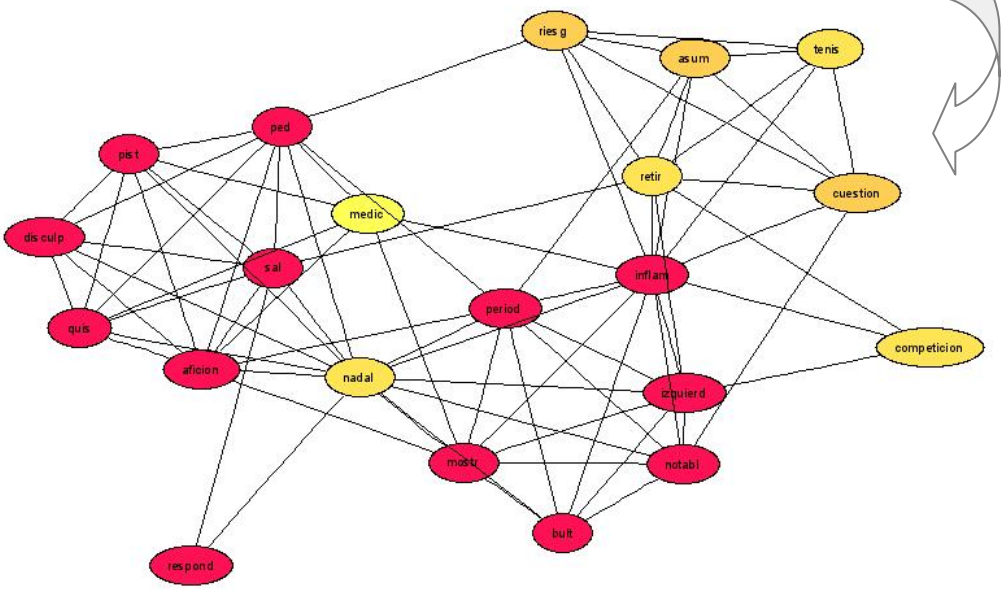
competicion=1 retir=1 nadal=1 riesg=0.9 cuestion=0.9
asum=0.9 sal=0.077 respond=0.07 quis=0.077 pist=0.077
period=0.077 ped=0.077 notabl=0.077 mostr=0.077 izquierd=0.077
inflam=0.077 disculp=0.077 bult=0.077 aficion=0.077

"No es cuestión de asumir más riesgos. Nadal se retira de la competición de **tenis**. *Su médico le aconsejó que no jugase."*



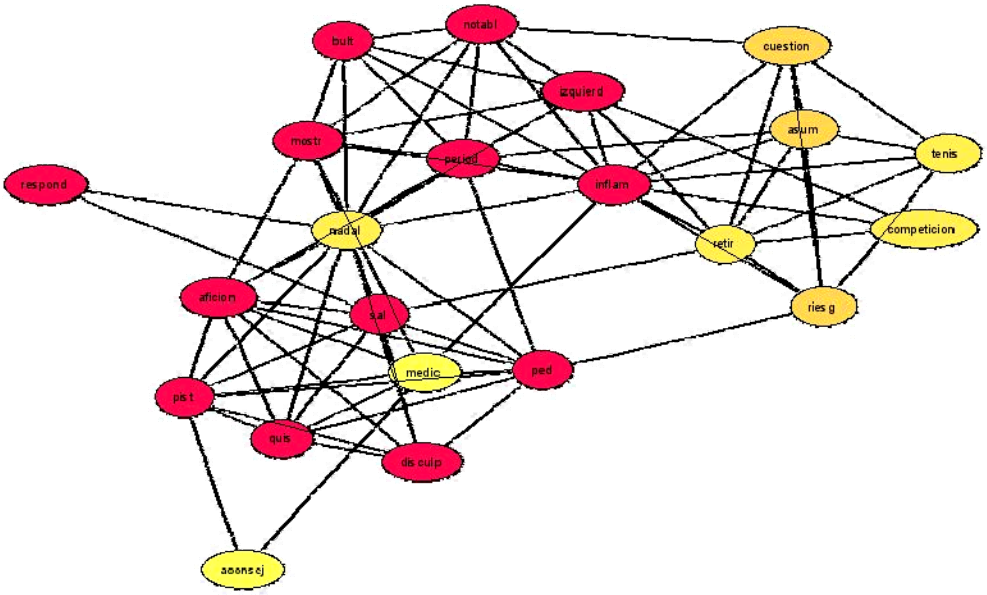
tenis=0.9	competicion=0.9	retir=0.9	nadal=0.9	riesg=0.81
cuestion=0.81	asum=0.81	sal=0.070	respond=0.070	quis=0.070
pist=0.070	period=0.070	ped=0.070	notabl=0.070	mostr=0.070
izquierd=0.070	inflam=0.070	disculp=0.070	bult=0.070	aficion=0.070

"No es cuestión de asumir más riesgos. Nadal se retira de la competición de tenis. Su **médico** *le aconsejó que no jugase."*



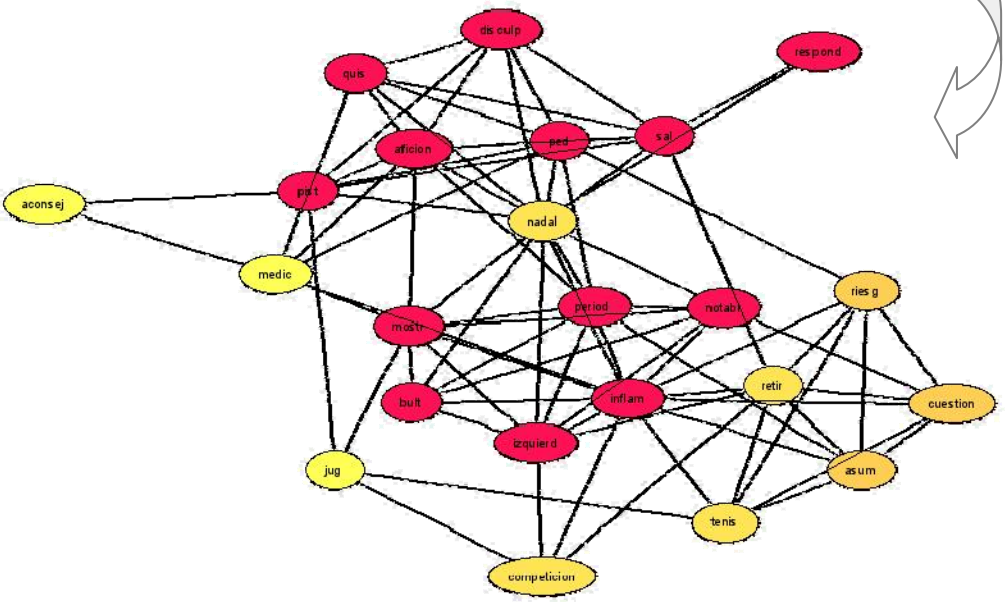
medic=1	tenis=0.9	competicion=0.9	retir=0.9	nadal=0.9
riesg=0.81	cuestion=0.81	asum=0.81	sal=0.070	respond=0.070
quis=0.070	pist=0.070	period=0.070	ped=0.070	notabl=0.070
mostr=0.070	izquierd=0.070	inflam=0.070	disculp=0.070	bult=0.070
aficion=0.070				

"No es cuestión de asumir más riesgos. Nadal se retira de la competición de tenis. Su médico le **<u>aconsejó</u>** *que no jugase."*

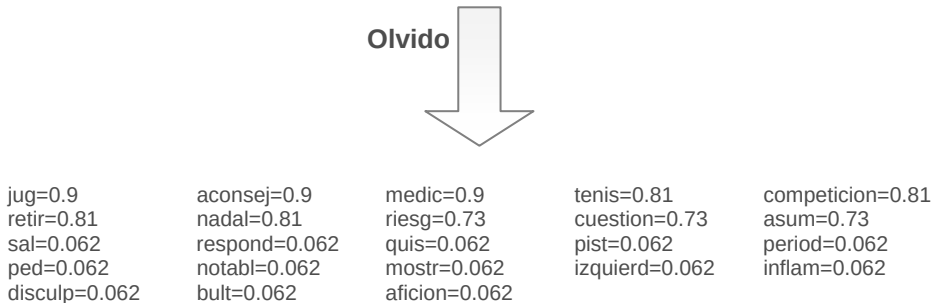


aconsej=1	medic=1	tenis=0.9	competicion=0.9	retir=0.9
nadal=0.9	riesg=0.81	cuestion=0.81	asum=0.81	sal=0.070
respond=0.070	quis=0.070	pist=0.070	period=0.070	ped=0.070
notabl=0.070	mostr=0.070	izquierd=0.070	inflam=0.070	disculp=0.070
bult=0.070	aficion=0.070			

"No es cuestión de asumir más riesgos. Nadal se retira de la competición de tenis. Su médico le aconsejó *que no* **<u>jugase</u>**."



jug=1	aconsej=1	medic=1	tenis=0.9	competicion=0.9
retir=0.9	nadal=0.9	riesg=0.81	cuestion=0.81	asum=0.81
sal=0.070	respond=0.070	quis=0.070	pist=0.070	period=0.070
ped=0.070	notabl=0.070	mostr=0.070	izquierd=0.070	inflam=0.070
disculp=0.070	bult=0.070	aficion=0.070		



La última red mostrada junto con los últimos valores de activación después del olvido conforma la representación semántica final del texto.

Como se desprende del diseño de los procesos del modelo, la forma y orden del discurso en el que están escritos los textos influye de manera directa en su representación semántica final, al igual que en los seres humanos [Meyer y Poon, 2001]. Esta es otra de las grandes diferencias con los sistemas de representación existentes, en los que sólo se considera la aparición de las palabras que componen los textos sin tener en cuenta el orden en el que lo hacen y cómo se distribuyen en el discurso.

4.4.5. Caracterización del modelo computacional de lectura en SILC

El modelo implementa pues un modelo híbrido "Bottom-Up-Down" (Abajo-Arriba-Abajo), donde las palabras se perciben e inducen una semántica del texto y esta semántica, a su vez, es la que otorga el nivel de significación a las siguientes palabras percibidas. Los niveles lingüísticos que contempla son el semántico y el sintáctico, interviniendo éste último sólo en la interpretación unitaria de las oraciones de los textos y en la asociación de conceptos. La representación que se genera es la de texto base y se almacena en un modelo de memoria de trabajo a largo plazo, como el propuesto por Ericsson y Kintsch [Ericsson y Kinstch, 1995], que contiene copias de los conceptos en la memoria a largo plazo con un nivel de activación que permiten la recuperación rápida e inferencia de conceptos asociados en la memoria a largo plazo. El mecanismo de inferencias se ajusta a la hipótesis Construccionista [Graesser et al., 1994], generando todos los tipos de inferencias contemplados al unísono de manera automática, y siendo el propio contexto del discurso el que las confirma o descarta posteriormente durante la lectura. Los tipos de inferencia implementados son las inferencias para mantener la coherencia semántica y las inferencias predictivas. Concretamente, para la

identificación del significado de las palabras el modelo sigue un enfoque basado en el modelo de Acceso Múltiple Ordenado [Neil et al., 1988], cobrando una mayor activación aquellos conceptos inferidos que tienen un nivel de asociación mayor con la palabra leída y que ya están activos debido al contexto definido por el texto leído hasta el momento.

En términos computacionales, el modelo de lectura descrito cumple los requisitos básicos, propuestos por Ram y Moorman [Ram y Moorman, 1999] en el capítulo anterior, para ser considerado como tal:

1) el modelo ha sido expresado en términos funcionales, es decir, de entrada-salida,
2) el modelo ha sido expresado en términos computacionales, ya que se han propuesto estructuras de datos y algoritmos para llevar a cabo en un ordenador las funciones especificadas,
3) el modelo ha sido expresado en términos estructurales, puesto que se han diseñado formalismos de representación y mecanismos para tratarlos,
4) y además, todo el diseño se ha guiado por la plausibilidad psicológica, justificando las funciones, procesos y representaciones, inspiradas en la estructura y funcionamiento de la mente y el cerebro humano, mediante evidencias experimentales.

Además, el modelo es capaz de tratar con términos novedosos o desconocidos, puesto que les asigna la semántica del concepto más fuertemente inferido en los momentos en los que se leen dichos términos. El modelo aporta también medidas para su evaluación "on-line", como los niveles medios de activación, el número de conceptos activos, el número de conceptos nuevos que se incorporan mediante las inferencias o el tiempo de propagación de la activación. En cuanto a medidas "off-line", la principal aportación es la propia representación estructural final generada por el modelo, ya que permite una comparación elaborada con resúmenes o síntesis realizadas por los seres humanos. El modelo también permite incorporar fácilmente la influencia de la percepción en la lectura mediante la variación, en función de los criterios deseados, del nivel de activación inicial que se le otorga a los conceptos leídos. Los intereses semánticos del lector son también fácilmente incorporables mediante el aumento de la activación inicial de los conceptos "interesantes" o partiendo de una representación del texto compuesta por oraciones representativas de dichos intereses. Para finalizar, el

modelo también contempla el nivel de atención o dedicación al proceso de lectura mediante la variación de parámetros como los umbrales de propagación y el factor e intervalo de olvido.

4.4.6. Parámetros del sistema

Recopilando todos los procesos que lleva a cabo el sistema tanto en la etapa de adquisición del conocimiento como en la de representación de textos, SILC se puede caracterizar mediante los siguientes parámetros:

- *Definición del contexto:* Determina si el contexto utilizado para relacionar los conceptos en la etapa de adquisición del conocimiento es una ventana deslizante de tamaño fijo o si viene dado por las oraciones de los textos.
- *Tamaño de ventana de contexto:* Número de palabras que forman la ventana en el caso de que el contexto se defina como tal.
- *Cuantificación de la percepción:* Determina el valor por defecto en el que se incrementa la activación de un concepto correspondiente a una palabra leída.
- *Tipo de inferencia:* Indica si la propagación de la activación durante el proceso de inferencia se realiza por niveles o en profundidad.
- *Nivel máximo de propagación:* Determina el número máximo de conceptos que puede atravesar la activación propagada desde el concepto origen.
- *Umbral mínimo de propagación:* Indica el nivel de activación mínimo que puede ser propagado a otros conceptos. Si la activación resultante es menor que el umbral no se propaga.
- *Intervalo de olvido:* Define el intervalo, como número de palabras consecutivas, a cuyo paso se disminuye la activación de los conceptos actuales en la memoria de trabajo. El intervalo puede ser de tamaño fijo o puede venir dado por las oraciones de los textos.
- *Factor de olvido:* Porcentaje de activación que pierden los conceptos a cada paso del intervalo de olvido.
- *Umbral de olvido:* Nivel de activación por debajo del cual los conceptos se eliminan de la representación actual en memoria del texto.

Todos estos parámetros han sido objeto de estudio en los experimentos, tanto en su optimización para la indexación de textos en aplicaciones prácticas como en su adaptación al modelado de la habilidad lectora de los seres humanos.

4.5. Caracterización del Sistema SILC

De acuerdo a las características generales de los sistemas de representación masiva de conocimiento lingüístico descritas en el capítulo dos, algunas de ellas propuestas por Lemaire [Lemaire y Denhière, 2004], el sistema SILC se define de la siguiente manera:

- Tipo de entrada: Colecciones de textos en lenguaje natural libre.
- Formalismo de representación del conocimiento: Red de asociación conceptual con arcos ponderados.
- Dinamismo Temporal: Permite su actualización directa y eficiente a lo largo del tiempo, sin necesidad de la reconstrucción del conocimiento existente.
- Definición del contexto: Contempla la venta deslizante de tamaño fijo o la oración como ámbito de contexto.
- Formalismo de representación de la semántica de los conceptos: Relaciones ponderadas con otros conceptos asociados.
- Formalismo de representación de la semántica de los textos: Subredes del conocimiento semántico general con niveles de activación para cada concepto. Esta representación en red elimina todos los inconvenientes que plantea la representación clásica como vectores.
- Concurrencias de orden superior: permite concurrencias del orden que se desee.
- "Composicionalidad": La representación de los textos es el resultado de un proceso de lectura secuencial de las palabras que lo componen. Durante dicho proceso se llevan a cabo inferencias y cambios en las estructuras, disponiendo de una representación del texto leído hasta el momento en cada instante. La representación semántica final puede contener conceptos que no aparecen representados por ninguna de las palabras del texto en lenguaje natural original.
- Tipos de relaciones contextuales: Recoge relaciones semánticas, de asociación y semánticas de asociación.
- Conocimiento a priori: No requiere ningún conocimiento experto.

- Conocimiento sintáctico: Recoge conocimiento sintáctico de manera implícita, aunque se persigue intencionadamente con el diseño de ciertos aspectos.
- Generación de lenguaje: Las representaciones de los textos resultantes son adecuadas como entrada a sistemas de generación de lenguaje.
- Aplicación objetivo: El sistema diseñado para utilizar las representaciones de los textos generadas en cualquier tipo de aplicación relacionada con el procesamiento de lenguaje natural. Así mismo, la parametrización del modelo pretende ser de utilidad para la aplicación del mismo a la detección y rehabilitación de trastornos del lenguaje en seres humanos.
- Plausibilidad psicológica: El sistema persigue la compatibilidad en todas las etapas del modelo de lectura, desde el diseño a la implementación, con evidencias psicológicas encontradas en seres humanos. La plausibilidad está también presente en ciertos aspectos de la etapa de adquisición del conocimiento.

4.6. Similitud Semántica en SILC

La representación de los textos obtenida por SILC puede ser considerada como un vector, como se verá en el capítulo siguiente, utilizando los valores de activación como las coordenadas del mismo. Esta transformación haría posible la aplicación práctica del sistema mediante cualquiera de los algoritmos de procesamiento de lenguaje natural existentes que admiten un vector como entrada. Sin embargo, la representación resultante del modelo de lectura contiene más información que los valores numéricos de activación, ya que se compone de una estructura de relaciones semánticas ponderadas entre los conceptos representativos. Con el fin de aprovechar la estructura e información adicional que produce el sistema con respecto a la representación vectorial, se han definido funciones para medir la similitud semántica entre las representaciones semántica-conceptuales que produce SILC. Así, los valores obtenidos de las funciones de similitud permitirán comparar textos en términos semánticos y harán posible la aplicación a tareas prácticas de clasificación, recuperación de información, evaluación, etc.

4.6.1. Similitud semántica entre conceptos

Las funciones de similitud entre textos propuestas para SILC están basadas en la similitud semántica entre conceptos individuales. Existen en la literatura diversas medidas de similitud entre conceptos de una red de asociación, concretamente en WordNet [Budanistky y Hirst, 2001]. Entre las medidas más populares se encuentran las denominadas Hirst-St-Onge [Hirst y St-Onge, 1998], Leacock-Chodorow [Leacock y Chorodow, 1998], Resnik [Resnik, 1995], Jiang-Conrath [Jiang y Conrath, 1997] y Lin [Lin, 1998]. Todas ellas están basadas en el tamaño del camíno mínimo que conecta en la red a los dos conceptos comparados. Además, todas hacen uso de la estructura jerárquica y de las relaciones específicas de la representación WordNet, aunque cada una de ellas introduce algún aspecto diferenciado. Por ejemplo, la medida Hirst-St-Onge utiliza el concepto de dirección. Así, dos conceptos están más asociados cuanto más corto sea el camino que los une y menos cambie de dirección (al igual que en SILC, las relaciones en WordNet tienen dirección). La medida Leacock-Chodorow utiliza la

noción de profundidad de una red como distancia media entre dos nodos cualesquiera de la red, normalizando el tamaño del camino mínimo con respecto a la misma. La medida de Resnik se basa en la idea de que dos conceptos similares comparten información similar. Así, mide la distancia al primer nodo más general común a ambos (recorriendo en orden inverso las relaciones ES-UN). En el caso de Jiang-Conrath se emplea la misma noción que en la medida anterior de Resnik ponderada por las probabilidades de encontrar nodos más generales comunes en la red. Ling emplea los mismos conceptos probabilistas que Jiang y Conrath pero combinados de manera diferente.

La medida de similitud entre conceptos propuesta para el conocimiento adquirido por el sistema SILC se sustenta simplemente en la noción de camino mínimo como mínima distancia semántica, ya que no se contemplan diferentes tipos de asociaciones. El tamaño del camino se define en términos de los pesos de las relaciones que lo forman y no en el número de estas últimas. Así pues, la similitud entre dos conceptos queda definida por la multiplicación de los pesos de las relaciones que forman el camino de mínima distancia semántica entre dichos conceptos, es decir, aquel que relaciona con más fuerza, en términos de ponderación, a los conceptos. En el ejemplo de la Figura 4.12, se pueden apreciar los caminos mínimos entre los nodos *A* y *D*. La similitud entre *D* y *A* es igual a 0.35, resultante de la multiplicación de los pesos de las relaciones que forman el camino de mínima distancia semántica entre ellos, 0.5 y 0.7 respectivamente. En el caso recíproco, la similitud entre *A* y *D* es igual a 0.9, ya que la relación directa entre los conceptos es la que obtiene la mayor ponderación.

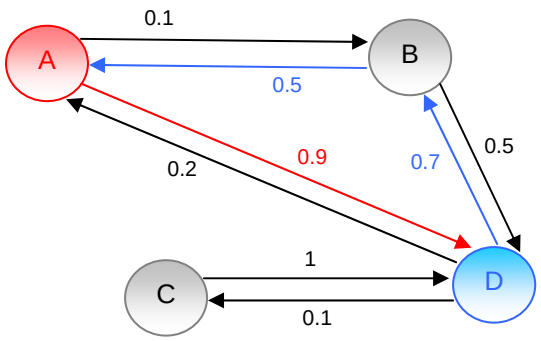


Figura 4.12. Caminos mínimos entre dos conceptos A y D en la red de conocimiento lingüístico semántico de SILC.

En la Figura 4.12 también se puede observar como la definición de distancia como ponderación de los pesos puede hacer que el camino de mínima distancia semántica no sea el más directo, como es el caso de la similitud entre *D* y *A*. Además, la similitud entre conceptos no es simétrica, es decir, no es igual de similar el concepto *A* con el *D* que el *D* con el *A*. Este aspecto que también es psicológicamente plausible [Tversky, 1977] no puede ser contemplando por la representación vectorial de la semántica empleada por la mayoría de los sistemas actuales de representación de textos. El cálculo del camino mínimo entre dos conceptos de la red se realiza mediante un algoritmo basado en el método de Dijskstra [Dijkstra, 1959], concretamente la implementación de Dial [Dial, 1969]. Puesto que el algoritmo de Dijkstra calcula el camino mínimo entre todos los pares de nodos de un grafo, la modificación utilizada termina cuando ya ha calculado el camino entre el nodo origen y el nodo destino, desechando los demás pares. De este modo se ahorra tiempo de procesamiento innecesario. Además, dado que los grafos a tratar pueden contener cientos de miles de nodos, la elección de la implementación elegida persigue la mejora sustancial de eficiencia con respecto al algoritmo de Dijsktra original [Hung y Divoky, 1988], [Cherkassky et al., 1993].

Dado el conocimiento lingüístico semántico utilizado en el ejemplo de la sección 4.4.5 de este capítulo, se muestra a continuación la similitud entre varios conceptos (en negrita) y el camino semántico mínimo entre los mismos:

fútbol → **deporte**, similitud = 0.071
deporte → **fútbol**, similitud = 0.023

petróleo → exportación → **economía**, similitud = 0.00012
economía → exportación → **petróleo**, similitud = 0.00006

fútbol → campeonato → **petróleo**, similitud = 0.00013
petróleo → campeonato → **fútbol,** similitud = 0.00002

lluvia → **nieve**, similitud = 0.009
nieve → **lluvia**, similitud = 0.018

lluvia → **ácida**, similitud = 0.013
ácida → **lluvia,** similitud = 0.004

nieve → lluvia → **ácida**, similitud = 0.00023
ácida → lluvia → **nieve,** similitud = 0.00004

nieve → humedad → **fría**, similitud = 0.0006
fría → humedad → **nieve,** similitud = 0.0001

Como se puede apreciar, la similitud entre un concepto general y uno más específico es menor que la similitud entre el específico y el más general, lo que es psicológicamente plausible. Igualmente, en el caso de los conceptos que caracterizan a otros (fría y nieve, lluvia y ácida), los conceptos caracterizados son más similares a sus complementos de lo que éstos son a los caracterizados. El ejemplo muestra además la similitud para los tres tipos de relaciones semánticas descritos en el Capítulo 2 obtenidas a partir de la concurrencia de conceptos: semántica (fútbol-deporte, petróleo-economía, lluvia-nieve), asociación (lluvia-ácida) y semántica-asociación (nieve-frío).

4.6.2. Similitud semántica entre textos

Dada la función de similitud entre conceptos definida anteriormente, la función de similitud entre dos textos está basada en una combinación de la similitud local entre los conceptos que forman la representación semántica de los mismos. La activación que posean los conceptos en las representaciones de los textos también toma parte en el cómputo de la similitud entre los mismos, considerando la significación o intensidad con la que se tratan los temas como otro criterio de similitud semántica.

4.6.2.1. Construcción del contexto de comparación

Puesto que el número de conceptos que comprende el conocimiento semántico lingüístico puede ser del orden de los cientos de miles, el algoritmo de caminos mínimos resulta muy costoso en términos de tiempo a pesar de la implementación eficiente utilizada. Teniendo en cuenta que la comparación entre las representaciones de dos textos implica el cómputo de la similitud entre cada concepto de un texto y todos los conceptos del otro, el proceso de comparación podría no ser factible. Para superar esta limitación, la red de conocimiento en la que los conceptos son comparados se reduce de manera significativa. Los caminos mínimos se calculan en una red constituida únicamente por el contexto que definen los textos involucrados en la comparación. Dicho contexto está formado por la unión de las subredes que representan a cada texto. En el caso de que la intersección de las dos subredes estuviera vacía al no tener ningún concepto en común es necesario rellenar el espacio semántico que existe entre ambas. Para ello se añaden a la unión de las subredes lo que se denominan puentes o nexos. Los puentes son caminos mínimos entre los conceptos más activados de las diferentes subredes y se obtienen de la red global de conocimiento adquirido. Así, los textos están

semánticamente conectados a través de sus conceptos más significativos, definiendo un contexto local a ambos. Además, los puentes aseguran que siempre existe un camino de cualquier concepto de un texto a cualquiera del otro en el contexto que ambos definen. De esta manera, se calculan caminos mínimos en redes compuestas por decenas de nodos en lugar de en redes compuestas por cientos de miles de nodos, lo que hace al cómputo de la similitud un proceso factible en el tiempo.

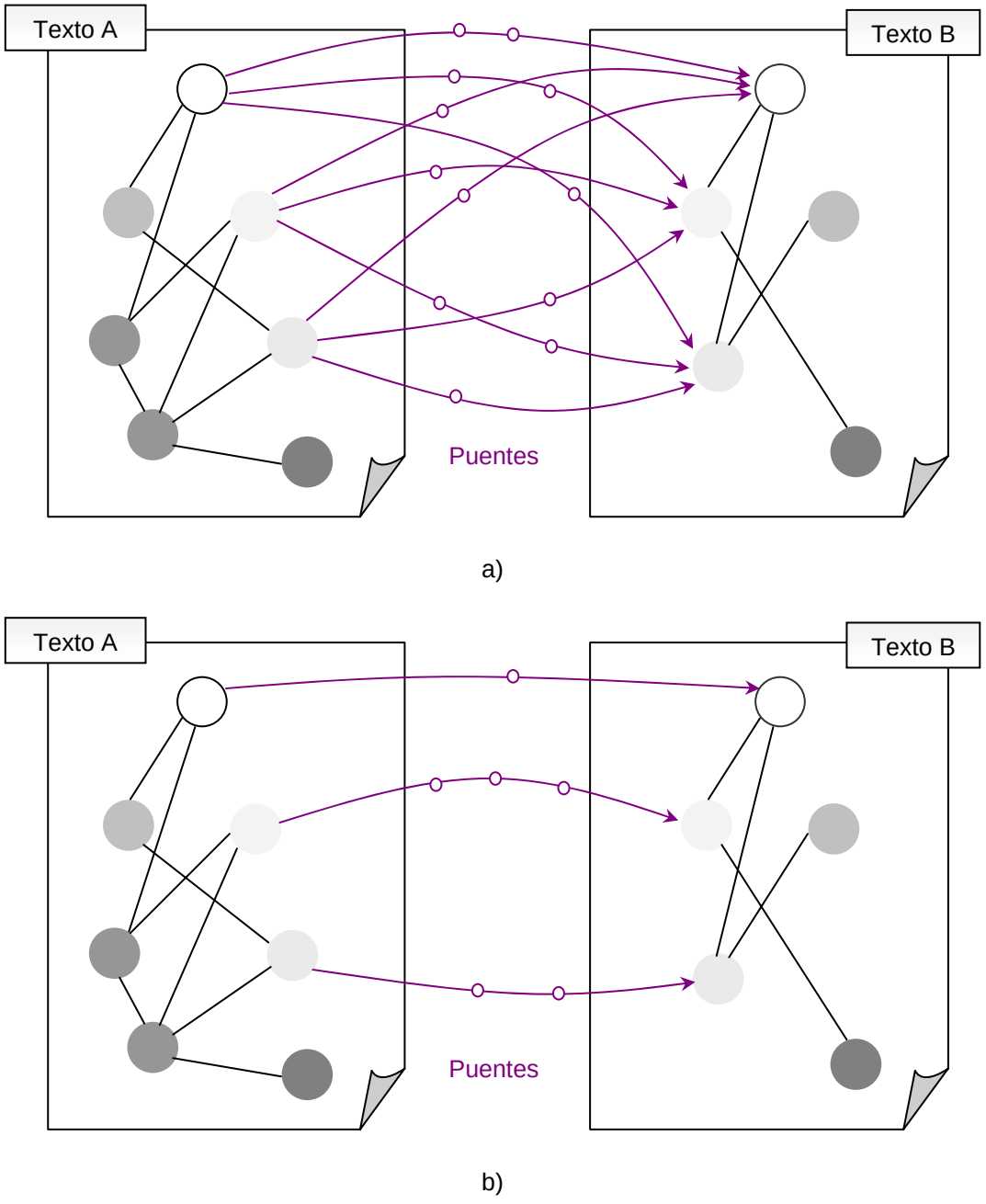


Figura 4.13. Construcción de puentes o nexos del contexto de comparación entre los nodos más activos a) de manera cruzada y b) por pares ordenados por activación.

La introducción de los puentes o nexos plantea ciertos detalles de implementación. El primero de ellos es cuántos de los conceptos más activos de cada texto, denominados "anclas" de ahora en adelante, se utilizan para construir los puentes. Este número es un parámetro que ha sido estudiado de manera experimental. El segundo es entre qué nodos se establecen puentes. Se pueden crear nexos cruzados entre cada par de los nodos más activos de ambos textos, como muestra la Figura 4.13a (un color más claro denota más actividad), o se pueden crear puentes por parejas en orden relativo de activación, como se muestra en la Figura 4.13b. La elección entre ambas opciones ha sido también objeto de experimentación.

4.6.2.2. Funciones de similitud semántica entre textos

Con el objetivo de cuantificar la similitud entre dos textos se proponen dos maneras de combinar las similitudes locales existentes entre los conceptos que componen los textos comparados, sim_T^1 y sim_T^2, respectivamente, que se detallan a continuación.

La primera forma consiste en calcular la media aritmética de la similitud entre cada par de conceptos de los dos textos que se comparan. Así, se calculará la similitud semántica local entre cada concepto de un texto y todos los conceptos del otro. Como ya se ha apuntado, el nivel de significación que los conceptos poseen en las representaciones de los textos también participa en el cómputo de la similitud. Así pues, la similitud del camino mínimo entre cada par de conceptos comparados se pondera, de manera inversa, con su diferencia de su activación, de tal forma que si dos conceptos son similares por sí mismos pero se tratan con distinta significación en los textos comparados, su similitud local contribuye en menor grado a la similitud global entre los textos, y viceversa. La función de similitud así definida, denominada sim_T^1, se presenta en la Ecuación 4.1, donde T_a y T_b son los textos que se comparan, sim_c es la similitud entre conceptos, definida como la ponderación del camino de mínima distancia semántica presentada en la sección anterior, c_{ia} es el concepto i-ésimo en orden de activación descendente del texto T_a y análogamente c_{jb} el concepto j-ésimo del texto T_b, siendo k es el número de conceptos más activos de cada texto que intervienen en la comparación. En esta ecuación *difact* es la función que pondera la similitud entre conceptos mediante la diferencia de la activación que poseen en las correspondientes representaciones de los textos comparados. La expresión analítica de esta ponderación se presenta en la Ecuación 4.2, donde *act* hace referencia al nivel de activación de un

concepto. La función de ponderación determina que cuanto mayor sea la diferencia de activación entre dos conceptos más se penaliza su valor de similitud en el cómputo de la similitud global entre los textos.

$$\mathrm{sim_T}^{1}(\mathrm{T_a}, \mathrm{T_b}) = \frac{\sum_{i,j}^{k} \mathrm{sim_c}(\mathrm{c_{ia}}, \mathrm{c_{jb}}) \cdot \mathrm{difact}(\mathrm{c_{ia}}, \mathrm{c_{jb}})}{k} \tag{4.1}$$

$$\begin{aligned} \mathrm{difact}(\mathrm{c_a}, \mathrm{c_b}) &= \frac{act(c_a)}{act(c_b)} \qquad \text{si } \mathrm{c_a} \leq \mathrm{c_b} \\ \mathrm{difact}(\mathrm{c_a}, \mathrm{c_b}) &= \frac{act(c_b)}{act(c_a)} \qquad \text{si } \mathrm{c_a} > \mathrm{c_b} \end{aligned} \tag{4.2}$$

En la medida sim_T^1 se comparan los *k* conceptos más activos de cada texto todos entre sí. Al igual que en la construcción de los puentes, se puede considerar una comparación por parejas por orden de activación. Para ello, se propone una variante de la medida de similitud sim_T^1, denominada $sim_T^{1'}$, que combina las similitudes de los conceptos emparejándolos por su orden relativo de activación y cuya expresión se presenta en la Ecuación 4.3.

$$\mathrm{sim_T}^{1'}(\mathrm{T_a}, \mathrm{T_b}) = \frac{\sum_{i}^{k} \mathrm{sim_c}(\mathrm{c_{ia}}, \mathrm{c_{ib}}) \cdot \mathrm{difact}(\mathrm{c_{ia}}, \mathrm{c_{ib}})}{k} \tag{4.3}$$

La segunda forma estudiada para calcular la similitud que se presenta está basada en la idea de que textos similares tratan temas comunes con la misma significación. Así, la similitud entre dos textos se calcula como el inverso de la diferencia media de activación entre las *k* parejas de conceptos más similares entre dichos textos, según la expresión de la Ecuación 4.4, donde *act* se refiere al nivel de activación de un concepto y sim_c es la similitud entre conceptos, análogamente a las ecuaciones anteriores.

$$\mathrm{sim_T}^{2}(\mathrm{T_a}, \mathrm{T_b}) = \frac{\sum_{i}^{k} \frac{1}{1 + |\,\mathrm{act}(\mathrm{c_{ia}}) - \mathrm{act}(\mathrm{c_{jb}})\,|_{\min_k(\mathrm{sim_c}(\mathrm{c_{ia}}, \mathrm{c_{jb}}))}}}{k} \tag{4.4}$$

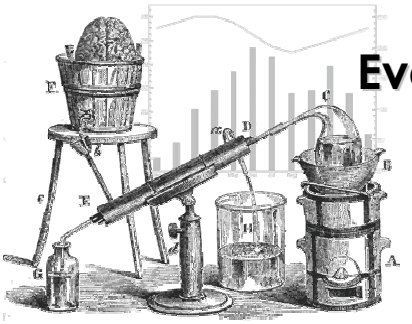
Puesto que la activación de los conceptos interviene en el cómputo de la similitud, se plantea también la posibilidad de normalizar los niveles de activación con respecto a las representaciones en las que aparecen, para dar homogeneidad al cálculo. De esta manera, la activación de todos los conceptos que componen la representación semántica

de un texto se dividen por la activación máxima en el mismo, es decir, la activación del concepto más significativo de dicho texto. Así, todos los niveles quedan comprendidos en el rango entre 0 y 1.

Todos los parámetros y opciones de implementación han sido objeto de estudio experimental, como se describe en el siguiente capítulo. Dichos aspectos comprenden la opción de normalización, el valor óptimo del parámetro *k* de las expresiones anteriores, es decir, el número de los conceptos más activos que intervienen en el cómputo de la similitud, el número óptimo de conceptos utilizados para construir puentes junto con la forma de construirlos, las diferentes formas de cómputo de la similitud propuestas y la evaluación de la reducción del espacio de conocimiento en el que se realiza la comparación de los textos.

Así pues, la definición de similitud semántica propuesta aporta una manera de aprovechar en aplicaciones prácticas la información estructural, no presente en la representación vectorial, de las representaciones obtenidas por el modelo de lectura en el sistema SILC.

CAPÍTULO 5



Evaluación Experimental de SILC

Un modelo computacional siempre plantea al menos una hipótesis de manera involuntaria, que consiste en la afirmación de que el mundo real opera de manera análoga a los mecanismos que implementa el modelo. La principal evaluación del modelo se centra en la validación de dicha hipótesis. En el caso de los modelos computacionales de lectura, el método deductivo aristotélico puro es difícilmente aplicable puesto que la formalización de los modelos es extremadamente compleja y, por tanto, no se puede llevar a cabo un razonamiento lógico para la validación de la hipótesis que plantea. Se puede decir que la investigación que conlleva el diseño de un modelo computacional de lectura sigue un método híbrido entre el inductivo y el hipotético-deductivo o de contraste de hipótesis. El carácter inductivo viene dado por la observación de la realidad, es decir, las evidencias experimentales psicológicas, de la que se extraen una serie de hipótesis, es decir, estructuras y mecanismos computacionales basados en dichas evidencias. Dichas hipótesis han de ser posteriormente contrastadas mediante experimentación, con el objetivo principal de constatar que no existen ejemplos que las contradigan, y de ahí el carácter hipotético-deductivo.

5.1. Objetivos de la Evaluación Experimental

En primer lugar, dado que el modelo presentado en esta tesis es paramétrico, es necesaria una optimización de dichos parámetros. Existen dos criterios de optimización: el primero es la búsqueda de la máxima eficacia en tareas prácticas y el segundo, la persecución de la máxima similitud con el ser humano. Según la hipótesis que ha motivado el trabajo de investigación propuesto en esta tesis, ambos criterios deberían conducir a los mismos valores para los parámetros. Entre los objetivos de optimización no sólo se encuentran los parámetros numéricos, sino también los diferentes mecanismos diseñados para una misma tarea como, en este caso, las dos clases de propagación de la activación o los distintos tipos de similitud entre palabras y textos propuestos en el Capítulo 4. En cualquier caso, la optimización de todas estas cuestiones se ha llevado a cabo mediante la experimentación.

Como se mencionó en el Capítulo 3, los criterios de evaluación para un modelo computacional de lectura son de diferente naturaleza [Fletcher, 1999]. Primeramente, se ha evaluado la eficacia del sistema SILC respecto a otros sistemas existentes en tareas de procesamiento de lenguaje natural, más concretamente en la clasificación automática de textos [Sebastiani, 2002], persiguiendo el objetivo de determinar la utilidad práctica del modelo. La plausibilidad psicológica es otro de los criterios contemplados en la evaluación del modelo. Aparte de las consideraciones de diseño, la plausibilidad del modelo se ha medido en términos de similitud con los seres humanos. Así pues, se han realizado también experimentos en los que se comparan las representaciones generadas por el modelo con representaciones generadas por sujetos humanos, tanto durante como al final del proceso de lectura, realizando así una evaluación "on-line" y "off-line" (ver Capítulo 3). El resto de criterios de evaluación propuestos por Fletcher no han sido comprobados mediante la experimentación ya que son de carácter metodológico y funcional, aunque sí se discuten en el siguiente capítulo.

Por supuesto, para todos los tipos de evaluaciones se han utilizado conjuntos de textos de diferente naturaleza e idioma expresados en lenguaje natural. Para la cuantificación de los criterios de evaluación perseguidos se han empleado medidas ya existentes ampliamente conocidas, además de otras nuevas medidas ideadas específicamente para algunos de los casos.

5.2. Optimización de los Parámetros y Elección Experimental de los Mecanismos del Sistema SILC

Como se ha comentado anteriormente, la optimización del sistema SILC no sólo consiste en la obtención de los valores óptimos de los parámetros en aras de la eficacia en clasificación de textos, sino también en la elección con base experimental entre los diferentes mecanismos propuestos para realizar un mismo proceso del modelo. De esta forma, se presenta en primer lugar la experimentación relativa a la optimización de los parámetros y mecanismos propios del modelo de lectura que intervienen en la generación de la representación semántica de los textos. A continuación, se describe el proceso experimental relativo a los parámetros y aspectos involucrados en la construcción del conocimiento semántico lingüístico previo sobre el que opera el modelo de lectura. Por último, se presentan los experimentos realizados en relación a los aspectos que operan posteriormente sobre la representación generada por el modelo, como la similitud semántica entre conceptos y entre textos.

5.2.1. Conjunto de datos

En primer lugar, para cualquier tipo de evaluación de un sistema de procesamiento de lenguaje natural es necesario definir el conjunto de datos, en este caso un conjunto de textos, con el que se alimenta al sistema. El conjunto de textos empleados para la optimización de parámetros consta de 500 noticias de actualidad recopiladas manualmente del servicio de noticias de Google en español (*Google News*[1]), durante un período de un mes. Los textos recopilados pertenecen a cinco categorías temáticas distintas según la clasificación del propio servicio de noticias de *Google*. Dichas categorías son: Ciencia y Tecnología, Cultura y Espectáculos, Deportes, Economía y Salud, y se distribuyen uniformemente y de manera disjunta en la colección. Así, la colección se compone de 100 textos de cada una de las cinco categorías descritas. Además, se ha empleado otra colección de 500 ensayos en español de una amplia

[1] http://www.news.google.es

variedad de temáticas, escritos y recopilados por el escritor Jorge Orellana Mora[2], que tratan de encerrar toda la cultura general media que posee un sujeto con estudios universitarios. Concretamente, los textos utilizados para la evaluación de SILC son los que se encuentran en los documentos accesibles en la página *web* del escritor desde Febrero de 2000 hasta Diciembre de 2006.

5.2.2. Medidas de evaluación

Dado que la optimización de los parámetros se ha realizado en base al criterio de la eficacia en la tarea de la clasificación de textos, se han empleado medidas típicas para cuantificar dicha eficacia. Teniendo en cuenta que la clasificación de textos consiste en la asignación automática de una categoría temática de entre un conjunto predefinido de las mismas a un texto no utilizado para el aprendizaje del sistema, las medidas empleadas se definen para cada categoría como [Sebastiani, 2002], [Sokolova et al., 2006]:

- Precisión. Porcentaje de textos que el sistema ha clasificado correctamente, es decir, la proporción de textos que pertenecen realmente a una categoría de entre todos los textos que el sistema ha clasificado como de dicha categoría. Su expresión analítica se recoge en la Ecuación 5.1:

$$\text{Precisión}(Categoría_i) = \frac{nº\_de\_textos\_clasificados\_como\_i\_que\_pertencen\_realmente\_a\_i}{nº\_de\_textos\_clasificados\_como\_i} \quad (5.1)$$

- Cobertura. Porcentaje de textos de una categoría que el sistema ha clasificado como tal. Analíticamente, la Cobertura se expresa como sigue:

$$Cobertura\ (Categoría_i) = \frac{nº\ de\_textos\_clasificados\_como\_i}{nº\_total\_de\_textos\_que\_pertencen\_a\_i} \quad (5.2)$$

- Medida-F. Esta medida es una combinación de las dos anteriores. Depende de un parámetro β, número natural mayor que cero, que determina la ponderación de

[2] http://j.orellana.free.fr/index.htm

una de las medidas sobre la otra. Su expresión analítica se presenta en la Ecuación 5.3:

$$Medida\_F(Categoría_i) = \frac{(1+\beta)\cdot \Pr ecisión(Categoría_i)\cdot Cobertura(Categoría_i)}{(\beta\cdot \Pr ecisión(Categoría_i)) + Cobertura(Categoría_i)} \quad (5.3)$$

Nótese que cuanto más aumenta β más peso adquiere la Precisión frente a la Cobertura. Con el objetivo de ponderar en igual grado a ambas medidas, el valor de β empleado en los experimentos es igual a 1.

