Un Método Automático para la Desambiguación Léxica de Nombres

Paolo Rosso * Francesco Masulli ** Davide Buscaldi ***

Resumen

Este artículo presenta un método completamente automático que resuelve la desambiguación léxica de nombres calculando la densidad conceptual de cada uno de los sentidos del nombre a desambiguar. La evaluación del método se ha realizado sobre el corpus SemCor con un contexto de sólo dos nombres, obteniendo una precisión de 81.5 % y un recall de 60.2 %.

Palabras clave: desambiguación léxica de nombres, densidad conceptual.

Abstract

This paper presents a completely automatic method which solves the noun sense disambiguation task calculating the conceptual density for each sense of the noun to be disambiguated. The evaluation of the method was carried out on the SemCor corpus; a precision of 81.5 % and a recall of 60.2 % were obtained with a context of only two nouns.

Keywords: noun sense disambiguation, conceptual density.

1. Introducción

Una de las tareas primordiales en cualquier aplicación de *Procesamiento del Lenguaje Natural (PLN)* es la desambiguación del sentido de las palabras (*Word Sense Disambiguation, WSD)* [4]. Cualquier sistema que pretenda trabajar sobre la estructura semántica de un documento, necesita utilizar conocimiento acerca de las estructuras del lenguaje, siendo este conocimiento de tipo morfológico, sintáctico, semántico y pragmático [9]. El conocimiento morfológico nos proporciona información cómo se construyen las palabras, el sintáctico de cómo combinar las palabras para formar frases, el semántico qué significan las palabras y cómo contribuye el significado de las mismas al significado completo de la frase, y por último, el pragmático de cómo el contexto afecta a la interpretación de las frases. Todas estas formas de conocimiento lingüístico, tienen el problema asociado de la ambigüedad. La ambigüedad léxica aparece cuando las palabras presentan diferentes significados [3]. La tarea de resolver la ambigüedad léxica, se le conoce como desambiguación del sentido de las palabras y cualquier sistema de PLN necesita un módulo de estas características.

*Dpto. de Sistemas Informáticos y Computación, Univ. Politécnica de Valencia, España, prosso@dsic.upv.es

**INFM-Genova and Dip. di Informatica, Università di Pisa, Italia, masulli@disi.unige.it

***IDip. di Informatica e Scienze dell'Informazione, Università di Genova, Italia, buscaldi@disi.unige.it

WSD es una tarea intermedia que sirve de ayuda cuando necesitamos conocer el sentido de las palabras en algunas aplicaciones del PLN (recuperación de la información, clasificación de textos, extracción de información, análisis del discurso, etc.). La tarea de desambiguación del sentido de las palabras consiste en la asociación de una palabra dada en un texto, con una definición o significado que la distingue de otros significados atribuíbles a esa palabra. La asociación de una palabra (un nombre, en nuestro caso) a un sentido, se cumple dependiendo de dos tipos de recursos de información: el contexto y los recursos léxicos de conocimiento externo [10]. El contexto del nombre a ser desambiguado, se define como el conjunto de nombres del mismo párrafo, que acompañan al nombre a desambiguar. Como recurso léxico de conocimiento externo, hemos utilizado *WordNet (WN)*, que ha sido desarrollado por la Universidad de Princeton [8]. El mismo combina las características de los diccionarios y de los tesauros con las relaciones semánticas (sinonimia, hiperonimia, hiponimia, etc.) entre palabras. La ontología de WordNet incluye, como diccionario, definiciones para sentidos individuales de palabras y, como tesauro, define grupos de sinónimos mediante synsets (*sets* of *syn*onyms), representando simples conceptos léxicos y organizándolo en una jerarquía conceptual.

Los métodos que hacen uso de la información estadística [5] [2], obtenida del corpus utilizado durante la fase de entrenamiento (*WSD corpus-driven*), pueden ser *supervisados* (si cada palabra del corpus tiene una etiqueta con información sintáctica y semántica, es decir, con su número de synset) o *no supervisados* (si vienen entrenados, como mucho, con palabras etiquetadas sólo como *Part-Of-Speech*, es decir, con su categoría sintáctica). Los métodos automáticos que no necesitan ningún proceso de aprendizaje, se basan sólo en el conocimiento léxico que proporciona un recurso externo como WordNet (*WSD knowledge-driven*). Es conveniente destacar, que WordNet no es en absoluto un recurso perfecto para desambiguar el sentido de las palabras, debido al problema de la granularidad fina para la distinción de los significados [4], y las divisiones de un sentido son demasiados finas para el propósito de muchos trabajos de PLN. Eso crea muchas dificultades a la hora de desambiguar el sentido de las palabras automáticamente, debido a que hay que hacer elecciones en cuanto al significado, que a veces es difícil de realizar inclusive manualmente.

