PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Introducción a la lingüística computacional

César Antonio Aguilar

Facultad de Lenguas y Letras

02/11/2017

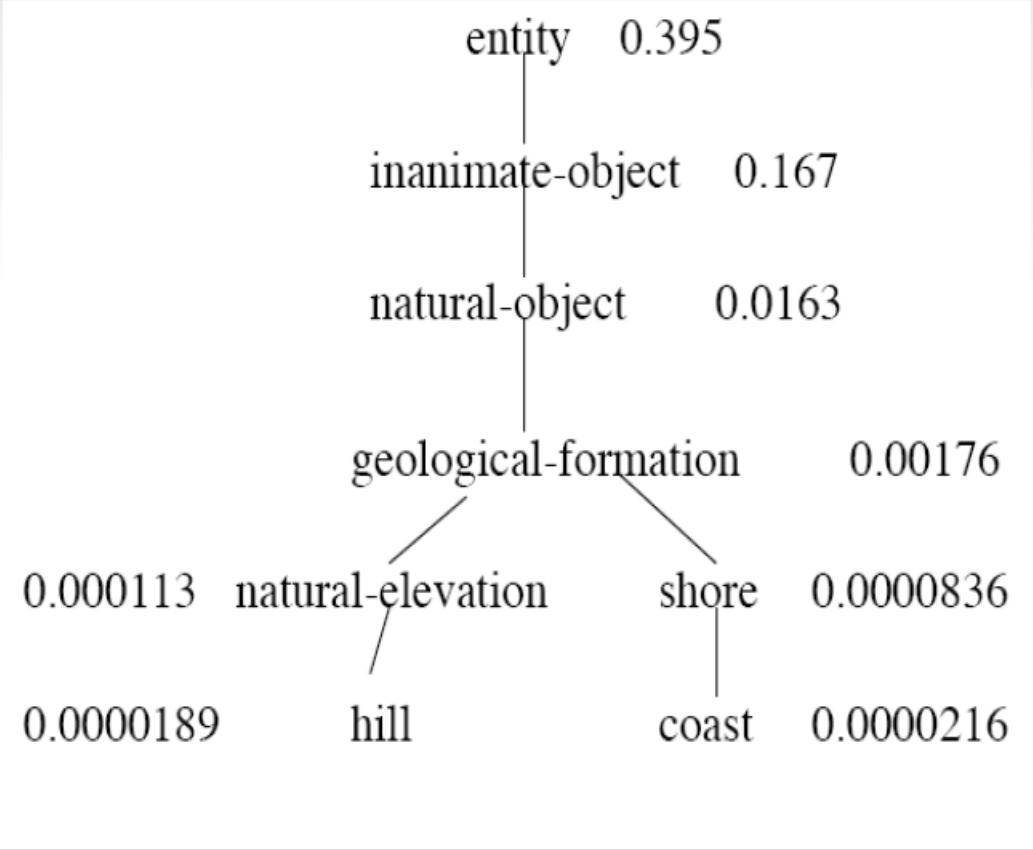
Cesar.Aguilar72@gmail.com

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:

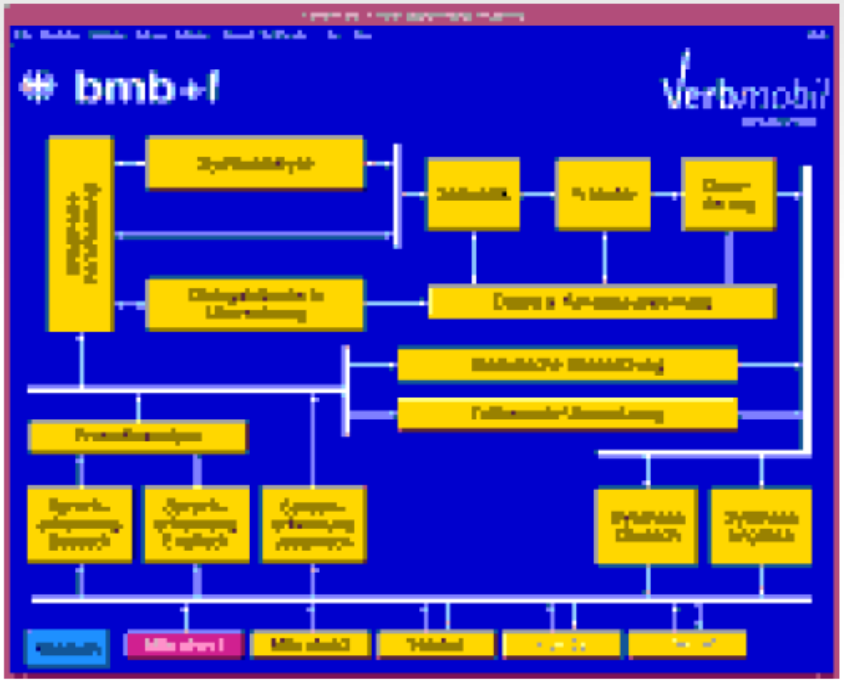
1. La semántica de una lengua natural es el fruto de las relaciones que se dan entre léxico y gramática.

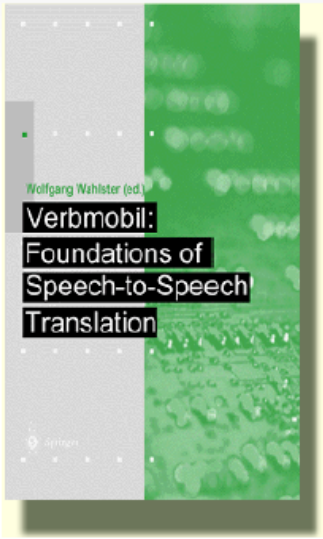
2. Una forma de explicar cómo opera la semántica es través de modelos formales, los cuales arrojan evidencia empírica tratable a partir de métodos estocásticos.