- Correlación de Pearson. Es un índice que mide la relación lineal entre dos variables cuantitativas. En este caso, dichas variables se corresponden con la categoría real de los textos y la categoría asignada por el algoritmo, respectivamente. Se calcula mediante la división de la covarianza de las dos variables por el producto de las desviaciones estándar de las mismas. Un índice de correlación de 1 indica una dependencia total directa. Por el contrario, un índice de -1 indica una dependencia total indirecta. Si el índice es igual a 0 entonces no existe correlación alguna. Lo deseable para la clasificación de textos es que el índice sea mayor que cero y se aproxime lo más posible a uno.

Puesto que las medidas descritas corresponden a cada categoría individual del conjunto de textos, salvo el índice de correlación, para evaluar la eficacia global del sistema se ha empleado la media aritmética de dichas medidas entre todas las categorías implicadas.

5.2.3. Procedimiento experimental

Los algoritmos de clasificación empleados pertenecen al paradigma de aprendizaje supervisado, es decir, necesitan textos etiquetados (con la categoría a la que pertenecen) para entrenarse y aprender un modelo de clasificación a partir de los mismos. Es necesario distinguir pues entre textos de entrenamiento y textos de test o evaluación. Un procedimiento clásico de evaluación es el denominado de "validación cruzada de subconjuntos" (*cross-fold validation*) [Mitchell, 1997], [Sebastiani, 2002]. En dicho procedimiento, el conjunto original de textos se divide en *n* subconjuntos de igual tamaño, generalmente con la misma distribución en categorías. A continuación, se

utilizan *n-1* subconjuntos para entrenar al algoritmo y el subconjunto restante para validarlo, obteniendo las medidas descritas anteriormente de dicha validación. En un siguiente paso, se toman otros *n-1* subconjuntos de entrenamiento y otro diferente de evaluación. De esta manera, el proceso de entrenamiento y validación se ejecuta *n* veces, actuando cada subconjunto una vez como conjunto de validación. Para cada categoría, las medidas de evaluación final se calculan realizando la media aritmética de las mismas entre las *n* ejecuciones que se han realizado.

Así pues, el valor de *n* seleccionado para los experimentos realizados es un valor típico de 3, utilizando el 67% de la colección compuesta por las noticias de *Google News* como entrenamiento (335 textos) y el 33% restante (165 textos) como validación en cada ejecución.

En cada una de las ejecuciones, el conocimiento semántico lingüístico sobre el que se aplica el modelo de lectura de SILC se construye con el 67% de textos de entrenamiento más los 500 textos de cultura general de Orellana. A continuación, se aplica el modelo de lectura y se representan los 500 textos de la colección de *Google News* (entrenamiento más validación), utilizando ciertos valores para los parámetros de lectura.

El propósito de evaluar al sistema SILC en la clasificación de textos es el de comprobar que la representación producida por el modelo contiene de manera precisa la semántica intrínseca del texto de entrada al que corresponde. Puesto que la mayoría de los algoritmos de clasificación requieren que los textos de entrada se presenten como vectores, todas las representaciones producidas por SILC se han representado, para los experimentos de optimización, en forma de vectores de la siguiente manera:

- El tamaño de los vectores es el número total de conceptos distintos que aparecen en todas las representaciones, tanto en el conjunto de entrenamiento como en el de validación.
- El vector correspondiente a una representación contiene el valor de activación de los conceptos de los que se compone en sus respectivas posiciones dentro el vector. El resto de posiciones del vector contendrá un valor de 0.

Una vez obtenidas las representaciones de los textos en forma de vector, éstas se presentan como entrada a tres algoritmos diferentes de clasificación:

- Naïve Bayes [Mitchell, 1997], [Sebastiani, 2002]. Es un algoritmo de clasificación probabilista basado en el cálculo de la probabilidad de Bayes. Calcula para cada palabra de los textos de entrenamiento su probabilidad de pertenecer a cada categoría y, a partir de estos valores, estima la probabilidad de un texto de pertenecer a cada categoría como la multiplicación de la probabilidad de las palabras que lo forman, asignando a dicho texto la categoría con la probabilidad más alta. A pesar de que asume la independencia entre las palabras que conforman los textos, este ha sido el algoritmo de la literatura más empleado en clasificación de textos debido a sus buenos resultados y capacidad para tratar con grandes cantidades de información de manera eficiente.

- Máquinas de Vectores de Soporte (*SVM, Support Vector Machines*) [Joachims, 2002], [Sebastiani, 2002]. Es un algoritmo de clasificación que trata de encontrar "hiperplanos" que separen a los vectores de las distintas categorías en el espacio de *m* dimensiones, siendo *m* el tamaño de los vectores de entrada. La búsqueda de dichos "hiperplanos" se realiza mediante iteraciones en las que se maximiza la distancia entre las fronteras de cada categoría, determinadas por los vectores de soporte, y los "hiperplanos". Las mejoras conseguidas de este algoritmo en los últimos tiempos le han hecho erigirse como la alternativa a Naïve Bayes.

- K Vecinos Cercanos (*K-NN, K-Nearest Neighbours*) [Aha et al., 1991], [Mitchell, 1997], [Sebastiani, 2002]. Es un algoritmo de clasificación basado en memoria o en ejemplos, es decir, no crea ningún modelo de clasificación por lo que no necesita ninguna fase de entrenamiento, aunque sí textos que conformen la memoria. El algoritmo compara un texto a clasificar con todos los que componen la memoria. A continuación selecciona los *K* textos más parecidos (vecinos) y asigna a dicho texto la categoría mayoritaria entre aquellos. La similitud entre los vectores que representan los textos es, generalmente, la distancia Euclídea [Black, 2004] entre los mismos. El valor de *K* empleado en los experimentos de optimización es de 5.

El hecho de emplear varios algoritmos de clasificación de diferente naturaleza aporta más consistencia a los resultados de optimización, suavizando la dependencia entre el conjunto de textos empleados, su representación y los algoritmos. Los algoritmos

utilizados son los implementados en el entorno *YALE*[3], versión 3.4, con los parámetros por defecto. Concretamente, se utilizaron las implementaciones *NaiveBayes* simple, *LibSVM Learner* e *IBk*, respectivamente.

Así, para cada ejecución de las tres llevadas a cabo en la validación cruzada, se representan todos los textos, tanto de entrenamiento como de validación, usando el modelo de lectura de SILC con unos determinados valores para los parámetros sobre el conocimiento semántico creado a partir de los textos de entrenamiento en cada caso más los 500 ensayos de cultura general. Las representaciones generadas se transforman en vectores de la manera descrita y se presentan a los tres algoritmos de clasificación mencionados, obteniéndose medidas de precisión, cobertura y medida-F para cada categoría, además la media de todas ellas y del índice de correlación. Una vez concluidas las tres ejecuciones se realiza la media aritmética de los resultados obtenidos en las mismas para cada algoritmo. A continuación se realiza el mismo proceso variando uno de los parámetros de lectura de SILC en determinados intervalos en un rango. Así, se procede de manera análoga con el resto de parámetros, empleando en cada optimización de un parámetro los valores de los parámetros anteriormente evaluados que han obtenido los mejores resultados medios de clasificación.

5.2.4. Optimización de los parámetros de lectura

5.2.4.1. Optimización del umbral mínimo de propagación y del factor de olvido

Siguiendo el procedimiento experimental descrito en la sección anterior, se variaron los valores de dos parámetros de manera conjunta. Dichos parámetros son el umbral mínimo de propagación, es decir, el valor por debajo del cual la activación no se propaga a los conceptos asociados y por debajo del cual un concepto es olvidado y se elimina de la memoria de trabajo (recuérdese que el umbral mínimo de propagación y el umbral de olvido toman el mismo valor), y el factor de olvido, es decir, el porcentaje del nivel de activación que pierden los conceptos activos en la memoria de trabajo a cada intervalo de olvido. Nótese en este caso que los valores del factor de olvido evaluados

[3] YALE (*Yet Another Learning Environment*): http://rapid-i.com/

son multiplicadores de la activación, y por tanto lo que denotan es el porcentaje de activación que los conceptos conservan, y no el que pierden, cada vez que se aplica.

El valor del factor del olvido se varió desde 0.9 hasta 0.3 en intervalos de 0.2. El umbral mínimo de propagación se varió desde 0.01 hasta el valor del factor de olvido menos 0.2 en cada caso, en intervalos de 0.2. Así, por ejemplo, cuando el factor de olvido toma el valor 0.5, el umbral mínimo de propagación se evaluó con los valores 0.01, 0.1 y 0.3. Dado que la activación inicial de un concepto leído es 1, si el umbral de propagación es mayor o igual que el factor de olvido, la activación no se propagará a ningún concepto asociado y el concepto leído será olvidado automáticamente en el momento en el que se le aplique el factor de olvido por primera vez. Los valores del resto de los parámetros son: propagación de la activación en profundidad, nivel máximo de propagación igual a 3 e intervalo de olvido variable (a cada frase leída). Así pues, la siguiente Tabla 5.1 muestra para cada categoría y la media de todas ellas, así como para cada algoritmo empleado y la media de todos ellos, los valores de precisión, cobertura, medida-F e índice de correlación medios de las tres ejecuciones llevadas a cabo en la validación de subconjuntos cruzados, empleando las variaciones descritas de los valores de los parámetros correspondientes al umbral mínimo de propagación y factor de olvido. La notación 'Pr' indica la precisión, 'Co' indica la cobertura y 'F' indica la medida-F.

Factor de Olvido = **0.9**				Umbral mínimo de propagación = **0.01**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	75.00	58.33	65.62	56.52	36.11	44.07	20.71	97.22	34.15	50.74	63.89	56.56
Cultura	77.14	75.00	76.05	52.27	63.89	57.50	100.00	5.56	10.53	76.47	48.15	59.09
Deportes	94.44	91.89	93.15	66.66	91.89	77.27	0.00	0.00	0.00	53.70	61.26	57.23
Economía	75.61	86.11	80.52	60.00	75.00	66.67	75.00	8.33	14.99	70.20	56.48	62.60
Salud	75.67	87.50	81.16	57.14	25.00	34.78	100.00	6.25	11.76	77.60	39.58	52.43
Media	*79.57*	*79.77*	*79.67*	*58.52*	*58.38*	*58.45*	*59.14*	*23.47*	*33.61*	*65.74*	*53.87*	*59.22*
Correlación	*0.695*			*0.335*			*0.118*			*0.383*		

Factor de Olvido = **0.9**				Umbral mínimo de propagación = **0.1**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	81.25	72.22	76.47	50.00	61.11	55.00	21.05	100.00	34.78	50.77	77.78	61.43
Cultura	90.90	83.33	86.95	58.14	69.44	63.29	100.00	5.56	10.53	83.01	52.78	64.53
Deportes	89.74	94.59	92.10	80.49	89.19	84.62	0.00	0.00	0.00	56.74	61.26	58.92
Economía	80.48	91.67	85.71	81.25	72.22	76.47	100.00	5.56	10.53	87.24	56.48	68.57
Salud	90.63	90.62	90.62	70.59	37.50	48.98	100.00	6.25	11.76	87.07	44.79	59.15
Media	*86.60*	*86.49*	*86.54*	*68.09*	*65.89*	*66.97*	*64.21*	*23.47*	*34.38*	*72.97*	*58.62*	*65.01*
Correlación	*0.800*			*0.418*			*0.157*			*0.458*		

Factor de Olvido = **0.9**				Umbral mínimo de propagación = **0.3**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.78	72.22	75.36	53.33	66.67	59.26	20.93	100.00	34.62	51.01	79.63	62.19
Cultura	87.87	80.56	84.06	57.77	72.22	64.19	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.11	94.59	93.33	79.07	91.89	85.00	0.00	0.00	0.00	57.06	62.16	59.50
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.84*	*67.54*	*69.63*	*64.19*	*22.92*	*33.78*	*74.44*	*59.19*	*65.95*
Correlación	*0.840*			*0.478*			*0.164*			*0.494*		

Factor de Olvido = **0.9**				Umbral mínimo de propagación = **0.5**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	80.64	69.94	74.91	54.28	52.78	53.52	20.81	100.00	34.45	51.91	74.24	61.10
Cultura	87.88	80.56	84.06	49.02	69.44	57.47	100.00	5.56	10.53	78.97	51.85	62.60
Deportes	87.18	91.89	89.47	72.34	91.89	80.95	0.00	0.00	0.00	53.17	61.26	56.93
Economía	78.57	91.67	84.62	82.76	66.67	73.85	100.00	2.78	5.41	87.11	53.71	66.45
Salud	93.75	93.75	93.75	73.33	34.38	46.81	100.00	3.12	6.05	89.03	43.75	58.67
Media	*85.60*	*85.56*	*85.58*	*66.35*	*63.03*	*64.65*	*64.16*	*22.29*	*33.09*	*72.04*	*56.96*	*63.62*
Correlación	*0.820*			*0.387*			*0.098*			*0.435*		

Factor de Olvido = **0.9**				Umbral mínimo de propagación = **0.7**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	71.43	69.44	70.42	45.45	41.67	43.48	20.69	100.00	34.29	45.86	70.37	55.53
Cultura	81.82	75.00	78.26	0.50	58.33	0.99	100.00	5.56	10.53	60.77	46.30	52.56
Deportes	86.84	89.19	88.00	54.84	91.89	68.69	0.00	0.00	0.00	47.23	60.36	52.99
Economía	78.57	91.67	84.62	77.77	58.33	66.66	100.00	2.78	5.41	85.45	50.93	63.82
Salud	96.55	87.50	91.80	76.92	31.25	44.44	0.00	0.00	0.00	57.82	39.58	47.00
Media	*83.04*	*82.56*	*82.80*	*51.10*	*56.29*	*53.57*	*44.14*	*21.67*	*29.07*	*59.43*	*53.51*	*56.31*
Correlación	*0.788*			*0.318*			*0.02*			*0.375*		

Factor de Olvido = **0.7**				Umbral mínimo de propagación = **0.01**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	74.07	55.56	63.49	40.00	16.67	23.53	20.81	100.00	34.45	44.96	57.41	50.43
Cultura	76.92	83.33	80.00	45.83	30.56	36.67	100.00	5.56	10.53	74.25	39.82	51.84
Deportes	86.11	83.78	84.93	33.03	100.00	49.66	0.00	0.00	0.00	39.71	61.26	48.19
Economía	76.19	88.89	82.05	80.95	47.22	59.65	100.00	5.56	10.53	85.71	47.22	60.90
Salud	81.82	83.38	82.59	80.00	12.50	21.62	0.00	0.00	0.00	53.94	31.96	40.14
Media	*79.02*	*78.99*	*79.00*	*55.96*	*41.39*	*47.59*	*44.16*	*22.22*	*29.57*	*59.72*	*47.53*	*52.93*
Correlación	*0.696*			*0.229*			*0.053*			*0.326*		

Factor de Olvido = **0.7**				Umbral mínimo de propagación = **0.1**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	68.96	55.56	61.54	45.45	13.89	21.28	20.81	100.00	34.45	45.07	56.48	50.14
Cultura	80.55	80.56	80.55	55.00	30.56	39.29	100.00	5.56	10.53	78.52	38.89	52.02
Deportes	84.61	89.19	86.84	28.46	100.00	44.31	0.00	0.00	0.00	37.69	63.06	47.18
Economía	76.19	88.89	82.05	80.00	33.33	47.06	100.00	5.56	10.53	85.40	42.59	56.84
Salud	87.10	84.38	85.72	100.00	3.12	6.05	0.00	0.00	0.00	62.37	29.17	39.75
Media	*79.48*	*79.72*	*79.60*	*61.78*	*36.18*	*45.64*	*44.16*	*22.22*	*29.57*	*61.81*	*46.04*	*52.77*
Correlación	*0.671*			*0.205*			*0.053*			*0.310*		

Factor de Olvido = **0.7**				Umbral mínimo de propagación = **0.3**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	65.62	58.33	61.76	44.44	11.11	17.78	20.81	100.00	34.45	43.62	56.48	49.23
Cultura	75.00	66.67	70.59	53.33	22.22	31.37	100.00	5.56	10.53	76.11	31.48	44.54
Deportes	81.58	83.78	82.67	26.81	100.00	42.28	0.00	0.00	0.00	36.13	61.26	45.45
Economía	78.05	88.89	83.12	85.71	33.33	48.00	100.00	5.56	10.53	87.92	42.59	57.39
Salud	79.41	84.38	81.82	100.00	3.12	6.05	0.00	0.00	0.00	59.80	29.17	39.21
Media	*75.93*	*76.41*	*76.17*	*62.06*	*33.96*	*43.89*	*44.16*	*22.22*	*29.57*	*60.72*	*44.20*	*51.16*
Correlación	*0.658*			*0.126*			*0.053*			*0.279*		

Factor de Olvido = **0.7**				Umbral mínimo de propagación = **0.5**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	62.50	55.56	58.83	41.67	13.89	20.84	20.93	100.00	34.62	41.70	56.48	47.98
Cultura	72.41	58.33	64.61	42.86	16.67	24.00	100.00	5.56	10.53	71.76	26.85	39.08
Deportes	86.49	86.49	86.49	26.28	97.30	41.38	0.00	0.00	0.00	37.59	61.26	46.59
Economía	65.96	86.11	74.70	92.31	33.33	48.98	100.00	8.33	15.38	86.09	42.59	56.99
Salud	75.00	75.00	75.00	100.00	3.12	6.05	0.00	0.00	0.00	58.33	26.04	36.01
Media	*72.47*	*72.30*	*72.38*	*60.62*	*32.86*	*42.62*	*44.19*	*22.78*	*30.06*	*59.09*	*42.65*	*49.54*
Correlación	*0.605*			*0.105*			*0.053*			*0.262*		

Factor de Olvido = **0.5**				Umbral mínimo de propagación = **0.01**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	56.67	47.22	51.52	50.00	8.33	14.28	20.81	100.00	34.45	42.49	51.85	46.71
Cultura	72.97	75.00	73.97	42.86	16.67	24.00	100.00	5.56	10.53	71.94	32.41	44.69
Deportes	85.29	78.38	81.69	24.49	97.30	39.13	0.00	0.00	0.00	36.59	58.56	45.04
Economía	65.31	88.89	75.30	80.00	22.22	34.78	100.00	5.56	10.53	81.77	38.89	52.71
Salud	77.78	65.62	71.18	0.00	0.00	0.00	0.00	0.00	0.00	25.93	21.87	23.73
Media	*71.60*	*71.02*	*71.31*	*39.47*	*28.90*	*33.37*	*44.16*	*22.22*	*29.57*	*51.75*	*40.72*	*45.57*
Correlación	*0.552*			*0.137*			*0.053*			*0.247*		

Factor de Olvido = **0.5**				Umbral mínimo de propagación = **0.1**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	60.71	47.22	53.12	0.00	0.00	0.00	20.93	100.00	34.62	27.21	49.07	35.01
Cultura	73.33	61.11	66.66	41.67	13.89	20.84	100.00	5.56	10.53	71.67	26.85	39.07
Deportes	75.61	83.78	79.49	24.03	100.00	38.75	0.00	0.00	0.00	33.21	61.26	43.07
Economía	68.08	88.89	77.11	77.78	19.44	31.11	100.00	8.33	15.38	81.95	38.89	52.75
Salud	74.19	71.88	73.02	0.00	0.00	0.00	0.00	0.00	0.00	24.73	23.96	24.34
Media	*70.38*	*70.58*	*70.48*	*28.70*	*26.67*	*27.64*	*44.19*	*22.78*	*30.06*	*47.76*	*40.01*	*43.54*
Correlación	*0.533*			*0.027*			*0.076*			*0.212*		

Factor de Olvido = **0.5**				Umbral mínimo de propagación = **0.3**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	59.26	44.44	50.79	0.00	0.00	0.00	21.18	100.00	34.96	26.81	48.15	34.44
Cultura	71.43	69.44	70.42	33.34	16.67	22.23	66.67	5.56	10.26	57.15	30.56	39.82
Deportes	74.36	78.38	76.32	24.16	97.30	38.71	0.00	0.00	0.00	32.84	58.56	42.08
Economía	68.08	88.89	77.11	87.50	19.44	31.81	100.00	8.33	15.38	85.19	38.89	53.40
Salud	68.96	62.50	65.57	0.00	0.00	0.00	0.00	0.00	0.00	22.99	20.83	21.86
Media	*68.42*	*68.73*	*68.57*	*29.00*	*26.68*	*27.79*	*37.57*	*22.78*	*28.36*	*45.00*	*39.40*	*42.01*
Correlación	*0.525*			*0.013*			*0.128*			*0.222*		

Factor de Olvido = **0.3**				Umbral mínimo de propagación = **0.01**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	51.72	41.67	46.15	0.00	0.00	0.00	20.83	97.22	34.31	24.18	46.30	31.77
Cultura	77.42	66.67	71.64	62.50	13.89	22.73	42.85	8.33	13.95	60.92	29.63	39.87
Deportes	72.97	72.97	72.97	23.57	100.00	38.15	0.00	0.00	0.00	32.18	57.66	41.31
Economía	61.22	83.33	70.58	77.78	19.44	31.11	100.00	5.56	10.53	79.67	36.11	49.70
Salud	58.06	56.25	57.14	0.00	0.00	0.00	0.00	9.00	0.00	19.35	21.75	20.48
Media	*64.28*	*64.18*	*64.23*	*32.77*	*26.67*	*29.40*	*32.74*	*24.02*	*27.71*	*43.26*	*38.29*	*40.62*
Correlación	*0.403*			*0.080*			*0.039*			*0.174*		

Factor de Olvido = **0.3**				Umbral mínimo de propagación = **0.1**								
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	51.72	41.67	46.15	0.00	0.00	0.00	20.71	97.22	34.15	24.14	46.30	31.74
Cultura	63.89	63.89	63.89	30.77	11.11	16.33	40.00	5.56	9.76	44.89	26.85	33.60
Deportes	69.44	67.57	68.49	22.78	97.30	36.92	0.00	0.00	0.00	30.74	54.96	39.43
Economía	63.26	86.11	72.94	80.00	11.11	19.51	100.00	8.33	15.38	81.09	35.18	49.07
Salud	70.37	59.38	64.41	0.00	0.00	0.00	0.00	0.00	0.00	23.46	19.79	21.47
Media	*63.74*	*63.72*	*63.73*	*26.71*	*23.90*	*25.23*	*32.14*	*22.22*	*26.28*	*40.86*	*36.62*	*38.62*
Correlación	*0.520*			*-0.005*			*0.064*			*0.193*		

Tabla 5.1. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados por SILC con diferentes combinaciones de valores del umbral mínimo de propagación y el factor de olvido.

En la Tabla 5.1 se puede apreciar como los resultados de clasificación son mejores cuanto mayor es el factor de olvido. Para apreciar mejor la relación entre los parámetros, la Figura 5.1a muestra los valores de medida-F medios entre todas las categorías y todos los algoritmos de clasificación para las distintas combinaciones de valores de los parámetros (umbral mínimo de propagación en el eje de abscisas y factor de olvido en líneas de colores). Análogamente, la Figura 5.1b muestra los valores de correlación. En la figura se aprecia de manera clara que los valores óptimos son de 0.3 para el umbral mínimo de propagación y de 0.9 para el factor de olvido. Así mismo, se aprecia el fenómeno comentado anteriormente de que cuanto mayor es el factor de olvido, es decir, cuanto más tiempo se recuerdan los conceptos, mejor son los resultados de clasificación. Las gráficas también indican que los mejores resultados se obtienen cuando los valores de los dos parámetros distan entre sí. A medida que dichos valores se asemejan los resultados de clasificación decaen. En el caso de un factor de olvido alto, un umbral mínimo de activación demasiado bajo hace que los resultados de clasificación empeoren ya que permite la inferencia de un gran número de conceptos que tardan en ser olvidados, por lo que se introduce ruido y ambigüedad en la representación de la semántica generada.

Los resultados de clasificación, tanto parciales como globales, no son demasiado altos pragmáticamente hablando. Sin embargo, son suficiente para poner de manifiesto la relación entre los parámetros y obtener su configuración óptima. Con seguridad, los resultados mejorarían notablemente si se aplicase un fase de reducción y selección de

características previa a la clasificación [Yang y Pedersen, 1997], [Sebastiani, 2002], [Del Castillo y Serrano, 2004]. Sin embargo, la intención de los experimentos es no sesgar al modelo de lectura permitiéndole, de este modo, operar en todo el espacio de conocimiento que construye y seleccionar por sí mismo los conceptos relevantes.

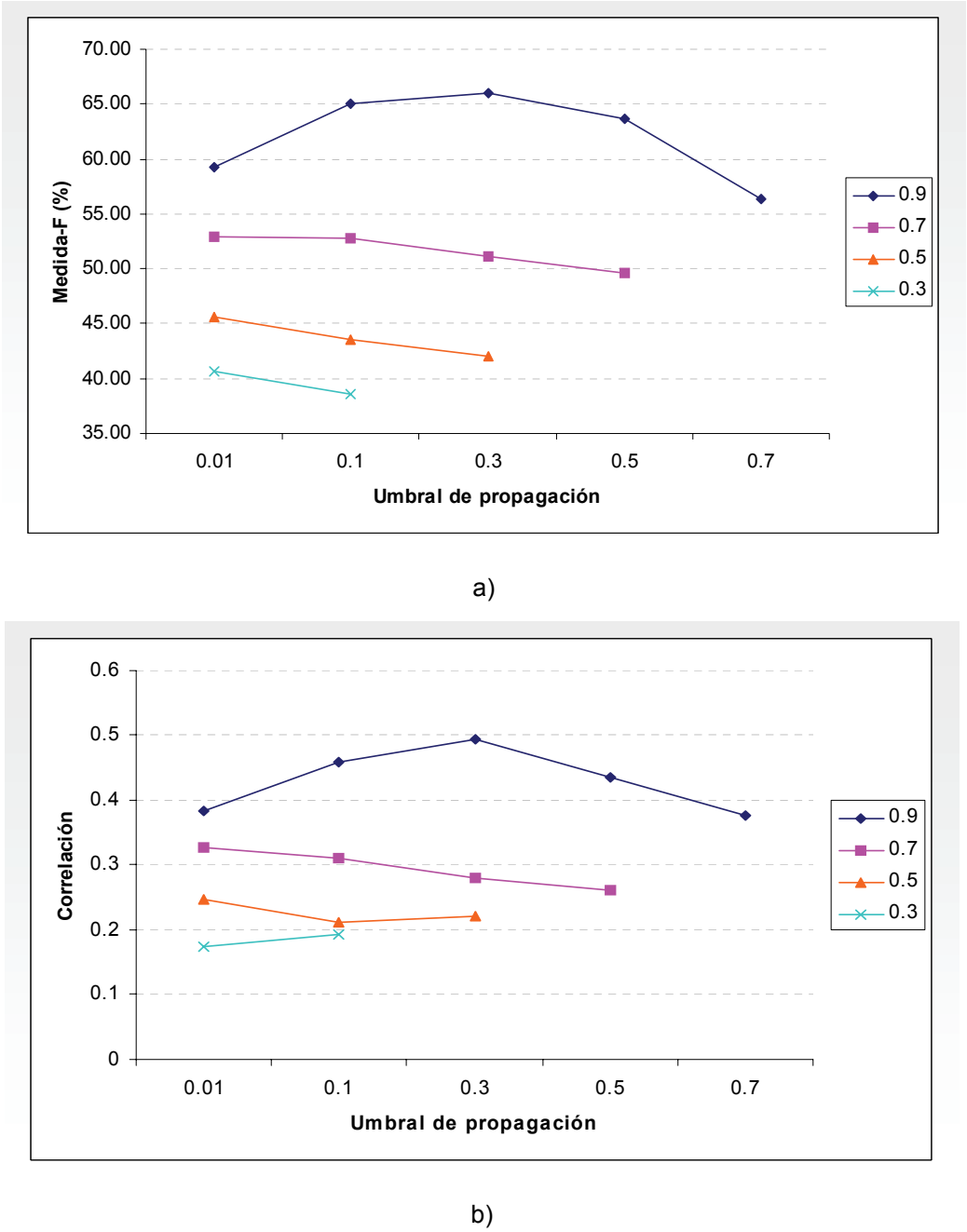


Figura 5.1. Valores medios de a) medida-F y b) correlación, en tareas de clasificación de textos representados por SILC empleando diferentes combinaciones de valores para los parámetros umbral mínimo de propagación y factor de olvido.

5.2.4.2. Optimización del tipo de inferencia y nivel de propagación

Siguiendo un procedimiento análogo al empleado para la optimización del umbral mínimo de propagación y del factor de olvido, se ha realizado la optimización del tipo y nivel máximo de propagación de la activación. Recuérdese que se contemplan dos tipos de propagación: por niveles y en profundidad. Recuérdese también que el nivel máximo de propagación se refiere a la cantidad máxima de conceptos que puede atravesar la activación propagada desde el concepto origen. Así pues, utilizando los valores óptimos para los parámetros evaluados en la sección anterior, se varió el nivel máximo de activación desde 1 hasta 3, en intervalos de 1, para los dos tipos de propagación o inferencia. La Tabla 5.2 muestra, para cada categoría y la media de todas ellas así como para cada algoritmo empleado y la media de todos ellos, los valores de precisión, cobertura, medida-F e índice de correlación medios de las tres ejecuciones llevadas a cabo en la validación de subconjuntos cruzados, empleando las variaciones descritas de los valores de los parámetros correspondientes al tipo y nivel máximo de propagación de la activación. Al igual que en la tabla anterior, la notación 'Pr' hace referencia a la precisión, 'Co' representa a la cobertura y 'F' indica la medida-F.

Tipo de propagación = **En Profundidad**							Nivel máximo de propagación = **1**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	51.11	63.89	56.79	20.93	100.00	34.62	50.28	78.70	61.36
Cultura	87.88	80.56	84.06	56.52	72.22	63.41	100.00	5.56	10.53	81.47	52.78	64.06
Deportes	92.10	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.68	62.16	59.84
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.52*	*66.99*	*69.18*	*64.19*	*22.92*	*33.78*	*74.34*	*59.01*	*65.79*
Correlación	*0.840*			*0.458*			*0.146*			*0.481*		

Tipo de propagación = **En Profundidad**							Nivel máximo de propagación = **2**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	53.33	66.67	59.26	20.93	100.00	34.62	51.02	79.63	62.19
Cultura	87.88	80.56	84.06	57.77	72.22	64.19	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.10	94.59	93.33	79.07	91.89	85.00	0.00	0.00	0.00	57.06	62.16	59.50
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.84*	*67.54*	*69.63*	*64.19*	*22.92*	*33.78*	*74.44*	*59.19*	*65.95*
Correlación	*0.840*			*0.478*			*0.146*			*0.488*		

Tipo de propagación = **En Profundidad**							Nivel máximo de propagación = **3**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.78	72.22	75.36	53.33	66.67	59.26	20.93	100.00	34.62	51.01	79.63	62.19
Cultura	87.87	80.56	84.06	57.77	72.22	64.19	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.11	94.59	93.33	79.07	91.89	85.00	0.00	0.00	0.00	57.06	62.16	59.50
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.84*	*67.54*	*69.63*	*64.19*	*22.92*	*33.78*	*74.44*	*59.19*	*65.95*
Correlación	*0.840*			*0.478*			*0.164*			*0.494*		

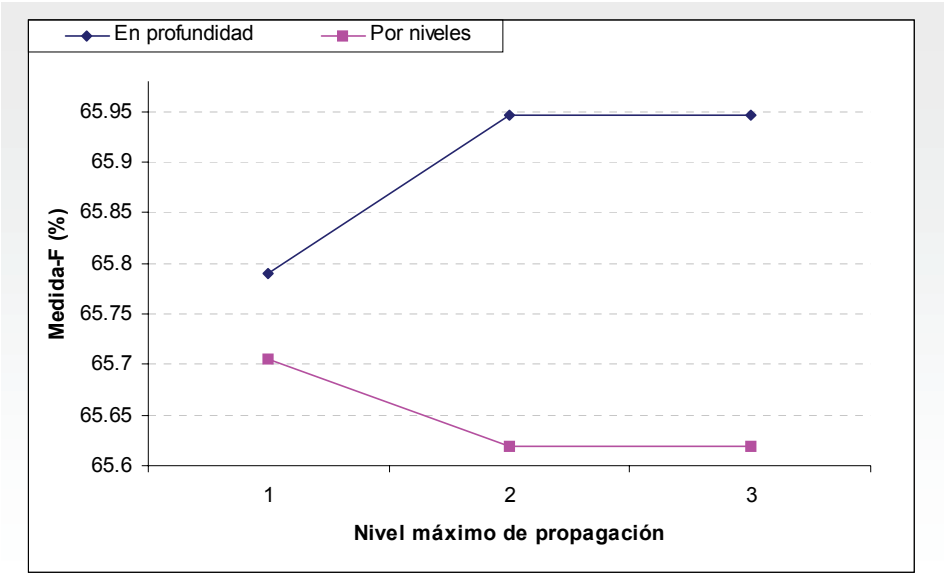
Tipo de propagación = **Por niveles**							Nivel máximo de propagación = **1**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	50.00	61.11	55.00	20.93	100.00	34.62	49.91	77.78	60.80
Cultura	87.88	80.56	84.06	57.78	72.22	64.20	100.00	5.56	10.53	81.89	52.78	64.19
Deportes	92.10	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.68	62.16	59.84
Economía	80.95	94.44	87.18	84.37	75.00	79.41	100.00	2.78	5.41	88.44	57.41	69.62
Salud	100.00	93.75	96.77	78.57	34.38	47.83	100.00	6.25	11.76	92.86	44.79	60.43
Media	*87.94*	*87.11*	*87.53*	*70.33*	*66.92*	*68.58*	*64.19*	*22.92*	*33.78*	*74.15*	*58.98*	*65.70*
Correlación	*0.840*			*0.419*			*0.146*			*0.468*		

Tipo de propagación = **Por niveles**							Nivel máximo de propagación = **2**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	50.00	61.11	55.00	20.93	100.00	34.62	49.91	77.78	60.80
Cultura	87.88	80.56	84.06	57.78	72.22	64.20	100.00	5.56	10.53	81.89	52.78	64.19
Deportes	92.10	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.68	62.16	59.84
Economía	80.95	94.44	87.18	84.37	75.00	79.41	100.00	2.78	5.41	88.44	57.41	69.62
Salud	96.77	93.75	95.24	78.57	34.38	47.83	100.00	6.25	11.76	91.78	44.79	60.20
Media	*87.30*	*87.11*	*87.20*	*70.33*	*66.92*	*68.58*	*64.19*	*22.92*	*33.78*	*73.94*	*58.98*	*65.62*
Correlación	*0.840*			*0.419*			*0.146*			*0.468*		

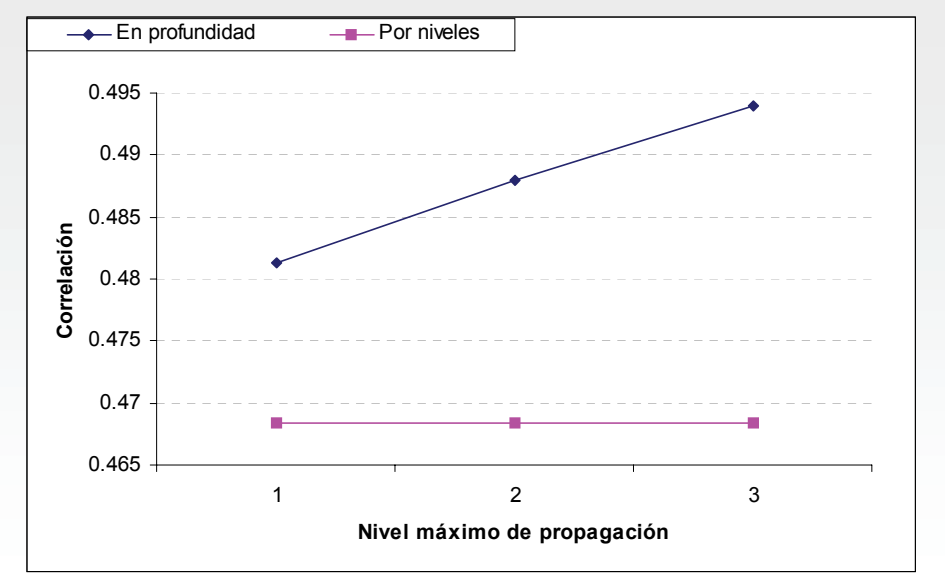
Tipo de propagación = **Por niveles**							Nivel máximo de propagación = **3**					
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	50.00	61.11	55.00	20.93	100.00	34.62	49.91	77.78	60.80
Cultura	87.88	80.56	84.06	57.78	72.22	64.20	100.00	5.56	10.53	81.89	52.78	64.19
Deportes	92.10	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.68	62.16	59.84
Economía	80.95	94.44	87.18	84.37	75.00	79.41	100.00	2.78	5.41	88.44	57.41	69.62
Salud	96.77	93.75	95.24	78.57	34.38	47.83	100.00	6.25	11.76	91.78	44.79	60.20
Media	*87.30*	*87.11*	*87.20*	*70.33*	*66.92*	*68.58*	*64.19*	*22.92*	*33.78*	*73.94*	*58.98*	*65.62*
Correlación	*0.840*			*0.419*			*0.146*			*0.468*		

Tabla 5.2. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados por SILC con diferentes combinaciones del tipo y valor del nivel máximo de propagación.

A pesar de que los valores de la tabla reflejan ciertas diferencias éstas son mínimas. La Figura 5.2 pone de relevancia la levedad de dichas diferencias mediante los valores medios entre todas las categorías y los algoritmos de la medida-F (Figura 5.2a) y del índice de correlación (Figura 5.2b), manifestándose la propagación en profundidad ligeramente mejor que la propagación por niveles.



a)



b)

Figura 5.2. Valores medios de a) medida-F y b) correlación, en tareas de clasificación de textos representados por SILC empleando diferentes combinaciones de valores para los parámetros tipo y nivel máximo de propagación de la activación.

La figura muestra también que los resultados de clasificación mejoran a medida que crece el nivel máximo de propagación en el caso de la inferencia en profundidad, y de manera inversa en el caso de la inferencia por niveles. Sin embargo, como ya se ha comentado, estos incrementos y decrementos son muy leves. El motivo es que el tipo de propagación afecta principalmente al nivel de activación de los conceptos en la representación final, y los algoritmos utilizados no son muy sensibles a pequeñas variaciones en los valores numéricos de los vectores de entrada. La levedad de la diferencia observada en la variación del nivel máximo de propagación se debe a que su efecto está en gran medida controlado por el valor del umbral mínimo de propagación, previamente fijado a su valor óptimo. Éste es el motivo por el que no se han considerado niveles de propagación mayores que tres, además de por no quebrantar la plausibilidad psicológica, ya que hay estudios que demuestran que la mente humana no procesa relaciones de más de tres órdenes, al menos de manera involuntaria [Perfetti, 1999].

A pesar de la escasa entidad numérica de las variaciones, los resultados de la experimentación han puesto de manifiesto la tendencia de ambos tipos de inferencia y su diferente relación con el parámetro referente al nivel máximo de propagación, permitiendo establecer los valores óptimos para dichos parámetros: propagación en profundidad con un nivel máximo de activación igual a 3.

5.2.4.3. Optimización del intervalo de olvido

Ejemplificando el sistema SILC con los valores óptimos de los parámetros anteriormente estudiados, se ha procedido a la optimización análoga del parámetro relativo al intervalo de olvido, es decir, al número de palabras consecutivas tras el cuál se aplica el factor de olvido a los conceptos presentes en la memoria de trabajo hasta ese momento. Puesto que ya se tienen resultados sobre el intervalo de olvido variable definido por el tamaño de las oraciones presentes de los textos, ya que ha sido utilizado en los experimentos descritos hasta el momento, se ha procedido a aplicar al modelo en tareas de clasificación utilizando una ventana de tamaño fijo, variando dicho tamaño desde 5 hasta 19 palabras en intervalos de una palabra. Así pues, la Tabla 5.3 presenta los resultados de clasificación, para las distintas categorías, algoritmos de clasificación y la media de todos ellos, en términos de precisión (‘Pr’), cobertura (‘Co’), medida-F (‘F’) y correlación para distintos tamaños fijos del intervalo de olvido.

