Creado el 27/01/2014
Última versión: 27/01/2014
Antoni Oliver, Universidat Oberta de Catalunya

Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Enfoques y recursos básicos de las TLH/APROXIMACIONES, RECURSOS Y HERRAMIENTAS DE LAS TLH

Basado en corpus y basado en conocimiento (½)

1. Introducción

En este tema vamos a ver dos estrategias para el tratamiento computacional de fenómenos lingüísticos: las **aproximaciones basadas en corpus** i las **aproximaciones basadas en conocimiento**. Como pasa en muchas otras clasificaciones sobre otras técnicas de las tecnologías del lenguaje, la frontera entre estas dos aproximaciones no es clara

Las **aproximaciones basadas en corpus** utilizan estos recursos para obtener la información que necesitan para resolver la tarea. En el siguiente apartado recordaremos el concepto de corpus y los diferentes niveles de anotación (es decir, de información que pueden contener).

Las **aproximaciones basadas en conocimiento** utilizan diferentes tipos de información para resolver la tarea. Esta información se podría considerar como *conocimiento del mundo* y se trata de información no contenida de manera implícita en los corpus. Un ejemplo de conocimiento sería por ejemplo que un *gato* es un animal pero que también puede ser una herramienta. Más adelante veremos diferentes fuentes de información que se pueden utilizar para resolver tareas lingüísticas.

Para finalizar este tema veremos cómo utilizar una y otra aproximación en una tarea concreta: la desambiguación de sentidos (WSD - *Word Sense Disambiguation*).

2. Aproximaciones basadas en corpus

2.1 ¿Qué es un corpus lingüístico?

Veamos una definición sencilla de corpus: [fuente Wikipedia]

Un **corpus lingüístico** es un conjunto, habitualmente muy amplio, de ejemplos reales de uso de una lengua. Estos ejemplos pueden ser textos (lo más común) o muestras orales (generalmente transcritas).

En el resto de este tema consideraremos corpus de textos.

2.2. Niveles de anotación de los corpus

Los corpus pueden tener diferentes niveles de anotación, es decir, diferentes tipos de información que puede estar asociada al corpus. La manera de asociar esta información al corpus puede ser tanto manual, como automática así como (opción muy frecuente) automática con revisión manual.

Así, los diferentes niveles de anotación pueden ser:

- **Ninguno (corpus sin anotar)**

  En este caso el corpus no dispone de ninguna información asociada. Incluso en este caso el corpus proporciona información sobre cómo es el corpus, el tipo de textos, si es lenguaje general o especializado. Esta mínima información puede ya ser útil para intentar resolver alguna tarea concreta.

- **Etiquetado morfosintáctico**

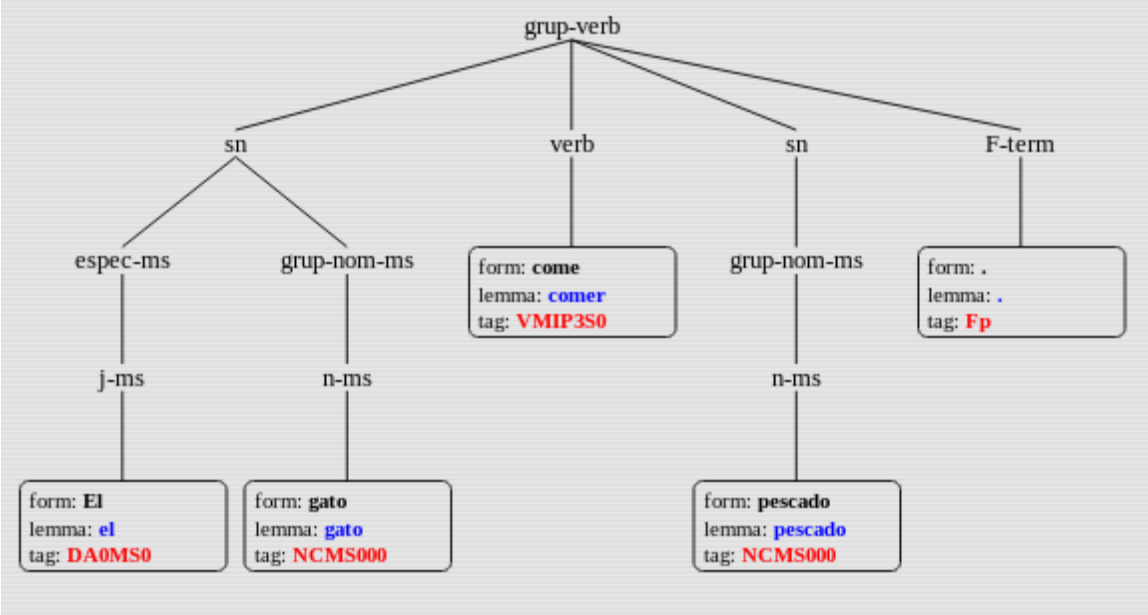
  Cada palabra del corpus tiene asociada su lema (forma base) y una etiqueta que indica su categoría gramatical y ciertas subcategorizaciones. En el siguiente ejemplo vemos una oración etiquetada a nivel morfosintáctico (la anotación se ha realizado automáticamente con la herramienta Freeling)