Semántica y PLN (2)

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

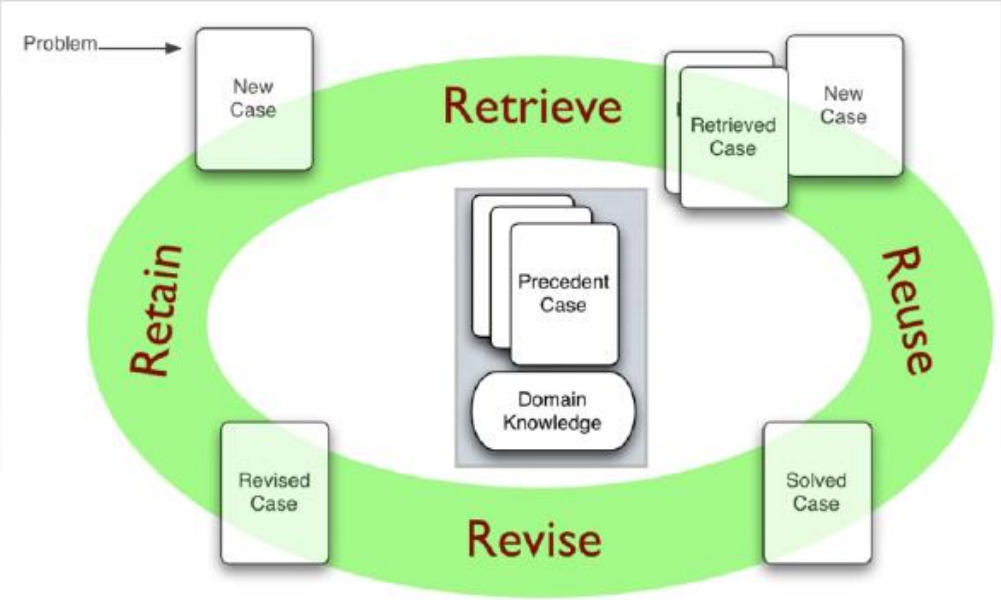




http://verbmobil.dfki.de/

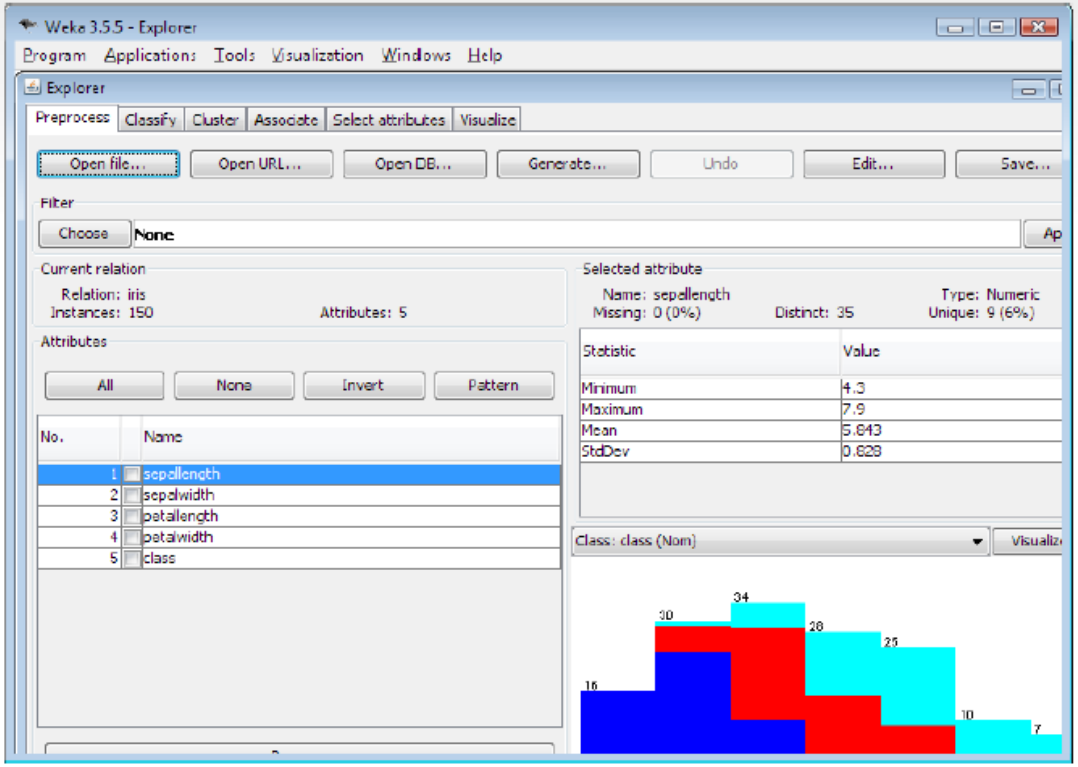
1. Traducción automática: originalmente, esta tarea se realizaba haciendo sustituciones directas entre palabras o frases, sin tomar en cuenta ningún criterio semántico. En los años 90 se contaban ya con recursos electrónicos que consideraban información léxico-semántica, p. e., el proyecto Verbmobil.

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)

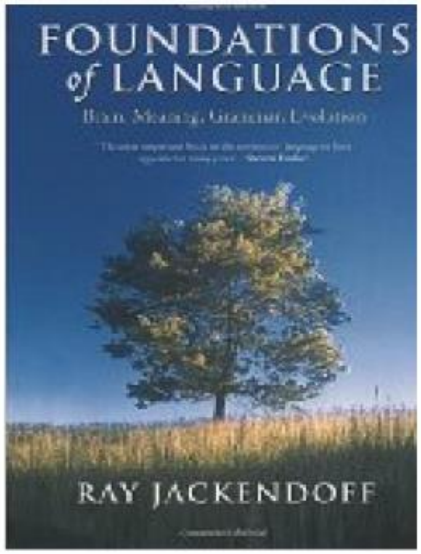


3. Búsqueda eficiente de información en Internet: dado el crecimiento exponencial de Internet, sabiendo además que es imposible regularlo, se ha pensado en sistemas de etiquetado y clasificación basados en el proceso que siguen los humanos para generar y comprender conceptos.