Tamaño del intervalo de olvido = **5 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	67.86	52.78	59.38	48.94	63.89	55.42	20.69	100.00	34.29	45.83	72.22	56.08
Cultura	76.47	72.22	74.28	62.50	55.56	58.83	100.00	5.56	10.53	79.66	44.45	57.06
Deportes	89.19	89.19	89.19	73.81	83.78	78.48	0.00	0.00	0.00	54.33	57.66	55.95
Economía	73.33	91.67	81.48	65.85	75.00	70.13	100.00	2.78	5.41	79.73	56.48	66.12
Salud	90.32	87.50	88.89	80.00	37.50	51.06	0.00	0.00	0.00	56.77	41.67	48.06
Media	*79.43*	*78.67*	*79.05*	*66.22*	*63.15*	*64.65*	*44.14*	*21.67*	*29.07*	*63.26*	*54.50*	*58.55*
Correlación	*0.707*			*0.398*			*0.020*			*0.375*		

Tamaño del intervalo de olvido = **7 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	75.76	69.44	72.46	57.57	52.78	55.07	20.57	100.00	34.12	51.30	74.07	60.62
Cultura	84.85	77.78	81.16	54.00	75.00	62.79	100.00	5.56	10.53	79.62	52.78	63.48
Deportes	89.47	91.89	90.66	91.18	83.78	87.32	0.00	0.00	0.00	60.22	58.56	59.38
Economía	78.57	91.67	84.62	76.47	72.22	74.28	0.00	0.00	0.00	51.68	54.63	53.11
Salud	96.77	93.75	95.24	81.48	68.75	74.58	0.00	0.00	0.00	59.42	54.17	56.67
Media	*85.08*	*84.91*	*84.99*	*72.14*	*70.51*	*71.31*	*24.11*	*21.11*	*22.51*	*60.45*	*58.84*	*59.63*
Correlación	*0.825*			*0.597*			*-0.073*			*0.450*		

Tamaño del intervalo de olvido = **9 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	79.41	75.00	77.14	57.14	66.67	61.54	20.57	100.00	34.12	52.37	80.56	63.48
Cultura	87.88	80.56	84.06	60.97	69.44	64.93	100.00	5.56	10.53	82.95	51.85	63.81
Deportes	91.89	91.89	91.89	96.87	83.78	89.85	0.00	0.00	0.00	62.92	58.56	60.66
Economía	76.74	91.67	83.54	77.78	77.78	77.78	0.00	0.00	0.00	51.51	56.48	53.88
Salud	100.00	93.75	96.77	84.61	68.75	75.86	0.00	0.00	0.00	61.54	54.17	57.62
Media	*87.18*	*86.57*	*86.88*	*75.47*	*73.28*	*74.36*	*24.11*	*21.11*	*22.51*	*62.26*	*60.32*	*61.28*
Correlación	*0.841*			*0.681*			*-0.073*			*0.483*		

Tamaño del intervalo de olvido = **11 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	75.00	75.00	75.00	61.11	61.11	61.11	20.69	100.00	34.29	52.27	78.70	62.82
Cultura	87.10	75.00	80.60	60.97	69.44	64.93	100.00	5.56	10.53	82.69	50.00	62.32
Deportes	91.89	91.89	91.89	96.87	83.78	89.85	0.00	0.00	0.00	62.92	58.56	60.66
Economía	76.19	88.89	82.05	74.36	80.56	77.34	100.00	2.78	5.41	83.52	57.41	68.05
Salud	96.77	93.75	95.24	75.86	68.75	72.13	0.00	0.00	0.00	57.54	54.17	55.80
Media	*85.39*	*84.91*	*85.15*	*73.83*	*72.73*	*73.28*	*44.14*	*21.67*	*29.07*	*67.79*	*59.77*	*63.53*
Correlación	*0.839*			*0.631*			*0.020*			*0.497*		

Tamaño del intervalo de olvido = **13 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	80.00	77.78	78.87	61.76	58.33	60.00	20.81	100.00	34.45	54.19	78.70	64.19
Cultura	90.32	77.78	83.58	60.97	69.44	64.93	100.00	5.56	10.53	83.76	50.93	63.34
Deportes	91.89	91.89	91.89	93.94	83.78	88.57	0.00	0.00	0.00	61.94	58.56	60.20
Economía	77.27	94.44	85.00	65.12	77.78	70.89	100.00	2.78	5.41	80.80	58.33	67.75
Salud	100.00	93.75	96.77	80.77	65.62	72.41	100.00	3.12	6.05	93.59	54.16	68.62
Media	*87.90*	*87.13*	*87.51*	*72.51*	*70.99*	*71.74*	*64.16*	*22.29*	*33.09*	*74.86*	*60.14*	*66.69*
Correlación	*0.855*			*0.641*			*0.098*			*0.531*		

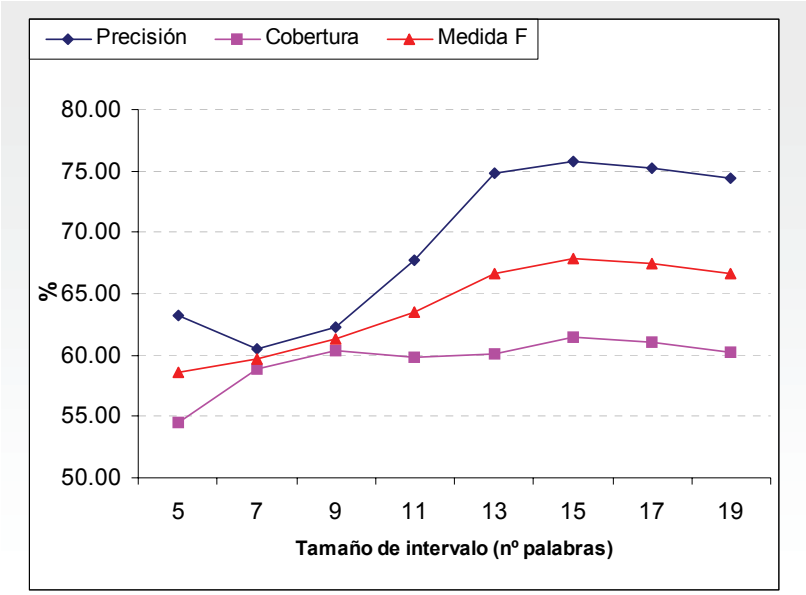
Tamaño del intervalo de olvido = **15 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	80.55	80.56	80.55	66.67	61.11	63.77	20.93	100.00	34.62	56.05	80.56	66.11
Cultura	90.32	77.78	83.58	63.41	72.22	67.53	100.00	5.56	10.53	84.58	51.85	64.29
Deportes	92.10	94.59	93.33	93.94	83.78	88.57	0.00	0.00	0.00	62.01	59.46	60.71
Economía	80.95	94.44	87.18	69.77	83.33	75.95	100.00	2.78	5.41	83.57	60.18	69.98
Salud	100.00	93.75	96.77	77.78	65.62	71.18	100.00	6.25	11.76	92.59	55.21	69.17
Media	*88.78*	*88.22*	*88.50*	*74.31*	*73.21*	*73.76*	*64.19*	*22.92*	*33.78*	*75.76*	*61.45*	*67.86*
Correlación	*0.870*			*0.662*			*0.146*			*0.559*		

Tamaño del intervalo de olvido = **17 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	82.86	80.56	81.69	64.70	61.11	62.85	20.93	100.00	34.62	56.16	80.56	66.18
Cultura	90.32	77.78	83.58	65.79	69.44	67.57	100.00	5.56	10.53	85.37	50.93	63.80
Deportes	92.10	94.59	93.33	91.18	83.78	87.32	0.00	0.00	0.00	61.09	59.46	60.26
Economía	80.95	94.44	87.18	65.91	80.56	72.50	100.00	2.78	5.41	82.29	59.26	68.90
Salud	96.77	93.75	95.24	77.78	65.62	71.18	100.00	6.25	11.76	91.52	55.21	68.87
Media	*88.60*	*88.22*	*88.41*	*73.07*	*72.10*	*72.58*	*64.19*	*22.92*	*33.78*	*75.29*	*61.08*	*67.44*
Correlación	*0.859*			*0.648*			*0.146*			*0.551*		

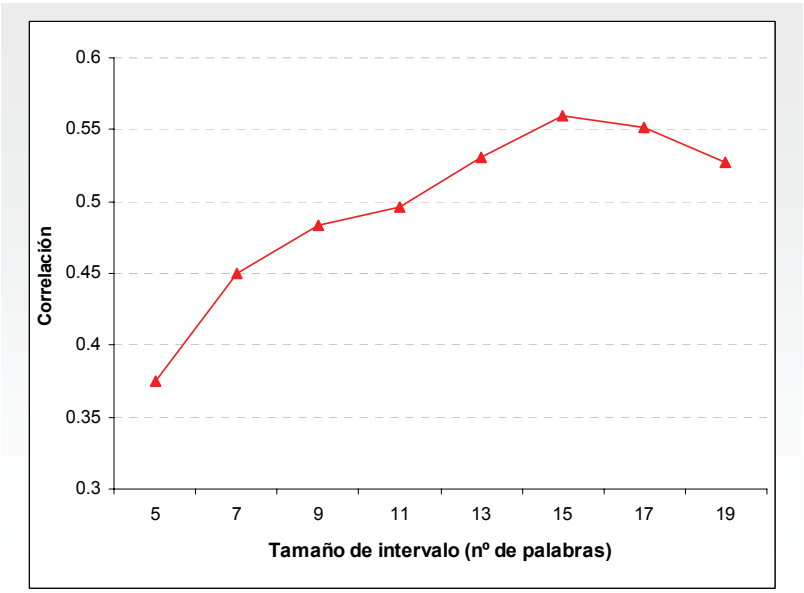
Tamaño del intervalo de olvido = **19 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	82.86	80.56	81.69	65.71	63.89	64.79	20.93	100.00	34.62	56.50	81.48	66.73
Cultura	90.32	77.78	83.58	61.54	66.67	64.00	100.00	5.56	10.53	83.95	50.00	62.68
Deportes	92.10	94.59	93.33	91.18	83.78	87.32	0.00	0.00	0.00	61.09	59.46	60.26
Economía	80.95	94.44	87.18	65.91	80.56	72.50	100.00	2.78	5.41	82.29	59.26	68.90
Salud	96.77	93.75	95.24	68.00	53.12	59.65	100.00	6.25	11.76	88.26	51.04	64.68
Media	**88.60**	**88.22**	**88.41**	**70.47**	**69.60**	**70.03**	**64.19**	**22.92**	**33.78**	**74.42**	**60.25**	**66.59**
Correlación	*0.859*			*0.578*			*0.146*			*0.528*		

Tabla 5.3. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados por SILC con diferentes tañanos fijos de intervalo de olvido.

Los valores de la Tabla 5.3 muestran una relación directa entre el tamaño del intervalo de olvido y la eficacia en la clasificación. Si se observan gráficamente los resultados medios entre todas las categorías y todos los algoritmos en la Figura 5.3 (precisión, cobertura, medida-F (a) y correlación (b)), se pone de manifiesto dicha tendencia creciente.



a)



b)

Figura 5.3. Valores medios de a) precisión, cobertura y medida-F y b) correlación en tareas de clasificación de textos representados por SILC empleando diferentes valores del tamaño fijo de intervalo de olvido.

La gráfica muestra cómo la eficacia de clasificación crece a la vez que el tamaño del intervalo hasta un valor de 11 palabras. A partir de ese punto el crecimiento es menor, alcanzando su máximo en 15 y experimentando un descenso de aquí en adelante, lo que denota que un intervalo de olvido demasiado extenso hace que se retengan en la memoria demasiados conceptos que introducen ruido y ambigüedad.

Los valores de medida-F y correlación máximos, con un intervalo de 15 palabras, son 67.86 y 0.559, respectivamente. Los valores máximos de dichas medidas para un intervalo de olvido de tamaño variable utilizando los mismos valores para el resto de parámetros del sistema, son 65.94 y 0.494, respectivamente (véanse los resultados de la sección 5.2.1.5 anterior). Estas cifras muestran que para valores altos del tamaño fijo del intervalo éste resulta ligeramente más eficaz que el intervalo de olvido de tamaño variable. El motivo principal es que el intervalo de tamaño fijo no se ve afectado por el estilo gramatical de la escritura de los textos, que puede en ocasiones llegar a confundir la semántica de los mismos. Sin embargo, el intervalo de tamaño variable sí refleja este aspecto, por lo que aporta plausibilidad psicológica al sistema y amplia las posibilidades del modelo por el precio de una pérdida leve de eficacia.

5.2.5. Optimización de los parámetros de construcción del conocimiento semántico

Como se explica en el Capítulo 3, el modelo de lectura opera sobre una red conceptual de conocimiento semántico previamente adquirido. Con respecto a la construcción de dicho conocimiento, dos son las cuestiones que se plantean para la experimentación: la definición del contexto en el que se asocian los conceptos que en él concurren, y la cantidad de textos a partir de los que se construye el conocimiento.

5.2.5.1. Optimización del tipo y tamaño de contexto de asociación

En los experimentos anteriores relativos a la optimización de los parámetros del modelo de lectura, el conocimiento fue construido a partir de la correspondiente porción de textos de entrenamiento en cada caso junto con los 500 textos de cultura general de Orellana. El contexto de asociación empleado es el de las oraciones que aparecen en los textos, asociando cada par de conceptos que aparecen en la misma oración. La alternativa es utilizar como contexto una ventana deslizante de palabras de tamaño fijo,

de tal forma que se asocian cada par de conceptos que concurren dentro de dicha ventana. Así pues, se ha repetido el procedimiento experimental utilizado en la optimización de los parámetros del modelo de lectura, descrito en la sección anterior, variando en esta ocasión el tamaño de la ventana de palabras utilizada para asociar los conceptos en el conocimiento semántico lingüístico. Los resultados de clasificación (precisión, cobertura, medida-F y correlación) se presentan en la Tabla 5.4 para las distintas categorías y algoritmos de clasificación, variando el tamaño de la ventana de contexto desde 5 hasta 19 palabras en intervalos de 2 palabras. Los parámetros del modelo de lectura utilizado han sido fijados a los valores óptimos obtenidos en los experimentos anteriores, utilizando un intervalo de olvido de tamaño variable (el factor de olvido se aplica al final de cada oración).

Tamaño de la ventana de contexto = **5 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	81.82	75.00	78.26	50.00	47.22	48.57	20.93	100.00	34.62	50.92	74.07	60.35
Cultura	87.88	80.56	84.06	58.54	66.67	62.34	100.00	5.56	10.53	82.14	50.93	62.88
Deportes	87.18	91.89	89.47	68.75	89.19	77.65	0.00	0.00	0.00	51.98	60.36	55.86
Economía	78.57	91.67	84.62	75.76	69.44	72.46	100.00	5.56	10.53	84.78	55.56	67.12
Salud	100.00	93.75	96.77	52.38	34.48	41.59	100.00	3.12	6.05	84.13	43.78	57.59
Media	*87.09*	*86.57*	*86.83*	*61.09*	*61.40*	*61.24*	*64.19*	*22.85*	*33.70*	*70.79*	*56.94*	*63.11*
Correlación	*0.878*			*0.377*			*0.114*			*0.456*		

Tamaño de la ventana de contexto = **7 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	79.41	75.00	77.14	46.67	58.33	51.85	20.93	100.00	34.62	49.00	77.78	60.12
Cultura	90.90	83.33	86.95	54.54	66.67	60.00	100.00	5.56	10.53	81.81	51.85	63.48
Deportes	92.10	94.59	93.33	78.57	89.19	83.54	0.00	0.00	0.00	56.89	61.26	58.99
Economía	78.57	91.67	84.62	76.67	63.89	69.70	100.00	2.78	5.41	85.08	52.78	65.15
Salud	96.67	90.62	93.55	70.59	37.50	48.98	100.00	6.25	11.76	89.09	44.79	59.61
Media	*87.53*	*87.04*	*87.29*	*65.41*	*63.12*	*64.24*	*64.19*	*22.92*	*33.78*	*72.37*	*57.69*	*64.20*
Correlación	*0.801*			*0.428*			*0.146*			*0.458*		

Tamaño de la ventana de contexto = **9 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	82.35	77.78	80.00	45.45	55.56	50.00	21.05	100.00	34.78	49.62	77.78	60.59
Cultura	90.91	83.33	86.96	54.17	72.22	61.91	100.00	5.56	10.53	81.69	53.70	64.81
Deportes	92.10	94.59	93.33	78.51	89.19	83.51	0.00	0.00	0.00	56.87	61.26	58.98
Economía	80.49	91.67	85.72	78.51	61.11	68.73	100.00	5.56	10.53	86.33	52.78	65.51
Salud	96.77	93.75	95.24	73.33	34.38	46.81	100.00	6.25	11.76	90.03	44.79	59.82
Media	*88.52*	*88.22*	*88.37*	*65.99*	*62.49*	*64.20*	*64.21*	*23.47*	*34.38*	*72.91*	*58.06*	*64.64*
Correlación	*0.839*			*0.406*			*0.157*			*0.467*		

Tamaño de la ventana de contexto = **11 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	79.41	75.00	77.14	53.66	61.11	57.14	20.93	100.00	34.62	51.33	78.70	62.14
Cultura	87.88	80.56	84.06	52.94	75.00	62.07	100.00	5.56	10.53	80.27	53.71	64.36
Deportes	92.10	94.59	93.33	82.50	89.19	85.71	0.00	0.00	0.00	58.20	61.26	59.69
Economía	78.57	91.67	84.62	80.00	66.67	72.73	100.00	2.78	5.41	86.19	53.71	66.18
Salud	100.00	93.75	96.77	80.00	37.50	51.06	100.00	6.25	11.76	93.33	45.83	61.48
Media	*87.59*	*87.11*	*87.35*	*69.82*	*65.89*	*67.80*	*64.19*	*22.92*	*33.78*	*73.87*	*58.64*	*65.38*
Correlación	*0.845*			*0.451*			*0.146*			*0.480*		

Tamaño de la ventana de contexto = **13 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	48.84	58.33	53.16	20.93	100.00	34.62	49.52	76.85	60.23
Cultura	87.88	80.56	84.06	57.45	75.00	65.06	100.00	5.56	10.53	81.78	53.71	64.83
Deportes	92.10	94.59	93.33	78.57	89.19	83.54	0.00	0.00	#¡DIV/0!	56.89	61.26	58.99
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	73.33	34.38	46.81	100.00	6.25	11.76	90.03	44.79	59.82
Media	*87.30*	*87.11*	*87.20*	*68.30*	*65.27*	*66.75*	*64.19*	*22.92*	*33.78*	*73.26*	*58.43*	*65.01*
Correlación	*0.840*			*0.416*			*0.146*			*0.467*		

Tamaño de la ventana de contexto = **15 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	53.66	61.11	57.14	20.93	100.00	34.62	51.13	77.78	61.70
Cultura	87.88	80.56	84.06	56.00	77.78	65.12	100.00	5.56	10.53	81.29	54.63	65.35
Deportes	92.10	94.59	93.33	78.57	89.19	83.54	0.00	0.00	0.00	56.89	61.26	58.99
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	78.57	34.38	47.83	100.00	6.25	11.76	91.78	44.79	60.20
Media	*87.30*	*87.11*	*87.20*	*70.03*	*66.38*	*68.15*	*64.19*	*22.92*	*33.78*	*73.84*	*58.80*	*65.47*
Correlación	*0.840*			*0.426*			*0.146*			*0.471*		

Tamaño de la ventana de contexto = **17 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	53.85	58.33	56.00	20.93	100.00	34.62	51.19	76.85	61.45
Cultura	87.88	80.56	84.06	53.85	77.78	63.64	100.00	5.56	10.53	80.58	54.63	65.12
Deportes	92.10	95.59	93.81	80.49	89.19	84.62	0.00	0.00	0.00	57.53	61.59	59.49
Economía	80.95	94.44	87.18	81.25	72.22	76.47	100.00	2.78	5.41	87.40	56.48	68.62
Salud	96.77	93.75	95.24	84.61	34.38	48.89	100.00	6.25	11.76	93.79	44.79	60.63
Media	*87.30*	*87.31*	*87.30*	*70.81*	*66.38*	*68.52*	*64.19*	*22.92*	*33.78*	*74.10*	*58.87*	*65.61*
Correlación	*0.840*			*0.438*			*0.146*			*0.475*		

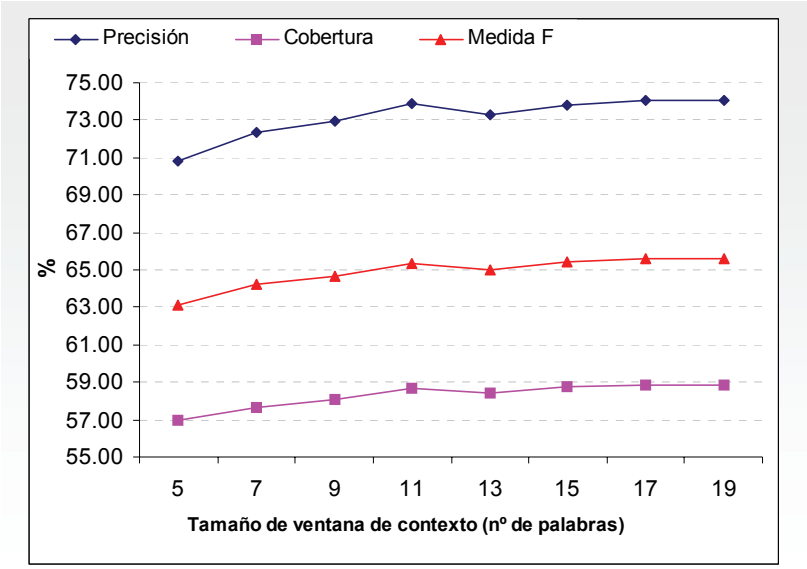
Tamaño de la ventana de contexto = **19 palabras**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.79	72.22	75.36	52.50	58.33	55.26	20.93	100.00	34.62	50.74	76.85	61.12
Cultura	87.88	80.56	84.06	54.90	77.78	64.37	100.00	5.56	10.53	80.93	54.63	65.23
Deportes	92.10	95.59	93.81	80.49	89.19	84.62	0.00	0.00	0.00	57.53	61.59	59.49
Economía	80.95	94.44	87.18	81.25	72.22	76.47	100.00	2.78	5.41	87.40	56.48	68.62
Salud	96.77	93.75	95.24	84.61	34.38	48.89	100.00	6.25	11.76	93.79	44.79	60.63
Media	*87.30*	*87.31*	*87.30*	*70.75*	*66.38*	*68.50*	*64.19*	*22.92*	*33.78*	*74.08*	*58.87*	*65.60*
Correlación	*0.840*			*0.430*			*0.146*			*0.472*		

Tabla 5.4. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados por SILC bajo un conocimiento semántico construido con diferentes tamaños fijos de ventana de contexto.

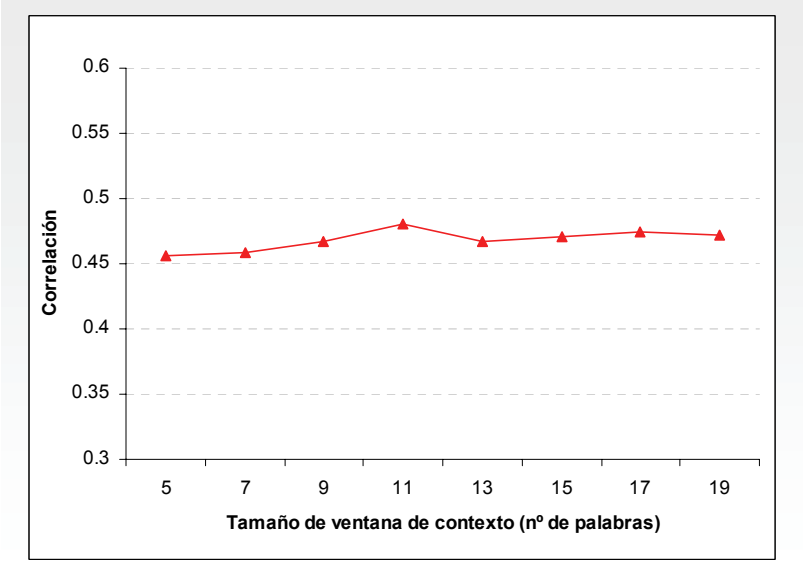
Los resultados de la tabla no parecen revelar ninguna dependencia entre el tamaño de la ventana de contexto y la eficacia en la clasificación. Aunque existen diferencias a cada variación de dicho tamaño éstas no son significativas. Para apreciar mejor la tendencia, la Figura 5.4a muestra los resultados medios de precisión, cobertura y medida-F, y la Figura 5.4b los valores medio de correlación de manera análoga.

En dicha Figura 5.4 se puede observar cómo, aunque existe una ligera tendencia creciente, el tamaño de la ventana de contexto apenas influye en la eficacia de la clasificación. Los experimentos corroboran los resultados obtenidos por [Lemaire y Denhière, 2004] que demostraban que el tamaño óptimo de la ventana de contexto para construir una red de asociación de conceptos se encuentra entre 11 y 15 palabras. Sin embargo, los experimentos citados mostraban también una relación directa de la eficacia con el tamaño de la ventana, contrariamente a los resultados presentes en la Figura 5.4, lo que pone de manifiesto que la utilización del modelo de lectura con intervalo de

olvido variable hace al sistema SILC independiente del contexto de asociación. El resultado máximo (ver Tabla 5.4) de medida-F es 65.60 con un tamaño de ventana de contexto igual a 17, y el de correlación 0.481 con un tamaño de ventana de contexto igual a 11.



a)



b)

Figura 5.4. Valores medios de a) precisión ,cobertura y medida-F y b) correlación en tareas de clasificación de textos representados por SILC empleando diferentes tamaños fijos de ventana de contexto.

Si se observan los resultados de los experimentos anteriores de optimización de los parámetros del modelo de lectura, donde se ha empleado una ventana de contexto de tamaño variable, los resultados máximos de medida-F y correlación, utilizando los mismos valores óptimos para el resto de parámetros, son 65.95 y 0.494 respectivamente, lo que demuestra que el contexto que definen las propias unidades gramaticales (oraciones) es apropiado por sí mismo para capturar las relaciones semánticas entre las palabras sin necesidad de definir el contexto de manera artificial.

5.2.5.2. Optimización de la cantidad de textos fuente

En lo que respecta a la construcción del conocimiento semántico lingüístico sobre el que opera el modelo de lectura del sistema SILC, la influencia del número de textos a partir de los cuales se construye dicho conocimiento ha sido también objeto de la experimentación. Una vez establecidos todos los parámetros con los valores óptimos hallados en los experimentos anteriores, se ha variado el conjunto fuente de textos para construir la red conceptual. En los experimentos anteriores de optimización dicha red se construyó siempre a partir de la porción de entrenamiento correspondiente junto con los textos de cultura general de Orellana. En esta ocasión, se ha aplicado el mismo procedimiento experimental utilizando, por un lado, sólo al subconjunto de entrenamiento y, por otro, sólo a los textos de cultura general para la construcción del conocimiento semántico lingüístico. La Tabla 5.5 muestra los resultados de precisión, cobertura, medida-F y correlación para todas las categorías y los tres algoritmos de clasificación, con distintos conjuntos de textos para la construcción de la red conceptual. Los resultados de la tabla muestran una ligera relación directa entre el tamaño del conjunto de entrenamiento y la eficacia en la tarea de clasificación.

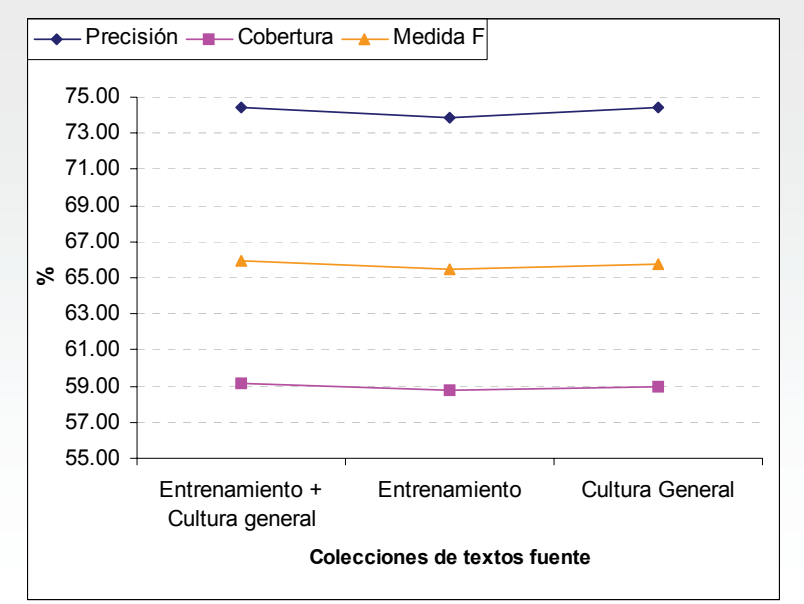
Conjunto de textos fuente = **Porción de entrenamiento + Colección de cultura general**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.78	72.22	75.36	53.33	66.67	59.26	20.93	100.00	34.62	51.01	79.63	62.19
Cultura	87.87	80.56	84.06	57.77	72.22	64.19	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.11	94.59	93.33	79.07	91.89	85.00	0.00	0.00	0.00	57.06	62.16	59.50
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.84*	*67.54*	*69.63*	*64.19*	*22.92*	*33.78*	*74.44*	*59.19*	*65.95*
Correlación	*0.840*			*0.478*			*0.164*			*0.494*		

Conjunto de textos fuente = **Porción de entrenamiento**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.78	72.22	75.36	48.89	61.11	54.32	20.93	100.00	34.62	49.53	77.78	60.52
Cultura	87.87	80.56	84.06	57.78	72.22	64.20	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.11	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.69	62.16	59.84
Economía	80.95	94.44	87.18	83.87	72.22	77.61	100.00	2.78	5.41	88.27	56.48	68.89
Salud	96.77	93.75	95.24	78.57	34.38	47.83	100.00	6.25	11.76	91.78	44.79	60.20
Media	*87.30*	*87.11*	*87.20*	*70.01*	*66.36*	*68.14*	*64.19*	*22.92*	*33.78*	*73.83*	*58.80*	*65.46*
Correlación	*0.840*			*0.408*			*0.146*			*0.465*		

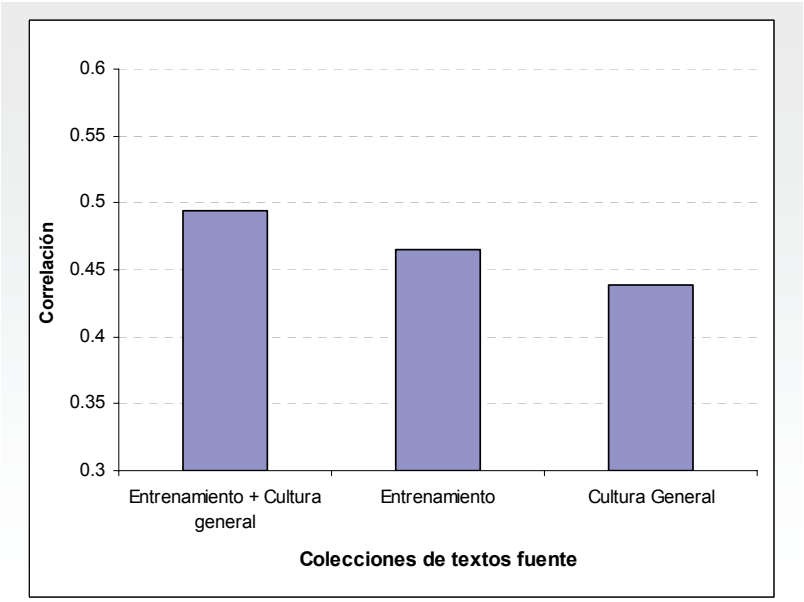
Conjunto de textos fuente = **Colección de cultura general**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	75.00	75.00	75.00	50.00	66.67	57.14	21.05	100.00	34.78	48.68	80.56	60.69
Cultura	90.62	80.56	85.29	59.57	77.78	67.47	100.00	5.56	10.53	83.40	54.63	66.02
Deportes	92.11	94.59	93.33	80.95	91.89	86.07	0.00	0.00	0.00	57.69	62.16	59.84
Economía	80.00	88.89	84.21	88.89	66.67	76.19	100.00	5.56	10.53	89.63	53.71	67.17
Salud	93.55	90.62	92.06	84.61	34.38	48.89	100.00	6.25	11.76	92.72	43.75	59.45
Media	*86.26*	*85.93*	*86.09*	*72.80*	*67.48*	*70.04*	*64.21*	*23.47*	*34.38*	*74.42*	*58.96*	*65.80*
Correlación	*0.778*			*0.382*			*0.157*			*0.439*		

Tabla 5.5. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados por SILC bajo un conocimiento semántico construido a partir de diferentes colecciones de textos.

La Figura 5.5 presenta un resumen de la tabla anterior con los valores medios absolutos de precisión, cobertura y medida-F (Figura 5.5a), e índice de correlación (Figura 5.5b). Efectivamente, en términos de medida-F, parece existir una relación directa entre el tamaño de las colecciones de textos fuente y la eficacia de clasificación, ya que el mejor resultado lo obtiene la colección formada por la porción de entrenamiento más los textos de cultura general (885 textos, 335 del 67% de entrenamiento más 500 de cultura general). A continuación, le sigue la colección formada sólo por los ensayos de cultura general (500 textos) y, finalmente, los peores resultados los obtiene la colección formada sólo por las porciones de entrenamiento (335 textos).



a)



b)

Figura 5.5. Valores medios de a) precisión ,cobertura y medida-F y b) correlación en tareas de clasificación de textos representados por SILC empleando diferentes colecciones de textos para construir el conocimiento semántico lingüístico.

Si se observa la Figura 5.5b, se aprecia que en términos de correlación no se da la relación directa entre el tamaño de la colección y la eficacia de clasificación. Esto prueba que no sólo el número de textos empleado en la construcción del conocimiento semántico lingüístico influye en el modelo, sino también los tipos de textos. En este

caso, los textos de entrenamiento están focalizados en las categorías implicadas en la tarea de clasificación por lo que obtienen un índice de correlación mayor que los textos de cultura general que, a pesar de superar en número a la porción de entrenamiento, son mucho más dispersos con respecto a las temáticas objetivo y pueden carecer de muchos de los conceptos necesarios para representar la semántica de los textos de manera precisa. Se puede concluir pues que los textos fuente no relacionados con los temas de clasificación son un buen complemento para los textos fuente de entrenamiento, pero por sí solos generalizan demasiado las representaciones obtenidas por el modelo de lectura.

Como resumen de todos los experimentos de optimización de parámetros y aspectos de diseño del sistema SILC, la Tabla 5.6 presenta todos los valores óptimos obtenidos. Dichos valores son los que serán empleados posteriormente por el sistema para comparar su eficacia en tareas de clasificación con la de otros sistemas existentes.

Parámetros	Valor óptimo
Umbral mínimo de propagación	0.3
Umbral de olvido	0.3
Factor de olvido	0.9
Intervalo de olvido	Variable (oraciones)
Método de inferencia	En profundidad
Nivel máximo de propagación	3
Ventana de contexto	Variable (oraciones)

Tabla 5.6. Valores óptimos en tareas de clasificación de textos para los parámetros del sistema SILC.

5.3. Evaluación y Optimización de las Medidas de Similitud Semántica Empleadas por SILC

La serie de experimentos descrita a continuación tiene dos objetos de evaluación. El primero es el tipo de medida de similitud y sus parámetros, de entre las tres propuestas en el Capítulo 4, que mejor discrimina la semántica principal de los textos. El otro aspecto a evaluar son las diferentes opciones para la construcción del contexto de comparación, como son el número de puentes y el tipo de los mismos.

5.3.1. Conjunto de datos

La colección de textos empleada para la experimentación con las medidas de similitud es la misma que la empleada en la optimización de los parámetros descrita en la Sección 5.2.1.1, compuesta por las 500 noticias del servicio *Google News* y de los 500 ensayos de cultura general de Orellana. Estas colecciones se han empleado para construir el conocimiento semántico lingüístico sobre el que opera el modelo de lectura. Para la validación propiamente dicha de los aspectos relacionados con las medidas de similitud, se han utilizado además 50 noticias de *Google News*, 10 de cada una de las cinco categorías escogidas (Ciencia y Tecnología, Cultura y Espectáculos, Deportes, Economía y Salud), no incluidas en las colecciones anteriormente citadas.

5.3.2. Medidas de evaluación

Se ha definido una nueva medida que permite evaluar el poder discriminatorio de las medidas de similitud. Dicha medida está basada en otros dos nuevos valores ideados para la evaluación llevada a cabo: la similitud "intraclase" y la similitud "interclase", que se definen para cada categoría. La similitud "intraclase" es la similitud media, según la medida de similitud evaluada, entre todos los textos de una misma categoría (Ecuación 5.4).

$$\text{Intraclase (Categoría}_k) = \frac{\sum_{i,j}^{n} \text{sim}(t_i, t_j)}{\binom{n}{2}} \qquad t_i, t_j \in \text{Categoría}_k \wedge i \neq j \tag{5.4}$$

donde *n* es el número de textos en la categoría *k*, t_i y t_j son textos de dicha categoría, y *sim* es la función de similitud evaluada. Nótese que el denominador de la expresión hace referencia a las combinaciones sin repetición de *n* elementos tomadas de dos en dos, es decir, a cada posible comparación entre textos diferentes de la misma categoría.

La similitud "interclase" es la similitud media entre cada uno de los documentos de una categoría y el resto de documentos de la colección correspondientes al resto de categorías (Ecuación 5.2).

$$\text{Interclase (Categoría}_k) = \frac{\sum_{i,j}^{n,m-n} \text{sim}(t_i, t_j)}{n \cdot (m-n)} \qquad t_i \in \text{Categoría}_k \wedge t_j \notin \text{Categoría}_k \tag{5.5}$$

donde *n* es el número de textos en la categoría *k*, *m* es el número de textos que componen la colección de evaluación, t_i son textos de dicha categoría y t_j textos del resto de categorías, y *sim* es la función de similitud evaluada.

Así, una medida de similitud será mejor cuanto mayor sea la similitud "intraclase" y menor sea la similitud "interclase". Para cuantificar dicha relación se ha definido la medida diferencia "intraclase-interclase", con respecto a una determinada categoría, como la similitud "intraclase" menos la similitud "interclase" normalizada por la similitud "intraclase", como indica la Ecuación 5.6.

$$\text{Diferencia (Categoría}_k) = \frac{Intraclase(Categoría_k) - Interclase(Categoría_k)}{Intraclase(Categoría_k)} \tag{5.6}$$

Dado que los valores absolutos pueden diferir mucho entre las diferentes categorías, la normalización de la diferencia pone de manifiesto la proporción en la que ambas medidas difieren con respecto a la similitud "intraclase".

5.3.3. Procedimiento experimental

Dado el conocimiento semántico lingüístico construido a partir de los textos de noticias de *Google* y de los ensayos de cultura general de Orellana, se representan las

otras 50 noticias mediante el modelo de lectura de SILC con los parámetros óptimos obtenidos previamente. Una vez representados dichos textos, se calcula para cada categoría la diferencia "intraclase-interclase" expresada por la Ecuación 5.6 y se calcula la media entre todas las categorías. El cálculo de la diferencia "intraclase-interclase" se realiza para cada tipo de contexto de comparación, para cada valor de los parámetros variados y para cada una de las medidas de similitud propuestas.

5.3.4. Evaluación de los aspectos de construcción del espacio de comparación

Las medidas de similitud propuestas en el Capítulo 4 están basadas en la similitud entre conceptos individuales. Para calcular dicha similitud local se construye un subconjunto del conocimiento semántico lingüístico total. Dicho subconjunto está formado por las subredes que representan a los textos que se comparan y por puentes o caminos de nodos conceptuales que unen los conceptos más relevantes de dichas representaciones. Los aspectos a tener en cuenta en la construcción del espacio o contexto de comparación, objeto de los siguientes experimentos, están relacionados con los mencionados puentes. El primero de ellos es el número de los conceptos más activos de cada una de las dos representaciones que se comparan entre los que se establecen dichos nexos, denominados anclas. Utilizando la medida de similitud sim_T^1, con un parámetro de k igual a 5, se ha variado el número de anclas estableciéndolo a 5, 15, 40 y 75 (casi ninguna de las representaciones de los textos de las colecciones descritas generadas con SILC, utilizando los parámetros óptimos, está compuesta por más de 75 conceptos). En este caso, se establecen puentes por pares ordenados por nivel de activación. Para cada valor, se han calculado la similitud "intraclase", "interclase" y su diferencia normalizada. La Tabla 5.7 muestra estos valores ("intraclase" en la diagonal) para cada categoría y la media de todas ellas.

Medida de similitud = sim_T^1 k = **5**					Número de anclas = **5**	
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0000351	0.0000549	0.0000276	0.0000533	0.0000303	*-0.183*
Cultura	0.0000372	0.0000318	0.0000214	0.0000313	0.0000180	*0.150*
Deportes	0.0000310	0.0000358	0.0000728	0.0000329	0.0000187	*0.593*
Economía	0.0000527	0.0000440	0.0000265	0.0000950	0.0000279	*0.602*
Salud	0.0000331	0.0000319	0.0000212	0.0000332	0.0000164	*-0.826*
					Media →	*0.067*

Medida de similitud = sim_T^1	k = **5**				Número de anclas = **15**
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0000394	0.0000549	0.0000279	0.0000534	0.0000307	*-0.059*
Cultura	0.0000376	0.0000319	0.0000214	0.0000313	0.0000180	*0.151*
Deportes	0.0000313	0.0000359	0.0000729	0.0000330	0.0000191	*0.591*
Economía	0.0000537	0.0000441	0.0000265	0.0000979	0.0000281	*0.611*
Salud	0.0000331	0.0000319	0.0000213	0.0000335	0.0000164	*-0.827*
					Media →	*0.093*

Medida de similitud = sim_T^1	k = **5**				Número de anclas = **40**
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0000396	0.0000564	0.0000284	0.0000543	0.0000310	*-0.073*
Cultura	0.0000381	0.0000335	0.0000218	0.0000326	0.0000206	*0.155*
Deportes	0.0000315	0.0000391	0.0000731	0.0000331	0.0000193	*0.579*
Economía	0.0000552	0.0000442	0.0000266	0.0000984	0.0000297	*0.604*
Salud	0.0000332	0.0000323	0.0000217	0.0000338	0.0000167	*-0.807*
					Media →	*0.092*

Medida de similitud = sim_T^1	k = **5**				Número de anclas = **75**
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0000398	0.0000565	0.0000286	0.0000553	0.0000311	*-0.078*
Cultura	0.0000381	0.0000338	0.0000218	0.0000332	0.0000208	*0.157*
Deportes	0.0000317	0.0000392	0.0000733	0.0000335	0.0000196	*0.577*
Economía	0.0000553	0.0000444	0.0000268	0.0000993	0.0000302	*0.606*
Salud	0.0000336	0.0000324	0.0000218	0.0000345	0.0000181	*-0.687*
					Media →	*0.115*

Tabla 5.7. Valores de similitud "intraclase", "interclase" y diferencia normalizada de ambas para distintos valores del número de conceptos ancla entre los que se construyen puentes, utilizando la medida de similitud sim_T^1.

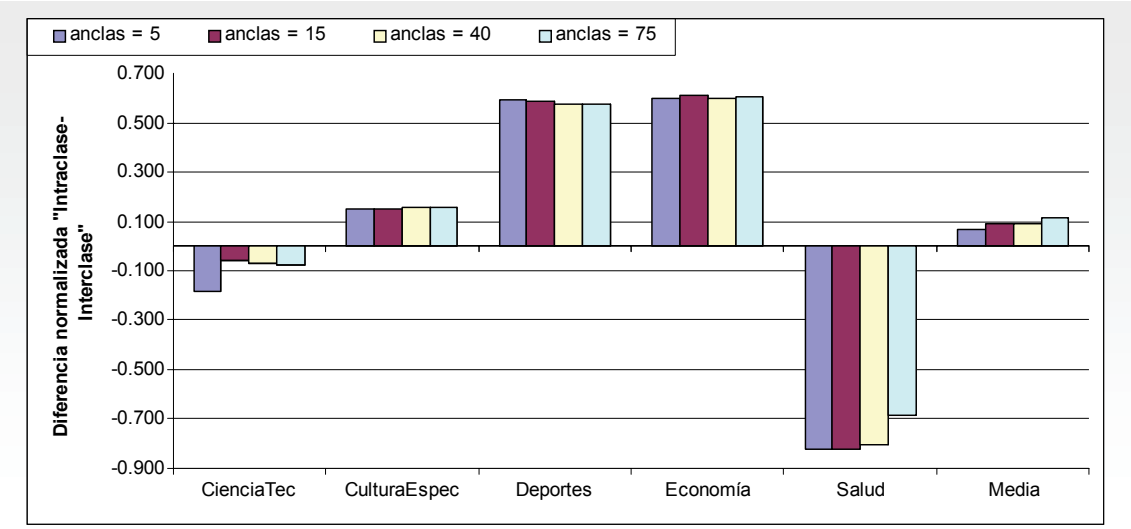


Figura 5.6. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para diferentes valores del número de conceptos ancla utilizando la medida de similitud sim_T^1.

La Figura 5.6 muestra de manera gráfica los resultados de la Tabla 5.7. Como se puede observar, la influencia del número de puentes es muy leve, aunque parece existir una ligera tendencia que indica que cuantos más puentes se utilicen mejor se discriminan las categorías.

Una vez determinado en número óptimo de conceptos ancla, el otro aspecto relacionado con la construcción de los puentes es la manera en la que se conectan dichas anclas. En el experimento anterior, los nodos anclas se conectaban por pares ordenados por su nivel de activación, de tal forma que se construyen el mismo número de puentes que anclas se determinan. La alternativa mencionada en el Capítulo 4 es la conexión de los anclas N a N de manera cruzada, es decir, todos los de una representación con todos los de la otra. Así, utilizando la misma medida de similitud sim_T^1 con el mismo valor del parámetro k igual a 5 y un número de anclas igual a 75, la Tabla 5.8 presenta los resultados análogos a la Tabla 5.7 para el establecimiento de los puentes de manera cruzada.

Medida de similitud = sim_T^1 k = **5** Número de anclas = **75** Puentes = **cruzados**						
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0072588	0.0123353	0.0058695	0.0103809	0.0087154	*-0.285*
Cultura	0.0091450	0.0093092	0.0058212	0.0071576	0.0062179	*0.239*
Deportes	0.0071617	0.0094327	0.0271421	0.0067226	0.0058844	*0.731*
Economía	0.0111129	0.0099672	0.0055112	0.0212175	0.0088508	*0.582*
Salud	0.0075786	0.0095555	0.0063075	0.0091475	0.0063548	*-0.282*
					Media →	*0.197*

Tabla 5.8. Valores de similitud "intraclase", "interclase" y diferencia normalizada utilizando la medida de similitud sim_T^1 estableciendo los puentes N a N de manera cruzada.