Yo	bajo	con	el	hombre	bajo	a	tocar	el	bajo	bajo	la	escalera	.
yo	*bajar*	*con*	*el*	*hombre*	*bajo*	*a*	*tocar*	*el*	*bajo*	*bajo*	*el*	*escalera*	*.*
PP1CSN00	VMIP1S0	SPS00	DA0MS0	NCMS000	AQ0MS0	SPS00	VMN0000	DA0MS0	NCMS000	SPS00	DA0FS0	NCFS000	Fp

  Fijaos que esta oración contiene una palabra muy ambigua a nivel morfosintàctico (*bajo*) que tanto puede ser sustantivo, verbo, adjetivo o preposición.

- **Etiquetado sintáctico**

  En este caso las oraciones del corpus están analizadas sintácticamente. A este tipo de corpus a menudo se denominan *Treebanks*. Veamos un ejemplo de análisis sintáctico (realizado también automáticamente con Freeling).



Esta misma estructura se puede representar a nivel textual de la siguiente manera:

```
+grup-verb_[
 sn_[
  espec-ms_[
   +j-ms_[
    +(El el DA0MS0 -)
   ]
  ]
  +grup-nom-ms_[
   +n-ms_[
    +(gato gato NCMS000 -)
   ]
  ]
 ]
 +verb_[
  +(come comer VMIP3S0 -)
 ]
 sn_[
  +grup-nom-ms_[
   +n-ms_[
    +(pescado pescado NCMS000 -)
   ]
  ]
 ]
]
```

- **Etiquetado semántico**

En este caso las palabras (si no todas, al menos algunas) tienen una etiqueta que indica el sentido de la palabra. Estas etiquetas pueden ser de diferentes tipos, pero una de las más extendidas son los llamados *synsets* de *WordNet*. Veremos un poco más adelante que es el WordNet. Pongamos unos ejemplos:

Necesito un gato para cambiar la rueda del coche.

Necesito necesitar VMIP1S0 01188725-v
un uno DI0MS0
gato gato NCMS000 03588951-n
para para SPS00
cambiar cambiar VMN0000 00126264-v
la el DA0FS0
rueda rueda NCFS000 04574999-n
de de SPS00
el el DA0MS0
coche coche NCMS000 02959942-n
. . Fp 1 -

Tengo un gato de color negro que come mucho pescado.

Tengo tener VMIP1S0 02205098-v:
un uno DI0MS
gato gato NCMS000 02121620-n
de de SPS00
color color NCMS000 04956594-n
negro negro AQ0MS0 00392812-a
que que PR0CN000
come comer VMIP3S0 01166351-v
mucho mucho DI0MS0
pescado pescado NCMS000 07775375-n
. . Fp 1 -

En la primera oración la etiqueta que ha recibido la palabra *gato* es el *synset* de WordNet 03588951-n correspondiente al sentido de *herramienta para elevar objetos pesados o para ejercer presiones elevadas*. En cambio en la segunda oración ha recibido la etiqueta 02121620-n, correspondiente al sentido *mamífero felino generalmente tienen piel gruesa y suave sin capacidad de rugir.*

Pueden haber otros tipos de etiquetados, pero los expuestos aquí son los más frecuentes.