Semántica y PLN (5)

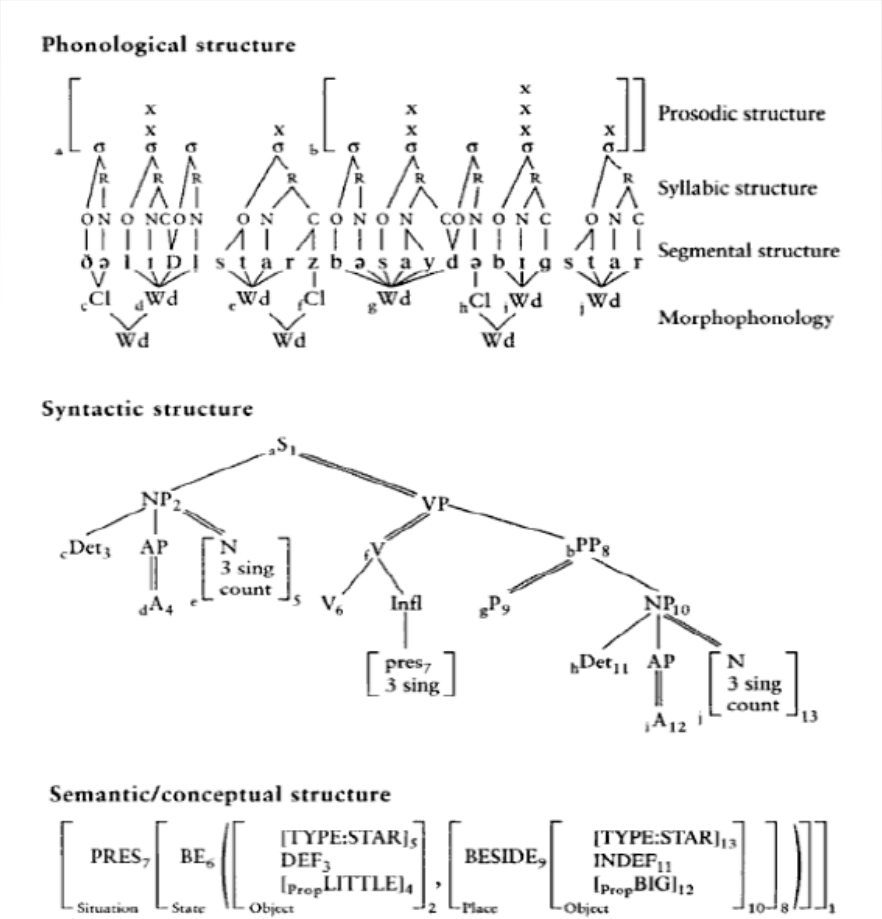
Este último punto, como hemos visto en clases previas, ha tenido relevancia en áreas como la extracción de información, la minería de textos la generación de ontologías, pues se parte de la base de que la información semántica contenida en un texto permite deducir descubrir nuevo conocimiento.

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:

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.

Semantic/conceptual structure

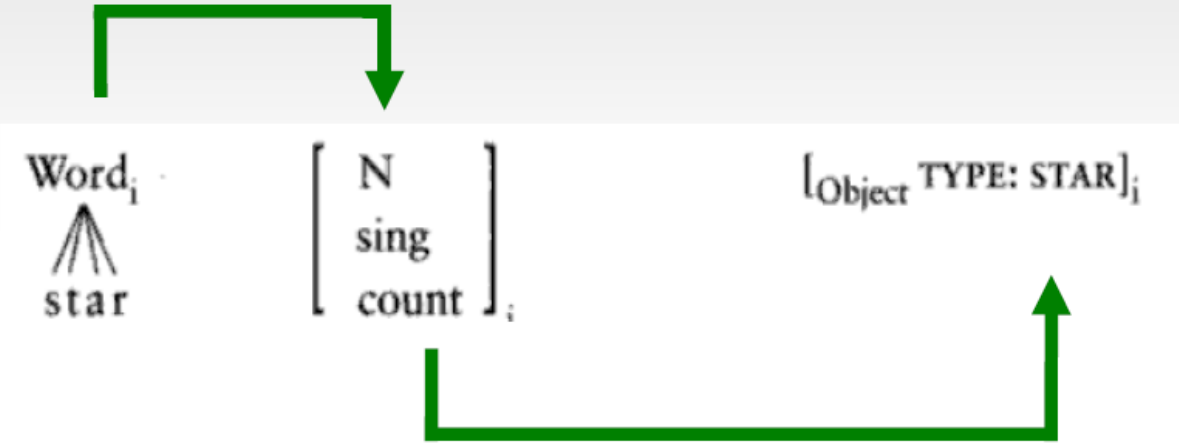
$$\left[\begin{array}{l} \text{PRES}_7 \left[\begin{array}{l} \text{BE}_6 \left(\left[\begin{array}{l} [\text{TYPE:STAR}]_5 \\ \text{DEF}_3 \\ [_{\text{Prop}}\text{LITTLE}]_4 \end{array}\right]_{\text{Object}}, \left[\begin{array}{l} \text{BESIDE}_9 \left[\begin{array}{l} [\text{TYPE:STAR}]_{13} \\ \text{INDEF}_{11} \\ [_{\text{Prop}}\text{BIG}]_{12} \end{array}\right]_{\text{Object}} \end{array}\right]_{\text{Place}} \right)_{\text{State}} \end{array}\right]_{\text{Situation}} \end{array}\right]_1$$

Spatial structure

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)