Como se aprecia en la Figura 5.7, que presenta una comparativa en términos de diferencia "intraclase-interclase" normalizada de los dos tipos de construcción de puentes, el establecimiento de puentes de manera cruzada obtiene mejores resultados de discriminación de las categorías, puesto que es más flexible que el establecimiento de puentes por pares ordenados por activación.

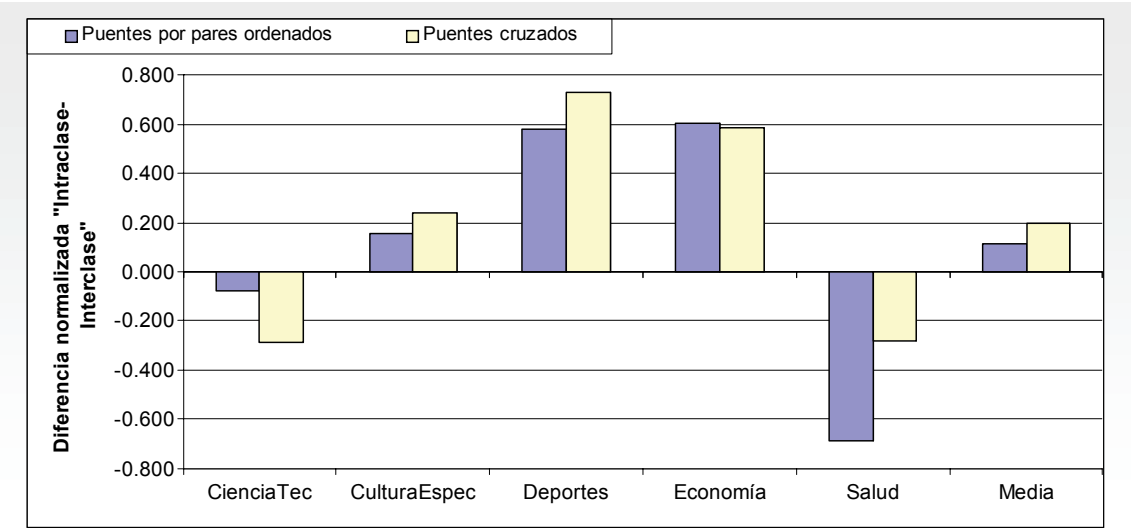


Figura 5.7. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para el establecimiento de puentes por pares ordenados por activación y de manera cruzada, utilizando la medida de similitud sim_T^1.

La reducción del contexto o espacio de comparación con respecto al conocimiento semántico lingüístico global supone, por definición, un beneficio en términos de coste computacional. El siguiente experimento trata de evaluar si esta reducción es también beneficiosa en términos de eficacia. Así pues, se han calculado nuevamente las similitudes "intraclase", "interclase" y su diferencia normalizada utilizando la medida de similitud sim_T^1, construcción de puentes cruzados con 75 anclas y empleando como espacio de comparación el conocimiento global. La Tabla 5.9 presenta los valores mencionados.

Medida de similitud = sim_T^1 k = **5** Número de anclas = **75** Puentes **cruzados** Contexto = **Global**						
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0008336	0.0027483	0.0009855	0.0040793	0.0018600	*-1.901*
Cultura	0.0028002	0.0005981	0.0006030	0.0006077	0.0007222	*-0.978*
Deportes	0.0004521	0.0004023	0.0205557	0.0006111	0.0005378	*0.976*
Economía	0.0043143	0.0005153	0.0005506	0.0154342	0.0032686	*0.860*
Salud	0.0016595	0.0003661	0.0006981	0.0031671	0.0009317	*-0.581*
					Media →	*-0.325*

Tabla 5.9. Valores de similitud "intraclase", "interclase" y diferencia normalizada utilizando la medida de similitud sim_T^1, estableciendo los puentes de manera cruzada y empleando como contexto de comparación el conocimiento semántico lingüístico en su totalidad.

El hecho de que el valor medio de la diferencia normalizada sea negativo implica que el uso del conocimiento global como ámbito de comparación no es adecuado. Si comparamos los resultados obtenidos para un contexto de comparación reducido (Figura 5.8) se observa como éste último hace que la medida de similitud discrimine positivamente en media y, desde luego, significativamente mejor que empleando el conocimiento semántico en su totalidad. Esto se debe a que la reducción del espacio de comparación, al estar restringido a los contextos definidos por las temáticas de los propios textos que se comparan y la relación entre las mismas, reduce la ambigüedad de los conceptos y por lo tanto se obtienen medidas de similitud local entre los textos más precisas.

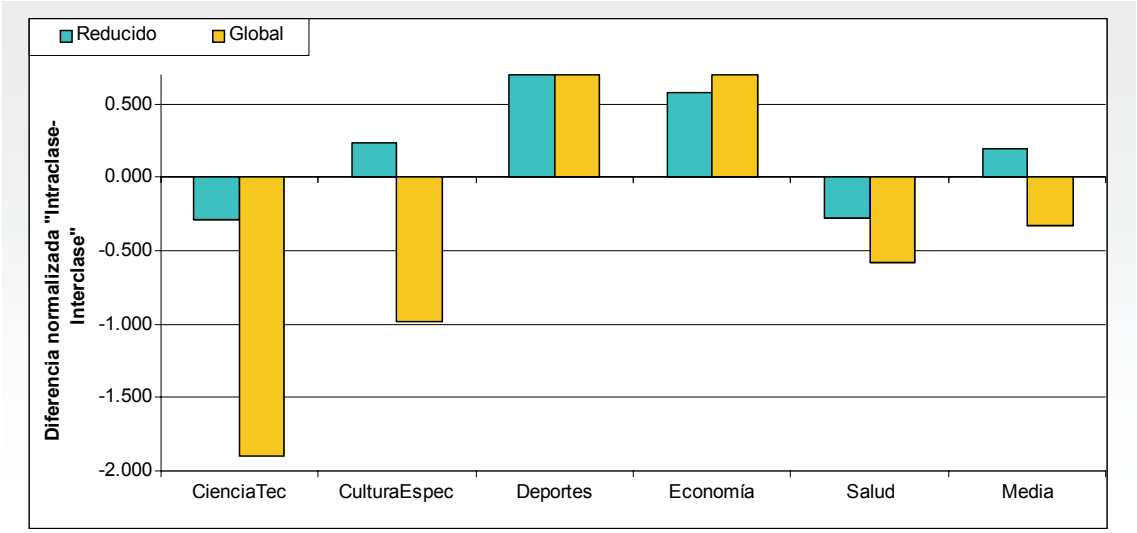


Figura 5.8. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para los espacios de comparación reducido y global, utilizando la medida de similitud sim_T^1.

Así pues, los resultados muestran que el contexto de comparación óptimo es el reducido, con el máximo número de conceptos ancla posible y estableciendo los puentes de manera cruzada entre los mismos.

5.3.5. Evaluación de los tipos de similitud y sus parámetros

Una vez determinado el contexto óptimo de similitud se han evaluado las diferentes medidas de similitud propuestas en el Capítulo 4 y los aspectos que las conciernen. En

primer lugar, para realizar una comparación del poder de discriminación de dichas medidas se realizó el cálculo de la similitud "intraclase" e "interclase" y su diferencia, para cada uno de los tres tipos de similitud, sim_T^1, $sim_T^{1'}$ y sim_T^2, variando el parámetro k o número de los conceptos más activos de las representaciones que intervienen en la comparación. Los valores asignados a dicho parámetro fueron 5, 15, 40 y 75 (como se mencionó en la sección inmediatamente anterior, apenas alguna de las representaciones producidas con los parámetros óptimos del modelo de lectura posee más de 75 conceptos). Se emplearon los cinco conceptos más activos de cada representación para construir puentes de manera cruzada, y no se utilizó normalización de la activación. La siguiente Tabla 5.10 muestra las similitudes "intraclase" (valores en las diagonales), "interclase" (resto de valores) y la diferencia normalizada entre ambas para los distintos valores de k utilizando la medida de similitud sim_T^1.

Medida de similitud = sim_T^1						$k = 5$
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0072588	0.0123353	0.0058695	0.0103809	0.0087154	*-0.285*
Cultura	0.0091450	0.0093092	0.0058212	0.0071576	0.0062179	*0.239*
Deportes	0.0071617	0.0094327	0.0271421	0.0067226	0.0058844	*0.731*
Economía	0.0111129	0.0099672	0.0055112	0.0212175	0.0088508	*0.582*
Salud	0.0075786	0.0095555	0.0063075	0.0091475	0.0063548	*-0.282*
					Media →	*0.197*

Medida de similitud = sim_T^1						$k = 15$
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0004477	0.0004548	0.0001916	0.0003070	0.0001737	*0.371*
Cultura	0.0003006	0.0003326	0.0001304	0.0002286	0.0001287	*0.408*
Deportes	0.0003608	0.0003575	0.0002094	0.0002585	0.0001520	*-0.347*
Economía	0.0004039	0.0004205	0.0001772	0.0003374	0.0001661	*0.135*
Salud	0.0003289	0.0003337	0.0001711	0.0002406	0.0001207	*-1.225*
					Media →	*-0.132*

Medida de similitud = sim_T^1						$k = 40$
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0026755	0.0025720	0.0014731	0.0019046	0.0016171	*0.293*
Cultura	0.0020211	0.0018711	0.0010412	0.0013641	0.0011203	*0.259*
Deportes	0.0023171	0.0020843	0.0013258	0.0015746	0.0013013	*-0.372*
Economía	0.0026725	0.0023564	0.0013795	0.0018200	0.0014779	*-0.083*
Salud	0.0022005	0.0019072	0.0010942	0.0014081	0.0012065	*-0.370*
					Media →	*-0.055*

Medida de similitud = sim_T^1						k = 75
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0045143	0.0043210	0.0030012	0.0036028	0.0030665	*0.225*
Cultura	0.0041835	0.0040358	0.0027309	0.0032628	0.0027579	*0.199*
Deportes	0.0042241	0.0039337	0.0032670	0.0034377	0.0028296	*-0.104*
Economía	0.0046816	0.0042520	0.0031377	0.0038934	0.0031382	*0.023*
Salud	0.0043060	0.0039243	0.0027407	0.0033154	0.0028805	*-0.240*
					Media →	*0.021*

Tabla 5.10. Valores de similitud "intraclase", "interclase" y diferencia normalizada de ambas para distintos valores del número k de conceptos más activos que intervienen en la comparación utilizando la medida de similitud sim_T^1.

La Tabla 5.10 puede verse resumida en la Figura 5.9, que muestra la diferencia normalizada de cada categoría y la media de todas ellas, para la medida sim_T^1 y los diferentes valores del parámetro k.

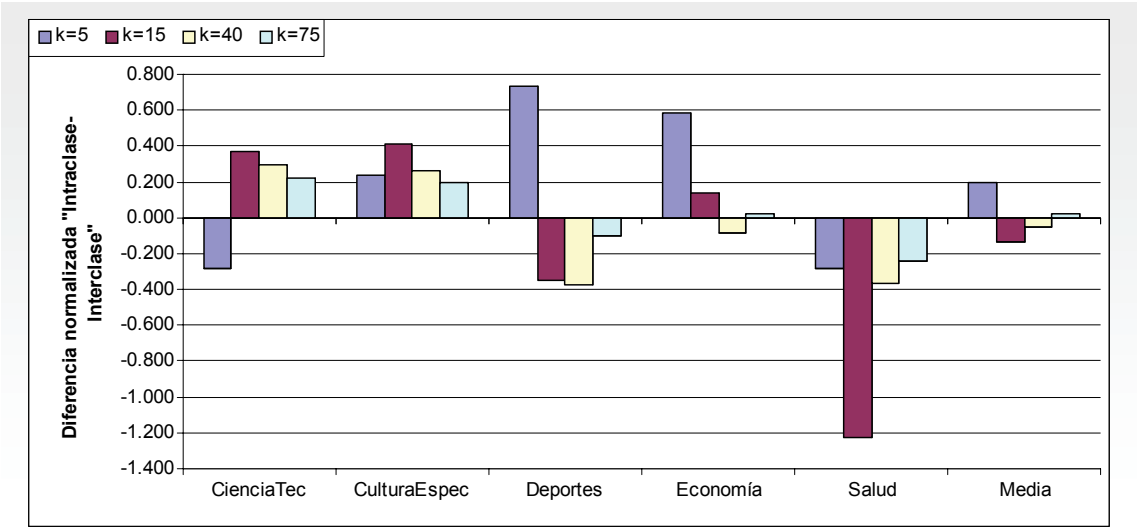


Figura 5.9. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para diferentes valores del parámetro k de la medida de similitud sim_T^1.

Como se aprecia en la Figura 5.9, el número óptimo de conceptos más activos que participan en el proceso de comparación es igual a 5. Salvo en las categorías "Ciencia y Tecnología" y "Salud", dicho valor de k hace que la medida sim_T^1 discrimine positivamente el resto de categorías. Como se aprecia, los resultados son pobres para la categoría "Salud" con cualquier valor del parámetro. Esto es debido a que los documentos de dicha categoría contienen una gran cantidad de términos técnicos que

determinan su temática y que no existen en el conocimiento semántico lingüístico construido, por lo que la similitud se calcula utilizando conceptos conocidos más generales y por lo tanto ambiguos. Lo mismo sucede en la categoría "Ciencia y Tecnología", aunque más levemente.

Medida de similitud = $sim_T^{1'}$						$k = 5$
	Similitud "intraclase" e "interclase"					*Diferencia "intra-inter"*
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.0013550	0.0030516	0.0010764	0.0035179	0.0012128	*-0.634*
Cultura	0.0023870	0.0019073	0.0011233	0.0013507	0.0011882	*0.207*
Deportes	0.0014118	0.0018630	0.0016415	0.0012975	0.0011075	*0.135*
Economía	0.0036002	0.0019361	0.0010691	0.0014455	0.0012285	*-0.355*
Salud	0.0013427	0.0018278	0.0011218	0.0012464	0.0011824	*-0.171*
					Media →	*-0.164*

Tabla 5.11. Valores de similitud "intraclase", "interclase" y diferencia normalizada utilizando la medida de similitud $sim_T^{1'}$ y los 5 conceptos más activos de cada representación de los textos.

Así pues, utilizando el valor 5 para el parámetro *k* se llevó a cabo el mismo proceso experimental para la variante de la medida anteriormente evaluada, en este caso $sim_T^{1'}$, que comparaba los conceptos más activos por pares ordenados por activación. La Tabla 5.11 muestra de manera análoga los valores de similitud "intraclase", "interclase" y diferencia normalizada para dicha variante.

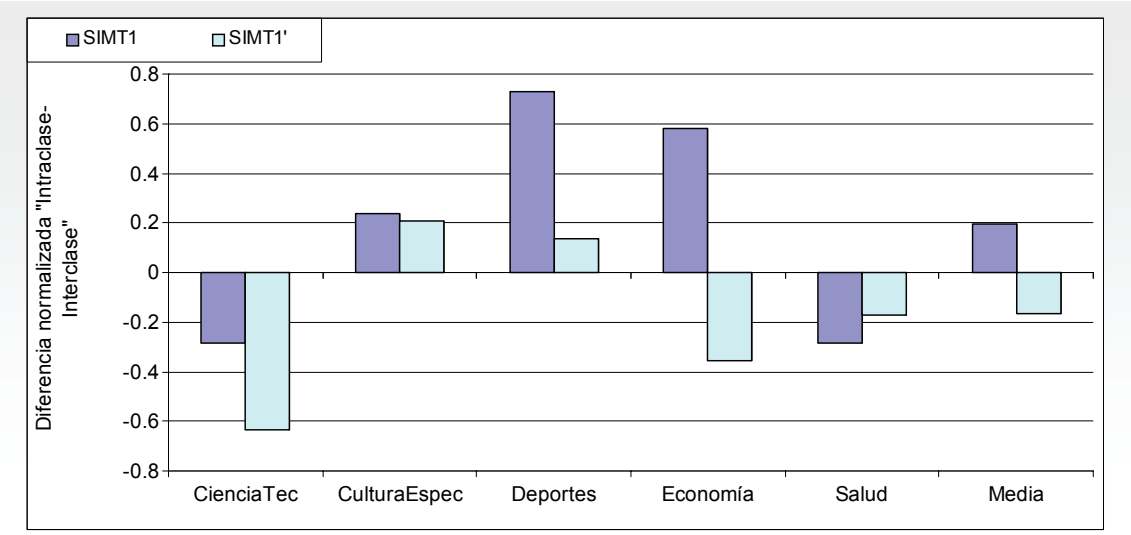


Figura 5.10. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para medida de similitud sim_T^1 y su variante $sim_T^{1'}$ con un utilizando los 5 conceptos más activos de la representaciones textuales.

En la Tabla 5.11 se observa cómo la variante de la medida sim_T^1 no supone una mejora de ésta. Esta diferencia se aprecia claramente en la Figura 5.10, donde se muestra una comparación de las dos medidas en términos de la diferencia normalizada. Los resultados demuestran que se obtiene una medida de similitud más precisa si se comparan todos los conceptos más activos de las representaciones entre sí que si se comparan por parejas ordenadas por el nivel de activación.

A continuación, se ha evaluado la medida de similitud sim_T^2 siguiendo el mismo procedimiento. Dicha medida cuantifica la diferencia de activación entre los pares de conceptos más similares de los textos comparados. La Tabla 5.12 presenta los valores "intraclase", "interclase" y su diferencia normalizada para la medida mencionada utilizando distintos valores del parámetro *k* (5, 15, 40 y 75).

Medida de similitud = sim_T^2					k = 5
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	1.0955717	0.9055827	0.6116044	0.9384926	0.9484483	*-0.223*
Cultura	0.7610125	0.9320553	0.5441136	0.6877058	0.7272930	*-0.270*
Deportes	0.8271252	0.7433538	1.0185267	0.4729538	0.7609248	*-0.312*
Economía	0.8880539	0.8407221	0.7241448	1.4561376	1.6057339	*-0.303*
Salud	0.9280229	0.6287310	0.8433707	1.1798141	0.8503347	*0.053*
					Media →	*-0.211*

Medida de similitud = sim_T^2					k = 15
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	1.0329839	0.8992877	1.2437563	1.2301756	1.0187514	*0.063*
Cultura	0.9774422	0.7673582	1.1015941	0.9853513	0.9446236	*0.306*
Deportes	0.9265007	0.7753845	0.8951247	1.0331056	0.9239098	*0.022*
Economía	1.3481489	1.1321525	1.3792425	1.1814982	1.2012995	*0.071*
Salud	1.0983724	0.9497231	1.2758398	1.2255002	0.9370616	*0.214*
					Media →	*0.135*

Medida de similitud = sim_T^2					k = 40
	Similitud "intraclase" e "interclase"				***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.5241084	0.5277579	0.7331466	0.9575319	0.5332733	*0.313*
Cultura	0.5258081	0.5793575	0.7867413	0.7534373	0.5523262	*0.130*
Deportes	0.4783719	0.4770678	0.7272086	0.7243482	0.4961126	*-0.252*
Economía	0.7093741	0.5808665	0.7923988	0.8778145	0.6181029	*-0.231*
Salud	0.5082704	0.5444494	0.7675176	0.8445588	0.5729300	*0.163*
					Media →	*0.024*

Medida de similitud = sim_T^2						k = 75
	Similitud "intraclase" e "interclase"					*Diferencia "intra-inter"*
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.4636051	0.5114240	0.5682253	0.8751686	0.4357378	*0.289*
Cultura	0.4370126	0.4882592	0.6211370	0.6936976	0.4317231	*0.118*
Deportes	0.4461035	0.4624787	0.5597224	0.6628286	0.4176514	*-0.112*
Economía	0.6618836	0.5807122	0.6189151	0.7623707	0.5182117	*-0.220*
Salud	0.4351188	0.4851889	0.5817001	0.7513311	0.4265246	*0.321*
					Media →	*0.079*

Tabla 5.12. Valores de similitud "intraclase", "interclase" y diferencia normalizada de ambas para distintos valores del número k de conceptos más activos que intervienen en la comparación utilizando la medida de similitud sim_T^2.

Los valores de la Tabla 5.12 indican que el valor óptimo de k es igual a 15 en este caso. Así pues, la medida sim_T^2 necesita un mayor número de los conceptos más activos que la medida sim_T^1 para alcanzar el nivel máximo de discriminación entre categorías. Sin embargo, dicho nivel es mayor para la medida sim_T^2, que no obtiene diferencia negativa para ninguna de las categorías, es decir, que la similitud "intraclase" es siempre mayor que la similitud "interclase", como se aprecia en la Figura 5.11, que muestra una vez más los valores de la diferencia normalizada para cada categoría y la media de todas ellas utilizando la medida sim_T^2 y los diferentes valores del parámetro k.

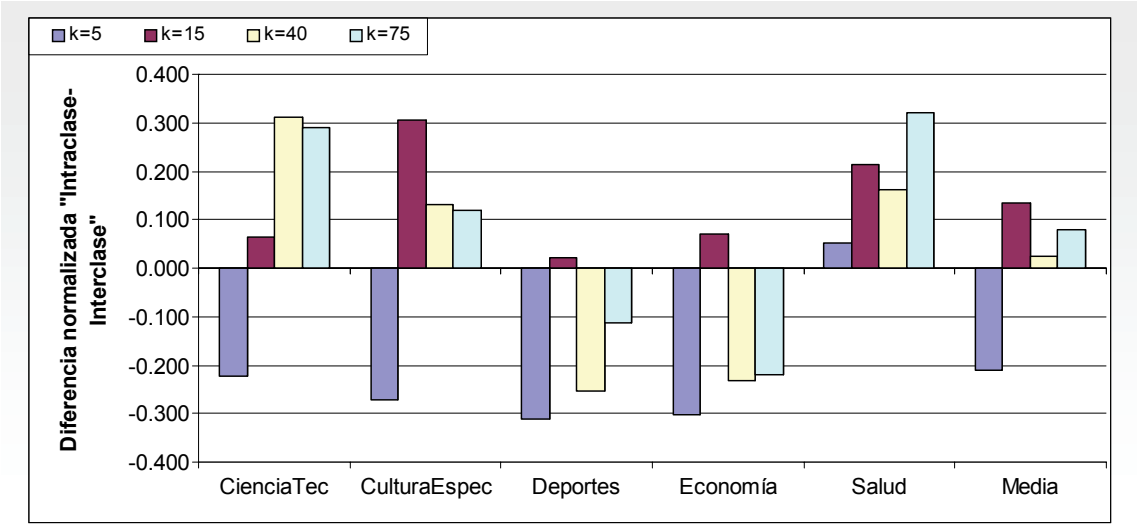


Figura 5.11. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para diferentes valores del parámetro k de la medida de similitud sim_T^2.

Aunque los resultados para valores de k elevados se muestran algo aleatorios, sí parece claro que valores muy reducidos (igual a 5) perjudican en gran medida el poder discriminatorio de la medida. En la Figura 5.11 se puede apreciar también que para cualquier valor de k distinto de 15, la medida sim_T^2 discrimina eficazmente algunas categorías pero no identifica en absoluto otras. Si se observa la Figura 5.12, donde se muestra la diferencia normalizada de la medida sim_T^1 con su valor óptimo de k igual a 5 y la de la medida sim_T^2 con su valor óptimo de k igual a 15, se aprecia como ambas medidas se complementan en cierto modo, una discriminado correctamente las categorías que la otra no es capaz de distinguir y viceversa.

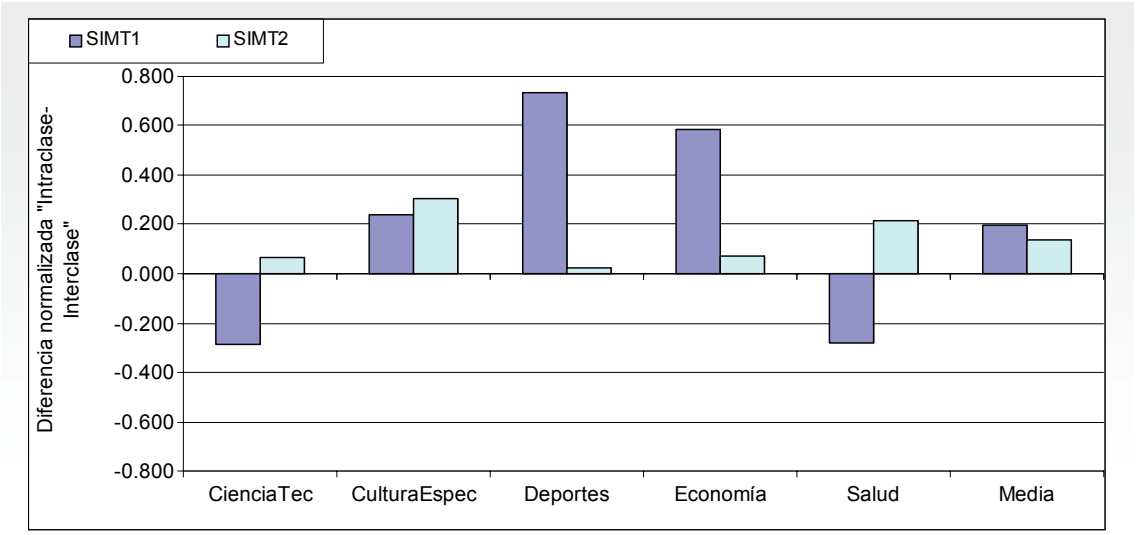


Figura 5.12. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para las medidas de similitud sim_T^1 y sim_T^2 utilizando sus valores óptimos de k igual a 5 y 15, respectivamente.

Como ya se ha comentado, la medida sim_T^2 obtiene una diferencia positiva para todas las categorías. La medida sim_T^1, por el contrario, es más dependiente del vocabulario representativo de las diferentes categorías, obteniendo diferencias negativas en algunas de ellas pero distinguiendo otras de manera mucho más precisa que sim_T^2, lo que la hace alcanzar una diferencia normalizada media mayor. Este hecho hace suponer que, con una colección de textos fuente relativamente extensa, la medida sim_T^1 tendrá la precisión suficiente para distinguir con claridad la semántica de los textos comparados a través de la misma.

Para finalizar, se ha evaluado la influencia de la normalización del nivel de activación de los conceptos en el cálculo de la similitud. Como se explica en el Capítulo 4, dicha normalización consiste en la división de los niveles de activación por el máximo nivel en cada representación. Para la experimentación se ha seguido el mismo procedimiento de comparaciones empleando la medida de similitud sim_T^2, puesto que dicha medida está esencialmente basada en el nivel de activación de los conceptos, con el valor óptimo del parámetro *k* obtenido en los experimentos anteriores. En dichos experimentos anteriores se empleaban los niveles de activación sin normalizar. En esta ocasión, la Tabla 5.13 muestra de manera análoga las similitudes "intraclase", "interclase" y su diferencia empleando niveles de activación normalizados.

Medida de similitud = sim_T^2	Activación normalizada					k = 15
	Similitud "intraclase" e "interclase"					***Diferencia "intra-inter"***
	Ciencia	**Cultura**	**Deportes**	**Economía**	**Salud**	
Ciencia	0.2819211	0.2348371	0.3088189	0.2901930	0.2776486	*-0.014*
Cultura	0.3157794	0.2904728	0.3802682	0.3414674	0.3413723	*0.187*
Deportes	0.3405697	0.3084431	0.3472079	0.3671349	0.3753570	*0.002*
Economía	0.2798897	0.2651235	0.3206308	0.3218855	0.3001491	*-0.095*
Salud	0.2851341	0.2561730	0.3238575	0.3090317	0.2895345	*0.014*
					Media →	*0.019*

Tabla 5.13. Valores de similitud "intraclase", "interclase" y diferencia normalizada utilizando la medida de similitud sim_T^2 , con *k* igual a 15 y niveles de activación normalizados.

Observando el valor medio de la diferencia se aprecia que la normalización del nivel de activación de los conceptos no beneficia la capacidad de discriminación de la medida sino que la perjudica gravemente. En la Figura 5.13 se puede apreciar la desventaja del uso de la normalización. Este decremento de la capacidad discriminatoria causado por la normalización se explica mediante la propia definición de la medida de similitud sim_T^2, ya que está basada en la diferencia de activación de los conceptos semejantes, es decir, en la diferencia de intensidad con la que se tratan temas parecidos, y la normalización desvirtúa dicha intensidad. Sin embargo, experimentos análogos sobre la normalización de la activación con la medida sim_T^1 no han revelado apenas influencia de la normalización en la capacidad de discriminación de dicha medida, puesto que en este caso la diferencia de activación tiene un papel secundario como ponderación de la similitud individual entre conceptos.

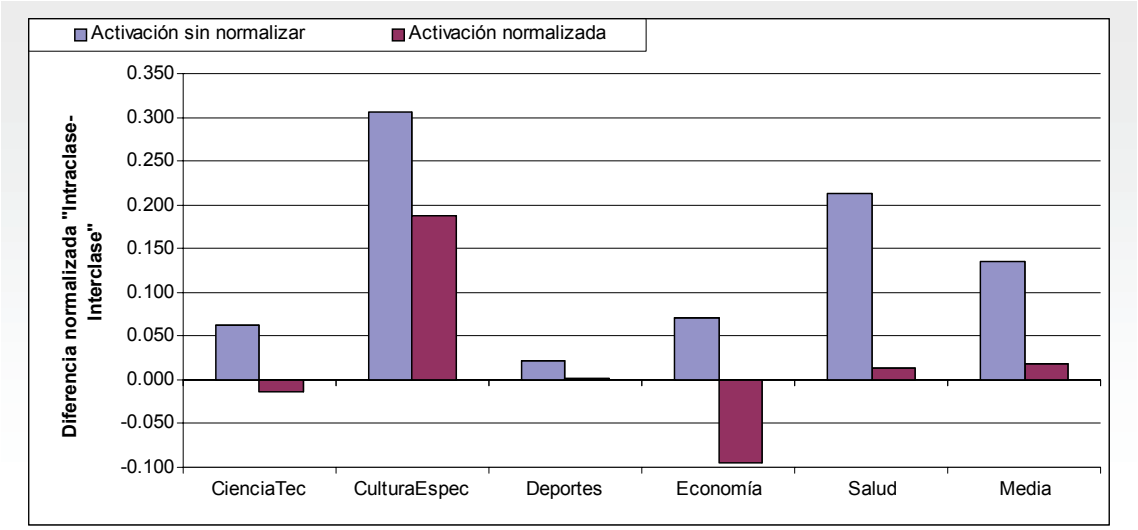


Figura 5.13. Valores de la diferencia "intraclase-interclase" normalizada de cada categoría y la media de todas ellas para de la medida de similitud sim_T^2 utilizando valores de activación normalizados y sin normalizar.

5.4. Comparación del Sistema SILC con otros Sistemas Existentes

La comparación de las representaciones generadas por SILC ha sido realizada una vez más en términos de eficacia en la tarea de clasificación de textos. Dichas representaciones han sido contrastadas con las generadas por otros dos sistemas de representación, ambos de tipo vectorial: la clásica "bolsa de palabras" y LSI (*Latent Semantic Indexing*) derivado del Análisis de Semántica Latente (LSA). Se han llevado a cabo dos tipos de experimentos. En el primero las representaciones generadas por SILC han sido evaluadas en forma vectorial, de igual forma que en los experimentos de optimización de parámetros de la Sección 5.2. En el segundo se han evaluando las representaciones de SILC aprovechando su forma estructural.

5.4.1. Conjunto de datos

La colección de textos empleada para el primer tipo de experimentos que evalúan la forma vectorial de las representaciones generadas por SILC es la misma colección que se utilizó en los experimentos de optimización de parámetros (descrita en la Sección 5.2.1). Dicha colección se compone de 500 textos correspondientes a noticias pertenecientes a cinco categorías distintas ("Ciencia y Tecnología", "Cultura y Espectáculos", "Deportes", "Economía" y "Salud") extraídas del servicio *Google News* junto con 500 ensayos de temas variados recopilados por el escritor Jorge Orellana.

Para el segundo tipo de experimentos, en los que se evalúa la representación estructural generada por SILC, se ha empleado un subconjunto de la colección estándar *20 NewsGroups*[4]. La colección completa fue recopilada por Ken Lang [Lang, 1995] y está compuesta por 20,000 documentos clasificados en 20 categorías semánticamente no disjuntas, con 1,000 documentos cada una de ellas. Las 20 categorías se pueden agrupar a su vez en seis categorías más generales que son "Ordenadores", "Entretenimiento", "Ciencia", "Debate", "Sociedad" y "Varios". Los textos están escritos en inglés y

[4] http://people.csail.mit.edu/jrennie/20Newsgroups/

pertenecen a un servicio de grupo de noticias e intercambio de opiniones a través de foros y correo electrónico, por lo que están escritos por una gran variedad de personas con un estilo libre y en ocasiones poco ortodoxo, lo que supone un reto complejo. El subconjunto de la colección empleado en los experimentos se resume en la Tabla 5.14. Se han seleccionado 8 categorías con 1,000 documentos cada una de ellas: Referente a "Ordenadores" se ha seleccionado la categoría "Hardware del PC", de la categoría "Varios" se ha seleccionado su única subcategoría "Anuncios de Alquiler", con respecto a "Entretenimiento" se han seleccionado las categorías "Coches" y "Béisbol", de la categoría "Sociedad" se han tomado las subcategorías "Religión" y "Ateísmo", de "Ciencia" se ha seleccionado la categoría "Electrónica" y, finalmente, referente a "Debate" se ha seleccionado la subcategoría "Política General". Como se intuye por su denominación, las categorías seleccionados siguen siendo semánticamente no disjuntas, para complicar aún más la tarea de clasificación.

Categoría	Subcategoría	Número de textos
Ordenadores	Hardware del PC	1,000
Entretenimiento	Coches	1,000
	Béisbol	1,000
Ciencia	Electrónica	1,000
Debate	Política General	1,000
Sociedad	Religión	1,000
	Ateísmo	1,000
Varios	Anuncios de Alquiler	1,000
	Total →	*8,000*

Tabla 5.14. Distribución y estructura del subconjunto de la colección de textos *20 NewsGroups* empleado en la evaluación de la representación estructural generada por SILC.

5.4.2. Medidas de evaluación

Puesto que la comparación se realiza en el ámbito de la clasificación automática de textos, las medidas empleadas son las mismas que se utilizaron en la optimización de

parámetros (ver Sección 5.2.2) relativas a la eficacia en la mencionada tarea para cada categoría:

- Precisión (*Pr*). Porcentaje de clasificaciones correctas (Ecuación 5.1).
- Cobertura (*Co*). Porcentaje del total de documentos de una categoría correctamente identificados (Ecuación 5.2).
- Medida-F (*F*). Combinación normalizada de las dos medidas anteriores (Ecuación 5.3).
- Coeficiente de Correlación de Pearson. Correlación estadística entre las variables categoría real y categoría asignada por el algoritmo de clasificación.

Estas son pues las medidas utilizadas en todos los experimentos descritos a continuación para comparar cuantitativamente la eficacia de las distintas representaciones en la clasificación automática de textos.

5.4.3. Procedimiento experimental

5.4.3.1. Procedimiento para las representaciones vectoriales

Para la evaluación de las representaciones generadas por SILC en forma de vector, el procedimiento es análogo al descrito en la Sección 5.2.3 para la optimización de parámetros. Se ha realizado una validación de subconjuntos cruzada, donde la colección de *Google News* se divide en tres partes iguales, manteniendo la distribución de las categorías y utilizando dos de ellas como entrenamiento de los algoritmos y la restante como validación. Así pues, se llevan a cabo tres ejecuciones, tomando las medidas descritas en cada una de las tres y calculando la media aritmética de todas ellas a la finalización de las mismas. En cada una de estas ejecuciones se han aplicado los tres algoritmos de clasificación de diferente naturaleza descritos anteriormente: Naïve Bayes, Máquinas de Vectores de Soporte y K-Vecinos más Cercanos (K-NN, con K igual a 5), obteniendo las medidas medias de evaluación para cada uno de ellos

Todo este proceso se lleva a cabo tres veces, una para cada tipo de representación utilizado:

- Representación tradicional. Se representan todos los documentos de la colección mediante el enfoque de "bolsa de palabras", donde cada documento es un vector

de tamaño igual al número de palabras del vocabulario. Dicho vector contendrá valores numéricos en las posiciones correspondientes a las palabras que forman el texto al que representa. Dichos valores corresponden a los valores *tf·idf* (*term frequency · inverse document frequency*) que se obtienen a partir de la división de la frecuencia total de la palabra en cuestión (proporción de veces que aparece con respecto a todas las apariciones de todas las palabras) entre la frecuencia en documentos de dicha palabra (porcentaje de documentos donde aparece la palabra). Tanto el vocabulario total como los valores *tf·idf* se calculan a partir de los textos de entrenamiento en cada ejecución más los 500 textos de cultura general.

- Representación LSI. Todos los documentos de la colección se representan como vectores que son combinación de los vectores correspondientes a las palabras que contienen [Deerwester et al., 1990]. Los vectores de cada palabra están representados en las filas de la matriz LSA (término por documento). Dicha matriz se construye, en cada ejecución, a partir de la porción de entrenamiento más los textos de cultura general. Se aplica una reducción SVD a 100 dimensiones, que es el valor óptimo de este parámetro LSI según el trabajo [Dumais, 1994]. La reducción SVD se realizó empleando el entorno *Matlab v14*[5]. Para la indexación LSI se empleó la librería software *LSI/SDD*[6].

- Representación SILC vectorial. Al igual que en los experimentos de optimización, se aplica el modelo de lectura con los valores óptimos de los parámetros a toda la colección. A continuación, cada representación generada se convierte en un vector de tamaño igual al número total de conceptos distintos presentes en la porción de entrenamiento en cada ejecución. En la posición correspondiente de cada concepto se introduce la activación que posea el mismo en la representación generada. El resto de posiciones del vector que no corresponden a ningún concepto presente en la representación del texto se establecen a cero. El conocimiento semántico lingüístico sobre el que opera el modelo de lectura se construye, en cada ejecución, a partir de la porción de entrenamiento más los textos de cultura general.

[5] MathWorks, Inc.: http://www.mathworks.com

[6] Software de tesis de maestría de Jason Dowling (QUT): http://members.pcug.org.au/~jdowling/

Así, para cada representación se obtienen las medidas medias de evaluación descritas para cada categoría, para cada algoritmo de clasificación empleado y la media de todos ellos.

5.4.3.2. Procedimiento para las representaciones estructurales

En este caso, el procedimiento experimental es similar al llevado a cabo para las representaciones vectoriales, con las salvedades de que la colección de datos utilizada es el subconjunto de *20 NewsGroups*. Además, las representaciones generadas por el sistema SILC no son transformadas en vectores. Así pues, dado que los algoritmos de clasificación empleados requieren que los textos que tratan se encuentren en forma vectorial, ha sido necesario utilizar y adaptar otro tipo de algoritmo para evaluar las representaciones estructurales en su forma original. Dicho algoritmo está basado en una medida de similitud semántica entre las representaciones de los textos. Entre los algoritmos más populares que requieren una medida de similitud, denominados "basados en memoria" [Aha et al., 1991], se encuentran el propio K-NN y los algoritmos basados en "centroides" [Serrano y Del Castillo, 2007], de los que el algoritmo de Rocchio es el más representativo, [Joachims, 1997]. Este tipo de algoritmos no realizan una fase de entrenamiento para construir un modelo de clasificación, aunque sí necesitan un conjunto de textos de referencia, generalmente la porción de textos de entrenamiento, que forman la "memoria" (de ahí su denominación). En el caso de K-NN, cada texto del conjunto de validación o test se compara mediante una medida de similitud con cada texto de la memoria, y se le asigna la categoría mayoritaria entre los *K* textos de la memoria más similares. Los algoritmos basados en "centroides" utilizan también una memoria de textos pero no los utilizan tal cual tal cual para la clasificación, sino que crean, a partir de los mismos, uno o varios "centroides" para cada categoría. Un "centroide" es un ejemplo representado de la misma manera que los textos tratados, que es representativo con respecto a la temática de una categoría. Aunque este proceso de construcción de "centroides" se puede ver como una fase de entrenamiento y los mismos "centroides" como modelos de clasificación, en realidad estos algoritmos son una especie de K-NN: una vez construidos los "centroides", el conjunto de estos actúa como memoria y los textos del conjunto de validación se clasifican como pertenecientes a la categoría a la que representa el "centroide/s" más similar/es a los mismos (aquí entra en juego de nuevo la medida de similitud).

Debido al alto coste computacional del algoritmo K-NN por el elevado número de comparaciones que requiere, el método elegido para la validación de las representaciones estructurales de SILC es un algoritmo basado en "centroides". El algoritmo ideado construye un "centroide" para cada categoría empleando las representaciones generadas por SILC para los textos de entrenamiento en cada ejecución. Un "centroide" está compuesto por la unión de todas las redes conceptuales de las representaciones de entrenamiento. La activación de los conceptos de dicha unión se establece como la activación media que poseen dichos conceptos entre las representaciones en las que aparecen. Las representaciones del conjunto de validación se clasifican pues como pertenecientes a la categoría del "centroide" al que más se asemejan según la medida de similitud empleada. Se ha experimentado con las medidas de similitud sim_T^1, con su valor óptimo del parámetro k igual a 5, y sim_T^2, análogamente con k igual a 15, utilizando un contexto de comparación reducido creando puentes cruzados entre 75 anclas (valores óptimos obtenidos experimentalmente).

Así pues, para las representaciones "bolsa de palabras", LSI y SILC estructural se ha aplicado una validación cruzada de tres subconjuntos. Para las dos primeras representaciones se han aplicado tres algoritmos de clasificación, Naïve Bayes, SVM y K-NN (K=5) en cada una de las tres ejecuciones. Para las representaciones de SILC sólo se ha aplicado el algoritmo basado en "centroides" descrito anteriormente. En cada una de las ejecuciones y para cada uno de los algoritmos se han obtenido las medidas de evaluación y la media de dichas ejecuciones.

5.4.4. Evaluación de las representaciones vectoriales

Como se ha descrito en el procedimiento experimental, para cada una de las tres representaciones "bolsa de palabras" *tf·idf*, LSI y SILC vectorial se ha calculado la precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, en cada una de las tres ejecuciones de la validación cruzada y para cada uno de los tres algoritmos de clasificación mencionados. Todos estos valores medios se muestran en la Tabla 5.15. El símbolo '-' indica que no ha sido posible calcular el valor puesto que ningún texto ha sido clasificado como de la categoría correspondiente.

Los valores de la Tabla 5.15 ponen de manifiesto la dependencia entre el algoritmo de clasificación y el tipo de representación. Utilizando los algoritmos Naïve Bayes y

SVM, las representaciones vectoriales de SILC obtienen los mejores resultados de clasificación, manifestando la mayor diferencia con el algoritmo SVM. Sin embargo, la representación LSI obtiene los mejores resultados con el algoritmo K-NN. Esto indica que la representación LSI es eficaz cuando se utiliza en el propio espacio de palabras definido por la matriz de semántica latente, como se hace en el cálculo de similitud K-NN.