3. Aproximaciones basadas en conocimiento

3.1. El *conocimiento del mundo*

Desde los orígenes del tratamiento informático del lenguaje se ha visto que la resolución de muchas de las tareas relacionadas con este ámbito no se podían resolver de manera satisfactoria sin disponer de lo que se ha denominado a menudo *conocimiento del mundo*; es decir, el conocimiento general que tienen los hablantes de una lengua para poder mantener una conversación. Este conocimiento del mundo, también denominado a menudo *conocimiento enciclopédico*, es pues la información que una persona tiene almacenada en su memoria a partir de lo que ha experimentado, vivido o aprendido.

3.2. Los diccionarios y enciclopedias como fuentes para el conocimiento del mundo

Los sistemas de Procesamiento del Lenguaje Natural a menudo han utilizado diccionarios y enciclopedias como fuentes para obtener este deseado *conocimiento del mundo*.

Si consultamos la entrada de *gato* del Diccionario de la Real Academia de la Lengua Española, obtenemos hasta 17 acepciones o significados diferentes:

gato[1].

(Del lat. *cattus*).

1. m. Mamífero carnívoro de la familia de los Félidos, digitígrado, doméstico, de unos cinco decímetros de largo desde la cabeza hasta el arranque de la cola, que por sí sola mide dos decímetros aproximadamente. Tiene cabeza redonda, lengua muy áspera, patas cortas y pelaje espeso, suave, de color blanco, gris, pardo, rojizo o negro. Es muy útil en las casas como cazador de ratones.

2. m. Bolso o talego en que se guardaba el dinero.

3. m. Dinero que se guardaba en él.

4. m. Instrumento de hierro que sirve para agarrar fuertemente la madera y traerla a donde se pretende. Se usa para echar aros a las cubas, y en el oficio de portaventanero.

5. m. Máquina compuesta de un engranaje de piñón y cremallera, con un trinquete de seguridad, que sirve para levantar grandes pesos a poca altura. También se hace con una tuerca y un husillo.

6. m. Trampa para coger ratones.

7. m. Instrumento que consta de seis o más garfios de acero, y servía para reconocer y examinar el alma de los cañones y demás piezas de artillería.

8. m. coloq. Ladrón, ratero que hurta con astucia y engaño.

9. m. coloq. Hombre sagaz, astuto.
10. m. coloq. Hombre nacido en Madrid.
11. m. *Carp.* Instrumento de hierro o de madera compuesto de dos planchas con un tornillo que permite aproximarlas de modo que quede fuertemente sujeta la pieza que se coge entre ambas.
12. m. *Zool.* Nombre aplicado a todos los félidos en general.
13. m. *Arg.* y *Ur.* Baile de movimientos rápidos, de pareja suelta que suele acompañarse de coplas cuya letra coincide con las distintas figuras.
14. m. *Arg.* Música que acompaña ese baile.
15. m. *C. Rica.* Variedad de pastel, cortado rectangularmente, compuesto de dos tapas unidas con miel o conserva.
16. m. *El Salv.* **bíceps braquial.**
17. m. despect. coloq. *El Salv.* y *Méx.* **servidor** (‖ persona que sirve como criado).

Cada una de estas acepciones están acompañadas de una definición. Esta definición puede contener muchas pistas para resolver algunas tareas de procesamiento del lenguaje.

Otra fuente de información clásica son las enciclopedias. En este sentido la enciclopedia libre Wikipedia (http://wikipedia.org) ha obtenido una gran relevancia. Se trata de un proyecto colaborativo, donde cualquier usuario puede crear o mejorar artículos. Existen versiones de la Wikipedia en muchísimas lenguas y además las diferentes versiones lingüísticas están enlazadas entre ellas a nivel de artículo. La WIkipedia se puede consultar por Internet y además se puede descargar en diferentes formatos, con lo que se facilitan las tareas de procesamiento informático,

Veamos que nos muestra la Wikipedia española si realizamos la consulta *gato*.

Felis silvestris catus

(Redirigido desde «Gato»)

«Gato» redirige aquí. Para otras acepciones, véase Gato (desambiguación).

