

Semántica española

César Antonio Aguilar
Facultad de Lenguas y Letras
07/06/2017

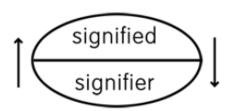
caguilara@uc.cl

La semántica de una lengua natural (1)



Uno de los temas que ha cobrado gran actualidad en PLN es el análisis semántico. No es un tema fácil, dada la complejidad que muestra explicar la semántica de una lengua natural, tanto para computólogos como para los mismos lingüistas.





Una perspectiva que se ha hecho más o menos popular para entender y explicar cómo funciona la semántica de una lengua natural es la que planteó Ferdinand de Saussure (1857-1913) en su famoso *Cours de Linguistique Générale* (1916).

Saussure concebía que la semántica y la fonología eran "las dos caras" que daban forma al signo lingüístico. De ahí vienen las nociones de *significado* ("la imagen mental que asociamos a una palabra") y *significante* (el sonido que también ligamos a dicha palabra).

La semántica de una lengua natural (2)

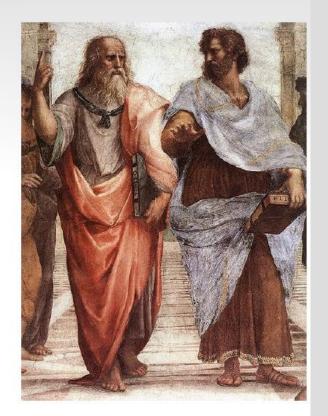


Sin embargo, este tema es mucho más viejo. De hecho, dos autores clásicos en cuestiones de semántica son Platón y Aristóteles.

Ninguno de los dos estaba realmente interesado en abordar este problema desde un punto de vista lingüístico. Empero, son conocidos sus aportes a la investigación semántica.

En el caso de Platón, su diálogo llamado *Cratilo* dio lugar a la discusión sobre si el lenguaje (y en particular, la semántica) es un sistema innato ("un don dado por los dioses a los mortales"), o si se trata de un sistema emergente relacionado con aspectos socio-biológicos ("un hacer de los hombres")

El caso de Aristóteles es más interesante: varios filósofos contemporáneos consideran incluso que se le puede ver como el primer semantista formal de la historia, pues su modelo lógico es justo eso: un sistema de análisis formal para el significado en lenguas naturales



Platón y Aristóteles, según Rafael

4

La semántica de una lengua natural (3)



En el caso particular del PLN, se ha priorizado sobre todo los estudios semánticos que tengan un corte lógico-matemático.

He aquí algunos autores relevantes hoy en día para PLN.



Charles S. Peirce (1839-1914)



Willam James (1842-1910)



Gottlob Frege (1848-1925)



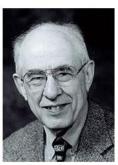
Bertrand Russell (1872-1970)



Ludwig Wittgenstein (1889-1951)



Willard V. O. Quine (1908-2000)



Hillary Putman (1926)



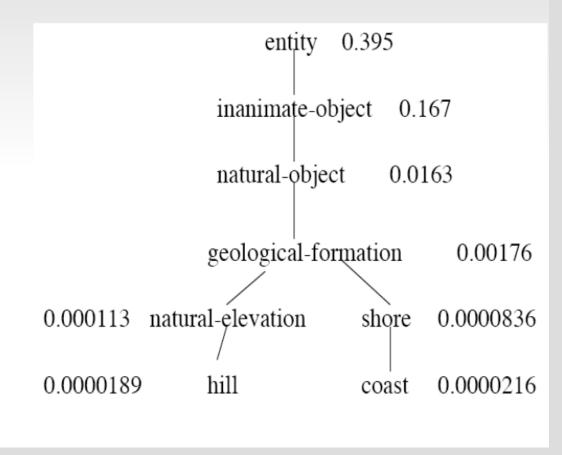
Richard Montague (1930-1971)

Semántica y PLN (1)



En PLN, hay por lo menos dos que son básicas para para todos los tratamientos computacionales que se hagan sobre la semántica:

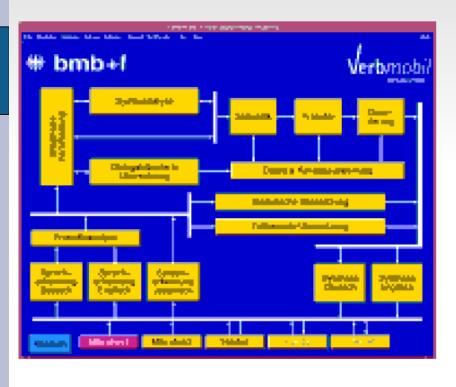
- 1. La semántica de una lengua natural es el fruto de las relaciones que se dan entre léxico y gramática.
- 2. Una forma de explicar cómo opera la semántica es través de modelos formales, los cuales arrojan evidencia empírica tratable a partir de métodos estocásticos.



Semántica y PLN (2)



PLN se ha tenido que involucrar en el análisis semántico debido a las siguientes necesidades:





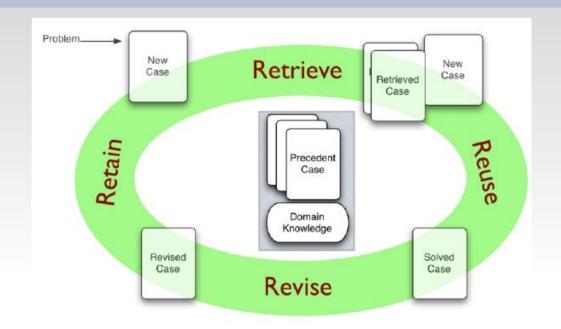
http://verbmobil.dfki.de/

1. Traducción automática:

originalmente, esta tarea se realizaba haciendo sustituciones directas entre palabras o frases, sin tomar en cuenta ningún criterio semántico. En los años 90 se contaban ya con recursos electrónicos que consideraban información léxicosemántica, p. e., el proyecto Verbmobil.

Semántica y PLN (3)





2. Robótica y agentes inteligentes: debido a la complejidad que hay en la interacción entre máquinas y humanos, un camino que se ha explorado para mejorar el desempeño de las máquinas es brindarles información léxica y semántica para que concreten tareas específicas, vía el aprendizaje de conceptos.

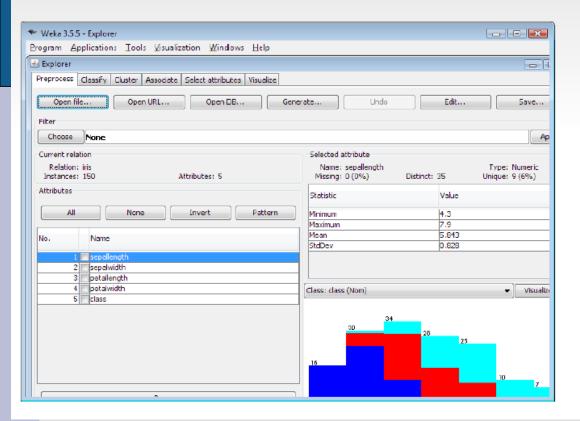


Semántica y PLN (4)









3. Búsqueda eficiente de información en Internet: dado el crecimiento exponencial de Internet, sabiendo además que es imposible regularlo, se ha pensado en sistemas de etiquetado y clasificación basados en el proceso que siguen los humanos para generar y comprender conceptos.

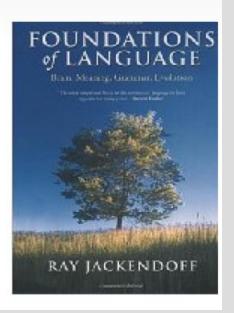
Semántica y PLN (5)



Este último punto, como hemos visto en clases previas, ha tenido relevancia en áreas como la extracción de información, la minería de textos la generación de ontologías, pues se parte de la base de que la información semántica contenida en un texto permite deducir descubrir nuevo conocimiento.

Tal idea ha sido desarrollada ampliamente por varios lingüistas. P. e., Ray Jackendoff (1948) delinea una arquitectura en donde el nivel semántico es la base en la cual se configuran conceptos.