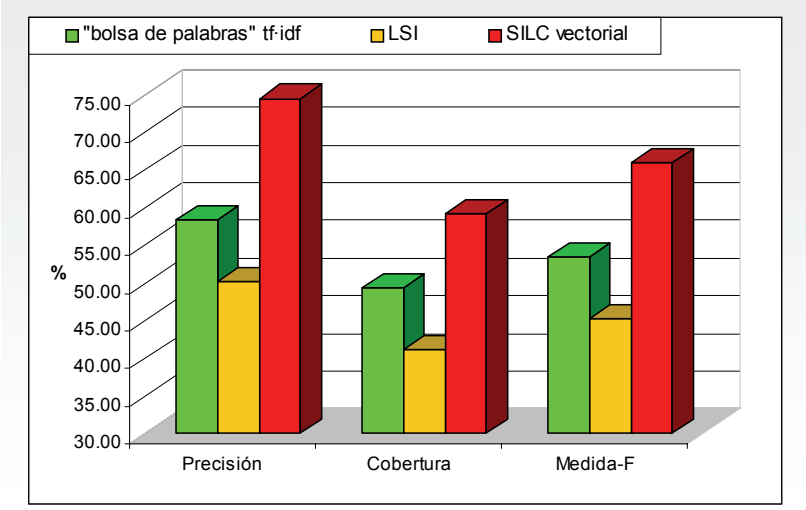
Representación = **"bolsa de palabras" tf·idf**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	75.00	75.00	75.00	27.27	25.00	26.09	21.64	80.56	34.12	41.30	60.19	48.99
Cultura	90.00	75.00	81.82	22.64	33.33	26.96	25.00	25.00	25.00	45.88	44.44	45.15
Deportes	94.59	94.59	94.59	43.21	94.59	59.32	0.00	0.00	0.00	45.93	63.06	53.15
Economía	78.57	91.67	84.62	100.00	27.28	42.87	100.00	8.33	15.38	92.86	42.43	58.24
Salud	96.87	96.88	96.87	0.00	0.00	0.00	100.00	12.50	22.22	65.62	36.46	46.88
Media	*87.01*	*86.63*	*86.82*	*38.62*	*36.04*	*37.29*	*49.33*	*25.28*	*33.43*	*58.32*	*49.32*	*53.44*
Correlación	*0.841*			*0.101*			*0.214*			*0.385*		

Representación = **LSI**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	66.67	33.33	44.44	0.00	0.00	0.00	25.00	33.33	28.57	30.56	22.22	25.73
Cultura	54.55	100.00	70.59	-	0.00	0.00	57.14	66.67	61.54	55.85	55.56	55.70
Deportes	57.14	66.67	61.54	-	0.00	0.00	71.43	83.33	0.00	64.29	50.00	56.25
Economía	75.00	50.00	60.00	-	0.00	0.00	25.00	16.67	20.00	50.00	22.22	30.77
Salud	60.00	50.00	54.55	14.81	66.67	24.24	75.00	50.00	60.00	49.94	55.56	52.60
Media	*62.67*	*60.00*	*61.31*	*7.41*	*13.33*	*9.52*	*50.71*	*50.00*	*50.35*	*50.12*	*41.11*	*45.17*
Correlación	*0.586*			*0.393*			*0.261*			*0.413*		

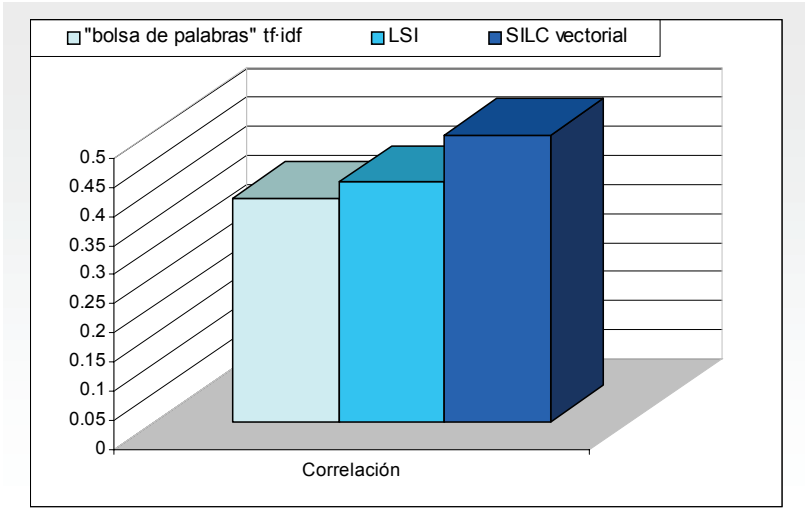
Representación = **SILC vectorial**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ciencia	78.78	72.22	75.36	53.33	66.67	59.26	20.93	100.00	34.62	51.01	79.63	62.19
Cultura	87.87	80.56	84.06	57.77	72.22	64.19	100.00	5.56	10.53	81.88	52.78	64.19
Deportes	92.11	94.59	93.33	79.07	91.89	85.00	0.00	0.00	0.00	57.06	62.16	59.50
Economía	80.95	94.44	87.18	83.33	69.44	75.75	100.00	2.78	5.41	88.09	55.55	68.14
Salud	96.77	93.75	95.24	85.71	37.50	52.17	100.00	6.25	11.76	94.16	45.83	61.66
Media	*87.30*	*87.11*	*87.20*	*71.84*	*67.54*	*69.63*	*64.19*	*22.92*	*33.78*	*74.44*	*59.19*	*65.95*
Correlación	*0.840*			*0.478*			*0.164*			*0.494*		

Tabla 5.15. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos representados mediante "bolsa de palabras", LSI y SILC vectorial.

La representación LSI aislada no aporta suficientes regularidades semánticas como para que los otros algoritmos de clasificación extraigan de las mismas un modelo preciso de las categorías implicadas. Por el contrario, los conceptos de las representaciones SILC, junto con los valores de activación vistos como niveles de significación de dichos conceptos dentro de los textos, sí parecen generalizar con precisión la temática de los mismos.



a)



b)

Figura 5.14. Valores medios de a) precisión ,cobertura y medida-F y b) correlación comparativos entre los sistemas de representación "bolsa de palabras", LSI y SILC vectorial en tareas de clasificación de textos.

Así, los valores medios totales posicionan a la representación SILC como el mejor enfoque para la clasificación, aunque los resultados sean equiparables con los de la

representación "bolsa de palabras" con el algoritmo Naïve Bayes. La Figura 5.14 muestra las medidas de evaluación medias totales (Figura 5.14a: precisión, cobertura y medida-F, y Figura 5.14b correlación) para cada una de las representaciones "bolsa de palabras", LSI y SILC vectorial.

5.4.5. Evaluación de las representaciones estructurales

Representación = **"bolsa de palabras" tf·idf**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ateísmo	29.50	63.66	40.31	12.50	33.33	18.18	12.55	100.00	22.29	18.18	65.66	26.93
Ordenadores	16.93	1.41	2.60	5.08	12.45	7.22	0.00	0.00	0.00	7.34	4.62	3.27
Alquiler	39.60	18.27	25.01	12.50	33.33	18.18	0.00	0.00	0.00	17.37	17.20	14.40
Coches	30.24	75.70	43.22	-	0.00	0.00	100.00	0.60	1.19	65.12	25.43	14.80
Béisbol	52.22	18.88	27.73	-	0.00	0.00	0.00	0.00	0.00	26.11	6.29	9.24
Electrónica	21.71	36.75	27.30	0.00	0.00	0.00	100.00	1.51	2.97	40.57	12.75	10.09
Política	35.82	18.47	24.38	-	0.00	0.00	100.00	0.60	1.19	67.91	6.36	8.52
Religión	11.06	4.42	6.31	-	0.00	0.00	0.00	0.00	0.00	5.53	1.47	2.10
Media	*29.64*	*29.69*	*29.67*	*7.52*	*9.89*	*8.54*	*39.07*	*12.84*	*19.33*	*25.41*	*17.47*	*20.71*
Correlación	*-0.009*			*-0.106*			*0.037*			*-0.026*		

Representación = **LSI**												
	Naïve Bayes			**SVM**			**K-NN**			**Media**		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***	***Pr***	***Co***	***F***
Ateísmo	23.36	30.52	26.47	8.79	33.33	13.91	23.82	45.58	31.29	18.66	36.48	23.89
Ordenadores	39.39	1.20	2.34	-	12.45	0.00	29.04	40.56	33.85	34.21	18.07	12.06
Alquiler	29.56	42.75	34.95	12.50	33.33	18.18	22.84	26.50	24.54	21.63	34.20	25.89
Coches	17.25	25.50	20.58	-	0.00	0.00	16.97	14.46	15.61	17.11	13.32	12.06
Béisbol	22.92	0.40	0.79	-	0.00	0.00	20.73	16.26	18.23	21.82	5.55	6.34
Electrónica	49.63	8.63	14.71	0.00	0.00	0.00	33.72	26.91	29.93	27.78	11.85	14.88
Política	18.17	57.43	27.61	-	0.00	0.00	25.18	14.26	18.21	21.68	23.90	15.27
Religión	18.88	4.22	6.89	-	0.00	0.00	24.96	13.66	17.65	21.92	5.96	8.18
Media	*27.39*	*21.33*	*23.99*	*7.10*	*9.89*	*8.26*	*24.66*	*24.77*	*24.72*	*19.72*	*18.67*	*19.18*
Correlación	*0.070*			*-0.106*			*0.048*			*0.004*		

Tabla 5.16. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante Naïve Bayes, Vectores de Soporte y K-NN, de textos de la colección *20 NewsGroups* representados mediante *tf·idf* y LSI.

De la misma manera que para la evaluación de las representaciones vectoriales de SILC, la Tabla 5.16 muestra los valores medios (de la validación cruzada) de precisión, cobertura, medida-F y correlación empleando los algoritmos Naïve Bayes, SVM y K-

NN (K=5) en el caso de las representaciones "bolsa de palabras" *tf·idf* y LSI, aplicados al subconjunto de textos de la colección *20 NewsGroups*.

Los valores de la tabla muestran de nuevo la dependencia entre representación y algoritmo. Al igual que en los experimentos anteriores con la colección de *Google News*, se aprecia que la representación LSI obtiene buenos resultados cuando la clasificación se basa en comparaciones dentro del espacio de palabras que él mismo define, es decir, al utilizar K-NN. Los resultados de clasificación con este algoritmo son los únicos que superan a los resultados de la representación *tf·idf* con el mismo algoritmo. Se vuelve a constatar pues que las representaciones LSI no contienen patrones sobre la temática del texto al que representan, por lo que los algoritmos que tratan de extraerlas no obtienen buenos resultados.

La Tabla 5.17 presenta los valores análogos para el algoritmo basado en "centroide" descrito anteriormente, utilizando las similitudes sim_T^1 y sim_T^2 con sus parámetros óptimos en el caso de la representación estructural generada por SILC.

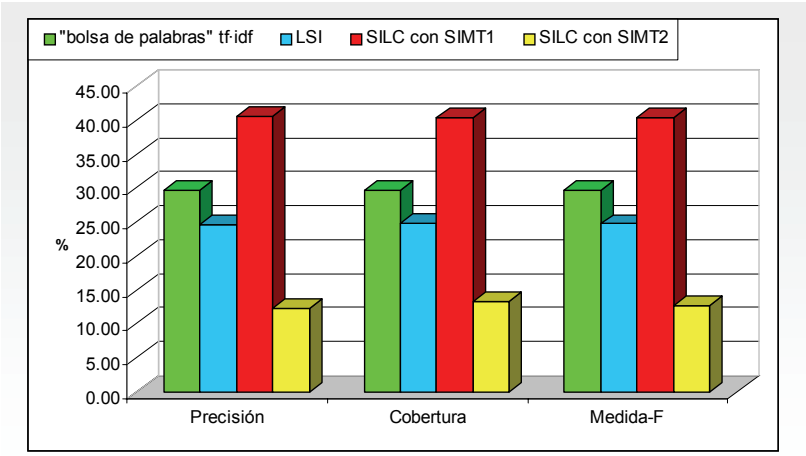
Representación = **SILC estructural**						
	sim_T^1			sim_T^2		
	Pr	***Co***	***F***	***Pr***	***Co***	***F***
Ateísmo	38.07	45.18	41.32	14.29	6.02	8.47
Ordenadores	46.21	40.36	38.00	15.32	11.45	13.11
Alquiler	39.75	58.43	60.25	20.76	36.14	26.37
Coches	42.54	34.34	32.84	6.90	2.41	3.57
Béisbol	62.18	58.43	35.17	9.81	12.65	11.05
Electrónica	32.00	33.73	17.55	6.52	3.61	4.65
Política	31.16	40.36	35.17	9.81	15.66	12.06
Religión	32.26	12.05	17.55	14.35	18.67	16.23
Media	*40.52*	*40.36*	*40.44*	*12.22*	*13.33*	*12.75*
Correlación	*0.224*			*-0.001*		

Tabla 5.17. Valores medios de precisión, cobertura, medida-F y correlación para cada categoría y la media de todas ellas, obtenidos de la clasificación mediante un algoritmo basado en "centroide", con diferentes medidas de similutud, de textos de la colección *20 NewsGroups* representados por SILC.

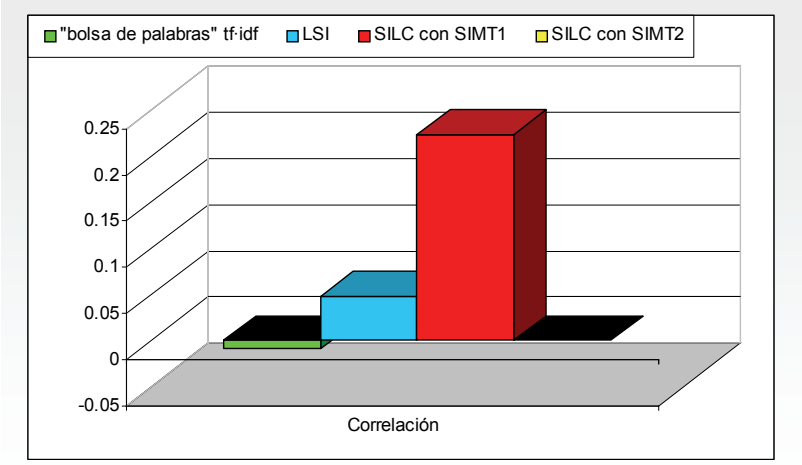
La Tabla 5.17 pone de manifiesto la diferencia del poder de discriminación de las dos medidas. Aunque los experimentos de optimización de las medidas de similitud han mostrado que la medida sim_T^2 optimizada siempre obtenía una discriminación positiva para todas las categorías, dicha medida no es en absoluto eficaz en tareas de clasificación duras, como lo es la clasificación de la colección *20 NewsGroups*. La

suposición en la que dicha medida se basa ("textos parecidos tratan temas parecidos con la misma intensidad") sólo es válida cuando se tratan textos correctamente escritos y estructurados, tanto en forma como en estilo, como son las noticias de prensa. Este hecho podría hacer pensar en la medida sim_T^2 como un indicador de la corrección del discurso de un texto, aunque este estudio se plantea como un trabajo futuro.

La medida sim_T^1, por el contrario, obtiene los mejores resultados de clasificación, corroborando la capacidad de discriminación mostrada en los experimentos de optimización cuando se dispone de un vocabulario suficientemente completo.



a)



b)

Figura 5.15. Valores medios máximos de a) precisión ,cobertura y medida-F y b) correlación comparativos entre los sistemas de representación "bolsa de palabras", LSI y SILC estructural en tareas de clasificación de textos.

La Figura 5.15 muestra los valores de precisión, cobertura y medida-F (Figura 5.15a) y de correlación (Figura 5.15b) del algoritmo Naïve Bayes para las representaciones "bolsa de palabras" y LSI, ya que es el que mejores resultados obtiene para ambas, y del algoritmo basado en "centroide" para las representaciones estructurales de SILC empleando las medidas sim_T^1 y sim_T^2. Como se aprecia, el empleo de la información estructural de las representaciones generadas por el sistema SILC junto con la medida de similitud semántica sim_T^1 y un algoritmo basado en "centroide" hacen al sistema SILC apto para ser aplicado a tareas reales de clasificación de textos, superando de manera clara los resultados obtenidos para las otras representaciones.

Es necesario remarcar que, en todos los experimentos llevados a cabo hasta el momento, los valores absolutos de las medidas de evaluación son anormalmente bajos para una tarea de clasificación. Esto es debido a que no se aplica ningún método de selección previa de características [Guyon y Elisseff, 2003]. La selección de características es una etapa que reduce el ruido y la ambigüedad descartando los términos que menos información aportan a la clasificación y mejorando así la eficiencia computacional. De hecho, existen estudios que demuestran que una reducción del 75% del vocabulario puede producir mejoras de precisión y cobertura de hasta el 60%. Dado el alto número de conceptos manejados, entorno a los 25,000, el ruido y ambigüedad presentes en las colecciones es de una magnitud considerable. Esta fase de selección de características es un mecanismo artificial que se aplica por motivos puramente prácticos. El objetivo de no aplicarla en estos experimentos es no desvirtuar con ello la naturaleza y capacidad de síntesis propia de cada uno de los tipos de representación evaluados.

5.5. Evaluación de la Similitud del Modelo de Lectura de SILC con los Seres Humanos

Como se mencionó en el Capítulo 3, existen dos maneras de evaluar la similitud de un modelo computacional de lectura con los seres humanos: evaluación "on-line" y evaluación "off-line". La primera de ellas mide la similitud del modelo durante el proceso de lectura. La segunda realiza una comparación del modelo con los humanos al finalizar la lectura de un texto. Ambos tipos de evaluación se basan en la comparación de ciertas medidas tomadas tanto de los seres humanos como del modelo de lectura. Las medidas registradas deben ser las mismas, o al menos con el mismo objetivo y naturaleza, en ambas fuentes objeto de la comparación.

5.5.1. Conjunto de datos

Una vez más, la colección de textos utilizada es la descrita en las secciones anteriores compuesta por las 500 noticias en español del servicio *Google News* junto con los 500 ensayos de cultura general recopilados y escritos por el autor Jorge Orellana. Además, se han empleado otras 10 noticias del servicio *Google News*, dos de cada una de las cinco categorías consideradas, no incluidas en las 500 noticias anteriores. Cinco de dichas noticias (una de cada categoría) han sido empleadas para la evaluación "on-line" del sistema SILC, y las otras cinco para su evaluación "off-line". Para que el lector pueda tener una idea clara del objetivo y naturaleza de la experimentación realizada, los textos correspondientes a estas 10 noticias se pueden encontrar en los Apéndices I y II.

5.5.2. Medidas de evaluación

Como se comentó anteriormente, la evaluación de un modelo computacional de lectura con respecto a los seres humanos se basa en una sencilla comparación de los valores de ciertas medidas tomadas tanto al modelo como a los sujetos humanos. Dichas medidas pueden cuantificar ciertos aspectos durante el proceso de lectura ("on-line") o aspectos al término de la misma ("off-line"). Para evaluar al sistema SILC se han ideado dos medidas, una de cada tipo mencionado:

- La primera de ellas es la medida "on-line". Dicha medida cuantifica el porcentaje de aciertos que obtienen tanto los humanos como el modelo en la predicción o inferencia de palabras durante la lectura de los textos. Dicho porcentaje de aciertos es el que se compara en diferentes circunstancias. Dado que las inferencias que realiza el modelo dependen en gran medida de la representación del texto que el propio modelo ha generado hasta el momento de realizar cada predicción, la comparación de los aciertos permite evaluar tanto el método de inferencia implementado como la representación intermedia que SILC va generando a lo largo del proceso de lectura.

- Para la evaluación "off-line" las medidas que se toman de cada fuente no poseen un valor numérico. En este caso, la información proveniente de cada fuente (sujetos humanos y modelo) que se compara es la propia representación del texto que cada una ha generado al final de la lectura del mismo. Así pues, se comparan las representaciones generadas por los sujetos humanos con las generadas por el modelo. La medida descrita a continuación cuantifica la similitud entre las representaciones de cada fuente y está basada en los coeficientes de Dice y Jaccard [Manning y Schutze, 2002], que miden la similitud entre cadenas de *tokens*. Dadas las representaciones de los sujetos humanos como listas de términos en orden decreciente de significación, y dadas las representaciones finales de SILC como listas de conceptos ordenadas de manera decreciente por el nivel de activación, la similitud entre listas de cada fuente se mide como la diferencia media de las posiciones relativas en cada lista de los términos o conceptos que se encuentran en su intersección, es decir, que aparecen en las dos. La expresión de dicha medida se presenta en la Ecuación 5.7.

$$Diferencia\ (A, B) = \frac{\sum_{i}^{n} \left| \frac{Pos(Concepto_i, A)}{Tamaño(A)} - \frac{Pos(Concepto_i, B)}{Tamaño(B)} \right|}{n} \qquad (5.7)$$

$$Concepto_i \in A \cap B,\ n = Tamaño(A \cap B)$$

donde $Pos(Concepto_i, A)$ es la posición del concepto i en la lista A, *Tamaño* indica el número de conceptos o términos de cada lista y n es el número de elementos en la intersección de las dos listas comparadas. Dado que las posiciones se normalizan con el tamaño de las listas, la diferencia para cada concepto de la

intersección, y por tanto la diferencia media, tomará un valor entre 0 y 1. Así, se ha calculado la similitud entre dos listas como 1 menos la diferencia entre las mismas, manejando así la semejanza como un concepto más intuitivo que la diferencia o la distancia.

La comparación "off-line" también se realiza con los títulos de las noticias, que al fin y al cabo son resúmenes de las mismas generadas por los autores. Así pues, otras medidas de evaluación utilizadas son la activación relativa media, con respecto a la máxima activación y la posición relativa media que ocupa una palabra del título en la representación generada por SILC, con respecto al total de conceptos en dicha representación del texto al que dicho título encabeza, en orden descendente de activación. Así, se observa la posición y activación relativa de cada palabra de cada título en la representación SILC del texto al que corresponde y se calcula la media aritmética de todas las palabras de todos los títulos.

Estos tres tipos de medidas son pues los obtenidos de la experimentación que se describe a continuación.

5.5.3. Procedimiento experimental

5.5.3.1. Evaluación "on-line"

El primer tipo de experimentos corresponde a la comparación "on-line" entre el modelo de lectura implementado por SILC y sujetos humanos. Para ello, se han utilizado cinco textos de los 10 mencionados en la sección anterior, uno de cada categoría considerada. Cada texto contiene 5 "huecos" excepto el perteneciente a la categoría Economía, que posee 8 (ver Apéndice I). Cada uno de esos "huecos" corresponde a una palabra del texto original que ha sido omitida. De cada texto existen a su vez cinco versiones. Cada versión contiene indicios sobre las palabras que están omitidas en cada hueco. Los indicios se corresponden con los fragmentos iniciales de cada palabra en cuestión. La longitud de los fragmentos varía desde 0 hasta 4. De esta manera, si en un texto se ha omitido la palabra "anchoa" se tienen cinco versiones del mismo, una de ellas con el "hueco" sin indicio alguno, y el resto con las pistas "a", "an", "anc" y "anch", respectivamente. Así pues, cada uno de los textos tendrá cinco versiones que se corresponden con el tamaño de las pistas.

Una vez confeccionado el marco experimental, se pidió a 50 sujetos humanos, todos ellos lectores habituales de prensa y con un nivel académico similar, que cada uno de ellos rellenara los huecos de los cinco textos en cuestión con un mismo tamaño de indicio. Así, cada versión de los textos fue completada por 10 sujetos. La asignación de las versiones a los sujetos fue totalmente aleatoria. Una vez completos, se calculó el porcentaje de aciertos con respecto a las palabras originales omitidas para las diferentes categorías temáticas y los diferentes tamaños de indicio.

De manera análoga, se presentaron al modelo de lectura de SILC las cinco versiones de los cinco textos. El conocimiento semántico lingüístico sobre el que operó el modelo se construyó a partir de la colección de *Google News* y los textos de cultura general de Orellana. Se calcularon los porcentajes de acierto de las palabras inferidas por el sistema con respecto a las palabras originales para cada versión de cada texto. Este proceso se realizó varias veces, variando en cada ejecución el valor de ciertos parámetros del proceso de lectura y utilizando los valores óptimos obtenidos experimentalmente para los parámetros no modificados. Por último, se realizó la comparación de los porcentajes de acierto de los sujetos humanos con los del modelo para las distintas categorías y distintos tamaños de indicio.

Así, el modelo lee los textos siguiendo su procedimiento habitual hasta que encuentra un "hueco", momento en el cual procede a rellenarlo. El relleno de "huecos" por parte del modelo es el producto de un proceso de inferencia o predicción de palabras basado en la representación conceptual del texto leído hasta el momento. Así pues, ha sido necesario modelar dicho proceso en el marco estructural y funcional que presenta SILC.

5.5.3.1.1. Mecanismos de inferencia o predicción de conceptos

El objetivo de la predicción implementada es la inferencia de un concepto para rellenar un término omitido, total o parcialmente (con indicio), en un determinado punto de un texto. Para ello, se empleará la palabra inmediatamente anterior al concepto omitido, es decir, la última palabra leída antes del "hueco" con el fin de contemplar las asociaciones semánticas y sintácticas locales, junto con la representación generada por el modelo hasta ese momento. Se han implementado dos formas distintas de inferencia. La primera de ellas está principalmente condicionada por el contexto semántico, es decir, por la representación semántica que se tiene del texto leído hasta el momento. Este tipo de inferencia se denomina inferencia "por contexto global". El segundo tipo de

inferencia está principalmente influenciado por la última palabra leída antes del concepto a inferir. Es la denominada inferencia "por asociación local".

La dinámica de ambos tipos de inferencia es simple. Llegado al punto de decisión, los mecanismos eligen el primer concepto que cumpla una serie de criterios. Así pues, los criterios de selección para la inferencia "por contexto global" son:

1. El concepto está en la representación actual.
2. El concepto cumple el indicio (si se dispone de él).
3. El concepto tiene la máxima activación posible.

Es decir, que se infiere el concepto más activo en la memoria de trabajo actual que cumpla el indicio. Si no existe ningún concepto que cumpla alguno de los criterios se ignora dicho criterio y se vuelve a realizar la selección. Así, por ejemplo, si no existiese ningún concepto en la representación actual de la memoria de trabajo que cumpliera el indicio, entonces se seleccionaría el concepto más activo sin más. Análogamente, los criterios para la inferencia "por asociación local" son:

1. El concepto es el más asociado a la última palabra leída en el conocimiento semántico lingüístico.
2. El concepto cumple el indicio (si se dispone de él).
3. El concepto está en la representación actual de la memoria de trabajo.
4. El concepto tiene la máxima activación posible.

Es decir, que se selecciona el concepto más asociado a la palabra inmediatamente anterior que cumpla el indicio y que esté lo más activo posible en la representación actual. Como en el caso anterior, si ningún concepto cumple alguno de los criterios éstos se ignoran y se realiza la selección de nuevo. En el caso de que el "hueco" o concepto a inferir corresponda al primer término de una oración se considera que no existe palabra previa asociada y, por lo tanto, se ignora el primer criterio.

De esta manera, los conceptos inferidos mediante ambos mecanismos se han comparado con los conceptos originales omitidos para obtener el porcentaje de aciertos para los textos de las distintas categorías y para las diferentes versiones con tamaños de indicios diferentes.

Esta definición e implementación de los mecanismos de predicción de conceptos es un ejemplo de la generalidad que aporta el sistema SILC, mostrándose como una plataforma experimental que permite la evaluación eficiente y sencilla de hipótesis sobre la estructura y funcionamiento de diversos aspectos cognitivos relacionados con el proceso de lectura.

5.5.3.2. Evaluación "off-line"

La evaluación "off-line" supone la comparación del producto del sistema una vez finalizada la lectura de los textos. Para ello, se han comparado las representaciones finales generadas tanto por sujetos humanos como por SILC. Se han empleado los cinco textos restantes de los 10 mencionados en la descripción del conjunto de datos (ver Apéndice II). Dichos textos han sido leídos por 20 sujetos humanos, lectores habituales de prensa y con un nivel académico similar. A cada uno de ellos se le pidió que leyese cada texto con atención una sola vez y que, inmediatamente después de haberlo leído, escribiese un resumen del mismo exponiendo los aspectos principales en orden de relevancia. A dichos resúmenes se les aplicó el mismo preprocesamiento que aplica SILC, consistente en la eliminación de palabras comunes vacías de significado, mediante una lista de corte, y la reducción de las palabras a una raíz morfológica. El resultado es una lista de conceptos ordenados de manera decreciente por nivel subjetivo de significación, ya que los resúmenes estaban escritos en ese orden por petición expresa.

A continuación, se aplicó el modelo de lectura de SILC a los 5 textos variando sus parámetros. Cada representación generada por SILC fue comparada con la representación correspondiente generada por cada uno de los sujetos humanos, obteniendo así los valores de los parámetros que hacen que el sistema SILC genere representaciones más parecidas en media a las que generan los seres humanos y permitiendo, además, observar la relación entre la similitud con los seres humanos y la eficacia en la clasificación de textos. La comparación entre las listas correspondientes a las representaciones generadas por ambas fuentes fue realizada mediante la diferencia entre listas descrita en la Sección 5.5.2. De nuevo, el conocimiento semántico lingüístico sobre el que operó SILC fue construido a partir de la colección *Google News* más los ensayos de Orellana empleando un contexto de asociación variable.

Por último se representaron los 10 textos mencionados, utilizando los parámetros de SILC que producían resultados "off-line" más similares a los seres humanos, y se calculó la posición y activación relativa media de cada palabra de los títulos en la representación SILC de sus correspondientes textos.

5.5.4. Evaluación de la similitud "on-line" con los seres humanos

La similitud "on-line" del modelo de lectura con respecto a los seres humanos se ha medido en términos de porcentaje de aciertos en la tarea de predicción de conceptos: utilizando la representación generada hasta el momento y la palabra inmediatamente anterior se han de inferir conceptos omitidos en los textos. Dichos conceptos omitidos pueden contener indicios consistentes en los primeros caracteres de los términos que los representan. Así, se han calculado los porcentajes de acierto para todos los textos y todos los tamaños de indicio variando el umbral mínimo de propagación desde 0.001 hasta 0.7, utilizando un factor de olvido de 0.9, un intervalo de olvido de tamaño variable y propagación de la activación en profundidad con un nivel máximo de 3. La Tabla 5.18 presenta el número de aciertos en las predicciones y los porcentajes medios para las distintas categorías y tamaños de indicio utilizando la inferencia de conceptos "por contexto global" y distintos valores del umbral mínimo de propagación.

Umbral mínimo de propagación = **0.001**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	2	1	3	0	0	*6*	*21.43*
2	3	2	3	3	1	*12*	*42.86*
3	3	3	3	3	2	*14*	*50.00*
4	4	3	4	5	3	*19*	*67.86*
Total	*13*	*10*	*13*	*11*	*6*	*53*	*37.86*
%	*52.00*	*40.00*	*52.00*	*27.50*	*24.00*		

Umbral mínimo de propagación = **0.01**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	0	1	3	0	0	*4*	*14.29*
2	2	3	3	2	0	*10*	*35.71*
3	2	3	4	3	2	*14*	*50.00*
4	2	3	4	4	2	*15*	*53.57*
Total	*7*	*11*	*14*	*9*	*4*	*45*	*32.14*
%	*28.00*	*44.00*	*56.00*	*22.50*	*16.00*		

Umbral mínimo de propagación = **0.1**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	0	1	3	0	0	*4*	*14.29*
2	1	3	3	3	0	*10*	*35.71*
3	1	3	4	3	2	*13*	*46.43*
4	1	3	4	4	2	*14*	*50.00*
Total	*4*	*11*	*14*	*10*	*4*	*43*	*30.71*
%	*16.00*	*44.00*	*56.00*	*25.00*	*16.00*		

Umbral mínimo de propagación = **0.3**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	0	1	3	0	0	*4*	*14.29*
2	1	3	3	3	0	*10*	*35.71*
3	1	3	4	3	2	*13*	*46.43*
4	1	3	4	4	2	*14*	*50.00*
Total	*4*	*11*	*14*	*10*	*4*	*43*	*30.71*
%	*16.00*	*44.00*	*56.00*	*25.00*	*16.00*		

Umbral mínimo de propagación = **0.5**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	0	1	3	0	0	*4*	*14.29*
2	1	3	3	3	0	*10*	*35.71*
3	1	3	4	3	2	*13*	*46.43*
4	1	3	4	4	2	*14*	*50.00*
Total	*4*	*11*	*14*	*10*	*4*	*43*	*30.71*
%	*16.00*	*44.00*	*56.00*	*25.00*	*16.00*		

Umbral mínimo de propagación = **0.7**					Inferencia =**"por contexto global"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	1	0	0	0	*2*	*7.14*
1	0	1	3	0	0	*4*	*14.29*
2	1	3	3	3	0	*10*	*35.71*
3	1	3	4	3	2	*13*	*46.43*
4	1	3	4	4	2	*14*	*50.00*
Total	*4*	*11*	*14*	*10*	*4*	*43*	*30.71*
%	*16.00*	*44.00*	*56.00*	*25.00*	*16.00*		

Tabla 5.18. Número de aciertos y porcentajes medios de los mismos en la predicción de conceptos con el modelo de lectura de SILC para diferentes valores del umbral mínimo de propagación utilizando inferencia "por contexto global".

Los valores medios de la tabla denotan que cuanto menor sea el umbral de propagación, es decir, cuantos más conceptos activos se posean en la memoria de trabajo, más correctamente se infieren los conceptos omitidos. Esta tendencia se aprecia claramente en la Figura 5.16, que muestra los porcentajes totales de aciertos (suma de los cinco textos) para los distintos valores del umbral mínimo de propagación y los distintos tamaños de indicios. La gráfica muestra como para cualquier valor del umbral mínimo de propagación se tiene una relación directa entre el número de aciertos y el tamaño de los indicios, es decir, que cuanta más información aportan los indicios mejor se predicen los conceptos omitidos. La gráfica pone también de manifiesto la superioridad, en cuanto a eficacia de predicción, del uso de umbrales de propagación reducidos, obteniendo éstos un mayor porcentaje de aciertos, sobre todo en presencia de los indicios de mayor tamaño. De hecho, todos los valores de umbrales por encima de 0.01 obtienen los mismos resultados sin diferencia alguna, lo que denota que la inferencia de conceptos es una tarea que precisa de una lectura profunda y reflexiva, cualidad modelada en este caso por un bajo umbral de propagación. Producto de esta lectura profunda se tiene en la memoria de trabajo una representación más amplia y específica del contexto semántico en el que se enmarca el texto que se lee.

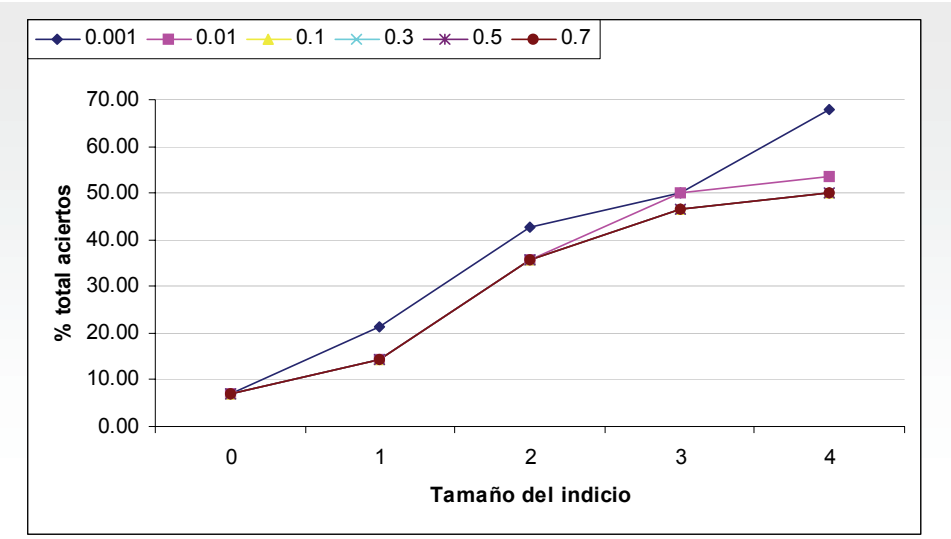


Figura 5.16. Porcentajes totales de aciertos de predicción de conceptos con el modelo de lectura de SILC para diferentes valores del umbral mínimo de propagación utilizando inferencia "por contexto global".

De manera análoga, la Tabla 5.19 muestra el número de aciertos y los porcentajes parciales y totales para diferentes valores del umbral mínimo de propagación, pero utilizando en esta ocasión la inferencia "por asociación local".

Umbral mínimo de propagación = **0.001**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	0	*1*	*3.57*
1	1	0	1	1	1	*4*	*14.29*
2	2	3	3	3	1	*12*	*42.86*
3	3	3	3	4	1	*14*	*50.00*
4	3	3	3	5	2	*16*	*57.14*
Total	*10*	*9*	*10*	*13*	*5*	*47*	*33.57*
%	*40.00*	*36.00*	*40.00*	*32.50*	*20.00*		

Umbral mínimo de propagación = **0.01**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	0	*1*	*3.57*
1	2	0	1	1	1	*5*	*17.86*
2	2	3	3	2	1	*11*	*39.29*
3	3	3	3	5	1	*15*	*53.57*
4	3	3	3	5	2	*16*	*57.14*
Total	*11*	*9*	*10*	*13*	*5*	*48*	*34.29*
%	*44.00*	*36.00*	*40.00*	*32.50*	*20.00*		

Umbral mínimo de propagación = **0.1**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	1	*2*	*7.14*
1	1	0	1	0	1	*3*	*10.71*
2	2	3	2	2	1	*10*	*35.71*
3	3	3	2	3	2	*13*	*46.43*
4	3	3	2	4	2	*14*	*50.00*
Total	*10*	*9*	*7*	*9*	*7*	*42*	*30.00*
%	*40.00*	*36.00*	*28.00*	*22.50*	*28.00*		

Umbral mínimo de propagación = **0.3**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	1	*2*	*7.14*
1	1	0	1	0	1	*3*	*10.71*
2	2	0	2	2	1	*7*	*25.00*
3	3	3	2	3	2	*13*	*46.43*
4	3	3	2	4	2	*14*	*50.00*
Total	*10*	*6*	*7*	*9*	*7*	*39*	*27.86*
%	*40.00*	*24.00*	*28.00*	*22.50*	*28.00*		

Umbral mínimo de propagación = **0.5**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	1	*2*	*7.14*
1	1	0	1	1	1	*4*	*14.29*
2	2	0	2	2	1	*7*	*25.00*
3	3	3	2	3	2	*13*	*46.43*
4	3	3	2	4	2	*14*	*50.00*
Total	*10*	*6*	*7*	*10*	*7*	*40*	*28.57*
%	*40.00*	*24.00*	*28.00*	*25.00*	*28.00*		

Umbral mínimo de propagación = **0.7**					Inferencia =**"por asociación local"**		
Tamaño indicio	**Categorías**						
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***	***%***
0	1	0	0	0	1	*2*	*7.14*
1	1	0	1	0	1	*3*	*10.71*
2	2	0	2	2	1	*7*	*25.00*
3	3	3	2	3	2	*13*	*46.43*
4	3	3	2	4	2	*14*	*50.00*
Total	*10*	*6*	*7*	*9*	*7*	*39*	*27.86*
%	*40.00*	*24.00*	*28.00*	*22.50*	*28.00*		

Tabla 5.19. Número de aciertos y porcentajes medios de los mismos en la predicción de conceptos con el modelo de lectura de SILC para diferentes valores del umbral mínimo de propagación utilizando inferencia "por asociación local".

En este caso, existe una menor dependencia entre el umbral mínimo de propagación y el porcentaje total de aciertos. Aunque los umbrales más pequeños obtienen mejor porcentaje, el valor máximo de aciertos se alcanza con un umbral de 0.01 y no con el más bajo de 0.001, contrariamente al caso de la inferencia "por contexto global". Así pues, en el caso de la inferencia "por asociación local", el hecho de tener muchos conceptos en la memoria de trabajo puede confundir la predicción ya que pueden existir varios conceptos que concuerden con el indicio. La Figura 5.17 presenta los porcentajes totales para cada valor del umbral y cada tamaño de indicio.

Al igual que en la inferencia "por contexto global", se aprecia una relación directa entre el porcentaje de aciertos y el tamaño de los indicios para cualquier valor del umbral mínimo de propagación. La gráfica de la Figura 5.17 muestra que éste es menos dependiente con respecto al tamaño de los indicios, puesto que según el valor de dicho tamaño algunos valores de umbral producen más inferencias correctas que otros.

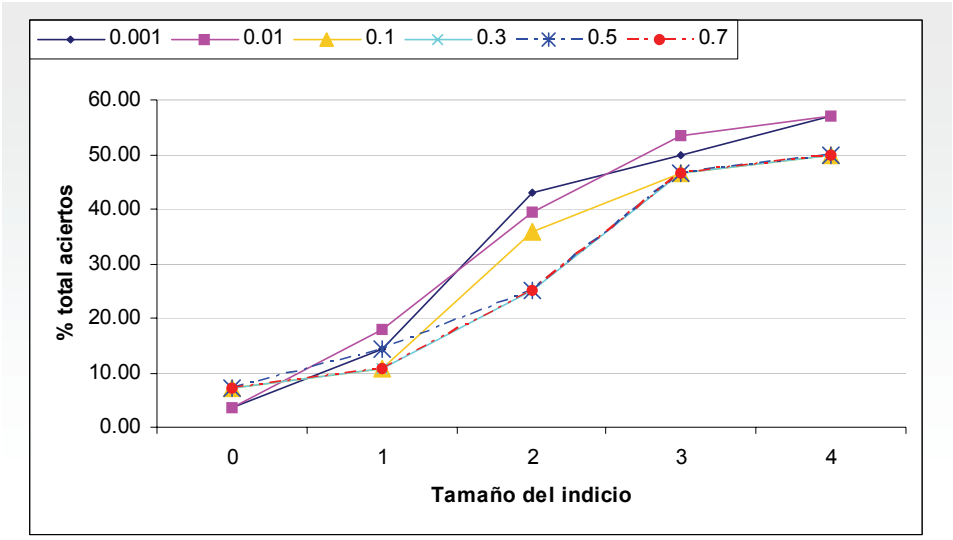


Figura 5.17. Porcentajes totales de aciertos de predicción de conceptos con el modelo de lectura de SILC para diferentes valores del umbral mínimo de propagación utilizando inferencia "por asociación local".

Aunque con cualquiera de los dos mecanismos de inferencia de conceptos se obtiene una relación de dependencia directa entre el tamaño de los indicios y el porcentaje de aciertos, la predicción "por contexto local" obtiene porcentajes de acierto superiores en cualquier caso. La Figura 5.18 muestra claramente esta diferencia entre ambos tipos de inferencia para todos los valores del umbral mínimo de propagación evaluados. Los porcentajes mostrados en la figura son los correspondientes a todos los textos con todos los tamaños de indicio.

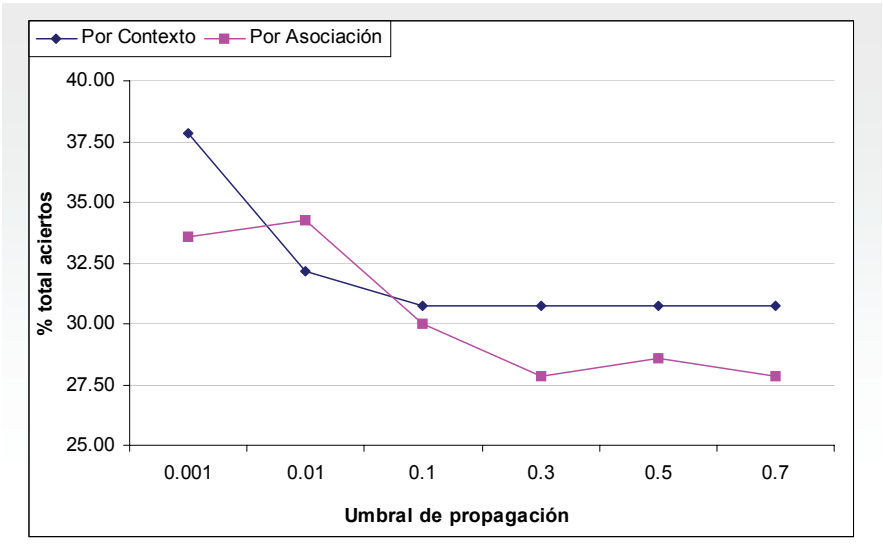


Figura 5.18. Comparación de los porcentajes totales de aciertos de predicción de conceptos con el modelo de lectura de SILC utilizando inferencia "por contexto global" y "por asociación local".