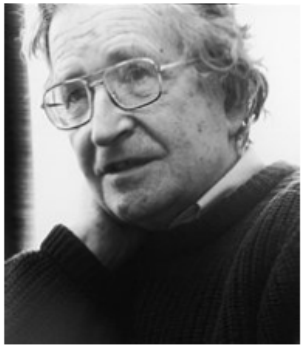
Decir que una palabra (o un conjunto de palabras) alude a un concepto (o a varios conceptos) es una obviedad que todos sabemos. Empero, lo difícil de esta obviedad no es pensarla, sino modelarla y explicarla de una forma clara:

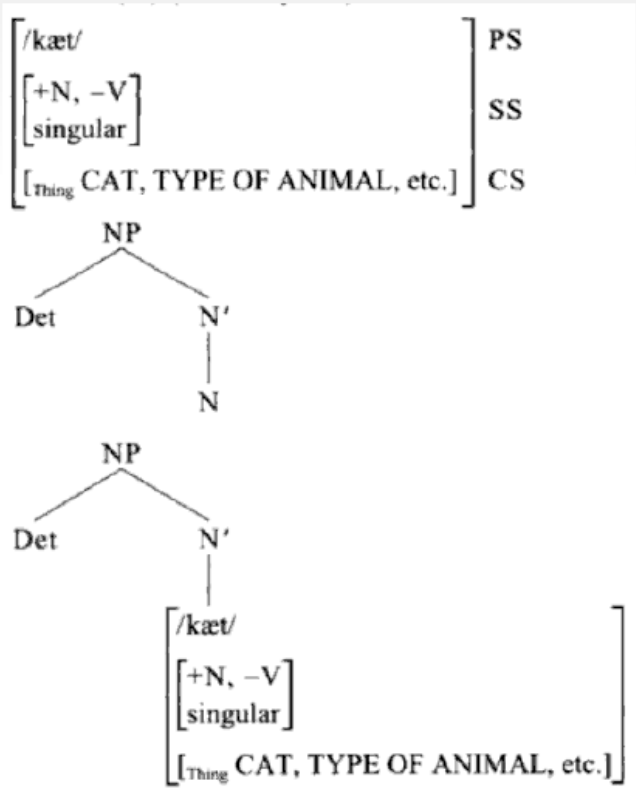


Entre la propuesta de Saussure y la de Jackendoff hay una larguísima serie de discusiones y posturas. Con todo, actualmente en lingüística se ha llegado a un consenso, y este consenso plantea que las lenguas naturales son sistemas compuestos por varios módulos, los cuales procesan en paralelo diferentes 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 no-lingüistas, 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:





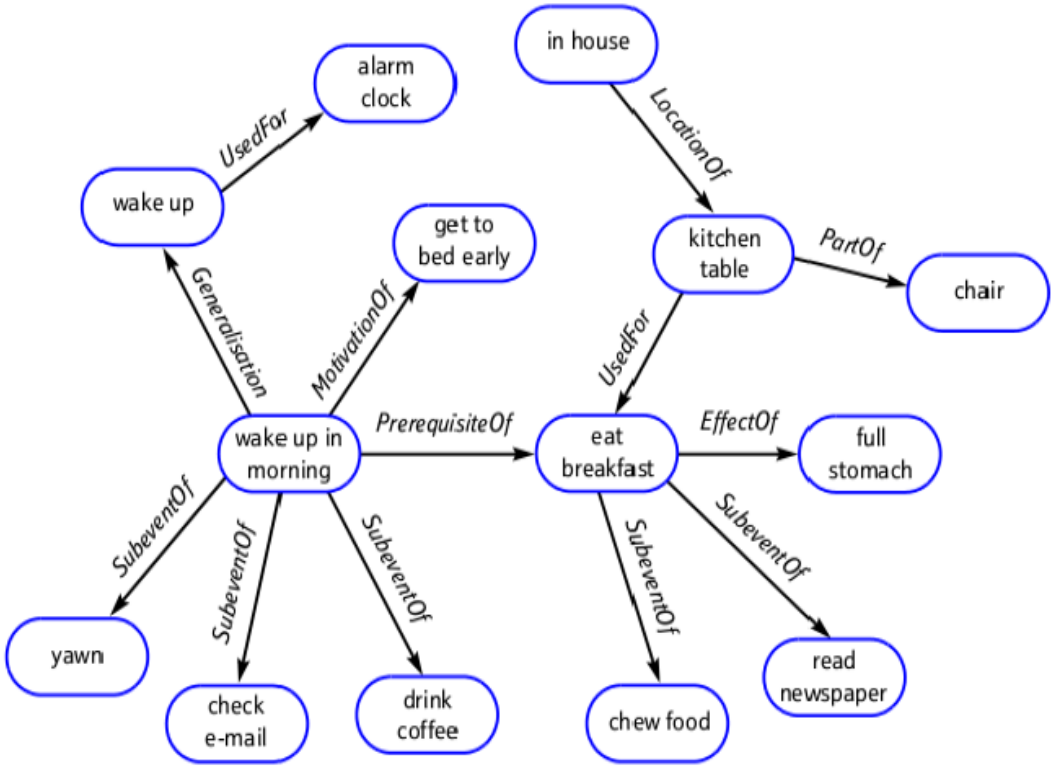
¿Qué es la palabra *cat*? Aparte de una secuencia de fonemas que asociamos con el inglés, aparte de ser un nombre, y aparte de ser el núcleo de frase nominal, es también la representación lingüística de **un tipo de animal**, del cual podemos decir que *es un mamífero de la familia de los felinos, el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