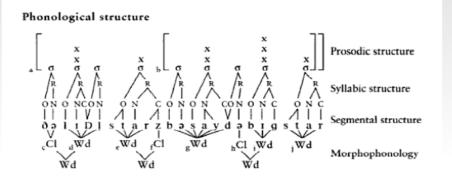




Una interfaz que trabaja eficientemente (1)

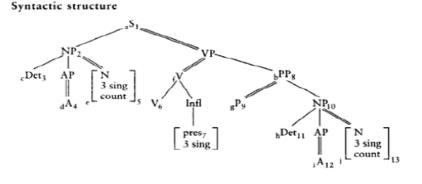


¿Cómo es que la semántica configura conceptos? Jackendoff nos propone que pensemos en una oración como:



The little star's beside the big star = ("la pequeña estrella cercana a la gran estrella").

Si analizamos la oración como lingüistas, estamos pensando en cosas como:





Semantic/conceptual structure

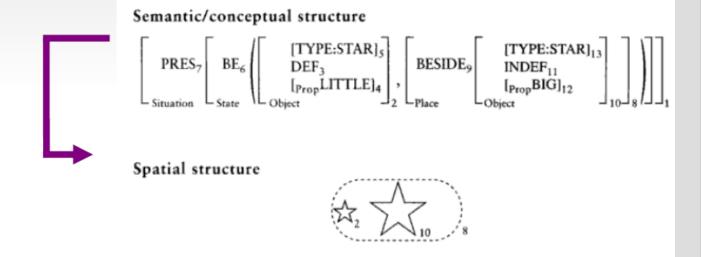
$$\begin{bmatrix} \text{PRES}_7 \\ \text{BE}_6 \\ \end{bmatrix} \begin{bmatrix} \text{ITYPE:STAR}]_5 \\ \text{DEF}_3 \\ \text{[p_{rop}LITTLE]}_4 \end{bmatrix}_2, \begin{bmatrix} \text{BESIDE}_9 \\ \text{INDEF}_{11} \\ \text{[p_{rop}BIG]}_{12} \\ \text{Object} \end{bmatrix}_{10} \end{bmatrix}_8 \end{bmatrix} \end{bmatrix}_1$$

Una interfaz que trabaja eficientemente (2)



Dejando de lado, por el momento, las estructuras fonológicas y sintácticas, enfoquémonos en la estructura semántica. Aquí hay una serie de datos que nos ayudan a modelar conceptos como "Situación", "Estado", "Lugar", "Objeto", etc., en un esquema más "cognitivo" (en este caso, se trata de una cuestión de ubicación de entidades en un espacio dado). Algo tan simple como:

Entre el proceso de manifestación de esta oración (= estructura fonológica + estructura sintáctica) y su plano de representación mental (= estructura espacial), hay una interfaz que conecta ambas partes: la estructura semántica.

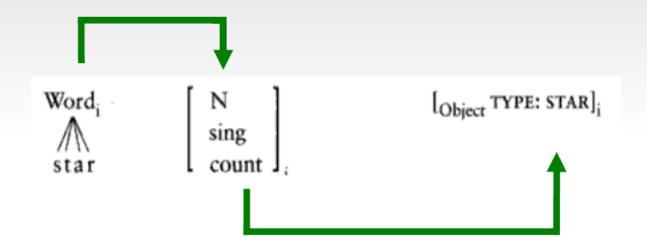


Nota importante: volviendo con la concepción que tenía Saussure sobre el signo lingüístico, Jackendoff retoma claramente tal idea, pero logra mostrar (con una mayor fineza) cómo se relaciona esa "imagen mental" con su representación física en palabras.

Una interfaz que trabaja eficientemente (3)



Decir que una palabra (o un conjunto de palabras) alude a un concepto (o a varios conceptos) es una obviedad que todos sabemos. Empero, lo difícil de esta obviedad no es pensarla, sino modelarla y explicarla de una forma clara:

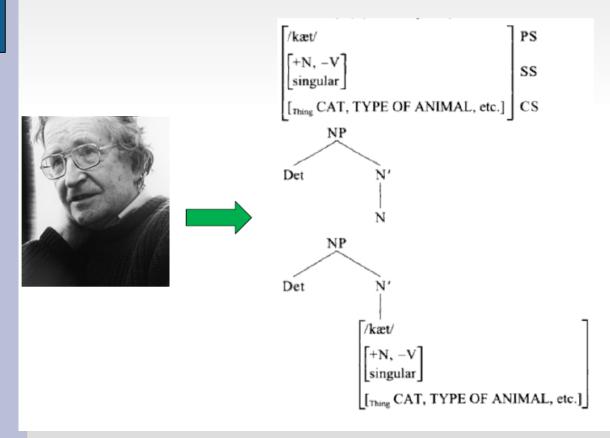


Entre la propuesta de Saussure y la de Jackendoff hay una larguísima serie de discusiones y posturas. Con todo, actualmente en lingüística se ha llegado a un consenso, y este consenso plantea que las lenguas naturales son sistemas compuestos por varios módulos, los cuales procesan en paralelo diferentes datos (desde sonidos hasta conceptos). Cuando un módulo termina un proceso y lo dirige a otro, este proceso de transferencia se hace a través de *interfaces*. Así visto, la semántica no es más que una interfaz entre "traduce" una imagen mental en palabras.

Lengua natural = gramática + lexicón (1)



Noam Chomsky fue quien propuso esta idea, y muchos lingüistas (y nolingüistas, también) la vieron como un "intento" de explicación reduccionista muy ramplón (es decir: era una tontería). Empero, siguiendo con Jackendoff, esta idea es pertinente, si la vemos del siguiente modo:



¿Qué es la palabra cat? Aparte de una secuencia de fonemas que asociamos con el inglés, aparte de ser un nombre, y aparte de ser el núcleo de frase nominal, es también la representación lingüística de un tipo de animal, del cual podemos decir que es un mamífero de la familia de los felinos, el cual es domesticable, etc.

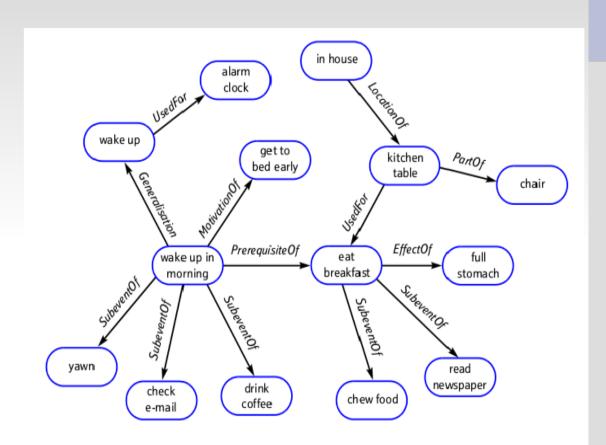
Supongamos que encapsulamos toda esta información dentro de la palabra *cat.* ¿Qué obtenemos? Una unidad léxica, que junto otras similares conforman un conjunto que llamamos **lexicón.**

Lengua natural = gramática + lexicón (2)



La idea de Chomsky dio lugar a lo que se conoce como semántica léxica, una línea de investigación que supone que, en nuestra mente, existe un súperbase de palabras (el citado lexicón), las cuales se combinan unas con otras para generar varias relaciones de sentido.

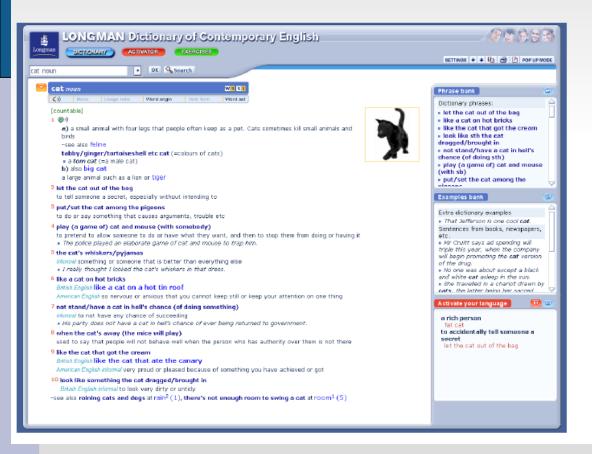
Estas combinaciones son controladas a partir de reglas, las cuales pueden ser tratables computacionalmente.



Lengua natural = gramática + lexicón (3)



Y si pasamos la semántica léxica a las investigaciones en PLN, entonces tenemos un tipo de semántica computacional que trata de reconocer y procesar información conceptual contenida en palabras.