En la Figura 5.18 se pone también de manifiesto la mayor dependencia entre el umbral mínimo de propagación y la eficacia de la predicción de los conceptos para la inferencia "por contexto de asociación". En el caso de la inferencia "por contexto global", el porcentaje de acierto se mantiene constante a partir del valor igual a 0.1, existiendo en todo momento una tendencia decreciente. Sin embargo, en el caso de la inferencia "por asociación local" la tendencia no es siempre decreciente, mostrando cierta estabilización a partir de un valor del umbral de 0.3.

Una vez obtenidos los aciertos de las inferencias realizadas por el modelo, se ha procedido al cálculo de los porcentajes de acierto de los sujetos humanos. La Tabla 5.20 muestra los porcentajes de acierto medios entre todos los sujetos para cada categoría y tamaño de los indicios.

	Inferencia ="sujetos humanos"					
Tamaño indicio	**Categorías**					
	Ciencia	**Cultura**	**Deportes**	**Econom.**	**Salud**	***Total***
0	40.00	35.00	60.00	34.38	35.00	40.18
1	80.00	73.33	73.33	54.17	73.33	69.05
2	80.00	100.00	90.00	93.75	80.00	89.29
3	80.00	80.00	100.00	87.50	80.00	85.71
4	80.00	100.00	100.00	87.50	80.00	89.29
Total	72.00	77.67	84.67	71.46	69.67	74.70

Tabla 5.20. Porcentajes medios de acierto en la predicción de conceptos por parte de sujetos humanos para textos de distintas categorías y con distintos tamaños de indicio.

Como era de esperar, los valores de la Tabla 5.20 muestran una relación directa entre el porcentaje de aciertos y el tamaño de los indicios. Los resultados medios para cada categoría indican que los sujetos evaluados infieren de manera más certera los conceptos en las categorías Ciencia, Cultura y Deportes, mientras que obtienen un menor porcentaje de acierto en las categorías Economía y Salud, puesto que son las que requieren un mayor nivel de especialización y poseen un vocabulario más específico. Las mismas relaciones entre el tamaño de los indicios, las categorías y los porcentajes de acierto se dan también en las predicciones realizadas por el modelo de SILC, ya sea mediante la inferencia "por contexto global" o "por asociación local". La Figura 5.19 muestra los porcentajes de acierto por tamaño de indicio, para cada categoría y la media de todas ellas, de los sujetos humanos y del modelo de lectura empleando los dos tipos de inferencia con los valores de umbral mínimo de propagación que mayor porcentaje

de acierto obtienen en cada caso, es decir, 0.001 para la inferencia "por contexto global" y 0.01 para la inferencia "por asociación local".

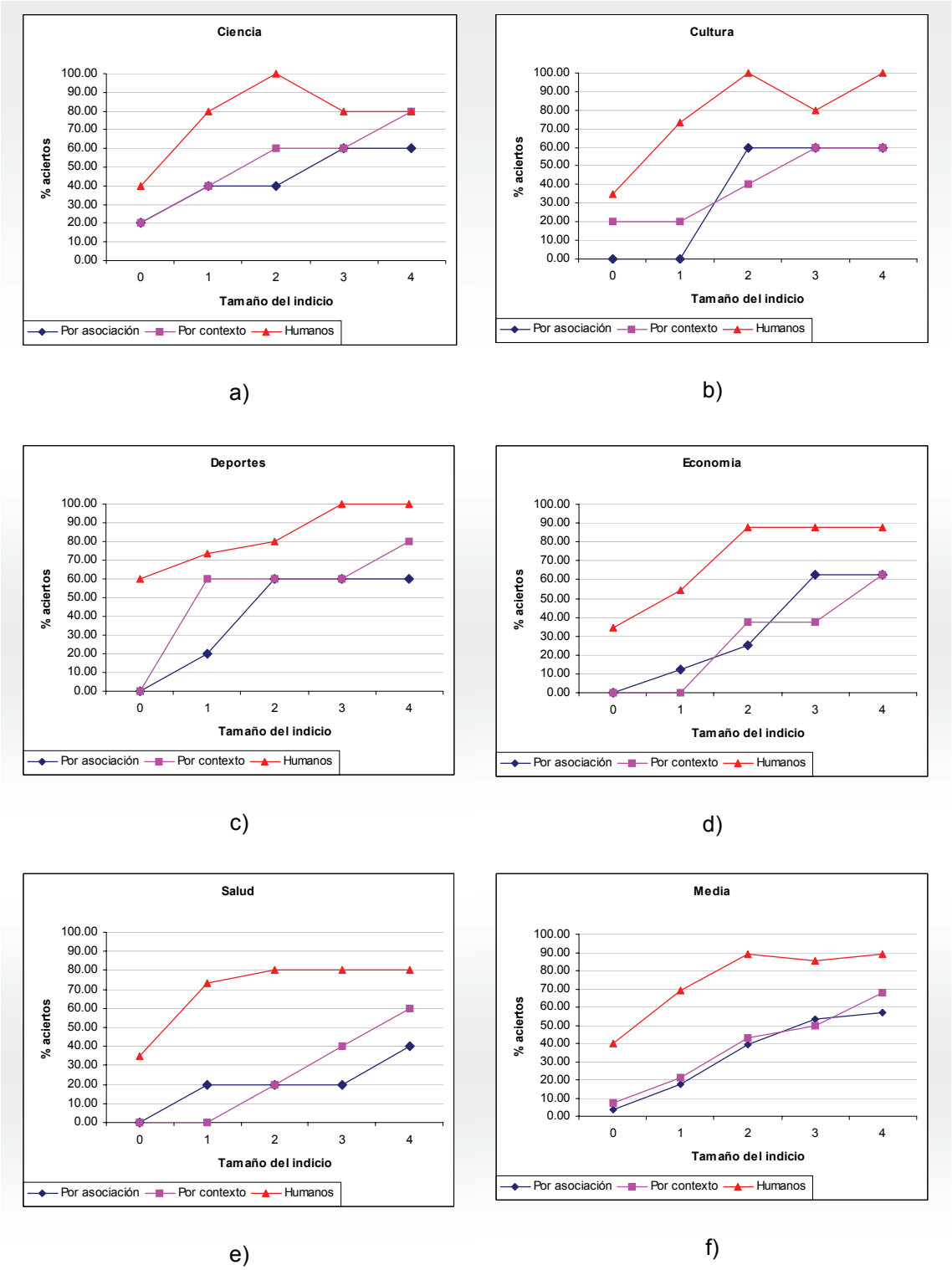


Figura 5.19. Porcentajes totales de aciertos de predicción de conceptos obtenidos por sujetos humanos y por el modelo de lectura de SILC mediante inferencia "por contexto global" y "por asociación local" para distintos tamaños de indicio y para las categorías a) Ciencia y Tecnología, b) Cultura y Espectáculos, c) Deportes, d) Economía, e) Salud y f) la media de todas ellas.

Como se aprecia en la figura, la inferencia "por contexto global" obtiene porcentajes de aciertos mayores, y por tanto más cercanos a los de los sujetos humanos, que la inferencia "por asociación local". Sin embargo, en cuanto al comportamiento con respecto al tamaño de los indicios determinado por la forma de las gráficas, este último tipo de inferencia se asemeja mucho más al de los sujetos humanos en cualquier categoría. De esta relación se puede deducir que los seres humanos infieren, en primera instancia, los conceptos asociados semántica o gramaticalmente al último concepto leído que sean coherentes con el contexto semántico. Nótese también que, aunque la inferencia "por contexto global" obtenga mayor porcentaje de aciertos que la inferencia "por asociación local" de manera general, esta última iguala o supera a la primera en las categorías más difíciles de inferir, como son Economía y Salud, sobre todo ante indicios de tamaño reducido, concretamente igual a 1. Una vez más, este hecho indica que en contextos semánticos confusos donde se desconocen muchos de los términos, las asociaciones locales, ya sean sintácticas o semánticas, son las que permiten al lector construir la semántica y la representación de texto base de las oraciones que lee.

Así pues, los resultados experimentales ponen de manifiesto la generalidad del modelo y su cualidad de marco experimental para la validación de hipótesis sobre aspectos de la lectura, como es la inferencia o predicción de conceptos. Además, los resultados demuestran que las representaciones intermedias generadas por el modelo son coherentes con las de los seres humanos, puesto que se han desarrollado mecanismos que, partiendo de la información que contienen las mismas, hacen que el sistema se comporte de manera similar a los sujetos evaluados.

5.5.5. Evaluación de la similitud "off-line" con los seres humanos

Como ya se ha descrito en el procedimiento experimental, las representaciones generadas por los seres humanos han sido comparadas con las representaciones generadas por el modelo de lectura de SILC utilizando diferentes valores de los parámetros. La variación de dichos valores ha sido análoga a la realizada en la optimización de parámetros. En primer lugar, se han comparado las representaciones de los cinco textos generadas por los 20 sujetos con las generadas por SILC para los mismos textos utilizando diferentes valores para el factor de olvido, desde 0.9 hasta 0.3 en intervalos de 0.2, y diferentes valores para el umbral mínimo de propagación, desde 0.01 hasta el valor del factor de olvido menos 0.2 en cada caso, en intervalos de 0.2. La

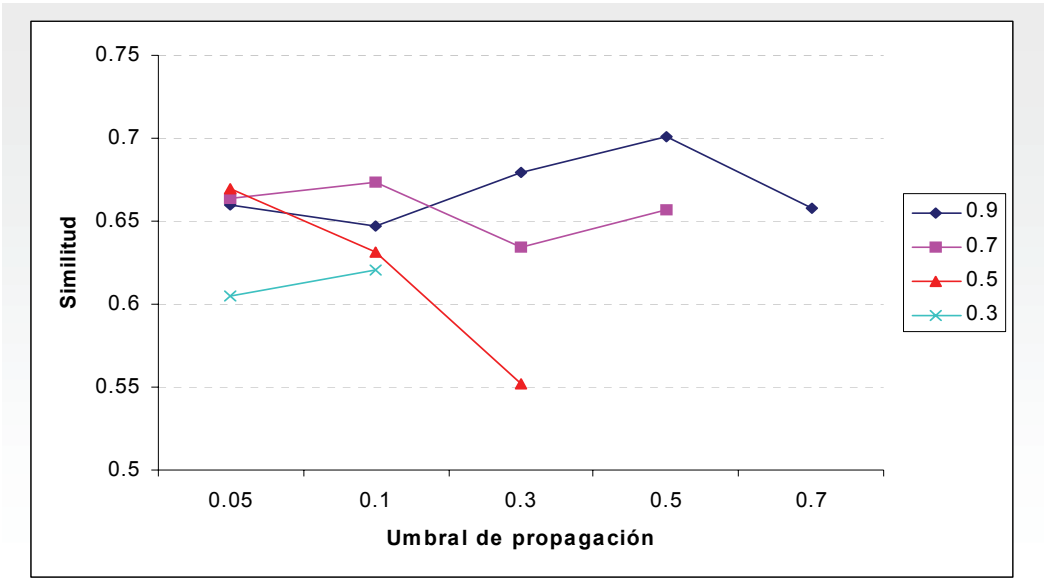
Tabla 5.21 muestra la similitud (1 menos la diferencia) media para todos los textos entre las representaciones generadas por todos los sujetos y las generadas por SILC con distintas combinaciones de valores de los parámetros anteriormente mencionados.

Umbral mínimo de propagación	**Factor de Olvido**			
	0.3	**0.5**	**0.7**	**0.9**
0.01	0.605	0.669	0.664	0.660
0.1	0.621	0.631	0.673	0.647
0.3	-	0.552	0.634	0.679
0.5	-	-	0.657	0.701
0.7	-	-	-	0.658

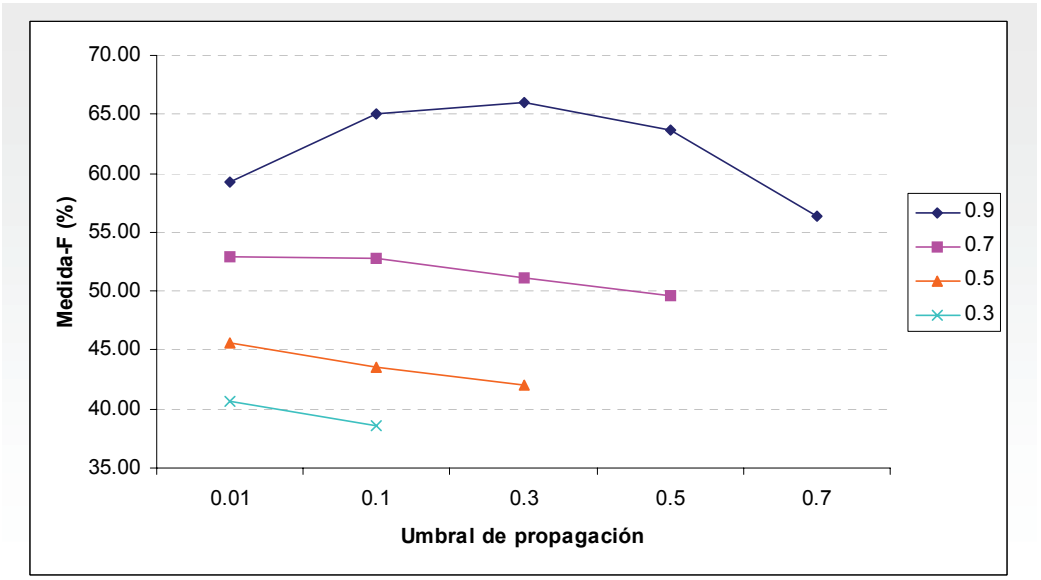
Tabla 5.21. Valores medios de similtud entre representaciones de textos generadas por sujetos humanos y las representaciones generadas por SILC con diferentes combinaciones de valores del umbral mínimo de propagación y el factor de olvido.

Como muestran los valores de la tabla, la máxima similitud del modelo de lectura de SILC con los sujetos humanos se alcanza con los valores 0.9 y 0.5 del factor de olvido y umbral mínimo de propagación, respectivamente. Recuérdese que los valores de estos parámetros que mejores resultados de clasificación inducen son 0.9 y 0.3, respectivamente (ver sección 5.2.4.1), lo que corrobora la hipótesis de que cuanto más se asemeja el modelo al ser humano mejor realiza las tareas de lenguaje natural, puesto que la representación de la semántica de los textos es más precisa.

La Figura 5.20a muestra la Tabla 5.21 de manera gráfica y la Figura 5.20b muestra los resultados análogos de optimización para la clasificación de textos de los mismos parámetros (ver Figura 5.1). La relación entre la similitud con los sujetos humanos y la eficacia en clasificación se pone de manifiesto si se comparan ambas gráficas. Como se observa, no sólo los resultados máximos coinciden sino que además las relaciones entre los parámetros y la medida-F, y los parámetros y la similitud "off-line" con los seres humanos muestran una tendencia similar. La figura muestra también cómo un factor de olvido mayor se asemeja más a los seres humanos y a su vez produce mejores resultados de clasificación. El umbral de propagación también presenta tendencias similares en ambos casos, salvo para un factor de olvido reducido igual a 0.3.



a)



b)

Figura 5.20. Valores medios de a) similitud "off-line" con los sujetos humanos y b) medida-F en tareas de clasificación de textos representados por SILC empleando diferentes combinaciones de valores para los parámetros umbral mínimo de propagación y factor de olvido.

Al igual que en los experimentos de optimización, el siguiente paso ha sido evaluar la similitud del modelo con los seres humanos en función de sus parámetros relativos al tipo y nivel máximo de propagación. Fijando el intervalo de olvido al valor que más similitud obtiene, igual a 0.9, y umbral mínimo de propagación igual a 0.01 para permitir la inferencia de un mayor número de conceptos, se ha variado el nivel máximo de propagación desde 1 hasta 3, en intervalos de 1, para los tipos de propagación, en

profundidad y por niveles. Las representaciones así obtenidas se han comparado una vez más con las generadas por los sujetos humanos. La similitud media para cada valor de los parámetros evaluado se muestra en la Tabla 5.22.

Tipo de propagación	**Nivel máximo de propagación**		
	1	**2**	**3**
En profundidad	0.666	0.665	0.660
Por niveles	0.664	0.654	0.654

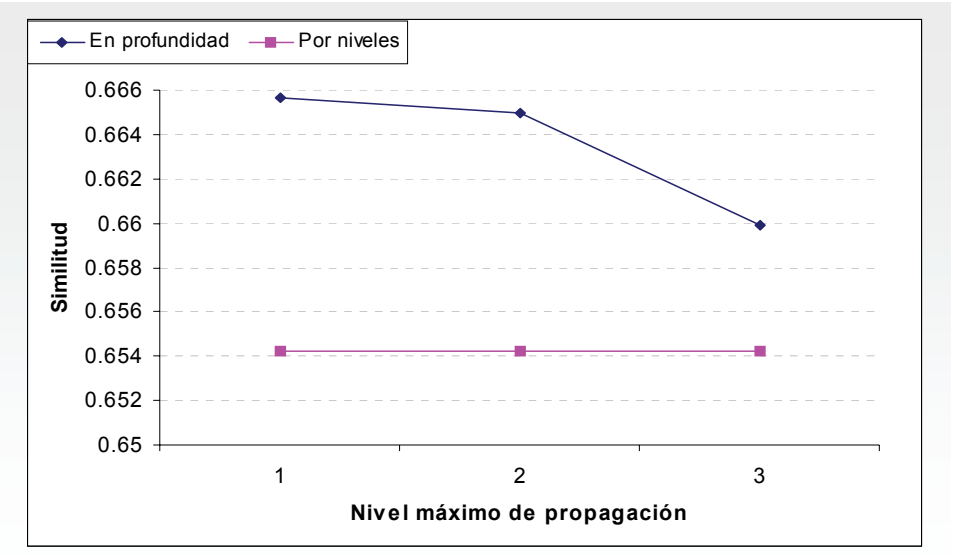
Tabla 5.22. Valores medios de similtud entre representaciones de textos generadas por sujetos humanos y las representaciones generadas por SILC con diferentes combinaciones del tipo de propagación y el valor del nivel máximo de propagación.

Los valores de la tabla muestran que las representaciones del modelo se asemejan más a las de los sujetos humanos cuando emplea la propagación en profundidad y un nivel de máximo de propagación de 1, es decir, cuando sólo infiere relaciones directas entre las palabras, aunque las diferencias son mínimas al igual que en los resultados de clasificación. Gráficamente, la Figura 5.21a presenta los mismos resultados de la Tabla 5.22, comparándolos con la tendencia de la relación entre los mismos y la eficacia obtenida en la clasificación de textos (medida-F) en la Figura 5.21b (ver Figura 5.2).

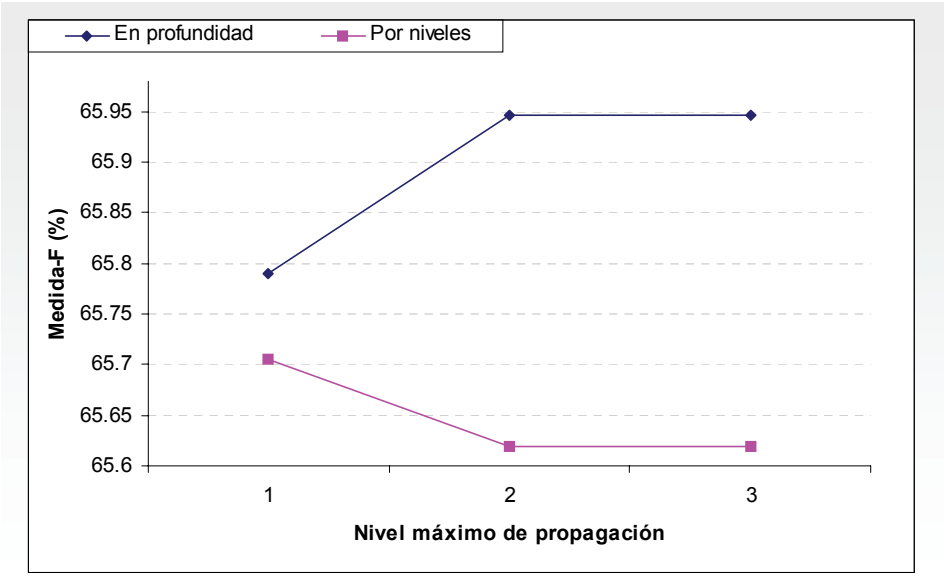
La comparativa muestra cómo efectivamente la propagación en profundidad, que es la que obtiene mejores resultados de clasificación, es también la que hace que las representaciones del modelo se asemejen más a los sujetos humanos. En cuanto al nivel máximo de propagación, se da la misma tendencia si se emplea la propagación por niveles (ver Figura 5.2b) aunque la tendencia es inversa si se emplea la propagación en profundidad. Esto se debe a que los algoritmos de clasificación de textos necesitan más información para encontrar regularidades o patrones que los seres humanos, y esa información la aporta un nivel de propagación mayor.

Por último, se han realizado experimentos análogos variando el tamaño del intervalo de olvido desde 5 hasta 19 palabras en intervalos de 2. Para el resto de parámetros se han empleado los valores que más similitud inducen: factor de olvido de 0.9, un umbral mínimo de propagación de 0.5, propagación en profundidad con un nivel máximo de 1.

La similitud media entre las representaciones del modelo y las provenientes de los sujetos humanos para cada tamaño del intervalo de olvido se presentan en la Tabla 5.23.



a)



b)

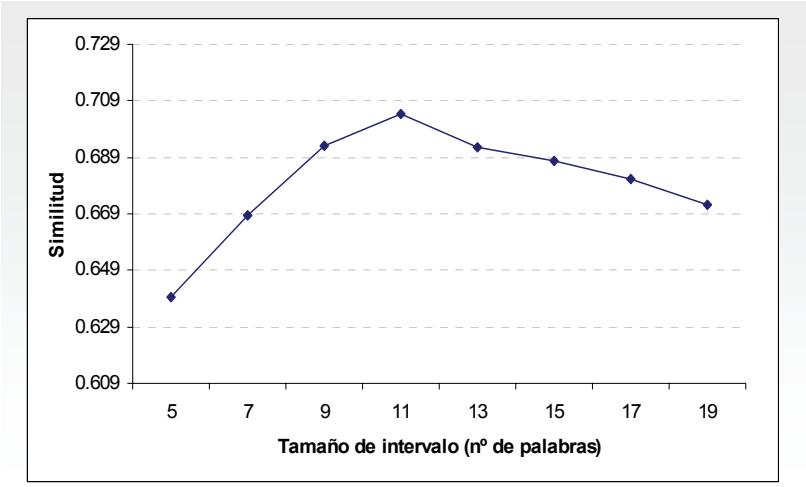
Figura 5.21. Valores medios de a) similitud "off-line" con los sujetos humanos y b) medida-F en tareas de clasificación de textos representados por SILC empleando diferentes combinaciones del tipo de propagación y el valor del nivel máximo de propagación.

Intervalo de Olvido	Similitud
5	0.640
7	0.668
9	0.693
11	0.704
13	0.693
15	0.688
17	0.681
19	0.672

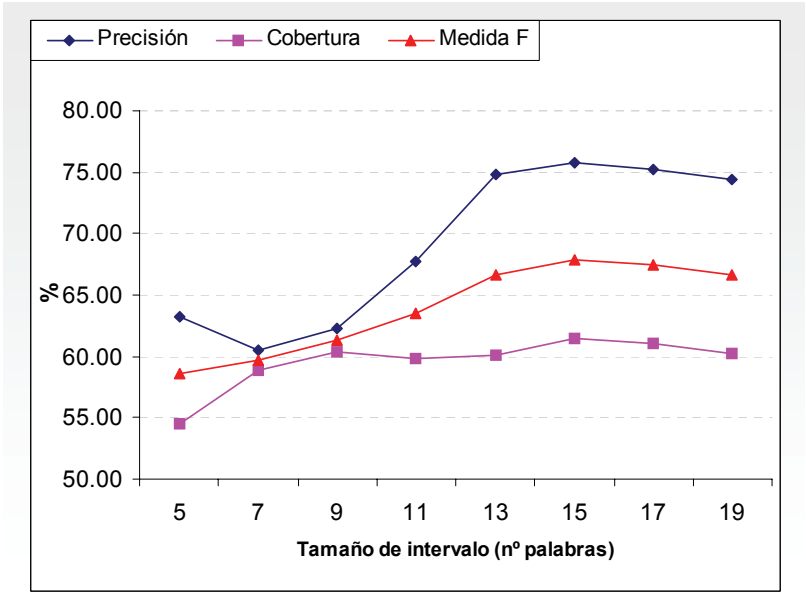
Tabla 5.23. Valores medios de similtud entre representaciones de textos generadas por sujetos humanos y las representaciones generadas por SILC con diferentes valores del tamaño fijo del intervalo de olvido.

La Figura 5.22a muestra la tendencia de la relación entre el tamaño del intervalo de olvido y la similitud "off-line". Como se puede apreciar, el valor del tamaño de intervalo que produce la máxima similitud es de 11 palabras. Si se observa la Figura 5.22b (ver Figura 5.3), la gráfica de la medida-F alcanza su máximo valor para un intervalo de 15 palabras. Esto hace que las gráficas se presenten desplazadas la una con respecto a la otra en el eje de ordenadas, aunque el trazado es prácticamente el mismo.

Así pues, una vez más se pone de manifiesto la relación directa entre la similitud del modelo con los humanos y la eficacia de discriminación de las representaciones que el mismo produce. Nótese además que en los experimentos anteriores el modelo empleaba las oraciones como intervalo de olvido (tamaño variable), alcanzando una máxima similitud de 0.701 (ver Tabla 5.21), y la similitud máxima obtenida utilizando un intervalo de tamaño fijo (11 palabras) es de 0.704. Al igual que en los experimentos de optimización, los resultados máximos con el intervalo tamaño fijo son ligeramente superiores que los obtenidos con intervalo de tamaño variable, manteniéndose de nuevo la relación entre la similitud "off-line" con los seres humanos y la eficacia que las representaciones de los textos obtienen cuando se utilizan en tareas de clasificación de textos.



a)



b)

Figura 5.22. Valores medios de a) similitud "off-line" con los sujetos humanos y b) medida-F en tareas de clasificación de textos representados por SILC empleando diferentes valores del tamaño fijo de intervalo de olvido.

Para finalizar, se representaron mediante SILC, con los valores de los parámetros que hacen al modelo más similar a los seres humanos, las 10 noticias del conjunto de datos sin incluir sus titulares. Una vez representadas, se calculó la posición y activación media de cada una de las palabras de dichos titulares que estaban recogidas en el conocimiento semántico lingüístico empleado, contenidas en las representaciones correspondientes ordenadas de manera decreciente por nivel de activación. La Tabla 5.24 muestra la posición relativa (con respecto al total de conceptos en la representación) media y el porcentaje medio (con respecto al máximo nivel) de activación, de las palabras de los

titulares en las representaciones de las noticias correspondientes, y el número y porcentaje medio de palabras de titulares que el sistema conocía, es decir, que estaban recogidas en el conocimiento semántico-lingüístico empleado.

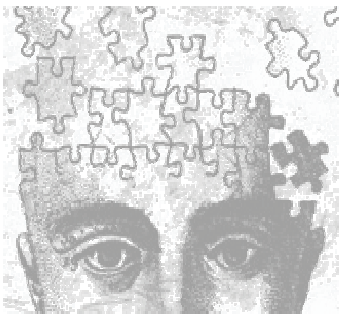
Posición relativa (%) media	% Activación media	Nº medio de palabras en cada título	% medio de palabras conocidas en cada título
22.32	51.60	5.1	73.98

Tabla 5.24. Posición relativa media y porcentaje medio de activación de las palabras en los títulos de las noticias, y número medio de palabras en títulos y porcentaje medio de palabras conocidas en cada título .

Los resultados de la Tabla 5.24 indican que, utilizando los parámetros indicados, cada palabra de un título se encuentra, en media, entre el 22.32% de los conceptos más activos de la representación generada por SILC de su texto correspondiente. Cada palabra de un título posee en media un 51.60% de la activación que posee el concepto más activo de la representación generada. Como se puede apreciar, el porcentaje de palabras conocidas por el sistema, 73.98%, es suficiente para validar los resultados anteriores.

En definitiva, la experimentación corrobora la hipótesis de que cuanto más se asemeje el modelo a los seres humanos mejores resultados obtendrán las representaciones generadas por el mismo en tareas prácticas de procesamiento de lenguaje natural, además de mostrar que dichas representaciones recogen la esencia semántica, según de los textos correspondientes.

CAPÍTULO 6



Conclusiones y Trabajo Futuro

Aunque parezca una paradoja, la finalización de un trabajo de investigación científica no debería concluirlo. En la ciencia, generalmente, el término de un trabajo da la solución a una pregunta previa que motivó al mismo, y cuya respuesta podía estar planteada en forma de hipótesis. Puesto que el conocimiento es un elemento vivo y en constante evolución, la resolución de sus enigmas hace que se generen nuevas preguntas. Así pues, uno puede estar orgulloso del resultado de una investigación si éste le induce más interrogantes y, por tanto, más inquietudes intelectuales de las que tenía antes de llegar a ella. Por esta razón, el cierre de un trabajo siempre tiene un sabor agridulce. Cada vez que se constata que se ha seguido el sendero del conocimiento correcto se llega a otra intersección del mismo con muchas más opciones que la anterior. La cuantificación del conocimiento que se posee depende del conocimiento del que se es consciente que no se posee. Es decir, el aprendizaje depende inversamente de lo que se conoce. Sin embargo, lejos de quitar el aliento, el hecho de encontrar encrucijadas de conocimiento más y más complejas hace que aumenten los ánimos de avanzar por ellas aunque sea sólo por la inherente curiosidad del ser humano. El hecho es que los interrogantes recursivamente complejos rinden un futuro más variado, permitiendo elegir entre muchas más opciones y, por ello, haciendo sentirse al ser humano algo más próximo al libre albedrío, más próximo a la libertad, aunque quizás también más lejos de la cordura. En definitiva, como dijo Sócrates: “Sólo sé que no sé nada; pero procuro saber un poco más” (nótese la segunda oración de la cita, comúnmente omitida).

6.1. Recapitulación

Los sistemas de procesamiento de lenguaje natural han sufrido diversos cambios de naturaleza a lo largo de su historia, motivados principalmente por la fe de la comunidad científica en la utilidad de los mismos. El lenguaje es una cualidad inherentemente humana. Al tratar de dotar a los ordenadores con esa cualidad se introducen limitaciones y simplificaciones con el fin de adaptar los mecanismos del lenguaje a una arquitectura física concreta. Esta adaptación hace que la dinámica de los sistemas se aleje del proceder de los seres humanos y, por tanto, se produzca una pérdida de confianza en los mismos y se desvanezca su capacidad de esbozar nuevo conocimiento sobre los aspectos cognitivos del lenguaje.

La especialización intensa de las tareas que conciernen al lenguaje ha hecho que muchos sistemas actuales, cuyos mecanismos están puramente basados en las matemáticas o la estadística, resulten realmente útiles en dichas tareas. Además, dichos sistemas son eficientes puesto que aprovechan al máximo los recursos de los ordenadores, por otra parte más amplios que los de los seres humanos. Sin embargo, a medida que avanza la tecnología y se extiende el uso de las máquinas, se requieren interfaces con los usuarios que les inspiren confianza y sencillez en su uso. El lenguaje natural aporta estos dos requisitos, pero los sistemas de procesamiento del mismo existentes están demasiado especializados y, dada la artificialidad y la complejidad de sus mecanismos, una extensión de su aplicación a tareas del lenguaje complejas se antoja realmente complicada. Por otro lado, los primeros sistemas de procesamiento de lenguaje natural, aunque con mecanismos basados en los aspectos psicológicos y cognitivos conocidos, no son en absoluto eficientes en tiempo y espacio computacional.

Así pues, con el objetivo de superar estas limitaciones junto con la propia avidez de conocimiento sobre la cognición humana, se ha presentado en esta tesis el diseño y desarrollo de un sistema de procesamiento de lenguaje natural para representar la semántica de textos escritos en el mismo, denominado SILC (Sistema para la Indexación de textos mediante un modelo de Lectura Cognitiva), cuyos mecanismos tratan de aproximarse lo más posible al proceder del ser humano. Según la hipótesis en la que se basa el trabajo realizado, esta proximidad al ser humano ha reflejado una mejor realización de las tareas relativas al lenguaje ya que el ser humano continúa

superponiéndose a las máquinas en dichas tareas aunque, claro está, de manera menos eficiente. Para superar las posibles limitaciones que el soporte físico informático impone se han definido estructuras, sobre las que el sistema opera, que simulan la fisonomía neuronal cerebral. Estas estructuras son las que dan la naturaleza conexionista al sistema propuesto y también las que permiten obtener un balance entre la plausibilidad psicológica y la eficiencia, además de aportar cierta validez a priori a las hipótesis funcionales que emergen del modelo de lectura que implementa el sistema.

En primer lugar, se ha diseñado una estructura de conocimiento semántico previo sobre el que opera el modelo de lectura que obtiene las representaciones semánticas de los textos. Se ha definido un formalismo de representación de dicho conocimiento en forma de red de nodos interconectados de manera bidireccional, donde cada nodo representa un concepto. La conexión de los nodos denota la asociación semántica o gramatical entre los conceptos que representan y se cuantifica con un valor. Se han diseñado también mecanismos para construir y modificar dicha red a partir de una colección de textos fuente, de los que se obtienen los conceptos, la relación entre los mismos como concurrencia en un mismo contexto y la cuantificación de dicha relación como la proporción de concurrencia. Este proceso trata de modelar la representación y adquisición de la semántica del lenguaje.

A continuación, se ha diseñado un modelo computacional de lectura de textos que opera sobre el conocimiento semántico lingüístico previamente adquirido. El modelo simula la lectura de un texto escrito en lenguaje natural y produce una representación de la semántica del mismo. El modelo representa dicha semántica en forma de red de conceptos asociados, cada uno de ellos con un nivel de activación que denota su significación dentro del texto. La dinámica del modelo procesa cada palabra de los textos de manera secuencial respetando el orden en el que aparecen. A cada lectura de una palabra se activa su concepto asociado y se generan inferencias de otros conceptos mediante la propagación de la activación del concepto leído a través de las asociaciones presentes en el conocimiento semántico lingüístico. Los conceptos leídos e inferidos, junto con sus asociaciones y su nivel de activación, se almacenan en la memoria de trabajo, por lo que el modelo siempre tiene una representación del texto leído en cada momento. El transcurso de la lectura hace que los conceptos presentes en la memoria pierdan activación y sean finalmente olvidados si no se reactivan mediante su lectura o su inferencia, modelando así la influencia del propio discurso y estilo de los escritos en

la comprensión de su semántica. Con todos estos aspectos del diseño del modelo se ha perseguido la plausibilidad psicológica y la proximidad con el proceder de los seres humanos.

Puesto que se trata de un modelo computacional, la realización informática de los mecanismos que lo componen ha dado lugar a diversas opciones de implementación y distintos parámetros. Se han realizado experimentos de optimización de dichos parámetros y de los aspectos de implementación en tareas de clasificación automática de textos. Se han obtenido los valores de los mismos que hacen que el modelo genere las representaciones que mejores resultados ofrecen en dicha tarea utilizando algoritmos de aprendizaje automático tradicionales. Estos resultados han superado incluso a otros sistemas de representación clásicos como son la "bolsa de palabras" *tf·idf* y LSI (sistema de indexación a partir del análisis de semántica latente LSA). También se han determinado en los experimentos de optimización los valores óptimos de los aspectos relacionados con la construcción del conocimiento semántico lingüístico, como son el tamaño y temática de la colección de textos fuente y la definición del contexto de asociación de conceptos, siendo las propias oraciones el contexto que mejores resultados produce, además de ser psicológicamente más plausible y de capturar la gramática latente de manera intencionada, al contrario que el resto de sistemas existentes.

Puesto que los algoritmos de clasificación empleados para los experimentos de optimización sólo admiten vectores como entrada, las representaciones generadas por el sistema SILC han sido transformadas a dicho formalismo, evaluando la precisión semántica de los conceptos contenidos en las representaciones junto con su nivel de activación. Dado que las representaciones generadas son estructuradas, es decir, están organizadas en forma de red y contienen información sobre cómo se asocian los conceptos presentes, se han ideado también mecanismos que hacen uso de la riqueza de dichas representaciones en tareas prácticas de procesamiento de lenguaje natural, como es, una vez más, la clasificación automática de textos. Así pues, se ha diseñado una medida de similitud semántica entre conceptos basada en la distancia del camino mínimo que une a dichos conceptos en el conocimiento semántico lingüístico. A partir de dicha medida de similitud entre conceptos, se han creado tres medidas de similitud entre textos o, más concretamente, entre las representaciones generadas por SILC a partir de la lectura de los textos. Se ha definido también un contexto o ámbito local para

realizar la comparación entre las representaciones de los textos, que supone una reducción significativa del conocimiento semántico lingüístico total y, por tanto, una amplia mejora de la eficiencia en el cálculo de la similitud. Además, la experimentación ha demostrado que esta reducción del contexto de comparación también aporta una mejora en la precisión del cálculo de la similitud, puesto que elimina la ambigüedad en gran medida. Se han ideado también mecanismos y expresiones para cuantificar el poder de discriminación de una medida de similitud. Estos mecanismos han sido empleados para optimizar los parámetros de las medidas de similitud propuestas, como son el número de los conceptos más activos que intervienen en la comparación, la normalización de los niveles de activación de los conceptos o los aspectos relacionados con la construcción del contexto o ámbito de comparación. Posteriormente, los mecanismos de evaluación de las medidas de similitud han permitido determinar qué medida con sus parámetros óptimos es la que mayor precisión semántica obtiene.

Dicha medida ha sido empleada posteriormente para la clasificación automática de textos representados mediante el modelo de lectura de SILC. Para dicha tarea se ha empleado un método basado en "centroide". Así pues, se ha diseñado e implementado un método que obtiene un "centroide" o ejemplo representativo de una categoría temática a partir de representaciones generadas por SILC de textos de dicha categoría. Se ha empleado la mejor medida de similitud mencionada para comparar representaciones de SILC con los "centroides" construidos y asignarles la categoría del "centroide" más similar a los mismos. Los resultados de clasificación obtenidos superan nuevamente a los obtenidos por las representaciones *tf·idf* y LSI, esta vez con una colección de textos que plantea un reto de clasificación más duro y ambiguo que la empleada en los experimentos de optimización. Los resultados obtenidos hacen al sistema SILC aplicable en tareas prácticas de procesamiento de lenguaje natural. Es necesario remarcar que se han empleado para la experimentación colecciones de textos en distintos idiomas, por lo que el sistema es independiente del mismo salvo en su etapa de preprocesamiento.

Por último, se han realizado experimentos para evaluar la similitud del modelo de lectura implementado por SILC con los seres humanos. Esta evaluación ha tenido dos fines: constatar la plausibilidad psicológica del sistema y corroborar la hipótesis de que cuanto más se asemeje un sistema de procesamiento de lenguaje general al ser humano más eficaz se mostrará en las tareas prácticas del mismo. En primer lugar, se ha

evaluado la similitud "on-line", es decir, la similitud con los seres humanos durante el proceso de lectura. Para ello se han diseñado e implementado sobre la estructura de SILC dos maneras de realizar inferencia o predicción de conceptos omitidos en un texto con posibles indicios: la inferencia "por contexto global" y la inferencia "por asociación local". Ambos mecanismos emplean la representación generada por el modelo hasta el momento de realizar la predicción, por lo que la experimentación también es útil para evaluar estas representaciones intermedias. Se han comparado los resultados de acierto en la predicción de los conceptos omitidos por parte del modelo, utilizando ambos mecanismos de predicción, y por parte de los seres humanos. Los resultados revelan que una de la dos formas de inferencia, la inferencia "por asociación local", aunque obtiene menor porcentaje de aciertos que la otra forma de inferencia, se comporta de manera análoga a los seres humanos ante la variación de los indicios proporcionados sobre los conceptos omitidos y ante las distintas categorías temáticas a las que pertenecen los textos en los que se han omitido los conceptos. De los resultados de esta experimentación se puede concluir que las representaciones intermedias son adecuadas, que la estructura y dinámica de SILC es psicológicamente plausible, y que el sistema aporta un buen marco experimental que permite una fácil incorporación y evaluación de hipótesis sobre aspectos cognitivos de la lectura, como es la predicción o inferencia de conceptos.

En segundo lugar, se ha evaluado la similitud "off-line", es decir, al término de la lectura, entre el modelo y los seres humanos. Para ello se ha ideado una medida de similitud entre las representaciones generadas por SILC y los resúmenes de los textos escritos por sujetos humanos. Dichos resúmenes tratan de recoger las ideas ordenadas por significación que los sujetos tienen acerca de los textos. De manera análoga a los experimentos de optimización, las representaciones generadas por SILC empleando distintos valores para sus parámetros han sido comparadas con los resúmenes generados por los sujetos humanos. Los resultados muestran que los valores de los parámetros que hacen que el modelo genere las representaciones más parecidas a las de los seres humanos son muy próximos a los valores que hacen que el modelo genere las representaciones que mejores resultados obtienen en la tarea de clasificación. Más aún, los resultados revelan una relación directa entre la similitud con las representaciones de los sujetos humanos y la eficacia en clasificación de textos para cualquiera de los parámetros. Esta relación hallada corrobora la hipótesis de partida de que a mayor

semejanza con los humanos mejor eficacia en las tareas prácticas. Los resultados arrojan también una prueba experimental de la plausibilidad psicológica de SILC y hacen pensar en aplicaciones futuras como la validación de textos o exámenes y el modelado del proceso cognitivo de lectura en sujetos individuales para la detección y tratamiento de posibles disfunciones. Se han comparado también las representaciones generadas por SILC con los títulos de los textos correspondientes, identificando si los conceptos de dichos títulos están recogidos en la representación y con qué nivel de activación. Los resultados muestran que las representaciones generadas por SILC contienen la semántica esencial, representada por el título, de los textos correspondientes con un nivel de activación significativo. Además, este procedimiento se propone como una nueva medida "off-line" de la eficacia de los sistemas de representación semántica.

Todas las estructuras, métodos y algoritmos descritos han sido implementados e integrados en un entorno software desarrollado específicamente para la evaluación experimental del sistema propuesto esta tesis doctoral. Todo el software ha sido desarrollado íntegramente en lenguaje C++, utilizando el *framework* Borland C++ Builder 6.0 para sistema operativo Windows. Ejemplos de la interfaz de dicho software se recogen en el Apéndice III.

6.1.1. Evaluación del modelo computacional de lectura

Según los criterios de evaluación propuestos por Fletcher (ver Capítulo 3, Sección 3.2.5), el modelo implementado por SILC:

- No requiere simplificaciones de los textos de entrada, salvo la homogeneidad del idioma. Admite cualquier tipo de texto escrito en lenguaje natural siempre y cuando todo él esté escrito en el mismo idioma de manera íntegra.
- Permite el aislamiento y evaluación individual de sus módulos a través de la variación de los valores de sus parámetros. Con ello, el modelo aporta una manera de descubrir la influencia de cada aspecto en el resultado final, evaluación que resulta imposible en experimentación con sujetos humanos.
- Provee de medidas de similitud con el comportamiento humano, tanto "on-line" como "off-line", mediante la comparación con resúmenes o mediante la predicción de conceptos omitidos. Así mismo, el modelo provee otras medidas implícitas que permiten compararlo con los seres humanos, como la

cantidad total y máxima de activación en cada instante, la cantidad máxima y total de conceptos en cada momento, o los tiempos de lectura o procesamiento de diversas unidades lingüísticas.

- Tiene utilidad y es aplicable a la resolución de tareas reales como, al menos, la clasificación automática de textos.

6.2. Cumplimiento de los Objetivos Propuestos

Se ha desarrollado un sistema que transforma un texto escrito en lenguaje natural en una represtación estructurada de la semántica que éste encierra, de tal forma que dicha representación puede ser empleada en diversas tareas prácticas de procesamiento de lenguaje natural.