En las siguientes secciones, se explica de forma detallada e intuitiva el método que resuelve la ambigüedad léxica de nombres basándose sólo en el conocimiento del recurso externo WordNet. El método completamente automático, calcula la densidad conceptual basándose también en la frequencia de cada uno de los sentidos del nombre a ser desambiguado (información disponible en WordNet). En la última sección, se presentan las conclusiones y los trabajos futuros para mejorar el método explicado.

2. El Método Automático: Densidad Conceptual y Frecuencia

Nuestro trabajo está basado en la idea de densidad conceptual. La *Densidad Conceptual (DC)* es una medida de correlación entre el sentido de una palabra y su contexto, y se basa a su vez, en la *Distancia Conceptual*, es decir, la longitud del camino más corto que conecta dos synsets en la taxonomía de nombres que utiliza WordNet. El método, que utiliza las relaciones jerárquicas (hiperonimia / hiponimia), es completamente automático y no necesita ningún proceso de entrenamiento.

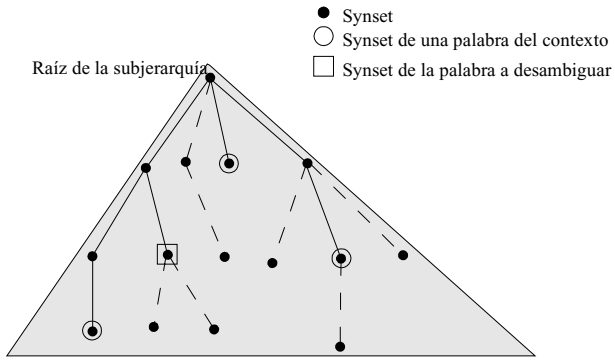

Figura 1: Subjerarquía y synsets relevantes

2.1. Densidad Conceptual

La ontología de WordNet está compuesta por tres taxonomías, asociadas con las categorías sintácticas de los nombres, de los verbos y, la última, de los adjetivos y de los adverbios. Cada jerarquía conceptual puede verse como un árbol n-ario donde cada nodo es un synset y está conectado con otro synset, a través de un arco que representa la relación de hiperonimia /hiponimia. Los n sentidos de la palabra a desambiguar caen en diferentes zonas de la jerarquía que puede verse como particionada en diferentes subjerarquías, cada una conteniendo uno de los n synsets. En este artículo, nos vamos a referir a estas particiones como *aglomeraciones* o *clusters*, si bien no se podrían considerar como tales en el sentido estrícto. Hay casos en los cuales no puede hacerse el particionamiento y, en consecuencia, tampoco puede desambiguarse la palabra. Esto ocurre cuando dos, o más, sentidos de una palabra son hipónimos el uno del otro.

La distancia conceptual fue introducida por Agirre-Rigau en [1] y se calcula con la fórmula:

$$DC(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \quad (1)$$

donde c es el synset raíz de la subjerarquía, m el número de sentidos de las palabras a desambiguar, h la altura de la subjerarquía y $nhyp$ el número medio de hipónimos por cada nodo (synset) de la subjerarquía.

Un ejemplo de aglomeración está ilustrado en la Figura 1, donde las líneas continuas representan los caminos del nombre a desambigar y de los nombres de su contexto. Los nodos terminales de estos caminos son los *synsets relevantes*, es decir, aquéllos para los cuales hay que calcular la densidad de la aglomeración.