Un ejemplo muy concreto: un diccionario electrónico como el Logman Dictionary of Contemporary English (o LODCE), en su versión electrónica es justo eso, un tipo de lexicón que presenta la definición de una palabra, ligándola con información relacionada con varios conceptos alternos: p. e., significado de frases hechas, categorías gramaticales, pronunciación para inglés americano y británico, etc.

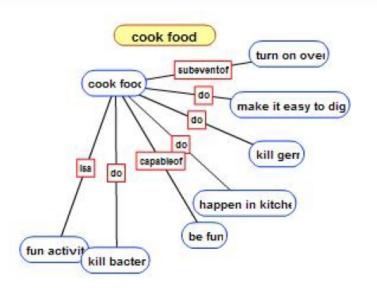
Lengua natural = gramática + lexicón (4)



Otro ejemplo mucho más complejo que el LODCE: **una base de conocimientos** (o *knowledge base*) que hace inferencias de forma automática para crear redes de conceptos representados por palabras. Esta base se llama *ConcepNet*.

The ConceptNet Project_{V2.1}

http://web.media.mit.edu/~hugo/conceptnet/



ConceptNet lo que hace es inferir conceptos, a partir de datos como opiniones, comentarios, textos, etc. Lo interesante es que ConceptNet opera usando criterios de semántica léxica, por lo que podríamos verla como una red de conocimientos léxicos basados en nuestro "sentido común".

Aunque no todo mundo esté de acuerdo (claro), podría ser que una palabra como *cocinar* nos haga pensar en conceptos como *actividad divertida*, *proceso para matar bacterias en los alimentos*, o *algo que me hace feliz cuando me meto a la cocina*.

Lengua natural = gramática + lexicón (5)

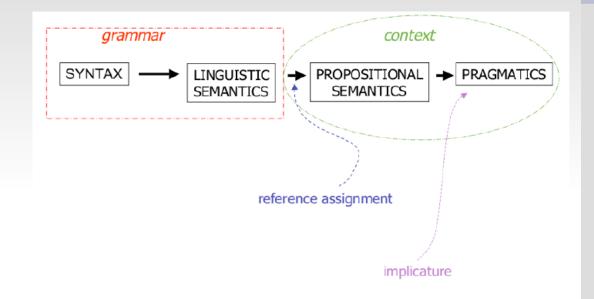


Cuando hablamos de "sentido común" en el marco del PLN, nos damos cuenta que información que suponemos se infiere de manera "simple", no lo es tanto.

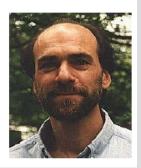
Veamos. Una cosa es decir: Juan barrió su casa hoy, y otra cosa es expresar que: el satélite barrió la superficie lunar ayer. Esto es porque:

- 1. En el primer caso nos referimos a un proceso de limpieza.
- 2. En el segundo caso, más bien describimos el recorrido que hizo el satélite sobrevolando la Luna.

Esta ambigüedad es explicable si decimos que el verbo *barrer* está semánticamente ligado a una noción como: *X hace un recorrido por Y espacio*.



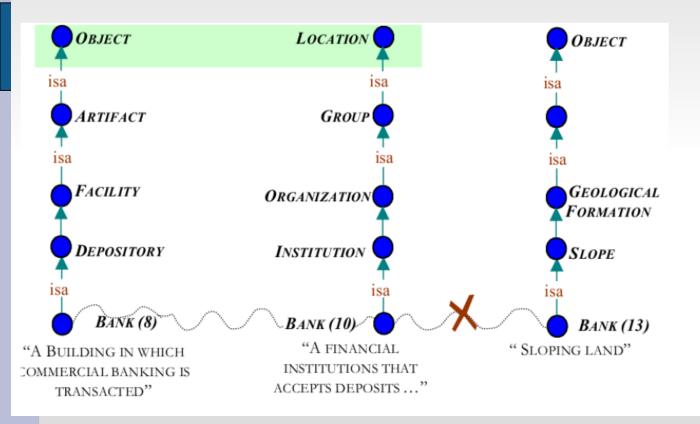
Nota importante: este esquema lo propone James Pustejovsky (Brandeis University). Con él, nos damos cuenta que *barrer* es una palabra cuyo contenido léxico se altera dependiendo del contexto pragmático en donde la aplico: si estoy en la casa de Juan, o si estoy leyendo el reporte de un satélite que orbita sobre la luna.