Según los objetivos preliminares expuestos en el Capítulo 1, la conclusión del trabajo de investigación y desarrollo de esta tesis doctoral constata su cumplimiento. En términos funcionales y estructurales:

- El sistema lleva a cabo una etapa de adquisición de la semántica automáticamente, ya que se ha diseñado e implementado un algoritmo que construye una red de asociación que representa el conocimiento semántico lingüístico a partir de una colección extensa de textos en lenguaje natural. Tanto el formalismo de representación empleado como los mecanismos que operan sobre el mismo permiten la actualización (ampliación, reducción o modificación) de dicho conocimiento de manera sencilla y eficiente.
- El sistema lleva a cabo una etapa de comprensión de manera automática, ya que se ha concebido e implementado un modelo computacional de lectura, compuesto por diversos mecanismos, que extrae la semántica de un texto y la representa de tal forma que permite su utilización en aplicaciones prácticas de manera sencilla.
- El modelo computacional de lectura que lleva a cabo la etapa de comprensión es psicológicamente plausible, es decir, se basa en evidencias neurológicas y psicológicas sobre el cerebro y la cognición humana:
    - Las estructuras sobre las que opera son de índole conexionista, simulando la fisonomía cerebral humana.
    - Los mecanismos de inferencia durante la lectura se realizan mediante la propagación de activación a través de las estructuras, lo que simula la dinámica funcional del cerebro humano.
    - Los textos se procesan palabra a palabra en orden secuencial, al igual que los seres humanos.

  - En cada momento se tiene una representación semántica del texto leído hasta el momento.
  - El orden y estilo de escritura de los textos influye en la representación final de su semántica.
  - Permite modelar la intensidad de la percepción de las palabras.
  - Implementa mecanismos de inferencia que simulan diversas teorías existentes sobre dicho aspecto.
  - Se contemplan dos tipos de memoria: a largo plazo y a largo plazo de trabajo. En la primera se almacena el conocimiento semántico lingüístico adquirido. En la segunda las representaciones de la semántica de los textos que se leen.
  - Se contempla el transcurso del tiempo a lo largo del proceso lectura mediante el olvido de elementos de la memoria.
  - No necesita conocer todas las palabras de un texto para generar una representación precisa de su semántica.

- Aporta un marco y herramienta experimental para la evaluación de hipótesis sobre aspectos cognitivos de la lectura, como muestra la sencilla integración y evaluación de los dos posibles métodos de predicción de conceptos implementados. En términos estructurales, el sistema permite también la evaluación de hipótesis mediante la modificación de los parámetros del mismo. Además, el sistema contiene medidas implícitas de similitud "on-line" con los seres humanos análogas a los mismos, como son la capacidad de la memoria o el tiempo de lectura. Las representaciones generadas por el sistema son también una fuente directa para la comparación "off-line", es decir, al término del proceso de lectura, con los seres humanos.

Entre los objetivos también se encontraba la evaluación del sistema en tareas prácticas de procesamiento de lenguaje natural, concretamente en la clasificación automática de textos:

- Se ha diseñado e implementado un método para la transformación de las representaciones generadas por el sistema SILC a vectores que sirvan como entrada a los algoritmos de clasificación automática de textos empleados.

- Se han realizado experimentos para optimizar los parámetros del modelo que hacen que las representaciones generadas por el mismo transformadas en vectores obtengan la eficacia máxima en la tarea de clasificación.
- Se han realizado experimentos para comparar la eficacia en clasificación de las representaciones generadas por el sistema propuesto, en forma vectorial, con sus parámetros optimizados y la eficacia de las representaciones "bolsa de palabras" *tf·idf* y LSI con diversos algoritmos de clasificación, obteniendo las primeras los mejores resultados.
- Se han ideado e implementado tres medidas distintas para cuantificar la similitud semántica entre representaciones estructurales generadas por el sistema.
- Se han realizado experimentos para la optimización de los parámetros de dichas medidas y para determinar cuál de ellas es la que posee el mayor poder de discriminación semántica.
- Se han evaluado las representaciones estructurales generadas por el sistema en tareas de clasificación empleando la mejor de las medidas de similitud utilizando un algoritmo de clasificación de textos basado en "centroides".
- Se ha diseñado e implementado un método de construcción de "centroides" que caracterizan la semántica de una categoría temática a partir de un conjunto de representaciones generadas por el sistema.
- Se han realizado experimentos para la comparación entre la eficacia en clasificación de textos obtenida por las representaciones estructurales junto con la mejor medida de similitud, y la obtenida por las representaciones "bolsa de palabras" *tf·idf* y LSI con diversos algoritmos de clasificación, obteniendo las primeras los mejores resultados.

En definitiva, los resultados de experimentación demuestran que SILC puede ser aplicado en tareas prácticas de procesamiento de lenguaje natural.

Por último, otro de los objetivos era comprobar la plausibilidad psicológica del modelo de manera experimental, y con ello corroborar la hipótesis de partida de que cuanto más se asemeje el modelo al ser humano mejor realizará las tareas prácticas:

- Se han ideado e implementado dos mecanismos de inferencia o predicción de conceptos omitidos en un texto, que utilizan la representación del texto generada hasta el momento.
- Se han realizado experimentos para medir la similitud "on-line" del sistema con los seres humanos, mediante la comparación del porcentaje de aciertos de ambos en la predicción de conceptos. Los resultados han mostrado que uno de los mecanismos, la inferencia "por asociación local", se comporta igual que los seres humanos ante la variación del tamaño de los indicios y ante las distintas categorías temáticas, lo que indica que las representaciones generadas y la dinámica del modelo son adecuadas para reflejar el proceder humano.
- Se ha ideado e implementado un método para comparar las representaciones finales generadas por el sistema con resúmenes generados por sujetos humanos.
- Se ha ideado un método para cuantificar la semántica esencial que recogen las representaciones generadas por SILC mediante la comprobación de la existencia y nivel de activación en las mismas de los conceptos que componen los títulos de los textos representados.
- Se han realizado experimentos para comparar "off-line", mediante los métodos anteriores, a sujetos humanos con el modelo, bajo distintas configuraciones de sus parámetros, de manera análoga a la optimización para la clasificación de textos. Los resultados han mostrado una relación directa entre la similitud con los sujetos humanos y la eficacia en clasificación de textos, corroborando así la hipótesis de partida. Además, se ha demostrado que el sistema es capaz de recoger la semántica esencial, descrita por los títulos de los textos, de manera significativa en las representaciones producidas correspondientes a dichos títulos.

Los resultados de comparación obtenidos de manera individual para cada sujeto presentan ligeras variaciones entre uno y otro, lo que esboza la posibilidad de emplear el sistema para modelar a sujetos individuales y estudiar tanto posibles disfunciones en la lectura como las maneras de ayudar a superarlas.

6.3. Aportaciones del Trabajo de Tesis Doctoral

Puesto que el tema principal del trabajo es de carácter multidisciplinar, se han realizado aportaciones que conciernen a diversas áreas de conocimiento:

- En primer lugar, se ha realizado un amplio estudio del estado del arte de los sistemas de representación e indexación semántica de textos hasta el momento no presente en la literatura.
- Respecto a la construcción de redes semánticas conceptuales, se ha propuesto un nuevo método de construcción de las mismas a partir de colecciones de textos que emplea un contexto de asociación de conceptos de tamaño variable y absolutamente determinado por los propios textos, como son las oraciones, y una nueva forma de cuantificar o ponderar las asociaciones semánticas entre conceptos.
- Respecto a la similitud semántica, se han propuesto tres nuevas medidas de similitud entre redes semánticas conceptuales basadas en los caminos óptimos entre los conceptos de las representaciones que se comparan.
- Respecto a la similitud semántica, se ha propuesto un nuevo método de reducción del ámbito o contexto de búsqueda de caminos óptimos en una red conceptual, consistiendo dicha reducción en la unión de las redes que se comparan más una serie de caminos de enlace entre los nodos más significativos de las mismas. El método de reducción hace que la búsqueda de caminos óptimos sea más eficiente y más eficaz, en términos de similitud semántica.
- Respecto a la similitud semántica, se ha propuesto un nuevo método y una nueva medida para cuantificar el poder de discriminación semántica de una medida de similitud mediante los conceptos “intraclase” e “interclase”.
- Respecto a la construcción de “centroides”, se ha propuesto un nuevo método para dicha tarea que parte de un conjunto de redes conceptuales de una misma categoría temática.

- Respecto a los modelos de lectura, se ha propuesto un nuevo modelo computacional que plantea hipótesis sobre cómo se lleva a cabo dicho proceso en los seres humanos.
- Respecto a los modelos de lectura, se han planteado dos hipótesis sobre la dinámica de la inferencia o predicción de conceptos en los seres humanos.
- Respecto a los modelos de lectura, se ha demostrado experimentalmente que los seres humanos realizan la predicción o inferencia de conceptos en base a la última palabra leída, usando el contexto semántico derivado de la lectura como complemento.
- Respecto a los modelos de lectura, se ha propuesto una nueva medida de similitud "off-line" entre el modelo y los sujetos humanos, mediante la comparación de las representaciones de los textos generadas por el modelo y los resúmenes de dichos textos generados por sujetos humanos.
- Respecto a los modelos de lectura, se ha propuesto un nuevo método de evaluar la capacidad semántica de las representaciones obtenidas mediante la comprobación de la existencia y cuantificación del nivel de significación de los conceptos que componen el título de los textos representados.
- Respecto a los modelos de lectura, se ha propuesto un marco y herramienta experimental que permite la incorporación y evaluación sencilla de hipótesis sobre aspectos cognitivos del proceso de lectura.
- Respecto a la indexación de textos, se ha propuesto un nuevo método que transforma un texto escrito en lenguaje natural en una representación estructurada de la semántica del mismo, donde el orden y el estilo de escritura influyen en la representación final, y donde la representación de la semántica de los conceptos y la de los textos son distintas.
- Respecto a la indexación de textos, se ha corroborado de manera experimental la hipótesis de que cuanto más se asemeje el sistema de indexación al ser humano, más eficaces serán las representaciones que genere en aplicaciones de lenguaje natural.

6.4. Trabajo Futuro

De igual manera que las aportaciones, el carácter multidisciplinar del trabajo presentado en esta tesis plantea líneas de investigación futura de diversa naturaleza.

En primer lugar, en cuanto a aplicaciones prácticas de lenguaje natural se refiere, se pretende utilizar las representaciones generadas por SILC en tareas de recuperación de información (búsqueda de documentos relacionados con una consulta expresada en lenguaje natural) y para la validación de textos. Además, se pretende utilizar dichas representaciones como fuente de aplicaciones de generación de lenguaje natural, y junto con éstas para la generación de resúmenes, tanto mono-documento como multi-documento, empleando en el último caso el método de construcción de "centroides".

En cuanto al formalismo de representación, se pretende incorporar asociaciones que denoten expresamente las relaciones sintácticas entre las palabras. Se pretende también estudiar algún método que emplee la información de las asociaciones, tanto semánticas como sintácticas, para cualificar a los nodos conceptuales con niveles de abstracción, o distancia de lo perceptible, profundidad, o cantidad de conocimiento, y generalidad a los que hace referencia su semántica. Así mismo, se estudiarán métodos para incorporar al conocimiento conceptos desconocidos que se encuentren durante el proceso de lectura, utilizando para ello el contexto donde se hallen dichos conceptos y la información de los conceptos que se infieran en su lugar para dotarlos de significado. Se presenta también como una tarea futura la construcción e integración homogénea en el sistema actual de otra red semántica de alto nivel que permita asociar los nodos conceptuales de la red de conocimiento semántico lingüístico mediante relaciones retóricas del discurso. Por ejemplo, si un nodo de la red de alto nivel es "Causalidad", los nodos de la red de conocimiento lingüístico "fuego" y "calor" estarían conectados a dicho nodo de alto nivel. De esta manera, se podrán utilizar las representaciones generadas por el modelo para responder (sistemas de pregunta-respuesta) o actuar consecuentemente con los textos leídos. Se estudiarán también mecanismos de representación de pensamientos propios o creencias y la influencia de éstas en el proceso de lectura y en la semántica extraída de los textos que el sistema procesa.

Respecto a la similitud semántica, se pretende idear nuevas medidas de similitud semántica entre conceptos y entre textos que empleen los niveles de abstracción, profundidad y generalidad anteriormente mencionados. Se pretende también utilizar las medidas de similitud para comparar las representaciones en puntos intermedios de los textos comparados, y utilizar dichas similitudes intermedias para componer un valor de similitud global entre dos textos.

En cuanto al modelado de aspectos cognitivos de la lectura, se pretende crear métodos que reflejen la influencia de la percepción de las palabras, ya sea por su formato o por su posición en el texto, empleando para ello el nivel de activación inicial con el que se incorporan a la memoria de trabajo. Se pretende también modelar la atención dedicada a la lectura mediante la variación de los parámetros del modelo. Otro objetivo futuro es representar e integrar en el sistema los intereses o inquietudes temáticas que posee cualquier lector, y modelar la influencia de dichos intereses en el proceso de lectura. Así, de manera general, se pretende variar los parámetros de lectura de manera dinámica, condicionando dicha variación a la interacción entre el estilo del texto, la cantidad de conceptos desconocidos, las cualidades de la representación semántica del texto leído en cada comento, los intereses temáticos previos y las creencias u opiniones a priori.

Por último, se estudiará la validez de características cuantitativas del modelo como medidas de similitud "on-line" con los seres humanos, como son la capacidad de la memoria de trabajo, el nivel máximo y total de activación, el número máximo y total de conceptos en memoria, etc. Empleando estas medidas junto con las medidas propuestas en esta tesis, se pretende utilizar el sistema para modelar el proceso y capacidad lectora de sujetos individuales, ajustando los parámetros del modelo que le hagan asemejarse en mayor medida a los mismos. El objetivo de dicho modelado es, en primer lugar, la detección o diagnóstico de limitaciones o disfunciones en la lectura. Se estudiará pues la bondad de los valores de los parámetros del modelo como indicadores de la existencia de dichas limitaciones. En segundo lugar, el modelado de individuos pretende ayudar al tratamiento y la corrección de las limitaciones o problemas detectados, estudiando la influencia de la variación de la forma, estilo y orden de los textos que se leen en la comprensión final de los mismos.

APÉNDICE I

Textos Empleados para la Similitud "on-line" con Sujetos Humanos

A continuación se muestran los textos de cada categoría empleados para la predicción o inferencia de conceptos, cuyo porcentaje de aciertos se utilizó como medida de similitud "on-line" entre el modelo de lectura de SILC y los sujetos humanos.

Categoría: Ciencia y Tecnología

Conceptos omitidos en orden de aparición:

ancha
ciento
tecnología
conexiones
zonas

Texto:

España supera los 4.600.000 usuarios de banda ancha.

En el mes de octubre se dieron de alta 166.620 nuevas líneas de banda _____ que llegaron a 4.613.835 líneas de ADSL y fibra óptica (cable), informaron fuentes de la Asociación de Internautas.

El ADSL se incrementa en octubre en 131.620 nuevos accesos a Internet El ADSL crece un 54 por _____ en los últimos 12 meses con un incremento de 1.266.222 de nuevas altas, y se consolida como el motor de la banda ancha en España.

Por su parte, los operadores de cable, según estimaciones de la Asociación de Internautas, aportaron 35.000 nuevos abonados y alcanzan los 996.000.

El ADSL de Telefónica sigue en cabeza

Sobre el total acumulado (4.613.835 líneas), el ADSL de Telefónica cuenta con 3.229.743 líneas, el desagregado suma 388.092. Además, esta _____ sigue siendo mayoritaria en gran parte de las Comunidades Autónomas, representando el 78 por ciento de la banda ancha de todas las tecnologías, alcanzando las 3.617.835 _____.

El 80 por ciento de las conexiones de banda ancha instaladas corresponden a _____ urbanas, pero se comienza a percibir un cierto cambio. Madrid y Barcelona que hasta ahora representaban el 40 por ciento de las líneas ADSL instaladas, ahora, suman el 35,76 por ciento.

Categoría: Cultura y Espectáculos

Conceptos omitidos en orden de aparición:

pintura
arte
disciplinas
antigüedad
público

Texto:

Feriarte 2005 se inaugura hoy con mayor presencia internacional.

La feria Internacional de Arte y Antigüedades, Feriarte, inaugurará esta noche su XIX edición, en la que contará con mayor presencia internacional gracias a la incorporación de anticuarios de Francia, Italia, Alemania, Portugal, Austria y Bélgica.

Más de 17.000 piezas, con al menos cien años de antigüedad, se exhibirán en Feriarte 2005, un salón que supone un encuentro con lo mejor del pasado, y donde se mostrarán desde muebles, arqueología, _____ , escultura, joyas, relojes, hasta arte religioso, porcelana, lámparas, plata, arte oriental o arte africano, entre otros.

Feriarte ofrece la posibilidad de adquirir directamente el mobiliario y los complementos de decoración que se exhiben y, según un comunicado de la feria, 'se confirma que el éxito en decoración sigue estando en emparejar pasado con presente, Oriente con Occidente y muebles antiguos con obras de _____.

Para garantizar la calidad de las piezas expuestas, una comisión integrada por especialistas en distintas _____ se encarga de revisar los objetos para asegurar que su _____, calidad y restauración son los permitidos por el certamen.

La Feria Internacional de Arte y Antigüedades permanecerá abierta hasta el 27 de noviembre y a partir de mañana podrá acceder el _____.

Categoría: Deportes

Conceptos omitidos en orden de aparición:

selección
vestuarios
clasificación
jugadores
suplente

Texto:

Turquía podría quedar excluida del M'2010.

El presidente de la FIFA anunció que el organismo ha abierto una investigación por los incidentes del miércoles.

El presidente de la FIFA, Joseph Blatter, amenazó ayer con excluir a la selección de Turquía de la fase de clasificación para el Mundial 2010, a raíz de los incidentes ocurridos el miércoles en el partido contra Suiza, en Estambul. En ese encuentro, de repesca para el Mundial del próximo verano, varios integrantes de la _____ fueron agredidos en el túnel de _____ tras eliminar a Turquía.

Blatter anunció una extensa investigación sobre los incidentes ocurridos en el estadio Surku Saracoglu. "No descarto sanciones a jugadores ni la suspensión de la Federación Turca de un gran torneo", dijo el presidente de la FIFA en referencia a la fase de _____ para el Mundial de 2010. "Algo así no lo había vivido nunca, hemos iniciado una investigación y sancionaremos a los culpables", añadió Blatter, quien podría anunciar sanciones antes de que se celebre el sorteo del Mundial de Alemania, el 9 de diciembre.

Según el dirigente suizo, "no puede ocurrir que un equipo no se pueda alegrar y sus _____ tengan que abandonar el campo como ladrones. En Estambul se le dio una patada al juego limpio". La expedición de Suiza no pudo abandonar el estadio hasta dos horas después de la conclusión del partido, y algunos de sus integrantes, así como el preparador de porteros Erich Burgener, recibieron el impacto de objetos. Asimismo, algunos jugadores de la selección turca agredieron a sus colegas suizos en el túnel de vestuarios e, incluso, el jugador _____ Stéphane Grichting debió ser trasladado a un hospital, al recibir una patada en los genitales.

Categoría: Economía

Conceptos omitidos en orden de aparición:

paro
económicas
inflación
tendencia
petróleo
comercial
exportaciones
crecimiento

Texto:

Bruselas advierte de que España empieza a crecer más despacio

La tasa de _____ española igualará a la de la UE en el 2006 por primera vez en la historia.

La economía española seguirá creciendo y generando empleo por encima de la media europea durante al menos los dos próximos años, pero su ciclo alcista empieza a mostrar algunos signos de agotamiento. Ésas son las perspectivas que maneja la Comisión Europea, cuyo responsable de Asuntos Económicos, el español Joaquín Almunia, presentó ayer en Bruselas el informe estacional sobre las previsiones _____ de la Unión y sus miembros para el período 2005-2007.

«Comparándola con otros países, la situación de España es envidiable», dijo Almunia, quien, sin embargo, advierte en su informe de que el elevado déficit exterior, el progresivo deterioro de la competitividad y la pérdida de tirón del sector de la construcción empiezan a lastrar la evolución de la economía nacional, que mantendrá tasas de _____ superiores a la media en todo el período analizado. Según los cálculos de la Comisión, el producto interior bruto (PIB) seguirá incrementándose en España por encima de la media de la UE y de la Eurozona, pero con una distancia cada vez menor, dado que su ritmo se irá desacelerando mientras el europeo irá aumentando progresivamente.

Según las previsiones del departamento de Almunia, el PIB español crecerá un 3,4% este año, frente al 1,5% de la UE; un 3,2% en el 2006, frente al 2,1% comunitario; y un 3% en el 2007, frente al 2,4% de la Unión. En los tres casos, Bruselas rebaja las perspectivas plasmadas en su anterior informe, publicado la primavera pasada.

Aunque de manera menos acusada, el empleo registrará una _____ similar al PIB, pues crecerá en España a ritmos del 3%, el 2,4% y el 2,2% esos tres años, respectivamente; mientras que en la Unión lo hará al 0,9% en el 2005, y al 1% en el 2006 y el 2007. Bruselas calcula que España llegará al 2006 con una tasa de paro del 8,5%; y al 2007, con el 8,1%. Así, por primera vez en su historia, el porcentaje de desempleados españoles será idéntico al de la media de los Veinticinco, que en

ese período, según la Comisión, crearán seis millones de puestos de trabajo.

El equipo de Almunia interpreta que la previsible recuperación del crecimiento europeo se deberá a la consolidación de la demanda interna «y al restablecimiento de la confianza de los agentes económicos», a pesar del riego, no descartado, de que se produzca una nueva escalada de los precios del _____ en los próximos meses.

Confianza

La demanda interna y la confianza son también los puntos fuertes de la economía española; frente a su principal debilidad, que el Ejecutivo comunitario ubica en la negativa tendencia de su balanza _____. Bruselas cree que ese desajuste no sólo se mantendrá en los próximos años, sino que se acelerará, pues el incremento anual de las _____ rondará el 6% de aquí al 2007, mientras que las importaciones sólo aumentarán entre un 1 y un 2% al año.

Por otro lado, los mayores índices de _____ se registrarán en los países recién incorporados, especialmente en Eslovaquia y en Letonia, cuyas economías crecerán a tasas anuales superiores al 7%. En cuanto a Alemania, el Reino Unido, Francia e Italia, mantendrán ritmos mucho más lentos y en ocasiones por debajo de la media, aunque la Comisión augura una moderada tendencia a la recuperación de esas economías a partir del año 2007.

Categoría: Salud

Conceptos omitidos en orden de aparición:

tratamiento
fármaco
terapéuticas
eficacia
enfermedad

Texto:

Betaferon evita el desarrollo de la esclerosis múltiple

Los últimos resultados realizados con Betaferon (interferón beta-1b) en esclerosis múltiple, procedentes del estudio BENEFIT, han mostrado que el _____ con este _____ reduce a la mitad el riesgo de desarrollo clínico de la enfermedad, comparado con placebo, según comunica la compañía germana Schering AG.

De acuerdo con Xavier Montalbán, miembro del comité internacional del estudio, "los resultados abren nuevas posibilidades _____ en el tratamiento precoz de la esclerosis múltiple, ya que nos permitirá atacar con mayor antelación y _____, logrando, quizá, retrasar la progresión de la misma".

El estudio BENEFIT ha contado con la participación de varios centros en nuestro país, donde se estima que 30.000 personas se encuentran afectadas por la _____.

APÉNDICE II

Textos Empleados para la Similitud "off-line" con Sujetos Humanos

A continuación se muestran los textos de cada categoría empleados para la comparación de las representaciones finales de los mismos generadas por SILC con los resúmenes de dichos textos realizados por los sujetos humanos, como medida de similitud "off-line" entre el modelo de lectura de SILC y los sujetos.

Categoría: Ciencia y Tecnología

Texto:

Las orcas noruegas son los animales más contaminados del Ártico.

Las 'ballenas asesinas' noruegas que viven en el océano Ártico han arrebatado a los osos polares el dudoso honor de ser considerados los animales más contaminados de esta remota región del planeta. Según un estudio elaborado por la organización ecologista WWF/Adena, el "sumidero tóxico" en que se ha convertido el Ártico es el culpable de la grave contaminación de su fauna y flora, particularmente grave en el caso de las orcas, animales que sirven como 'indicadores' de la salud del ecosistema marino.

El estudio lo han realizado los expertos del Instituto Polar Noruego Hans Wolkers, que precisaron que es "alarmante la cantidad de contaminantes registrados en esos cetáceos". Los biólogos tomaron muestras de grasa de algunas orcas capturadas en Tysfjord, fiordo de la región ártica de Noruega, donde esos cetáceos se congregan en el invierno para alimentarse.

El Fondo precisó que la abundancia de bifenilos poliriclorados (PCB), pesticidas y de los retardantes de llama bromados hallados otorgan a estas 'ballenas asesinas' el "dudoso honor" de situar a esos animales como los más contaminados del Artico.

"La aparición de los retardantes es particularmente preocupante, porque pueden afectar a los funciones neurológicas, reproductivas y de comportamiento de los animales y los más peligrosos no están prohibidos en la actualidad", afirmó la organización ecologista. Los PCB, por otra parte, son uno de los productos químicos más tóxicos que se utilizan principalmente en equipos eléctricos y plásticos.

Las orcas, también conocidas como 'ballenas asesinas', se alimentan de casi todo tipo de peces y pequeños mamíferos, como focas, leones marinos, pinguinos, tortugas de mar, tiburones, calamares e incluso algunos tipos de ballenas más pequeñas. Cazan en grupos, por lo que son muchos los que las denominan los 'lobos del mar'.

Pero en el Ártico noruego, estas ballenas se alimentan fundamentalmente de pescados capturados cerca de la costa. Tan cerca, que además atraen a miles de turistas, que se adentran en el mar en pequeños botes para ver el espectáculo de las orcas cazando en directo, un negocio que se ha convertido en gigantesco en algunas zonas de la costa de Noruega. Y es precisamente ésta la

zona que congrega más contaminación, y donde los peces que sirven de susternto a las ballenas también están altamente contaminados.

La ministra noruega de Medio Ambiente, Helen Bjornoy, manifestó que el uso de ese tipo de sustancias "es una de las mayores amenazas medioambientales a escala global". Por ello pidió a la Unión Europea que fortalezca la legislación de sustancias químicas en Europa. "Es imperativo que la normativa REACH sirva para dejar de usar los elementos más contaminantes", dijo Bjornoy.

El REACH tiene por objeto someter todos los productos químicos actualmente producidos o importados a Europa a un sistema de registro, evaluación y autorización y será sometido a votación por el Consejo Europeo de Ministros el próximo 13 de diciembre.

Categoría: Cultura y Espectáculos

Texto:

Rafael Argullol recoge en un libro 25 años de pensamiento.

'Enciclopedia del crepúsculo' reúne 500 artículos publicados desde 1980.

Rafael Argullol, filósofo, narrador, poeta y ensayista, ha recogido parte de lo que ha sido su mundo en los últimos 25 años en un libro al que ha dado carácter enciclopédico. Unos 500 textos contenidos en 90 entradas han dado lugar a Enciclopedia del crepúsculo, título que no debe engañar: no se trata de artículos relativos a la decadencia de nada ni de nadie, sino textos escritos por la tarde, cuando la luz del día empieza a declinar.

Argullol (Barcelona, 1949), profesor de Estética en la Universitat Pompeu Fabra, es autor de más de 20 libros; debutó como novelista con Lampedusa (1981) y obtuvo el Premio Nadal en 1993 con La razón del mal. Es colaborador habitual en la prensa diaria, trabajo que, dijo ayer, le ha facilitado "la escritura literaria y de reflexión". Enciclopedia del crepúsculo (Acantilado) es una selección de artículos publicados desde 1980 sobre literatura, arte, filosofía, cine y fotografía, temas que, casi siempre, son un pretexto para análisis más amplios. El libro es una referencia cultural de los últimos 25 años, pero también el contrapunto crítico a los acontecimientos ocurridos en ese tiempo.

Las 90 voces del libro son, casi, conceptos alegóricos que agrupan los artículos por su contenido. Alegría, Infierno, Mal, Misterio, Pasión, Sentimiento y Utopía son algunas de esas voces; Fuego, por ejemplo, incluye desde una reflexión sobre los frecuentes incendios que han sufrido los grandes teatros hasta la quema del Palacio de la Sabiduría de Bagdad. El índice onomástico final resume, en cierta manera, las obsesiones, preferencias y preocupaciones del autor en los últimos 25 años. Bush y Aznar son dos de los nombres más citados porque, "hay personajes que dan para mucho, aunque eso no implica que sean los que mejor valoración merezcan".

Argullol explicó que nunca ha sido cronista político, pero sí ha comentado las consecuencias del ejercicio malsano de la política. Afirmó que escribe desde la libertad y la reflexión --"he sido crítico con todos, con Maragall, Pujol, González y Aznar"-- y no se considera un dogmático. "No formo parte de ningún partido político ni secta". Comentó que, a lo largo de los 25 años que resume el libro, una parte de su pensamiento --"los criterios básicos"-- ha permanecido anclado, no ha variado, y otra parte se ha mantenido flotante, se ha modificado con el análisis de la situación de cada momento: "Los últimos años de Felipe González, por ejemplo, fueron una asfixia, así que la llegada de Aznar supuso una cierta satisfacción, por lo menos por el cambio que suponía; pero la asfixia se repitió en el segundo mandato de Aznar".

El filósofo expuso la necesidad de que cada uno tenga su propia verdad, "revestida con la percepción que nos llega del mundo". Hay que confiar en el sistema democrático, añadió, "pero eso no implica que no nos irriten las decisiones que toman algunos políticos". Argullol también habló ayer del terror, palabra que juzga desgastada por un uso excesivo y para la que nuestros tiempos no han encontrado antídoto: "Cada época tiene su bestia y su ángel y hoy no hay ideas fuertes que se opongan a la bestia".

Categoría: Deportes

Texto:

Clemente llama "pardillos" a sus jugadores.

El técnico del Athletic Club, Javier Clemente, calificó hoy a sus jugadores de "pardillos" al referirse a la inocencia y a la falta de picardía que está mostrando su equipo, y que quedó patente el pasado sábado en el gol marcado por el valencianista Vicente en Mestalla.

"Ese es un gol de pardillos. Somos más inocentes que... De haber ganado, hubiera estado bien, pero habríamos tenido un poco de suerte porque el equipo trabajó muy bien, pero lo que hicimos fue aguantar al Valencia, ya que sólo llegamos dos veces arriba", recordó.

El técnico rojiblanco cree que el mejor Athletic de Mestalla se pudo ver durante "un cuarto de hora muy bonito en la primera parte", mientras en los 75 minutos restantes, su equipo defendió "muy bien".

Clemente ahondó en la falta de experiencia y oficio de sus jugadores, recordando lo que pasó la pasada semana con los jugadores del Atlético de Madrid en San Mamés tras el lanzamiento del petardo. "Aquel día dos tíos del Atlético se tiraron cuatro minutos a la chumbera. En cambio, a nosotros nos pegan una patada y seguimos de pie diciendo: 'si no nos han matado, para qué vamos a tirarnos'".

"Nosotros no nos sabemos esas triquiñuelas y esas trampas. De los veintitrés de la plantilla, sólo dos tienen esa picardía. El resto, dejará el fútbol y hay cosas que no habrán aprendido nunca", añadió.

A raíz de este reproche, el de Barakaldo fue cuestionado por la alineación inicial con la que sorprendió en Mestalla, donde dejó en el banquillo a jugadores de experiencia como Urzaiz, Iraola y Tiko, apostando de salida por otros que, o bien habían jugado poco, como Tarantino, o bien no lo habían hecho, como Endika Bordas.

El entrenador señaló que lo que pretende es utilizar el máximo de jugadores repartiendo el tiempo de juego, sin perder de vista "un mes de enero que viene guapo", y en el que el Athletic deberá disputar al menos seis partidos, sin olvidar que el domingo se enfrentan al Betis, un rival directo por la permanencia.

A Javier Clemente no le gusta denominar este sistema como rotaciones porque "la gente puede preguntarse si chavales de 20 años están cansados o no pueden", y prefiere decir que lo que busca es que su equipo mantenga "una alta intensidad durante todos los partidos".

OCHOS EQUIPOS PELEANDO POR LA SALVACION.

Pese a haber puntuado en Mestalla, el Athletic se halla a dos puntos de la zona de permanencia, aunque también es cierto que la clasificación se ha comprimido en exceso, y equipos que parecían estar en una posición cómoda, ahora no están lejos de los últimos clasificados.

Es el caso de la Real Sociedad, que "va a estar en el grupo de los diez últimos porque está teniendo dificultades con la baja de Kovacevic y, ahora, la de Mikel (Aranburu)". "Somos equipos con plantillas muy justas que notamos mucho las bajas de los jugadores titulares", subrayó.

En este sentido, Clemente auguró que habrá "ocho equipos en la pelea" por la salvación, y la criba tendrá lugar "en el mes de febrero, que es cuando llega la alta competición".

KARANKA, LESIONADO DE NUEVO

Respecto al entrenamiento matinal en Lezama, eminentemente físico, tanto Julen Guerrero como Ibon Gutiérrez lo completaron con normalidad. Este último viajará el miércoles en Alemania para ser operado de pubalgia.

Por otro lado, el club informó que Aitor Karanka está de nuevo lesionado. El gasteiztarra, que el pasado 12 de abril fue operado de una rotura del ligamento cruzado anterior de la rodilla izquierda y disponía del alta médica desde el 19 de septiembre, padece ahora una tendinitis en el Aquiles de su pierna izquierda.

Categoría: Economía

Texto:

OPEP frena escalada en precios del petróleo.

Los precios del petróleo cedieron hoy terreno a la apertura, apoyados en pronunciamientos de la OPEP que calificaron de exagerada la reacción de los mercados a reportes sobre posibles cambios en las cuotas de producción.

En declaraciones a la prensa, el presidente de la Organización de Países Exportadores de Petróleo (OPEP), Ahmad Fahd al-Sabah, indicó que esta previsto un encuentro el 31 de enero, lo cual pudo ser interpretado como una señal de que el cártel reducirá la extracción si es necesario.

Por su parte, el ministro de Energía de Qatar, Abdullah bin Hamad Al-Attiyah, manifestó que las cotizaciones descenderán tras un repunte inicial que llego al tres por ciento el lunes.

A su vez, analistas del banco corporativo Mizuho indicaron que era razonable un análisis de su estrategia productiva, pues por lo general con la llegada del segundo trimestre del año ocurre una caída de la demanda por el fin de la temporada invernal en el hemisferio norte.

La última reunión ministerial de la OPEP dejó sin cambios las entregas actuales de crudo, situadas en 30,3 millones de barriles diarios, en un esfuerzo por aportar tranquilidad a los consumidores.

Con esas señales, el precio del West Texas Intermediate (WTI), de referencia en el New Cork Mercantile Exchange, sufrió un descenso de 43 centavos (0,7 por ciento) y se situó en 60,87 dólares el barril.

A corto plazo, los pronósticos apuntan a un mayor consumo de gasóleo en Estados Unidos, pues se espera un notable descenso de las temperaturas en extensas áreas del país, el principal consumidor mundial de energéticos.

En el ámbito corporativo, el consorcio Royal Dutch Shell, tercero del mundo en el sector petrolero, anunció un incremento del 27 por ciento en los gastos de prospección y equipamiento el próximo año, como parte de la política para aumentar la extracción.

La cifra prevista por la compañía oscila en torno a los 19 mil millones de dólares, unos cuatro mil millones mas en comparación con las previsiones anteriores para el 2006.

Categoría: Salud

Texto:

Un estudio sugiere que el té podría combatir cáncer de ovario.

Un grupo de investigadores suecos ha hallado evidencias atractivas, pero no concluyentes, de que beber dos tazas de té cada día podría ayudar a reducir el riesgo de desarrollar cáncer de ovario.

En el estudio participaron 61.057 mujeres suecas que respondieron a un cuestionario acerca de sus dietas y a las cuales se les hizo un seguimiento durante unos 15 años, hasta el 2004.

En ese período, 301 mujeres fueron aquejadas de cáncer de ovario. Aquellas que bebían dos o más tazas de té por día eran un 46% menos proclives a desarrollar la enfermedad que las mujeres que no bebían té. Beber menos de dos tazas por día también parecía ayudar, pero no demasiado.

Los investigadores no clasificaron los resultados de acuerdo a los tipos de té empleados, pero la mayoría de las bebedoras consumían té negro. Tanto el té negro como el verde contienen polifenoles, sustancias que se cree impiden daños en las células que podrían causar cáncer.

Previos estudios sobre los efectos del té en la prevención del cáncer han ofrecido resultados contradictorios.

Las investigadoras Susanna Larsson y Alicja Wolk, del Instituto Karolinska en Estocolmo, dijeron que se requieren ulteriores investigaciones para ofrecer un diagnóstico más acertado.

El estudio fue publicado el lunes en la revista especializada Archives of Internal Medicine.

"Si los hallazgos son verdaderos, eso podría ser importante, pues el cáncer de ovario es la cuarta causa más importante de muerte por cáncer entre las mujeres", dijo Marji McCullough, de la Sociedad Oncológica de Estados Unidos.

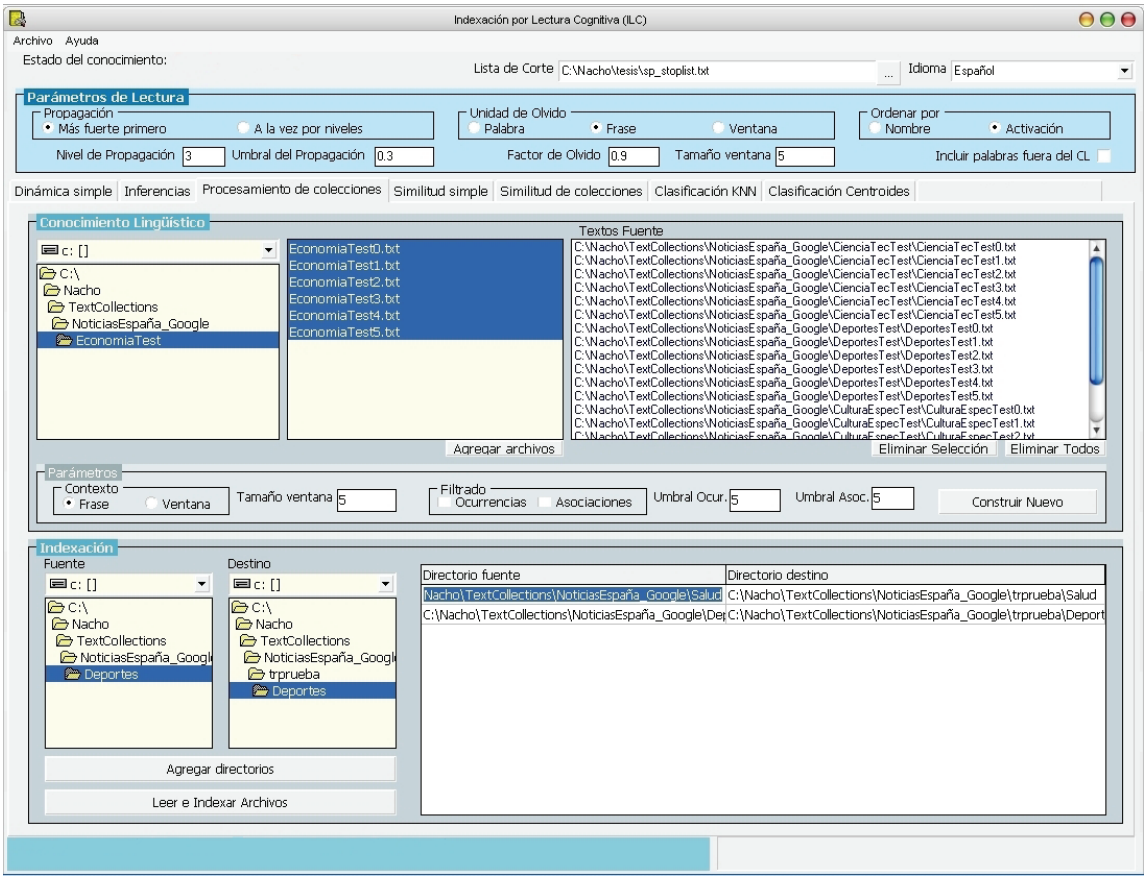
El cáncer de ovario es diagnosticado anualmente en más de 20.000 mujeres estadounidenses. Es difícil de detectar porque sus síntomas, entre ellos hinchazón abdominal, indigestión y deseos de orinar con frecuencia pueden ser vagos e imitar condiciones menos graves.

APÉNDICE III

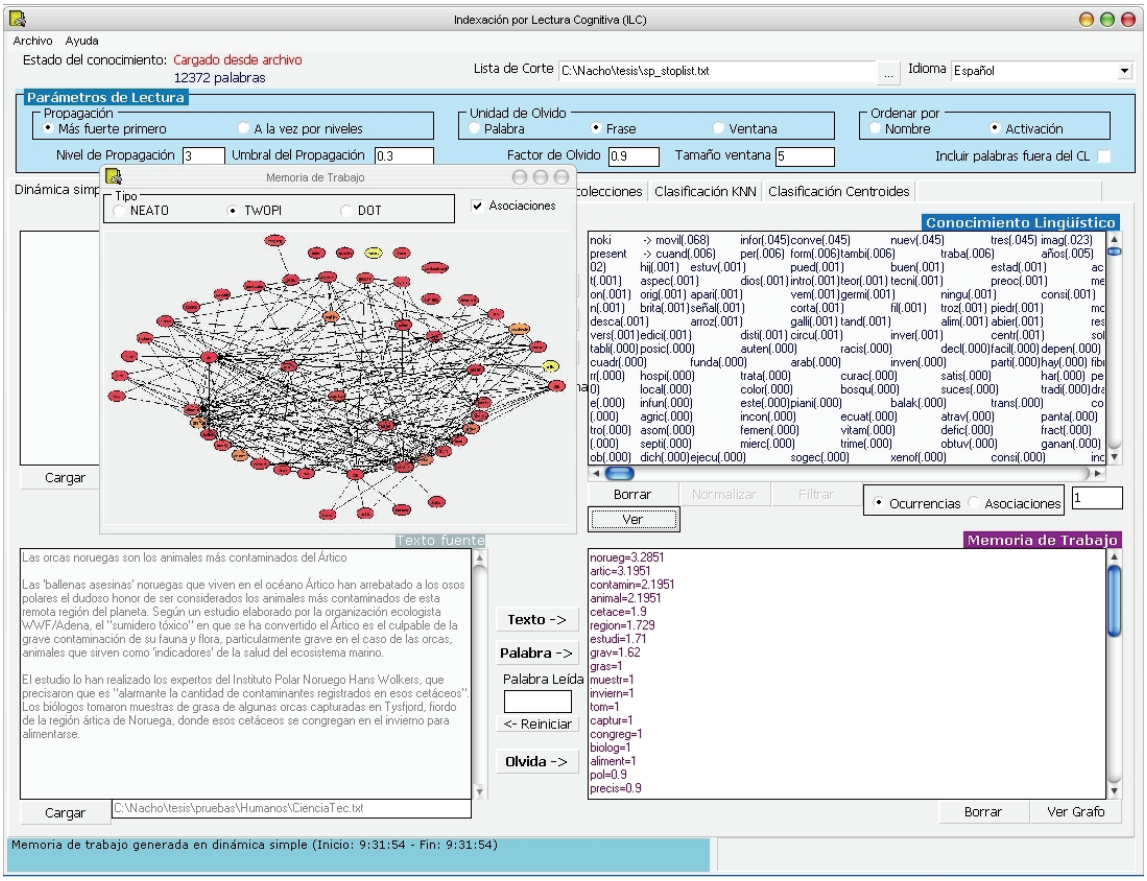
Interfaz de la Aplicación Experimental que Implementa a SILC

A continuación se muestran diversas capturas de pantallas de la aplicación software que implementa el sistema propuesto en esta tesis. Las pantallas se corresponden con diversos aspectos metodológicos y experimentales del trabajo realizado.