2.2. Combinación de Densidad Conceptual y Frecuencia

El número medio de hipónimos por cada synset de la subjerarquía, puede verse como una medida de dispersión de la aglomeración. En [1] los experimentos se llevaron a cabo utilizando la versión 1.4 de la ontología de WordNet. La versión 1.6 de WordNet tiene una granularidad mucho más fina para la distinción de los sentidos de una palabra y, por tanto, el número medio de hipónimos por cada synset de la subjerarquía resulta más grande que en la version 1.4. En nuestro estudio de investigación, decidimos introducir una fórmula diferente para

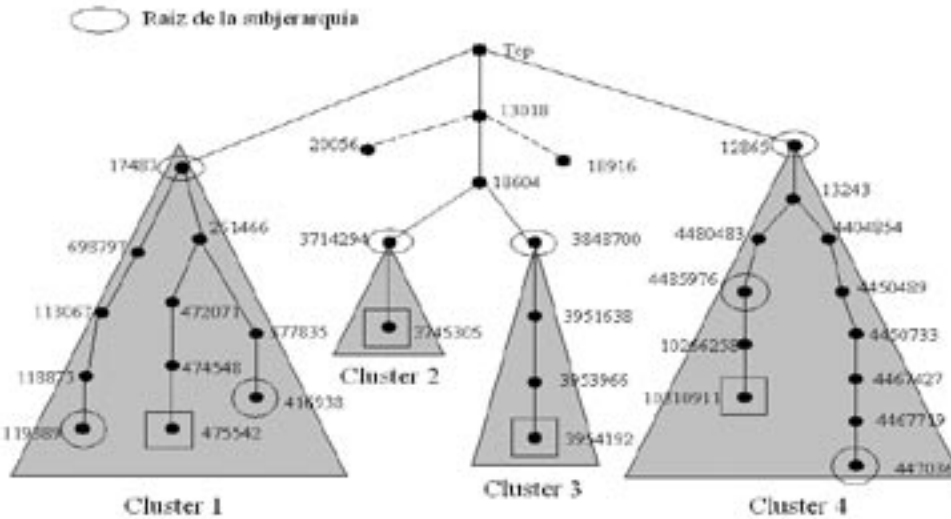

Figura 2: Subjerarquías para la desambiguación del nombre "irregularity"

el cálculo de la DC que no estuviera basada en el número medio de hipónimos por nodo, sino en el número M de synsets relevantes dividido el número total nh de synsets de la aglomeración. Para el cálculo de la DC, consideramos como no relevantes aquellos caminos que no llevan a ningún synset relevante de la subjerarquía (líneas puntedas de la Figura 1).

$$DC(M, nh) = \frac{M}{nh} \tag{2}$$

Desde un punto de vista práctico, esto significa considerar más bien las distancias conceptuales entre synsets, como se puede apreciar en el ejemplo de la Figura 2.

La figura muestra las subjerarquías obtenidas durante la desambiguación del nombre *irregularity* de la siguiente frase del fichero br-a01 (utilizado en [1]) del *Sem*antic con*Cor*dance (SemCor) [6] :

Fulton-County-Grand-Jury said Friday an investigation of Atlanta's recent primary-election produced no evidence that any irregularities took-place.

Los nombres del contexto de "irregularity" son: investigation, Atlanta, primary-election, evidence. Las raíces de las subjerarquías corresponden a los siguientes synsets (el offset es el número que identifica de manera unívoca un synset): act (offset 17487), quality (offset 3714294), property (offset 3848700) y psychological-feature (offset 12865).

El sentido correcto es aquél que pertenece a la primera aglomeración. Si se consideran aquellas aglomeraciones que contienen, en su interior, palabras del contexto (es decir, la primera y la cuarta), sus densidades conceptuales, si son evaluadas con nuestra fórmula (Fórmula 2), resultan ser 0.27 y 0.25. La segunda aglomeración y la cuarta, aunque no contengan ninguna palabra del contexto, tienen un valor de DC, respectivamente, de 0.5 y 0.25. Así por tanto, en función de los resultados obtenidos se seleccionaría, erróneamente, la segunda aglomeración. Esto significa que el contexto de la palabra a desambiguar no es bastante significativo para permitir su correcta desambiguación. Si hubiésemos considerado el número medio de hipónimos por cada nodo (Fórmula 1), el resultado obtenido habría sido todavía peor y habría sido elegida la aglomeración que contiene el sentido menos frecuente del nombre irregularity. Decidimos entonces, incluir también en la fórmula, la información

disponible en WorNet sobre la frecuencia de cada sentido (los sentidos de una palabra vienen enumerados con respecto a su frecuencia de aparición en los ficheros de SemCor):

$$DC(M, nh, f) = M^{\alpha}\Big(\frac{M}{nh}\Big)^{\log f} \tag{3}$$

donde α es una constante (cuyo valor se ha definido empíricamente igual a 0.25) y f es un entero (entre 1 y 25) que representa la información sobre la frecuencia (1 significa que es el sentido más frecuente, 2 el segundo más frecuente, etc.).