Infiriendo el mundo de forma automática (1)



Este problema de inferir conceptos es muy interesante para la semántica léxica computacional: más allá de la cuestión de "encapsular" toda esta información dentro del lexicón, el reto es: ¿cómo activar esta información latente cuando es necesaria?



Un nuevo ejemplo: el término banco, aparemente tiene un referente claro. Empero, ¿por qué sabemos que son cosas distintas el banco de la cocina, el banco de arena, y el banco de peces?

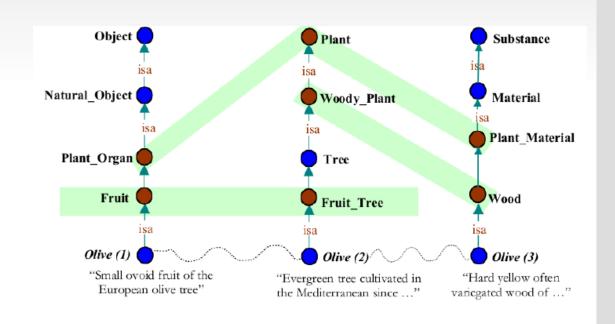
Infiriendo el mundo de forma automática (2)



Lo anterior nos lleva entonces a lo siguiente: suponiendo que nuestro cebrero, al momento de procesar palabras y conceptos, no tiene problemas para almacenar estas relaciones (para eso tenemos el lexicón), lo que queremos es entender cómo se dan esas relaciones.

En otras palabras: nuestro problema no es saber cuánta capacidad tiene la mente humana para guardar datos, sino explicar cómo se organizan esos datos.

Si hablamos de organización, ello implica pensar en cosas como: taxonomías, vocabularios, terminologías, redes, ontologías, etc.



WordNet (1)



El problema de saber cómo es que el cerebro asocia y organiza conceptos fue el punto de partida para un proyecto de investigacción, realizado en Princeton University en 1985, el cual fue coordinado por un psicolingüista llamado George Armitage Miller (1920).



PRINCETON UNIVERSITY

WordNetA lexical database for English

A grandes rasgos, Miller estaba interesado en modelar y reproducir a través de sistemas computacionales la forma en como la mente procesa y organiza conceptos referidos por palabras. Retomando la idea de Chomsky, podemos decir que Miller desarrolló un lexicón electrónico, organizado a partir de relaciones léxicas.

Con el paso del tiempo, lo que Miller y sus colegas lograron concretar fue una valiosa herramienta muy usada en PLN para realizar muchas tareas. Claro está, entre ellas destaca la búsqueda y clasificación de información semántica y conceptual en documentos.

WordNet (2)



Como todo buen lexicón, WordNet es una enorme base de datos que agrupa y clasifica palabras, con base en 4 categorías sintácticas: nombres, verbos, adjetivos y adverbios.

Con base en estas 4 categorías, WordNet determina una serie de relaciones entre palabras a partir de su contenido léxico. A este contenido que WordNet liga a cada palabra se le denomina *synset*, y equivale a un descriptor s que permite identificar y jerarquizar a una palabra dentro de una red léxica.

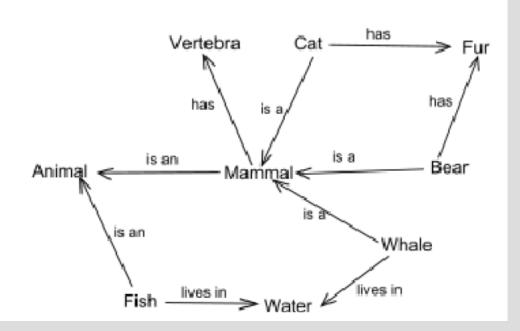
WordNet (3)



En síntesis, a partir del uso de relaciones léxicas WordNet logra crear una red de conceptos, organizada de manera jerárquica.

- Holonimia
- Polisemia
- Sinonimia
- Antonimia
- Hipernimia
- Hiponimia
- Meronomia







Gracias por su atención

Blog del curso:

http://cesaraguilar.weebly.com/clase-del-jueves-20032014.html