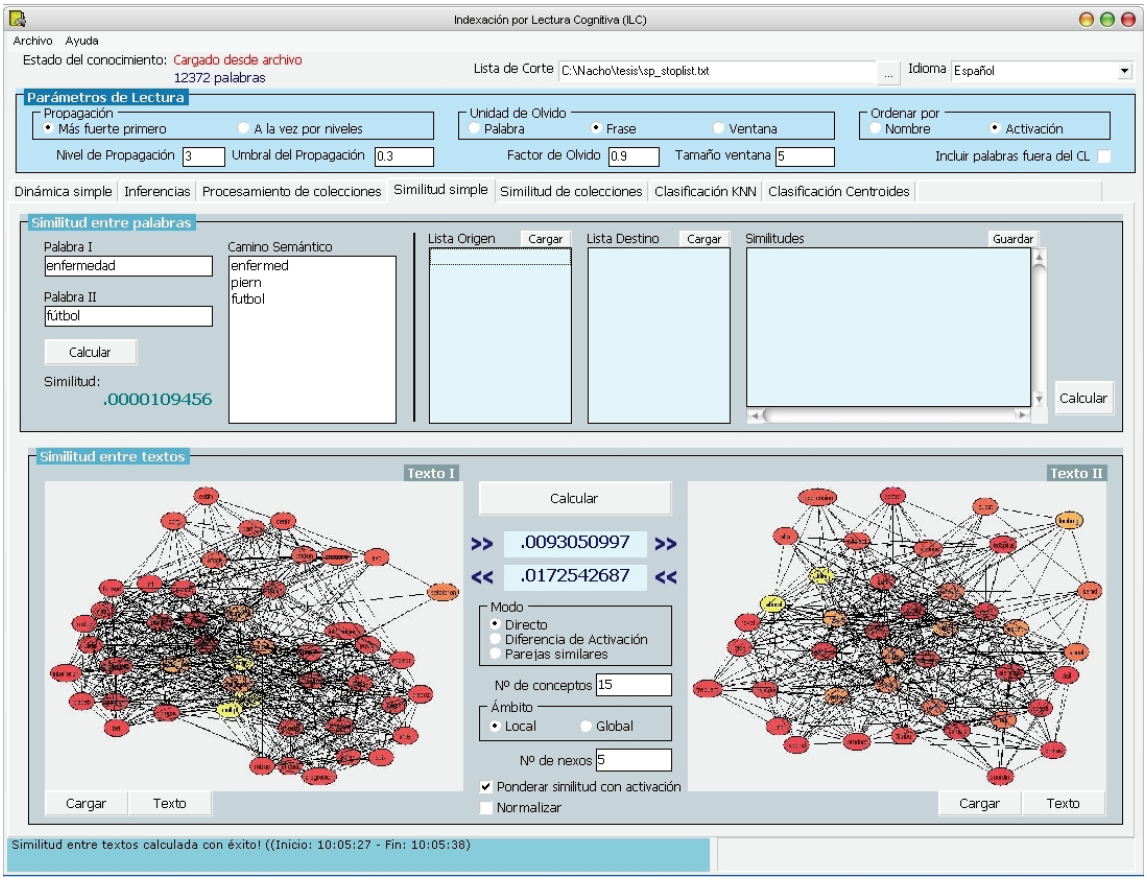
Construcción del conocimiento semántico lingüístico e indexación de colecciones de textos



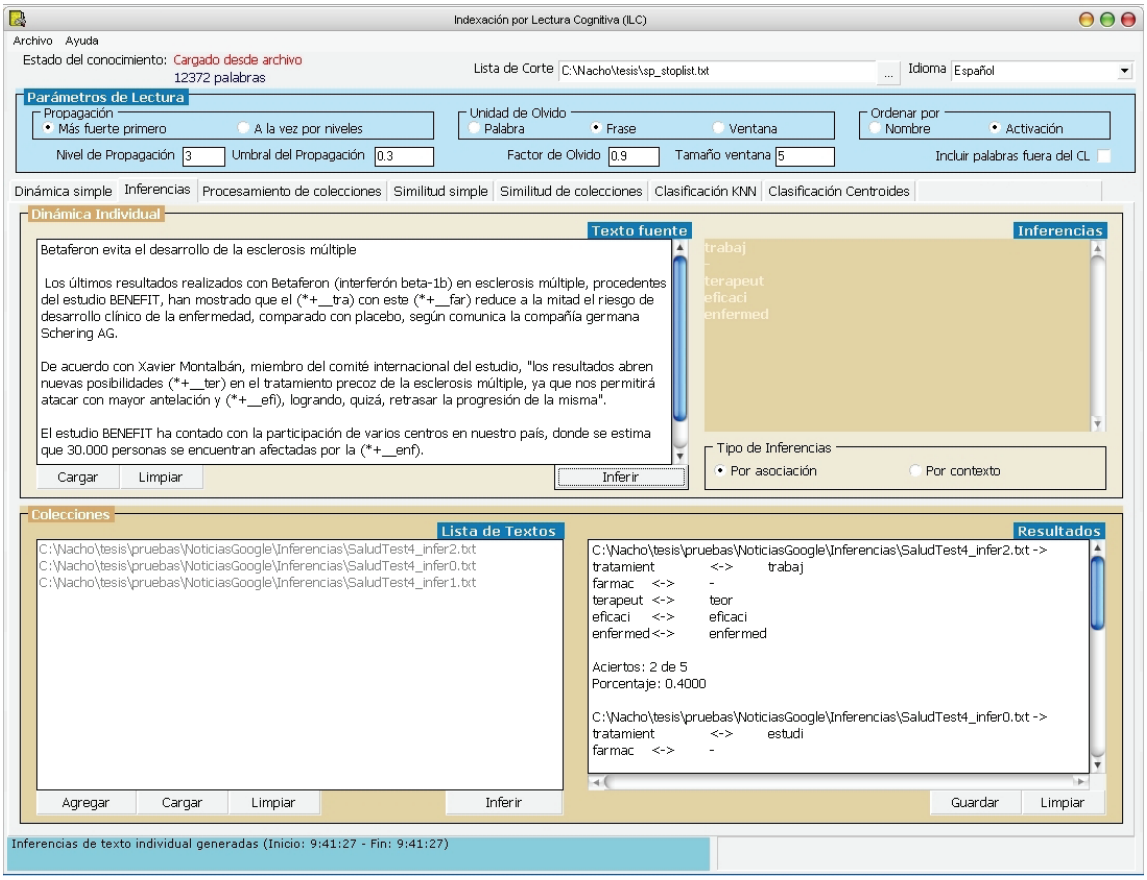
Dinámica del modelo de lectura



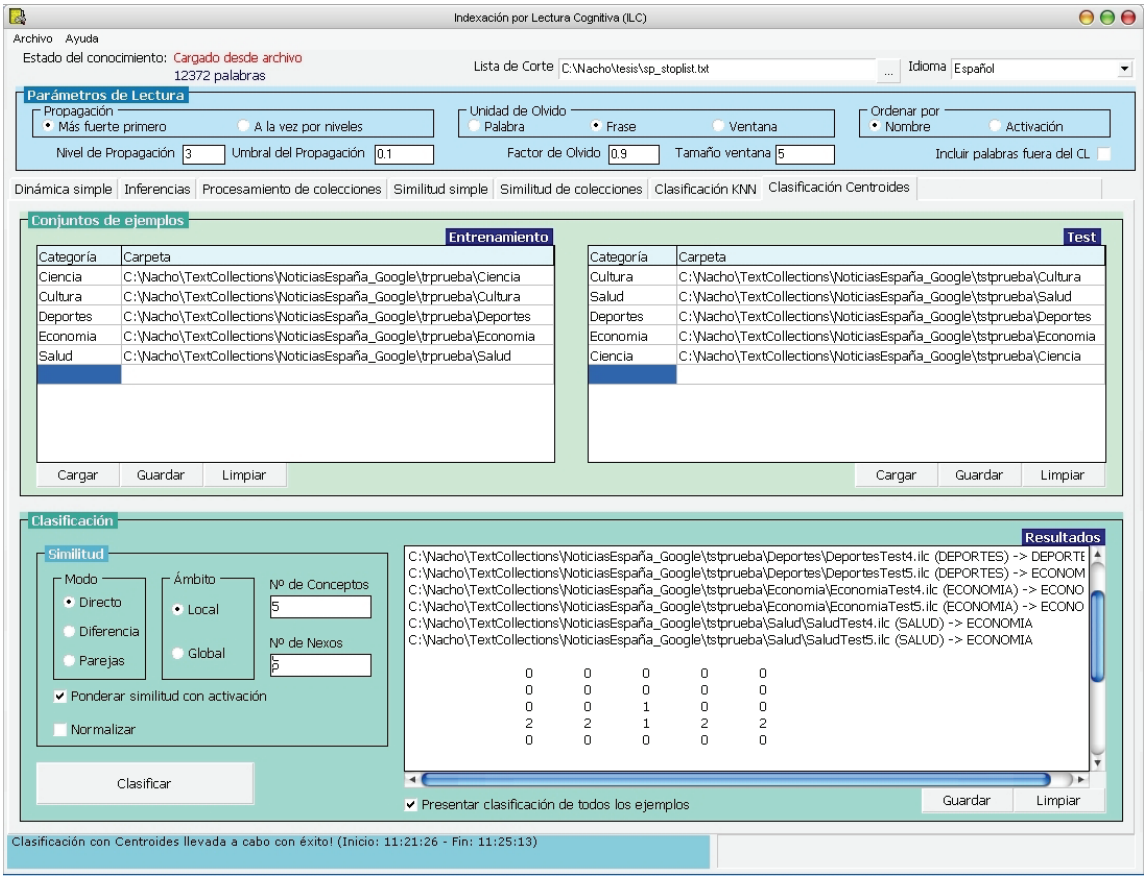
Similitud entre conceptos y representaciones semánticas



Predicción o inferencia de conceptos



Clasificación de textos basada en "centroides"



BIBLIOGRAFÍA

[Aaronson y Scarborough, 1977] Aaronson, D. y Scarborough, H. (1977). *Performance Theories for Sentence Coding: Some Quantitative Models*, Journal of Verbal Learning and Verbal Behavior, 16, pp. 277-304.

[Aha et al., 1991] Aha, D. W., Kibler, D. y Albert, M. K. (1991). *Instance-based Learning Algorithms*, Machine Learning, 6, pp. 37-66.

[Akmajian et al., 1995] Akmajian, A., Demers, R. A., Farmer, A. K. y Harnish, R. M. (1995). *Linguistics: An Introduction to Language and Communication (4th edition)*, Cambridge (Mass.): The MIT Press.

[Allen, 1995] Allen, J. F. (1995). *Natural Language Understanding*, Computer Science, 2nd ed., Benjaming Cummings.

[Altmann y Steedman, 1988] Altmann, G. y Steedman M. (1988). *Interaction with Context During Human Sentence Processing*, Cognition, 30, pp. 191-238.

[Anderson, 1983] Anderson, J. R. (1983). *The Architecture of Cognition*, Cambridge, Mass.: Harvard University Press.

[Angluin, 1982] Angluin, D. (1982) *Inference of Reversible Languages*, Journal of the ACM, 29(3), pp. 741-765.

[Atkinson y Shiffrin, 1968] Atkinson, R. C., y Shiffrin, R. M. (1968). *Human Memory: A Proposed System and its Control Processes*, K. Spence and J. Spence (Eds.), The Psychology of Learning and Motivation, New York: Academic Press, 2, pp. 89-195.

[Baddeley, 1986] Baddeley, A. D. (1986). *Working Memory*, New York: Oxford University Press.

[Bickerton, 1981] Bickerton, D. (1981). *Roots of Language*, Karoma Publishers.

[Bickerton, 1990] Bickerton, D. (1990). *Language and Species*, University of Chicago Press.

[Black, 2004] Black, P. E. (2004). *Euclidean Distance*, Dictionary of Algorithms and Data Structures [online], Paul E. Black (Ed.), U.S. National Institute of Standards and Technology.

[Blei et al., 2003] Blei, D., Ng, A. y Jordan, M. I. (2003). *Latent Dirichlet allocation*, Journal of Machine Learning Research, 3, pp. 993-1022.

[Bobrow y Norman, 1975] Bobrow, D. G. y Norman, D. A. (1975). *Some Principles of Memory Schemata*, Daniel G. Bobrow y Allan Collins (Eds.), Representation and Understanding. New York: Academic Press.

[Bowerman, 1973] Bowerman, M. (1973). *Structural Relations in Children's Utterances: Syntactic or Semantic?*, T. M. Moore (Ed.), Cognitive Development and the Acquisition of Language. New York: Academic Press, pp. 197-213.

[Braine, 1976] Braine, M. S. (1976). *Children's First Word Combinations*, J. D. Bransford, A. L. Brown y R. R. Cocking (Eds.), Monographs of the Society for Research in Child Development, 41 (1, serial nº 164).

[Briscoe, 1994] Briscoe, E. J. (1994). *Prospects for Practical Parsing of Unrestricted Text: Robust Statistical Parsing Techniques*, N. Oostdijk and P. De Haan (Eds.), Corpus-Based Research into Language.

[Brown y Fraser, 1963] Brown, R. y Fraser, C. (1963). *The Acquisition of Syntax*, C. N. Cofer y B. S. Musgrave (Eds.), Verbal Behavior and Learning: Problems and Processes, McGraw-Hill, New York, pp. 158-197.

[Brown, 1973] Brown, R. (1973). *A first language: The early stages*, Harvard University Press, Cambridge, MA.

[Budanitsky y Hirst, 2001] Budanitsky, A. y Hirst, G. (2001). *Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures*, Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh.

[Burgess, 1998] Burgess, C. (1998). *From Simple Associations to the Building Blocks of Language: Modeling Meaning in Memory with the HAL Model*, Behavior Research Methods, Instruments & Computers, 30, pp. 188-198.

[Cai et al., 2004] Cai, Z., McNamara, D.S., Louwerse, M., Hu, X., Rowe, M. y Graesser, A.C. (2004). *NLS: A Non-latent Similarity Algorithm*, Proc. of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004), pp. 180-185.

[Carbonell, 1970] Carbonell, J. R. (1970). *AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction*, IEEE Transactions on Man-Machine Systems, 11, pp. 190-202.

[Colheart et al., 1993] Colheart, M., Curtis, B., Atkins, P. y Haller, M. (1993). *Models of Reading Aloud: Dual-Route and Pararell-Distributed-Processing Approaches*, Psychological Review, 100, pp. 589-608.

[Cowan, 1988] Cowan, N. (1988). *Evolving Conceptions of Memory Storage, Selective Attention and their Mutual Constraints within the Human Information-Processing System*, Psychological Bulletin,104, pp. 163-191.

[Cox y Ram, 1999] Cox, M. T. y Ram, A. (1999). *On the Intersection of Story Understanding and Learning*, A. Ram & K. Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 397-434.

[Cherkassky et al., 1993] Cherkassky, B. V., Goldberg, A. V., y Radzik, T. (1993). *Shortest Paths Algorithms: Theory and Experimental Evaluation*, Technical Report 93-1480, Computer Science Department, Stanford University.

[Chomsky, 1986] Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*, New York: Praeger.

[Chomsky, 2006] Chomsky, N. (2006). *Language and Mind*, Cambridge University Press, 3ª edición.

[Damásio y Damásio, 1992] Damasio, A. R. y Damasio, H (1992). *Brain and Language*, Scientific American, September, pp. 73-71.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T, Landauer, T. K., Furnas, G. W. y Harshman, R. A. (1990). *Indexing by Latent Semantic Analysis*, Journal of the Society for Information Science, 41(6), pp. 391-407.

[Del Castillo y Serrano, 2004] Del Castillo, M. D. y Serrano, J. I. (2004). *A Multistrategy Approach for Digital Text Categorization from Imbalanced Documents*, ACM SIGKDD Explorations, 6, pp. 70-79.

[Del Castillo y Serrano, 2006] Del Castillo, M. D. y Serrano, J. I. (2006). *An Interactive Hybrid System for Identifying and Filtering Unsolicited E-mail*, E. Corchado et al. (Eds.), Intelligent Data Engineering and Automated Learning IDEAL 2006, Lecture Notes in Computer Science, 4224, Springer-Verlag Berlin Heidelberg, pp. 779-788.

[Démonet y Thierry, 2001] Demonet, J. F. y Thierry, G. (2001). *Language and Brain: What is up? What is coming up?*, Journal of Clinical and Experimental Neuropsychology, 23, pp. 96-120.

[Dempster, 1981] Dempster, F. N. (1981). *Memory Span: Sources of Individual and Developmental Differences*, Psychological Bulletin, 89, pp. 63-100.

[Dennis y Harrington, 2001] Dennis, S. y Harrington, M. (2001). *The Syntagmatic Paradigmatic Model: A Distributed Instance-based Model of Sentence Processing*, Proc. of the Second Workshop on Natural Language Processing and Neural Networks, Tokyo.

[Dewey, 1919] Dewey, J. (1916). *Essays in Experimental Logic*, Chicago: University of Chicago Press.

[Dewey, 1938] Dewey, J. (1938). *Logic: The Theory of Inquiry*, New York: Holt.

[Dial, 1969] Dial, R. (1969). *Algorithm 360: Shortest Path Forest with Topological Ordering*, Communications of the ACM, 12, pp. 632-633.

[Dijkstra, 1959] Dijkstra, E. W. (1959). *A Note on Two Problems in Conexion with Graphs*, Numerische Math, 1, pp. 269-271.

[Domeshek et al., 1999] Domeshek, E., Jones, E. y Ram, A. (1999). *Capturing the Contents of Complex Narratives*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 73-105.

[Doyle y McDermott, 1980] Doyle, J. y McDermott, D. (1980). *Nonmonotonic Logic I*, Artificial Intelligence, 13, pp 41-72.

[Dummet, 1993] Dummet, M. (1993). *Frege: Philosophy of Language*, Harvard University Press, 2ª edición.

[Dumais, 1994] Dumais, S. T. (1994). Latent *Semantic Indexing (LSI) and TREC-2*, D. Harman (Ed.), The Second Text Retrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215, pp. 105-116.

[Dupont, 2002] Dupont, P. (2002). *Inductive and Statistical Learning of Formal Grammars*, Reseach talk, Departement D'ingenerie Informatique, Universite Catholique de Louvain.

[El-Iman, 2005] El-Iman, Y. A. (2005). *Rules and Algorithms for Phonetic Transcription of Standard Malay*, IEICE Transactions on Information and Systems, E88-D(10), pp. 2354-2372.

[Ellis et al., 2001] Ellis, D., Singh, R. y Sivadas, S. (2001). *Tandem Acoustic Modeling In Large-Vocabulary Recognition*, Proc. ICASSP-2001, I, pp. 517-520.

[Engle et al., 1992] Engle, R. W., Cantor, J., y Carullo, J. J. (1992). *Individual Differences in Working Memory and Comprehension: A Test of Four Hypotheses*, Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, pp. 972-992.

[Ericsson y Kintsch, 1995] Ericsson, K. A. y Kintsch, W. (1995). *Long-Term Working Memory*, Psychological Review, 102, pp. 211-245.

[Farnham-Diggory y Gregg, 1975] *Farnham-Diggory, S., y Gregg, L. (1975). Short-Term Memory Function in Young Readers*, Journal of Experimental Child Psychology, 19, pp. 279-298.

[Feldman y Sanger, 2006] Feldman, R. y Sanger, J. (2006). *The Text Mining Handbook: Advances Approaches in Analysing Unstructured Data*, Cambridge University Press.

[Fellbaum, 1998] Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*, the MIT Press.

[Ferrer et al., 2004] Ferrer, R., Solé, R. V. y Köhler R. (2004). *Patterns in Syntactic Dependency Networks*, Physical Review, 69, pp. 051915-(1-8).

[Fischler y Goodman, 1978] Fischler, I. y Goodman, G. O. (1978). *Latency of Associative Activation in Memory*, Journal of Experimental Psychology: Human Perception and Performance, 4, pp. 455-470.

[Fletcher, 1999] Fletcher, C. R. (1999). *Computational Models of Reading and Understanding: What Good Are They?*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 483-490.

[Fodor, 1983] Fodor, J. A. (1983). *Modularity of Mind: An Essay in Faculty Psychology*, The MIT Press, Cambridge, MA.

[Frakes, 1992] Frakes, W. B. (1992). *Stemming algorithms, Information retrieval: data structures and algorithms*, Prentice-Hall, Inc., Upper Saddle River, NJ.

[Frazier y Rayner, 1982] Frazier, L. y Rayner K. (1982). *Making and Correcting Errors During Sentence Comprehension: Eye Movements in the Analysis of Structurally Ambiguous Sentences*, Cognitive Psychology, 14, pp. 178-210.

[Fu y Booth, 1975] Fu, K., y Booth, T. (1975). *Grammatical Inference: Introduction and Survey. Parts I and II*, IEEE Transactions on Systems, Man and Cybernetics, 5:303-309, pp. 409-423.

[García y Vidal, 1987] García, P. y Vidal, E. (1987). *Local Languages, the Sucesor Meted, and a Step towards a General Methodology for the Inference of Regular Grammars*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, 9(6), pp. 841-845.

[García y Vidal, 1990] García, P. y Vidal, E. (1990). *Inference of K-testable Languages In the Strict Sense ans Application to Syntactic Pattern Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, 12(9), pp. 920-925.

[Glanzer et al., 1984] Glanzer, M., Fischer, B., y Dorfman, D. (1984). *Short-Term Storage in Reading*, Journal of Verbal Learning and Verbal Behavior, 23, pp. 467-486.

[Gleitman y Newport, 1995] Gleitman, L. y Newport, E. (1995). *The Invention of Language by Children: Environmental and Biological Influences on the Acquisition of Language*, L. R. Gleitman y M. Liberman (Eds.), An Invitation to Cognitive Science, 1, Language ed. Cambridge, MA: MIT Press, pp. 1-24.

[Glucksberg et al., 1986] Glucksberg, S., Kreuz, R. J. y Rho, S. (1986). *Context Can Constrain Lexical Access: Implications for Models of Language Comprehension*, Journal of Experimental Psychology: Learning, Memory and Cognition, 12, pp. 323-335.

[Goldman et al., 1980] Goldman, S. R., Hogaboam, T. W., Bell, L. C. y Perfetti, C. A. (1980). *Short-term Retention of Discourse During Reading*, Journal of Educational Psychology, 72, pp. 647-655.

[Golub y Van Loan, 1996] Golub, G. H. y Van Loan, C. F. (1996). *Matrix Computation*, Johns Hopkins (Ed.), University Press, Baltimore.

[Goodman, 1970] Goodman, K. S. (1970). *Reading a Psycholinguistic Guessing Game*, Harry Singer y Robert B. Ruddell (Eds.), Theoretical Models and Processes of Reading, Newark, Delaware: International Reading Association.

[Graeser et al., 1994] Graesser, A. C., Singer, M. y Trabaos, T. (1994). *Construction of Inferences During Narrative Comprehension*, Psychological Review, 101, pp. 371-395.

[Guyon y Elisseff, 2003] Guyon, I. y Elisseff, A. (2003). *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, 3, pp. 1157-1182.

[Hasher y Zacks, 1988] Hasher, L. y Zacks, R. T. (1988). *Working Memory, Comprehension, and Aging: A Review and a New View*, The Psychology of Learning and Motivation, 22, pp. 193-225.

[Hirst y St-Onge, 1998] Hirst, G. y St-Onge, D. (1998). *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*, C. Fellbaum (Ed.), Wordnet: An Electronic Lexical Database, MIT Press, pp. 305-332.

[Hofmann, 1999] Hofmann, T. (1999). *Probabilistic Latent Semantic Analysis*, Proc. of Uncertainty in Artificial Intelligence.

[Hofmann, 2001] Hofmann, T. (2001). *Unsupervised Learning by Probabilistic Latent Semantic Analysis*, Machine Learning Journal, 42(1), pp. 177-196.

[Hull, 1996] Hull, D.A. (1996). *Stemming Algorithms – A Case Study for Detailed Evaluation*, Journal of the American Society for Information Science, 47(1), pp. 70-84.

[Hung y Dovoky, 1988] Hung, M. H. y Divoky, J. J. (1988). *A Computational Study of Efficient Shortest Path Algorithms*, Computers & Operations Research, 15, pp. 567-576.

[Jensen, 2001] Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*, Springer.

[Jiang y Conrath, 1997] Jiang, J. y Conrath, D. (1997). *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*, Proc. of International Conference on Research in Computational Linguistics, pp. 19-33.

[Joachims, 1997] Joachims, T. (1997). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*, Proc. 14th International Conference on Machine Learning (ICML-97), pp. 143-151.

[Joachims, 2002] Joachims, T. (2002). *Learning to Classify Text using Support Vector Machines*, Dissertation, Kluwer.

[Johnson-Laird, 1983] Johnson-Laird, P.N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*, Cambridge, MA: Cambridge University Press.

[Kintsch y Mross, 1985] Kintsch, W. y Mross, F. (1985). *Context Effects in Word Identification*, Journal of Memory and Language, 30, pp. 580-602.

[Kintsch, 1988] Kintsch, W. (1988). *The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model*, Psychological Review, 95(2), pp. 163-182.

[Kohonen, 1995] Kohonen, T. (1995). *Self-Organizing Maps*, Springer-Verlag, Heidelberg.

[Kolda y O'Leary, 1998] Kolda, T. G. y O'Leary, D. P. (1998). *A Semidiscrete Matrix Decomposition for Latent Semantic Indexing Information Retrieval*, ACM Transactions on Information Systems, 16(4), pp. 322-346.

[Kulkarni y Pedersen, 2005] Kulkarni, A. y Pedersen, T. (2005). *SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts*, Proc. of 43rd Annual Meeting of the Association for Computational Linguistics, University of Michigan, USA.

[Landauer y Dumais, 1997] Landauer, T. y Dumais, S. (1997). *A solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge*, Psychological Review, 104, pp. 211-240.

[Landauer et al., 1998] Landauer, T., Foltz, P. y Laham, D. (1998). *Introduction to Latent Semantic Analysis*, Discourse Processes, 25, pp. 259-284.

[Lang y Wharton, 1993] Lange, T. E. y Wharton, C. M. (1993). *Dynamic Memories: Analysis of an Integrated Comprehension and Episodic Memory Retrieval Model*, Proc. of International Joint Conferences on Artificial Intelligence (IJCAI'93), pp. 208-216.

[Lang, 1995] Lang, K. (1995). *Newsweeder: Learning to Filter Netnews*, Proc. of International Conference on Machine Learning, pp. 331-339.

[Langston et al., 1999] Langston, M. C., Trabasso, T. y Magliano, J. P. (1999). *A Connectionist Model of Narrative Comprehension*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 181-225.

[Lari y Young, 1991] Lari, K. y Young, S. (1991). *Application os Stochastic Context Free Grammars using the Inside-Outside Algorithm*, Computer Speech and Language, 5, pp. 237-257.

[Leacock y Chorodow, 1998] Leacock, C. y Chodorow, M. (1998). *Combining Local Context and Wordnet Similarity for Word Sense Identification*, C. Fellbaum (Ed.), Wordnet: An Electronic Lexical Database, MIT Press, pp. 265-283.

[Lemaire y Denhière, 2004] Lemaire, B. y Denhière, G. (2004). *Incremental Construction of an Associative Network from a Corpus*, Proc. of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004), pp. 825-830.

[Lin, 1998] Lin, D. (1998). *An Information-Theoretic Definition of Similarity*, Proc. of International Conference on Machine Learning, Madison, Wisconsin, July.

[Liu y Motoda, 1998] Liu, H. y Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*, The Springer International Series in Engineering and Computer Science, 454.

[Livescu et al., 2003] Livescu, K., Glass, J. y Bilmes, J. (2003). *Hidden Feature Models for Speech Recognition Using Dynamic Bayesian Networks*, Proc. of Eurospeech '03, Geneva, pp. 2529-2532.

[Lukatela y Turvey, 1998] Lukatela, G. y Turvey, M. T. (1998). *Learning to Read in Two Alphabets*. American Psychologist, 53, pp.1057-1072.

[Lund y Burgess, 1995] Lund, K. y Burgess, C. (1995). *Semantic and Associative Priming In High-Dimensional Semantic Space*, J. D. Moore y J. F. Lehman (Eds.), Proc. of Seventeenth Annual Conference of the Cognitive Science Society, pp. 660-665.

[Lund y Burgess, 1996] Lund, K. y Burgess, C. (1995). *Dissociating Semantic ans Associative Word Relationships Using High-Dimensional Semantic Space*, G. W. Cottrell (Ed.), Proc. of Eighteenth Annual Conference of the Cognitive Science Society, pp. 603-608.

[MacDonald et al., 1994] MacDonald, M. C., Pearlmutter, N. J. y Seidenberg, M. S. (1994). *The Lexical Nature of Syntactic Ambiguity Resolution*, Psychological Review, 101, pp. 676-703.

[Mahesh et al., 1999] Mahesh, K., Eiselt, K. P. y Holbrook, J. K. (1999). *Sentence Processing in Understanding: Interaction and Integration of Knowledge Sources*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 27-71.

[Manning y Schutze, 2002] Manning, C. D. y Schutze, H. (2002). *Foundations of Statistical Natural Language Processing*, London: The MIT Press.

[McKoon y Ratcliff, 1992] McKoon, G. y Ratcliff, R. (1992). *Inference During Reading*, Psychological Review, 99, pp. 440-466.

[McLachlan y Krishnan, 1997] McLachlan, G. y Krishnan, T. (1997). *The EM Algorithm and Extensions*, John Wiley & Sons (Eds.), Wiley series in Probability and Statistics.

[McRae et al., 1997] McRae, K., de Sa, V. R., y Seidenberg, M. S. (1997). *On the Nature and Scope of Featural Representations of Word Meaning*, Journal of Experimental Psychology: General, 126(2), pp. 99-130.

[Meyer y Poon, 2001] Meyer, B. J. F. y Poon, L. W. (2001). *Effects of Structure Strategy Training and Signaling on Recall of Text*, Journal of Educational Psychology, 93, pp. 141-159.

[Minsky, 1975] Minsky, M. (1975). *A Framework for Representing Knowledge*, W. P. Henry (Ed.), The Psychology of Computer Vision. New York et al.: McGraw-Hill, pp. 211-277.

[Miller, 1956] Miller, G. A. (1956). *The Magical Number Seven, Plus or Minus two: Some Limits of our Capacity for Processing Information*, Psychological Review, 63, pp. 81-97.

[Miller et al., 1990] Miller, G.A., Beckwith, R.T., Fellbaum, C.D., Gross, D. y Miller, K. (1990). *WordNet: An On-line Lexical Database*, International Journal of Lexicography, 3(4), pp. 235-244.

[Miller y Charles, 1991] Miller, G. A. y Charles, W. G. (1991). *Contextual correlates of Semantic Similarity*, Language and Cognitive Processes, *6(1)*, pp. 1-28.

[Mira y Delgado, 1995] Mira, J. y Delgado, A. E. (1995). *Computación Neuronal Avanzada: Fundamentos Biológicos y Aspectos Metodológicos*, Barro, S. y Mira, J. (Eds.), Redes Neuronales Naturales y Artificiales. Universidad de Santiago de Compostela.

[Mitchell, 1997] Mitchell, T. (1997). *Machine Learning*, McGraw-Hill.

[Moorman y Ram, 1999] Moorman, K. y Ram, A. (1999). *Creativity in Reading: Understanding Novel Concepts*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 359-433.

[Morris y Harris, 2004] Morris, A. L., y Harris, C. L. (2004). *Repetition Blindness: Out of Sight or Out of Mind?*, Journal of Experimental Psychology: Human Perception and Performance, 30, pp. 913-922.

[Mosterín, 2006] Mosterín, J. (2006).

[Narayanan y Jurafsky, 2002] Narayanan, S. y Jurafsky, D. (2002). *A Bayesian Model Predicts Human Parse Preference and Reading Time in Sentence Processing*, T. G. Dietterich, S. Becker y Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems, 14. MIT Press.

[Neil et al., 1988] Neil, W. T., Hilliard, D. V. t Cooper, E. (1988). *The Detection of Lexical Ambiguity: Evidence for Context-Sensitive Parallel Access*, Journal of Memory and Language, 27, pp. 279-287.

[Newell y Simon, 1972] Newell, A., y Simon, H. A. (1972). *Human Problem Solving*, Englewood Cliffs, N. J.: Prentice-Hall.

[Nigam et al., 2000] Nigam, K., McCallum, A., Thrun, S. y Mitchell, T. (2000). *Text Classification from Labeled and Unlabeled Documents Using EM*, Machine Learning, 39(2/3), pp. 103-134.

[Nosofsky, 1991] Nosofsky, R. M. (1991). *Stimulus Bias Asymmetric Similarity and Classification*, Cognitive Psychology, 23, pp. 94-140.

[O'Leary y Peleg, 1983] O'Leary, D, y Peleg, S. (1983). *Digital Image Compression by Outer Product Expansion*, IEEE Transaction on Communications, 31, pp. 441-444.

[Oncina, 1991] Oncina, J. (1991). *Aprendizaje de Lenguajes Regulares y Funciones Subsecuenciales*, Phd. Thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

[Ortony et al., 1985] Ortony, A, Vondruska, R. J., Foss, M. A., y Lawrence, E. J. (1985). *Salience, Similes, and the Asymmetry of Similarity*, Journal of Memory and Language, 24, pp. 569-594.

[Perfetti y Tan, 1998] Perfetti, C. A. y Tan L. H. (1998). *The Time-Course of Graphic, Phonological and Semantic Activation in Chinese Character Identification*. Journal of Experimental Psychology: Learning, Memory and Cognition, 24, pp. 1-18.

[Perfetti, 1998] Perfetti, C. A. (1998). *The Limits of Co-Occurrence: Tools and Theories in Language Research*, Discourse Processes, 25(2-3), pp. 363-377.

[Perfetti, 1999] Perfetti, C. A. (1999). *Comprehending Written Language: A Blue Print of the Reader*, Brown & Hagoort (Eds.), The Neurocognition of Language Processing, Oxford University Press, pp. 167-208.

[Peterson y Billman, 1999] Peterson, J y Billman, D. (1999). *Semantic Correspondence Theory*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 299-358.

[Piaget, 1970] Piaget, J. (1970). *Piaget's Theory*, P. Mussen (Ed.), Handbook of Child Psychology, 1, New York: Wiley.

[Pinker, 1991] Pinker, S. (1991). *Rules of Language*, Science, 253, pp. 530-535.

[Pinker, 1994] Pinker, S. (1994). *El Instinto del Lenguaje. Cómo Crea el Lenguaje la Mente*, Alianza Editorial.

[Pinker, 2005] Pinker, S. (2005). *So How Does the Mind Work?*, Mind and Language, 20(1), pp. 1-24.

[Pla, 2000] Pla., F. (2000). *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos*,. Phd. Thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

[Plaut et al., 1996] Plaut, D. C., McClelland, J. L., Seidenberg, M. S. y Patterson, K. (1996). *Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains*, Psychological Review, 103, pp. 56-115.

[Porter, 1980] Porter, M. F. (1980). *An Algorithm for Suffix Stripping*, Program, 14(3), pp. 130-137.

[Potter, 1983] Potter, M. C. (1983). *Representational Buffers: The Eye-Mind Hypothesis in Picture Perception, Reading, and Visual Search*, K. Rayner (Ed.), Eye Movements in Reading: Perceptual and Language Processes, New York: Academic Press, pp. 413-437.

[Quillian, 1968] Quillian, M. R. (1968). *Semantic Memory*, Semantic Information Processing, The MIT Press, Cambidge, MA.

[Ram y Moorman, 1999] Ram, A. y Moorman, K. (1999), *Introduction: Toward a Theory of Reading and Understanding*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 1-10.

[Ram, 1999] Ram, A. (1999). *A Theory of Questions and Question Answering*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 253-298.

[Rapaport y Shapiro, 1999] Rapaport, W. J. y Shapiro, S. C. (1999). *Cognition and Fiction: An Introduction*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 11-25.

[Rapp y Gerrig, en prensa] Rapp, D. N. y Gerrig, R. J. (en prensa) Predilections for Narrative Outcomes: The Impact of Story Contexts and Reader Preferences, Journal of Memory and Language.

[Reiter, 1980] Reiter, R. (1980). *A Logic for Default Reasoning*, Artificial Intelligence, 13, pp. 81-132.

[Resnik, 1995] Resnik, P. (1995). *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, Proc. of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453.

[Riloff, 1999] Riloff, E. (1999). *Information Extraction as a Stepping Stone Toward Story Understanding*, Ashwin Ram & Kenneth Moorman (Eds.), Understanding Language Understanding: Computational Models of Reading, Cambridge, MA: MIT Press, pp. 435-460.

[Romaine, 1992] Romaine, S. (1992). *The Evolution of Linguistic Complexity in Pidgin and Creole Languages*, John A. Hawkins y Murray Gell-Mann (Eds.), The Evolution of Human Languages. SFI Studies in the Sciences of Complexity, Addison-Wesley.

[Rubia, 2007] Rubia, F. J. (2007). *El Cerebro nos Engaña*, Temas de Hoy Ed.

[Rulot y Vidal, 1987] Rulot, H. y Vidal, E. (1987). *Modelling (sub)String-length-based Constraints through a Grammatical Inference Method*, Pattern Recognition: Theory and Applications, 1987.

[Rulot, 1992] Rulot, H. (1992). *ECGI. Un algoritmo de Inferencia Gramatical mediante Corrección de Errores*, Phd Thesis, Facultad de Ciencias Físicas, Universidad de Valencia.

[Rumelhart, 1977] Rumelhart, D. E. (1977). *Toward an Interactive Model of Reading*, S. Dornic (Ed.), Attention and Performance VI. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[Sahlgren, 2005] Sahlgren, M. (2005). *An Introduction to Random Indexing*, Proc. of Methods and Applications of Semantic Indexing Workshop, 7th International Conference on Terminology and Knowledge Engineering (TKE'05), Copenhagen, Denmark.

[Sapir, 1921] Sapir, E. (1921). *Language: An Introduction to the Study of Speech*, New York: Harcourt Brace.

[Schank, 1982] Schank, R.C. (1982). *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*, Cambridge University Press.

[Sebastiani, 2002] Sebastiani, F. (2002). *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, 34(1), pp. 1-47.

[Selfridge, 1986] Selfridge, M. (1986). *A computer Model of Child Language Learning*, Artificial Intelligence, 29 (2), pp.171-216.

[Serrano y Del Castillo, 2007] Serrano, J. I. y Del Castillo, M. D. (2007). *Evolutionary Learning of Document Categories*, Information Retrieval, 10, pp. 69-83.

[Shapiro, 1979] Shapiro, S. C. (1979). *The SNePS Semantic Network Processing System*, Associative Networks: Representation and Use of Knowledge by Computers, Findler N. V. (Ed.), Academic Press, New York

[Shastri, 1992] Shastri, L. (1992). *Structured Connectionist Models*, L. Fritz (Ed.), Semantic Networks in Artificial Intelligence. Pergamon Press, Oxford, pp. 293-328.

[Simon, 1973] Simon, H. A. (1973). *The Structure of Ill Structured Problems*, Artificial Intelligence, 4, pp. 181-201.

[Sokolova et al., 2006] Sokolova, M., Japkowicz, N. y Szpakowicz, S. (2006). *Beyond Accuracy, F-score and ROC: A Family of Discriminant Measures for Performance Evaluation*, Proc. of the 19th ACS Australian Joint Conference on Artificial Intelligence (AI'2006), pp. 1015-1021.

[Sowa, 1991] Sowa, J. F. (Ed.) (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, Morgan Kaufmann, San Mateo, CA.

[Smolensky, 1988] Smolensky, P. (1988). *On the Proper Treatment of Connectionism*, Behavioral and Brain Sciences, 11, pp. 1-74.

[Steyvers et al., 2004] Steyvers, M., Shiffrin, R. M. y Nelson, D. L. (2004). *Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory*, A. Healy (Ed.), Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer, Washington DC: American Psychological Association.

[Stillings et al., 1998] Stillings, N. A., Weisler, S. E., Chase, C. H., Feinstein, M. H., Garfield, J. L. y Rissland, E. L. (1998). *Cognitive Science: An Introduction (2nd edition)*, a Bradford Book, The MIT Press.

[Stone et al., 1997] Stone, G. O., Vanhoy, M. y Van Orden, G. C. (1997). *Perception is a Two-Way Street: Feedforward and Feedback Phonology in Visual Word Recognition*, Journal of Memory and Language, 36, pp. 337-359.

[Till et al., 1988] Till, R., Mross, E. F. y Kintsch, W. (1988). *Time Course of Priming for Associative and Inference Words in a Discourse Context*, Memory and Cognition, 16(4), pp. 283-298.

[Turing, 1950] Turing, A. M. (1950). *Computing Machinery and Intelligence*, Mind, 59, pp. 433-460.

[Turney, 2001] Turney, P. (2001). *Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL*, De Raedt, Luc and Flach, Peter (Eds.), Proc. of the Twelfth European Conference on Machine Learning (ECML-2001), pp. 491-502.

[Tversky, 1977] Tversky, A. (1977). *Features of Similarity*, Psychological Review, 84(4), pp. 327-352.

[Van Orden y Goldinger, 1994] Van Orden, G. C. y Goldinger, S. D. (1994). *The Interdependence of Form and Function in Cognitive Systems Explains Perception of Printed Words*, Journal of Experimental Psychology: Human Perception and Performance, 20, pp. 1269-1291.

[Ventura et al., 2004] Ventura, M., Hu, X., Graesser, A., Louwerse, M. y Olney, A. (2004). *The Context Dependent Sentence Abstraction Model*, Proc. of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004), pp. 1387-1392.

[Vigliocco et al., 2004] Vigliocco, G., Vinson, D. P, Lewis, W. y Garrett, M. F. (2004). *Representing the Meanings of Object and Action Words: The Featural and Unitary Semantic System (FUSS) Hypothesis*, Cognitive Psychology, 48, pp. 422-488.

[Walker, 1987] Walker, C. H. (1987). *Relative Importance of Domain Knowledge and Overall Aptitude on Acquisition of Domain-Related Information*, Cognition and Instruction, 4, pp. 25-42.

[Watts y Strogatz, 1998] Watts, D. J. y Strogatz, S. H. (1998). *Collective Dynamics of 'Small-World' Networks*, Nature, 393, pp. 440-442.

[Weiss et al., 1978] Weiss, S. M., Kulikowski, C. A., Amarel, S., y Safir, A. (1978). *A Model-based Method for Computer-Aided Medical Decision-Making*, Artificial Intelligence, 11, pp. 145-172.

[Wood, 1993] Word, D. (1993). *Data Structures, Algorithms and Performance*, Addison-Wesley, pp. 367-375.

[Yang y Pedersen, 1997] Yang, Y. y Pedersen, J. O. (1997). *A Comparative Study on Feature Selection in Text Categorization*, Proc. of 14th International Conference on Machine Learning.

[Zakaluk, 1998] Zakaluk, B. L. (1998). *Theoretical Overview of the Reading Process: Factors Which Influence Performance and Implications for Instruction*, National Adult Literacy Database.

[Zhao y Karypis, 2003] Zhao, Y. y Karypis, G. (2003). *Hierarchical Clustering Algorithms for Document Datasets*, Technical Report 03-027, Dept. of Computer Science, University of Minnesota.

"Es tan ligera la lengua como el pensamiento, que si son malas las preñeces de los pensamientos, las empeoran los partos de la lengua"

Don Quijote de la Mancha, "El Ingenioso Hidalgo Don Quijote de la Mancha", Miguel de Cervantes Saavedra (1547-1616)

"El auténtico problema no es si las máquinas piensan, sino si lo hacen los hombres"

Frederic Burrhus Skinner (1904-1990)