Con esta fórmula, la aglomeración que contiene el primer sentido de la palabra tiene una DC de por los menos 1. Las otras aglomeraciones, que se refieren a sentidos menos frecuentes, pueden venir seleccionadas, lógicamente, sólo si su DC es mayor que la de la primera aglomeración. Se ha introducido el factor M^{α} para poder dar más peso a las aglomeraciones con un número mayor de synsets relevantes, en el caso de obtener la misma densidad para diferentes aglomeraciones. Con estos ajustes (Fórmula 3), para el ejemplo de la Figura 2 obtenemos una densidad de 1.19 para la primera aglomeración y una de 0.19 para la cuarta y el nombre irregularity puede desambiguarse correctamente.

3. Resultados de la Evaluacion

El ejemplo de la Figura 2 muestra la importancia de la elección de los nombres del contexto. Esta elección puede afectar fuertemente las prestaciones del método. En nuestra investigación, decidimos componer el contexto de párrafo en párrafo, ya que a veces una frase puede ser demasiado corta (por ejemplo cuando, el único nombre es el sujeto) mientras todo el texto puede ser poco homogéneo. La Figura 3 muestra los valores de precisión y recall obtenidos para diferentes tamaños de la ventana de contexto. La *precisión* se ha definido como el resultado del cociente entre los sentidos desambiguados correctamente y el número total de sentidos desambiguados, mientras el *recall* es el resultado del cociente entre los sentidos desambiguados correctamente y el número total de sentidos [9].

Los experimentos se han llevado a cabo considerando 19 ficheros del Brown corpus que han sido escogidos al azar [1]. El mejor resultado se ha obtenido para la ventana de contexto más pequeña (igual a 2), confirmando que cuanto más cerca están las palabras es más fácil poder desambiguarlas correctamente. Implícitamente, esto demuestra que la elección de trabajar de párrafo en párrafo es correcta (por ejemplo, los valores obtenidos para ventanas mayores que 6 han debido calcularse considerando todo el texto, debido a que hay casos para los cuales en un párrafo hay menos de 6 nombres).

Los valores de recall obtenidos en la Figura 3 (alrededor del 60 % aunque cuando se consideren más nombres en el contexto) se deben al hecho que muchos nombres tienen sentidos que difieren por muy poco el uno del otro. En la jerarquía, este hecho se refleja cuando hay aglomeraciones a una profundidad bastante grande que contienen sólo un synset: el de la palabra a desambiguar. Para mejorar el recall, hemos empleado diferentes correcciones.

Un primer intento fue, por cada sentido de la palabra a desambiguar, extender su contexto con las palabras de su glosa (es decir, de su definición). Esto ha llevado a peores resultados, debido a que no se ha tenido en consideración la categoría sintáctica de cada palabra, considerando cada lexema como posible nombre. Aunque se han considerado sólo palabras monosémicas, la precisión obtenida ha decrecido en promedio en un 2 % respecto a la obtenida sin glosa. Se quiere investigar qué resultados se obtendrían lematizando previamente la glosa y etiquetándola sintácticamente.

[1] br-a01,b13,c01,d02,e22,r05,g14,h21,j01,k01,k11,l09,m02,n05,p07,r04,r06,r08,r09