El **gato** o **gato doméstico** (***Felis silvestris catus***) es una subespecie de mamífero carnívoro de la familia Felidae. El gato está en convivencia cercana al ser humano desde hace unos 9500 años,[1] periodo superior al estimado anteriormente, que oscilaba entre 3500 y 8000 años.

En las lenguas romances los nombres actuales más generalizados derivan del latín vulgar *catus*, palabra que aludía especialmente a los gatos salvajes en contraposición a los gatos domésticos que, en latín, eran llamados *felis*.

Hay docenas de razas, algunas sin pelo o incluso sin cola, como resultado de mutaciones genéticas, y existen en una amplia variedad de colores. Son depredadores por naturaleza, siendo sus posibles presas más de cien especies diferentes de animales para alimentarse. También son animales que pueden asimilar algunos conceptos, y ciertos ejemplares pueden ser entrenados para manipular mecanismos simples.

Se comunican con gemidos, gruñidos y alrededor de un centenar de diferentes vocalizaciones,[2] además del lenguaje corporal.

Se creía que el gato salvaje africano (*Felis silvestris lybica*) es su ancestro más

Gatos de diferentes razas

Vemos que nos muestra la entrada correspondiente al significado más habitual, pero en la primera línea ya nos indica que existen otros significados que se pueden consultar en la entrada de desambiguación. Si accedemos a esta entrada veremos muchos otros significados que puede tener esta palabra:

Gato (animal)[

- **Gato de Schrödinger** (*paradoja de Schrödinger*), gato empleado como ejemplo en un experimento imaginario.
- **Gato negro**, gato de color negro, relacionado con la superstición.
- **Signo del Gato**, en el horóscopo chino, denominación de los vietnamitas al llamado signo del conejo.

Plantas

- El nombre común de la planta *Calluna vulgaris*.
- El nombre común de la planta *Corylus avellana*.
- El nombre común de la planta *Antirrhinum majus*.
- El nombre común de la planta *Muscari neglectum*.

Motes de personas

- Carlos Alberto Dumas ("Gato" Dumas), cocinero argentino.
- Eduardo Alquinta ("Gato" Alquinta), músico chileno.
- Gastón *"el gato"* Gaudio, tenista argentino.
- Leandro *"Gato"* Barbieri, saxofonista argentino de jazz.
- Xavier *"Gato"* Pérez, músico de origen argentino establecido en Cataluña (España).

Otros

- **Almohadilla**, símbolo #
- **Gato**, una danza del folclore argentino y uruguayo.
- **Gato**, clase de submarinos de la Armada estadounidense utilizados en la Segunda Guerra Mundial.
- **Gato**, habitante de Madrid
- **Gato mecánico**, máquina empleada para la elevación de cargas que puede ser mecánica o hidráulica.
- **Gato de cable**, un tipo de gato mecánico hidráulico.
- **Gato fedorento**, programa humorístico portugués.
- **Gato y ratón**, el juego infantil.
- **Los Gatos**, banda argentina de música rock.
- **Gatos de Madrid**, franquicia madrileña de rugby.
- **Gato Rojo**, nombre que se le da a un tostel en Costa Rica.

3.3. La dBpedia

El contenido de la Wikipedia puede ser de gran interés para diversas aplicaciones de Procesamiento del Lenguaje Natural. El contenido de la Wikipedia es básicamente textual (aunque contiene imágenes y otros elementos multimedia) y la estructura de los artículos es libre. Existen

otros elementos que pueden ser de gran ayuda para obtener información de la Wikipedia, como pueden ser los enlaces con otros artículos y los enlaces interlingüísticos. No obstante, es deseable disponer de esta misma información de una manera que sea más efectiva desde el punto de vista del procesamiento informático. La dBpedia es uno de estos intentos de estructurar y ofrecer la información de la Wikipedia de una manera más fácil de procesar.

En la versión inglesa de esta base de datos se describen 3,77 millones de entidades, entre ellas al menos 764 mil personas, 563 mil lugares, 112 mil álbumes de música, 72 mil películas y 18 mil videojuegos.