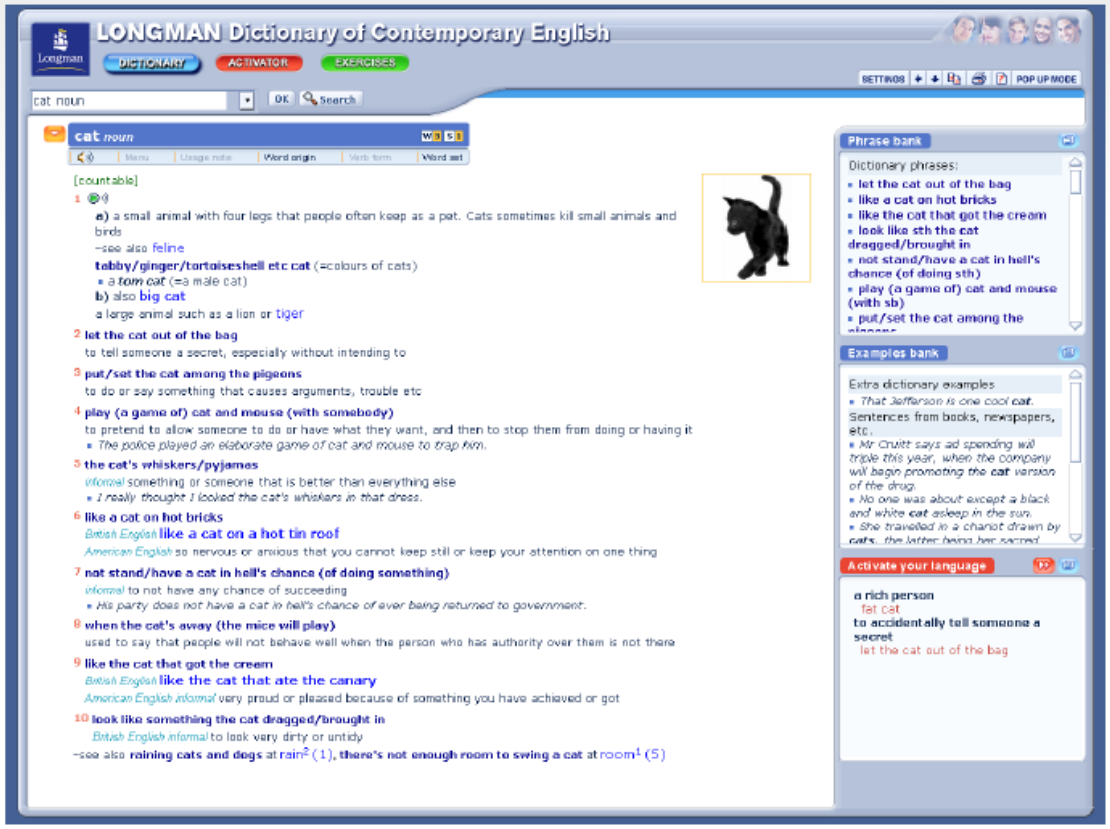
La idea de Chomsky dio lugar a lo que se conoce como **semántica léxica**, una línea de investigación que supone que, en nuestra mente, existe un súper-base de palabras (el citado lexicón), las cuales se combinan unas con otras para generar varias relaciones de sentido.

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.

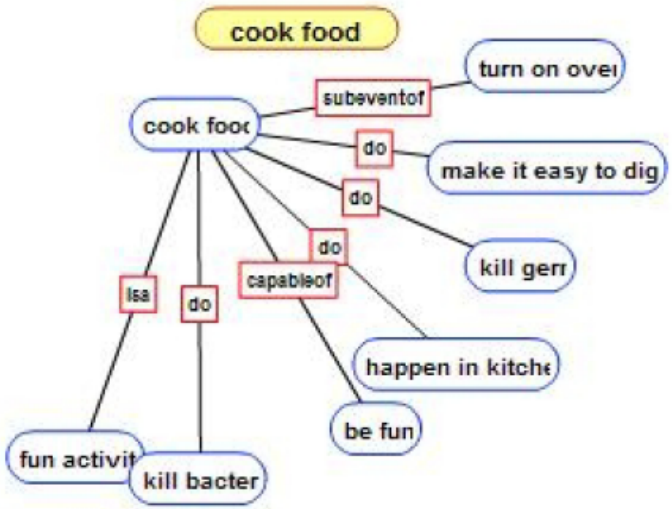
Lengua natural = gramática + lexicón (4)

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

A Very-Large Semantic Network of Common Sense Knowledge

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)

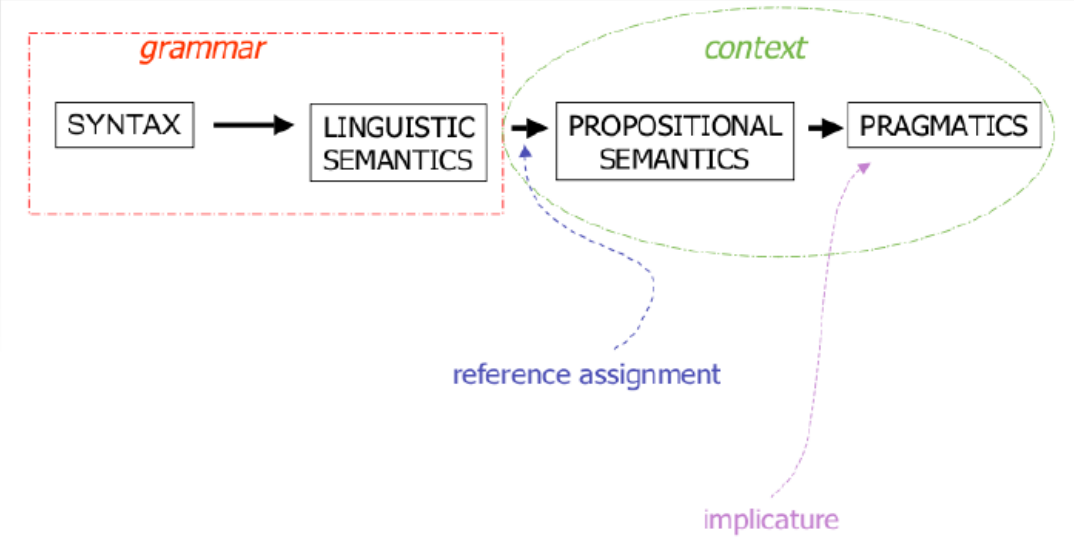
Cuando hablamos de "sentido común" en el marco del PLN, nos damos cuenta que información que suponemos se infiere de manera "simple", no lo es tanto.