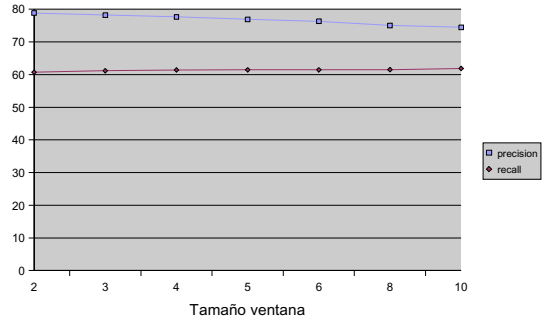

Figura 3: Precisión y recall en función del tamaño de la ventana de contexto

Para mejorar el valor de recall sin perder mucho en precisión, se ha probado elegir la aglomeración con (una o) más palabras del contexto: técnica de correccion del *contexto específico (specific context, sc)*. Debido a que los resultados no han sido muy satisfactorios, hemos dado más peso a las aglomeraciones más profundas en la jerarquía (y entonces con un significado más específico de la palabra a desambiguar). Cuando una aglomeración se encuentra más abajo de una cierta profundidad media (si $depth(cl) > avgdepth$), su densidad conceptual viene aumentada proporcionalmente al número de synsets relevantes que contiene según la fórmula:

$$DC * (depth(cl) - avgdepth + 1)^{\beta} \tag{4}$$

donde:

- cl es la aglomeración (o cluster) a estudiar;
- $depth(cl)$ devuelve la profundidad de cl respecto a la raíz de WN;
- $avgdepth$ es la profundidad media de todas las aglomeraciones en las subjerarquías obtenidas utilizando las frases de SemCor; su valor ha sido determinado empíricamente igual a 4;
- β es una constante; en la fórmula se ha utilizado β=0.20.

En la Figura 4, se muestran los resultados obtenidos al variar el tamaño de la ventana sobre todo el SemCor, para precisión, recall y cobertura con esta técnica, llamada *cluster depth correction (cdc)*, y el resto de correcciones. La *cobertura* se ha definido como el resultado del cociente entre el número total de sentidos desambiguados y el número total de sentidos [9]. La precisión más alta (81.5 %) se ha obtenido con el método base empleando una ventana de tamaño 2. Utilizando una ventana de contexto de tamaño 6 (sobre todo el texto, véase (1) en la Figura 4), y las diferentes correcciones, se ha conseguido mejorar el recall hasta un 61.4 %, aunque en detrimento de la precisión.

4. Conclusiones y Trabajos Futuros

El método propuesto en este artículo para desambiguar el sentido de los nombres, se basa en el concepto de densidad conceptual. Para el cálculo de la densidad conceptual de una subjerarquía, se ha introducido una fórmula que no se basa en el número medio de hipónimos por

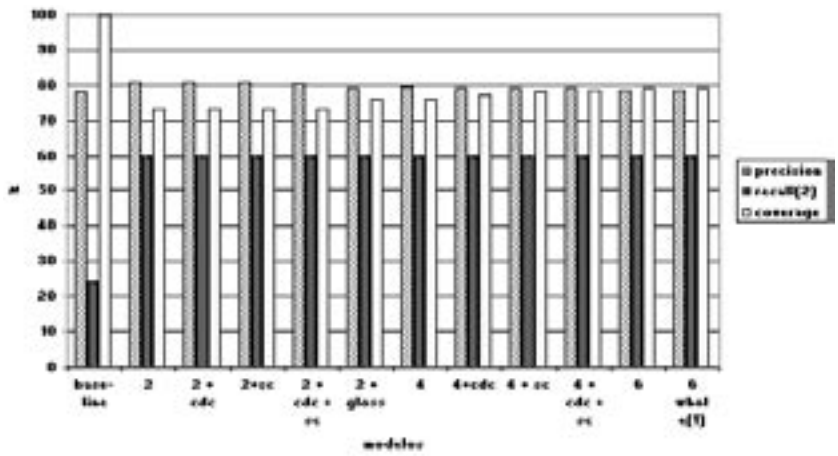

Figura 4: Subjerarquías para la desambiguación del nombre "irregularity"

synset sino que tiene en consideración sólo los caminos que llevan a los synsets relevantes, es decir, aquéllos que llevan a los synsets del nombre a desambiguar y, posiblemente, de los nombres de su contexto. El método tiene la ventaja de no necesitar procesos de entrenamiento, ni etiquetado manual. Por lo tanto, a partir de un texto de cualquier dominio se obtendrán de forma automática, los sentidos de las palabras cuyas categorías léxicas sean nombres. La salida de este método serán los nombres con el sentido correspondiente de WordNet.