Veamos la información de *Gato* en la dBpedia española:

Property	Value
dbpedia-owl:wikiPageID	1003999 (xsd:integer)
dbpedia-owl:wikiPageRedirects	dbpedia-es:Felis_silvestris_catus
dbpedia-owl:wikiPageRevisionID	31041550 (xsd:integer)
dbpedia-owl:wikiPageWikiLink	dbpedia-es:Categoría:Nombres_vernáculos_de_mamíferos dbpedia-es:Felis_silvestris_catus
dcterms:subject	dbpedia-es:Categoría:Nombres_vernáculos_de_mamíferos
rdf:type	skos:Concept
rdfs:label	Gato
http://www.w3.org/ns/prov#wasDerivedFrom	wiki-es:Gato?oldid=31041550
foaf:isPrimaryTopicOf	wiki-es:Gato
is dbpedia-owl:ethnicity of	dbpedia-es:Garfield_(personaje) dbpedia-es:Figaro_(Disney) dbpedia-es:Azrael_(Los_Pitufos)
is dbpedia-owl:wikiPageWikiLink of	dbpedia-es:Gatuna dbpedia-es:Morro dbpedia-es:Quarantine_2:_Terminal …. dbpedia-es:Criador
is prop-es:raza of	dbpedia-es:Figaro_(Disney) dbpedia-es:Azrael_(Los_Pitufos)
is prop-es:tipo of	dbpedia-es:Chantilly-Tiffany dbpedia-es:Gato_Skookum
is skos:subject of	dbpedia-es:Categoría:Gatos

is foaf:primaryTopic of	wiki-es:Gato

3.4. WordNet

WordNet es una base de datos de conocimiento léxico del inglés en la que los sustantivos, verbos, adjetivos y adverbios se organizan en conjuntos de sinónimos que reciben el nombre de *synsets*. Cada *synset* representa un concepto lexicalizado en inglés, y se conecta con los otros synsets mediante relaciones semánticas. Las principales relaciones que ofrece el WordNet son la *hiponimia* o relación de especificidad entre una palabra (el hipónimo) y otra de significado más genérico (el hiperónimo), la *antonimia* o relación entre palabras que tienen un significado directamente opuesto, la meronimia o la relación entre una parte y un todo, y la *troponímia* o implicación léxica, una relación que se da entre verbos y que se puede considerar en cierto modo equivalente a la relación de hiponimia para los sustantivos.

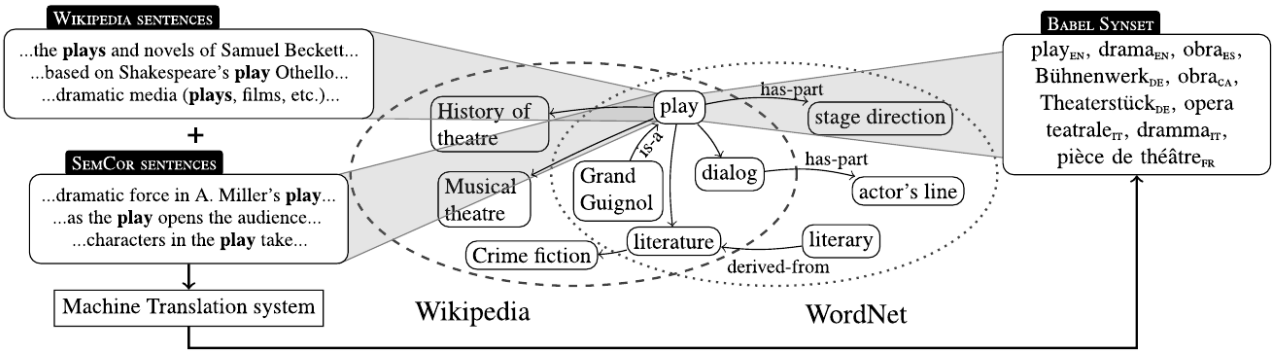
Por ejemplo, el *synset* 02121620-n del WordNet 3.0 español tiene asociada una *variant* o palabra: *gato* y tambié tiene asociada una glosa o definición: *mamífero felino generalmente tienen piel gruesa y suave sin capacidad de rugir: gatos domésticos, gatos monteses.* Este *synset* tiene dos hipónimos: el 02121808-n (*gato doméstico*) y el 02124623-n (*gato salvaje*). A su vez, tiene un hiperónimo, el 02120997-n (*felino*).

Por otro lado el *synset* 035889851-n (*gato*) (en el sentido de herramienta) tiene un hiperónimo, el 04451818-n (*instrumento, utensilio, herramienta*).