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:

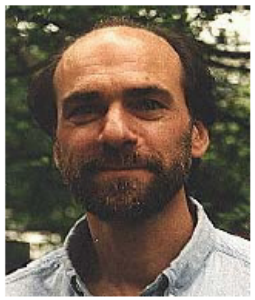
1. En el primer caso nos referimos a un proceso de limpieza.

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.



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

WordNet
A 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.

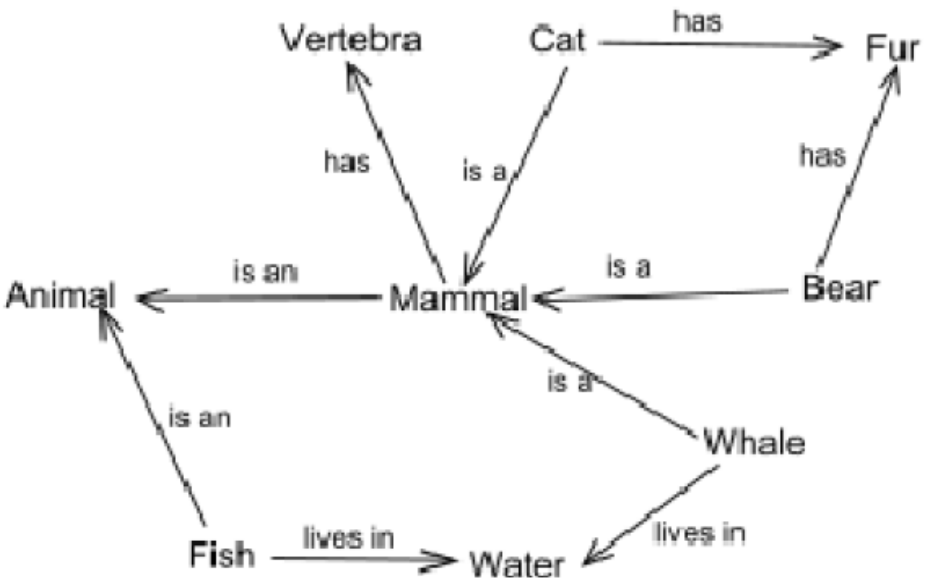
```
dog, domestic dog, Canis familiaris
   => canine, canid
      => carnivore
        => placental, placental mammal, eutherian, eutherian mammal
          => mammal
            => vertebrate, craniate
              => chordate
                => animal, animate being, beast, brute, creature, fauna
                  => ...
```

WordNet (3)

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

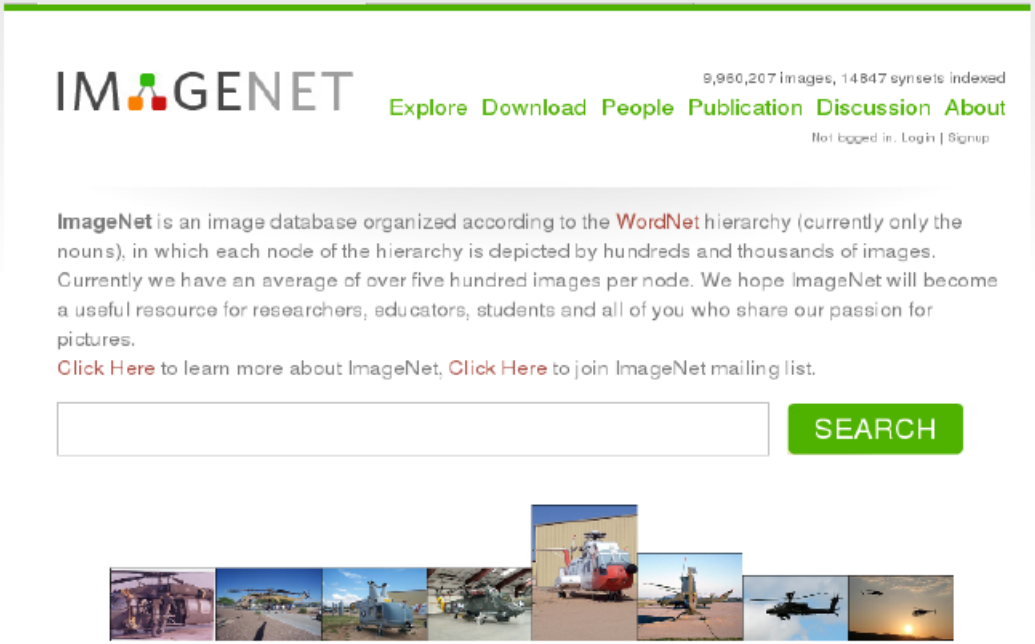
- Holonimia
- Polisemia
- Sinonimia
- Antonimia
- Hipernimia
- Hiponimia
- Meronomia

WordNet



ImageNet (1)

Más adelante analizaremos a detalle *WordNet* y otras redes léxico-semánticas relevantes. Por lo pronto, veamos algunas tecnologías relacionadas (o derivadas) de la creación de ontologías.