La fórmula para el cálculo de la densidad conceptual, ha sido afinada considerando también la información disponible en la ontología de WordNet acerca de la frecuencia de cada sentido de los nombres a desambiguar (los sentidos de cada nombre están ordenados en WordNet por su frecuencia en SemCor). El método ha sido evaluado sobre todos los ficheros de Brown1 y Brown2 del SemCor, obteniendo una buena precisión (81.5 %) y unos valores discretos de recall (60.2 %) y cobertura (73.8 %). Estos valores se han obtenido necesitando sólo 2 nombres como contexto mientras, por ejemplo en [1], para poder obtener la mejor precisión (calculada sólo sobre el fichero br-a01 de brown1), era necesario un contexto de 15 nombres.

La combinación de las diferentes técnicas de corrección y de un contexto más grande (6 nombres), no ha producido resultados muy notables ya que si se ha obtenido una mejora para recall (61.4 %) y cobertura (78.2 %), esto ha occurrido en detrimento de la precisión (78.4 %). Esto se debe probablemente a que las aglomeraciones afectadas por las diferentes correcciones, no son las mismas. Mayor investigación es necesaria, también para afinar los parámetros α, de la fórmula de la DC, y β, de la tecnica cdc, sobre todo el SemCor (el valor de α que se ha utilizado ha sido calculado empíricamente sobre el fichero br-a01 para el cual se han obtenido los mejores resultados de precisión, 82 %, y recall, 69 %, aproximadamente).

Como trabajo futuro se quiere aplicar el método automático para resolver la ambigüedad léxica de los nombres no sólo en documentos en lenguaje natural, sino también en documentos en un lenguaje más estructurado como, por ejemplo, XML [7]. Se pretende también modificar la fórmula del método para considerar la información de la frecuencia de cada synset, no como exponente, sino dando un peso a cada synset (es decir a los synsets más frecuentes se asigna un peso mayor). También se quiere investigar qué resultados se obtendrían extendiendo el contexto del nombre a desambiguar con su glosa una vez lemátizada

y etiquetada sintácticamente, así como dando más peso a las palabras monosémicas, que es lo que normalmente hacemos cuando intentamos interpretar un texto, debido que que estas palabras, al contrario de las polisémicas, no son ambiguas. Finalmente, se pretende añadir más categorías léxicas a la hora de desambiguar: los verbos, los adjetivos y los adverbios. Esto hará que se tenga más información de contexto y mejor relacionada.

Agradecimientos

Este trabajo es parte de las acciones integradas CIAO SENSO entre España y Italia (MCYT HI 2002-0140). El trabajo de investigacion de Paolo Rosso se enmarca en el proyecto de I+D TUSIR (CICYT TIC2000-0664-C02).

Referencias

[1] E. Agirre, G. Rigau. A Proposal for Word Sense Disambiguation using Conceptual Distance. En: *Memorias de la Conferencia Internacional Recents Advances in Natural Language Processing*, 1996.

[2] G. Escudero, L. Márquez, G. Rigau. A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. *Memorias de la Conferencia Internacional CoNNL-2000 and LLL-2000*, 2000.

[3] J. Gonzalo, A. Peñas, F. Verdejo. Lexical Ambiguity and Information Retrieval revised. *Memorias de la Conferencia Internacional SIGDAT: Empirical Methods in NLP and very Large Corpora*, 1999.

[4] N. Ide, J. Veronis. Introduction on the Special Issue of Word Sense Disambiguation. *Computational Linguistic*, 24, 1998.

[5] D. Jurafsky, J. Martin, Speech and Language Processing, Prentince Hall, 2000.

[6] S. Landes, C. Leacock, R.I. Tengi. Building Semantic Concordance. En: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, USA, 1998.

[7] M. Mesiti, P. Rosso, M. Merlo. A Bayesian Approach to WSD for the Retrieval of XML Documents, En: *Memorias de la Conferencia JOTRI*, España, 2002.

[8] A. Miller. WordNet: A Lexical Database for English, *Communications of the ACM*, 38 (11): 39-41, 1995.

[9] A. Montoyo. Método basado en Marcas de Especificidad para WSD, *Procesamiento del Lenguaje Natural*, 24, 2000.

[10] M. Stevenson, Y. Wilks. The Interaction of Knowledge Sources in Word Sense Disambiguation, *Computational Linguistic*, 3(27): 321-349, 2001.