Creado el 07/03/2014
Última versión: 07/03/2014
M. Antònia Martí i Mariona Taulé, Universitat de Barcelona
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Enfoques y recursos básicos/NIVELES DE ANÁLISIS DEL LENGUAJE

Semántica y pragmática 2

Maria Antònia Martí nos va hablar de semántica y pragmática como niveles de análisis del lenguaje. Esta lección la ha realizado conjuntamente con Mariona Taulé y ambas son doctoras de la Universitat de Barcelona.

Anteriormente hemos visto una introducción a qué es la semántica léxica. Vamos a continuar ahora describiendo los recursos y fenómenos lingüísticos que cubre su análisis desde el punto de vista de la lingüística computacional y el procesamiento del lenguaje natural.

El sistema de representación más común y con un nivel de formalización mínimo, son los **diccionarios**, actualmente disponibles en soporte electrónico. Las definiciones constituyen una manera sistemática de representar el significado de las palabras. Siguen unas pautas preestablecidas que garantizan su comprensión. Así, las definiciones nominales suelen empezar con el término genérico o hiperónimo de las palabras que se definen. Los adjetivos se definen generalmente, mediante sinónimos, y los verbos mediante sinónimos o bien hiperónimos.

En el caso de partido,

- La definición 1 tiene como genérico su sinónimo, 'dividido'
- La 2, organización
- La 3, Provecho
- La 4, Territorio
- La 5, Competición

Como se puede observar, el término genérico clasifica los diferentes sentidos según el campo semántico al que pertenecen. Además de la definición, en los diccionarios se suele incluir para cada sentido, sus sinónimos y ejemplos de uso.

Un paso más en la formalización del significado léxico lo tenemos en las redes semánticas, como **WordNet** (véase la presentación de Antoni Oliver). En este caso los sentidos se representan mediante las relaciones de sinonimia, hiperonimia, hiponimia, meronimia, etc.

Finalmente, desde la **semántica teórica** y computacional tenemos una gran variedad de propuestas para la representación del significado léxico, siendo las estructuras de rasgos el tipo de formalización más extendido. Las estructuras de rasgos se basan en la idea de que el significado de las palabras se puede analizar y descomponer en primitivas (o átomos de significado).

Como hemos podido observar en el ejemplo de 'partido', las palabras pueden tener más de un sentido. Una palabra es **monosémica** cuando tiene un único sentido (es el caso de 'cronómetro'), y **polisémica** cuando tiene dos o más sentidos que guardan una relación entre sí, como es el caso de 'sierra'.

En teoría SEMÁNTICA la relación entre forma (fónica o gráfica) y significado léxico presenta una casuística que debe ser tenida en cuenta.

Son **homónimas** aquellas formas que tienen la misma categoría y forma gráfica (y fónica) pero cuyos significados no tienen ningún tipo de relación, como vemos en el ejemplo de atracar-1 y atracar-2. En atracar-2 tenemos, a su vez, un caso de polisemia, dos sentidos diferentes pero relacionados etimológicamente, es decir con un origen común.

La resolución de la polisemia es uno de los problemas claves de los sistemas de PLN: la asignación del sentido correcto a una forma es la tarea fundamental de la **Desambiguación semántica automática** (Word Sense Disambiguation, WSD), tema sobre el que volveremos más adelante.

Son **homógrafas** aquellas formas que coinciden casualmente en la forma gráfica pero que no tienen la misma categoría y cuyo significado es totalmente distinto. Es el caso de 'sobre' preposición y 'sobre' sustantivo masculino.

Formas **homófonas** son aquellas que presentan una misma forma fónica y diferente forma gráfica y significado. Es el caso de 'haya' del verbo 'haber' y 'aya', sustantivo femenino.

Unidades en lexicografía: Un problema básico de la semántica léxica es la delimitación sintagmática de sus unidades. Veamos unos ejemplos:

- Juan **ha llegado** tarde
- No **se acuerda de** ti
- Lo cree **a pies juntillas**
- El **clamor popular** lo catapultó a la fama

'Ha llegado' es una forma compuesta del verbo llegar y debería considerarse como una única unidad. **'Acordarse_de'** tiene un sentido totalmente diferente de 'acordar', lo que nos permite concluir que 'acordarse_de' constituye una única unidad léxica. El significado de **'a pies juntillas'** no es compositivo y va asociado a la expresión en su conjunto. Finalmente, **'clamor popular'** si bien se puede interpretar de manera compositiva, es decir, el significado del conjunto es la suma del significado de las partes, la frecuencia de coaparición es alta lo que es indicio de su proceso de lexicalización, es decir, de constituirse como una única unidad.

La mayoría de los ejemplos que acabamos de ver corresponden a lo que en LC Y EN PLN se denominan expresiones o **unidades multipalabra** (*Multi Word Expressions*, MWE). La importancia de su detección e interpretación automática queda demostrada por el creciente número de workshops y grupos de discusión para su tratamiento.

Actualmente, la mayoría de sistemas de PLN resuelve estas expresiones tratándolas en una lista con la información idiosincrática asociada a cada una de ellas. Dada la gran cantidad y variedad de expresiones multipalabra de una lengua, se están realizando esfuerzos en su

detección o extracción automática. A pesar de todo, de momento, su tratamiento requiere un componente importante de trabajo manual.