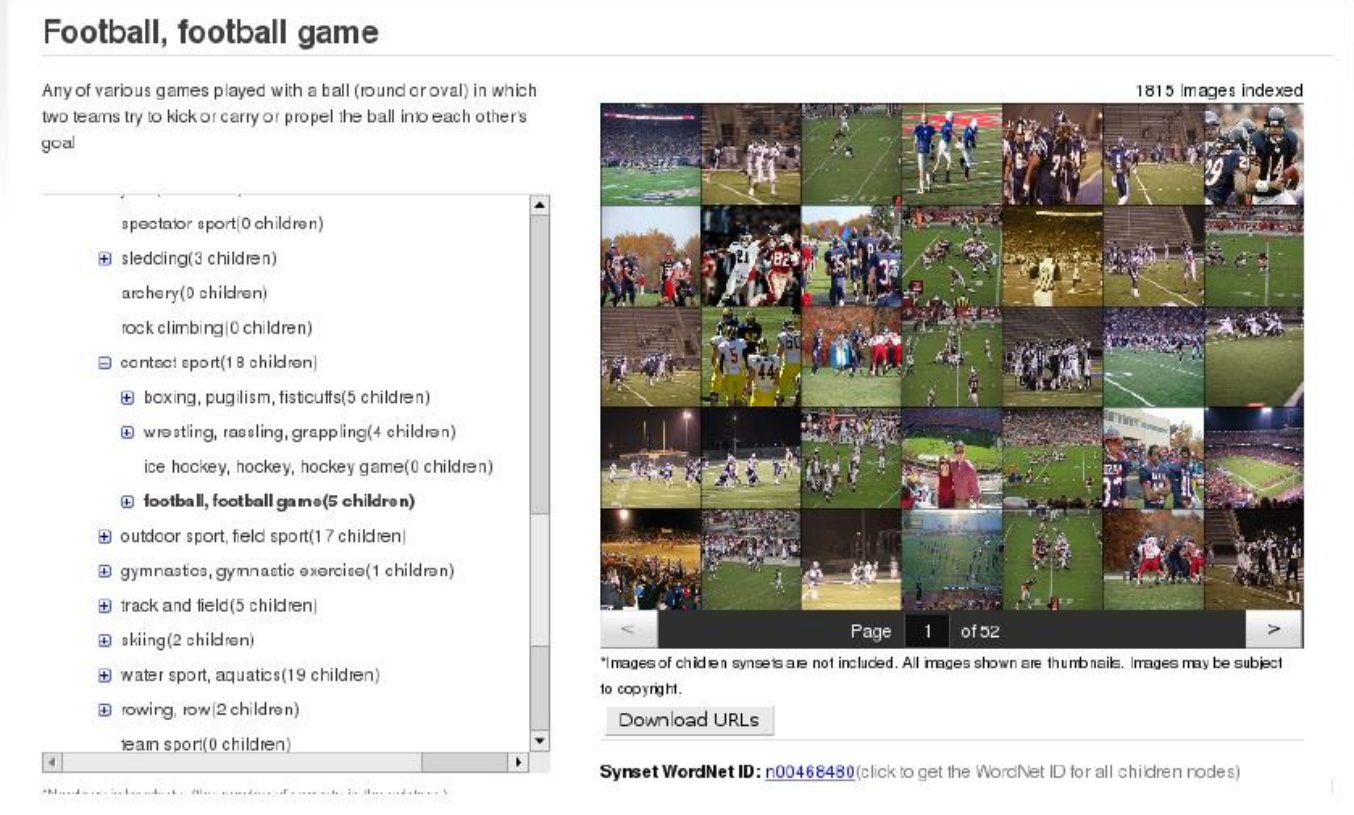


Un proyecto derivado de *WordNet* es *ImageNet*, un banco de imágenes (cerca de un millón), las cuales van asociadas a cada uno de los *synsets* de *WordNet*. La gran ventaja es que las imágenes están organizadas de acuerdo con una red jerárquica de conceptos (precisamente *WordNet*).

http://www.image-net.org

ImageNet (2)

ImageNet es un sistema que se ubica dentro de un paradigma de de identificación y clasificación de información llamado *búsquedas facetadas*. Esta clase de búsquedas parten de la idea de emplear diferentes formatos de datos (p. e., texto e imagen) para resolver consultas.



Por ello, *ImageNet* es más que una enciclopedia electrónica: la apuesta sería que este sistema fuese capaz de presentar esta información *en tiempo real*.

FrameNet (1)

En PLN, tras la aparición de WordNet se desarrollaron otras redes léxicas, más o menos similares a ésta, pero teniendo en mente siempre aportar información léxica que no aparece en WordNet.

Una de las más conocidas hoy en día es *FrameNet*.

FrameNet es un proyecto que se inició a mediados de los noventa, bajo la dirección de Charles Fillmore (1929), un lingüista que desarrolló una teoría léxico-semántica conocida como **Frame Semantics** (no hay traducción al español, pero podríamos hablar de "marcos de roles semánicos referenciales"). Esta teoría se plantea hacer descripciones semánticas de eventos. más que de conceptos asociados a palabras o términos.

https://framenet.icsi.berkeley.edu/fndrupal/

FrameNet (2)

¿Qué ofrece FrameNet? Justo una descripción semántica de los participantes o actores que intervienen en un evento o situación dada. Para la semántica lingüística en general, tales eventos o situaciones son referidos por **verbos**, por ejemplo:

La frase verbal *working_on* ("trabajar_en/con_X") proyecta todo un escenario o "marco referencial semántico", que permite asociar a un conjunto de actores y/o participantes que juegan un rol específico. FrameNet describe a todos estos actores, participantes y roles, dándonos una descripción conceptual.

Working_on

Definition:

An Agent expends effort towards acheiving a Goal. Alternatively, a Salient_entity involved in the Goal can be expressed in place of a Goal expression.

In addition, our organization will WORK on improving accountability in every department.

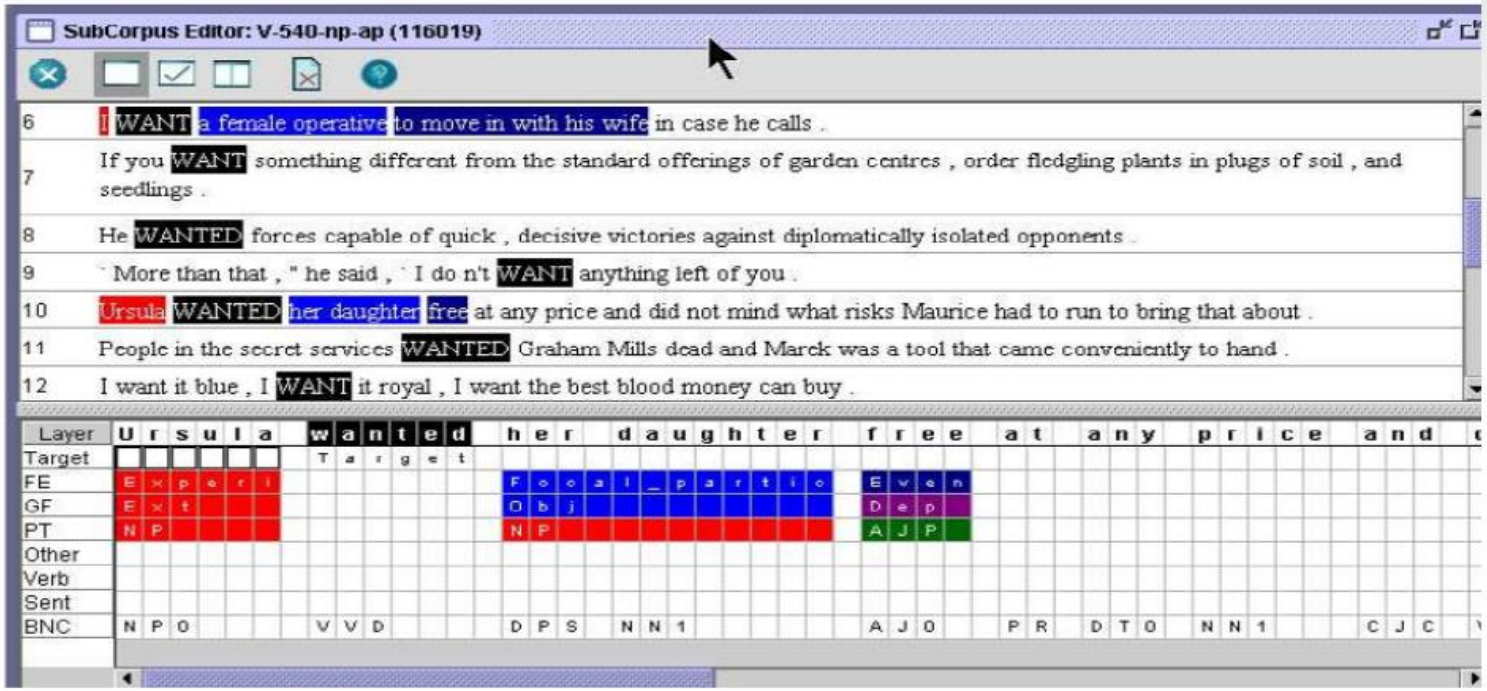
If somebody is WORKING on the pipes of your house and you turn on the faucet hard, you get sand, junk.

You could see several students PLUGGING AWAY at the problem.

We have benefited greatly from their WORK on this problem.

FrameNet (3)

Para lograr esto, FrameNet desarrolla todo un sistema de etiquetado sumamente fino, lo cual es un buen ejemplo de cómo funciona la **anotación semántica** para configurar esta clase de *frames*.



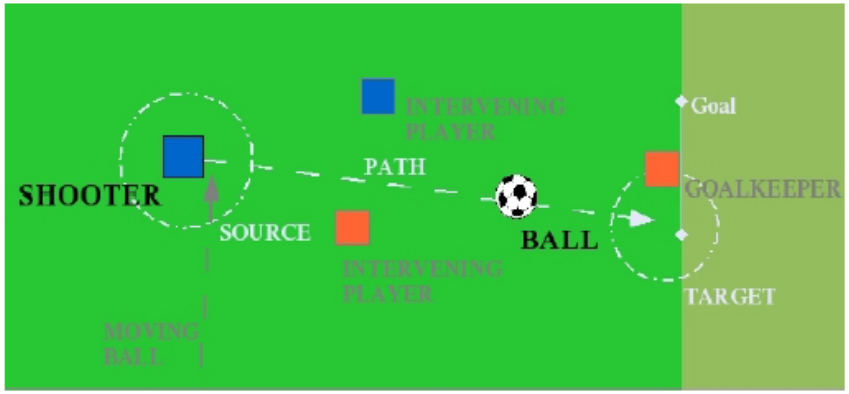
Nota importante: Al respecto, conviene quizá recordar un poco lo que vimos sobre anotación lingüística: con FrameNet, además de saber que *working_on* es una frase verbal (VP), ahora entedemos que esta VP proyecta un *Agente* y un *Paciente*, que los dos actuan en un *Proceso*, que el *Agente* tiene una *Meta*, para la cual el *Agente* puede ocupar un *Instrumento*, etc.

Kictionary

Así explicado, FrameNet parece un recurso casi de uso exclusivo para lingüistas. Empero, ¿se les hubiera ocurrido crear un diccionario de fútbol, basado en los *frames* semánticos que vimos?

Kicktionary

The multilingual electronic dictionary of football language.



Para *Kicktionary*, *to shot* ("disparar/tirar"), es una acción u evento que ocurre en el fútbol, y puede ser descrita a través de sus participantes y sus acciones: para "hacer un disparo", necesitamos un "tirador" que "le pegue" con el pie a un "balón", el cual debe recorrer una "trayectoria", hacia un "objetivo" y producir un "gol".

www.kicktionary.de/index.html

En resumen...

La semántica léxica de corte computacional es una muy buena herramienta que nos ayuda a identificar y clasificar información lingüística ya sean documentos especializados (como en el caso del DiCoInfo), o ya sean de cualquier clase de texto (por ejemplo, WordNet, que es un recurso para lengua en general).

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

Lo más importante para nosotros, más que profundizar en sus aspectos lingüísticos o computacionales, es imaginar cómo podemos aprovecharlas para obtener y clasificar información de forma eficiente y rápida.

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/introduccioacuten-a-la-linguumliacutestica-computacional.html