Así pues, WordNet nos proporciona una información muy útil sobre los sentidos de las palabras y sus relaciones con otros conceptos y palabra.

3.5. Babelnet

BabelNet es tanto un diccionario enciclopédico multilingüe con una cobertura lexicográfica y enciclopédica de los términos, como una ontología que conecta los conceptos y las palabras en una extensa red de relaciones semánticas construida a partir de más de 9 millones de entradas, denominadas *synsets Babel.* Cada *synset Babel* representa un significado dado y contiene todos los sinónimos que expresan este significado en una gran cantidad de lenguas.



Las versiones iniciales de BabelNet se construyeron combinando el conocimiento lexicográfico de WordNet y el conocimiento enciclopédico de la Wikipedia.

BabelNet 2.0 cubre 50 lenguas y se ha obtenido mediante la integración automática de los siguientes recursos:

- WordNet (version 3.0),
- Open Multilingual WordNet, una coleccion de WordNets disponibles en diferentes lenguas
- Wikipedia
- OmegaWiki, un diccionario colaborativo multilingüe

A continuación podemos observar la entrada de Babelnet para *gato* en el sentido de animal:

```
bn:00016606n      WN    02121620n   CONCEPT     111   WN:EN:cat   WN:EN:true_cat
      WN:ZH:喵     WN:DA:kat   WN:DA:mis   WN:DA:missekat    WN:FI:kissa
WN:FI:kissaeläin  WN:HE:חָתוּל  WN:IT:gatto WN:JA:にゃんにゃん  WN:JA:キャット
WN:JA:ネコ   WN:JA:猫     WN:CA:gata  WN:CA:gat   WN:EU:katu  WN:GA:gato
WN:ES:felis silvestris catus   WN:ES:gata  WN:ES:gato  WN:ID:kucing
WN:MS:kucing      WN:NO:katt  WN:NO:pus   WN:NO:pusekatt
WNTR:SV:sann_katt:1.0_0_0      WNTR:HE:0_0_1.0:חתול_אמיתי
WNTR:GA:cat_fíor:1.0_0_0       WNTR:ID:benar_kucing:1.0_0_0   WNTR:ZH:真正的猫:
1.0_0_0     WNTR:DA:sand_kat:1.0_0_0       WNTR:IT:vero_gatto:1.0_0_0     WNTR:FA:
0_0_1.0:گربه_واقعی       WNTR:SL:res_mačka:1.0_0_0
WNTR:CA:gat_veritable:1.0_0_0  WNTR:EL:αληθινή_γάτα:1.0_0_0
WNTR:FR:vrai_chat:1.0_0_0      WNTR:SK:pravda_mačka:1.0_0_0
WNTR:PL:prawdziwy_kot:1.0_0_0  WNTR:RO:adevărat_pisica:1.0_0_0      WNTR:JA:真の猫:
1.0_0_0     WNTR:ES:gato_verdadero:1.0_0_0       WNTR:IS:sannur_köttur:1.0_0_0
WNTR:BG:вярно_котка:1.0_0_0    WNTR:CS:pravda_kočka:1.0_0_0
WNTR:AF:ware_kat:1.0_0_0       WNTR:NL:echt_cat:1.0_0_0
WNTR:LT:tiesa_katė:1.0_0_0     WNTR:UK:істинна_кішка:1.0_0_0
WNTR:SW:kweli_paka:1.0_0_0     WNTR:TR:gerçek_kedi:1.0_0_0
WNTR:ET:tõsi_kass:1.0_0_0      WNTR:DE:wahre_cat:1.0_0_0
WNTR:SR:истина_мачка:1.0_0_0   WNTR:HU:true_macska:1.0_0_0    WNTR:AR::صحيح_القط
0_0_1.0     WNTR:MS:kucing_benar:1.0_0_0   WNTR:RU:истинная_кошка:1.0_0_0
WNTR:FI:true_kissa:1.0_0_0     WNTR:PT:gato_verdadeiro:1.0_0_0
WNTR:HR:istina_mačka:1.0_0_0   WNTR:CY:cat_gwir:1.0_0_0
WNTR:MT:vera_qattus:1.0_0_0
WNTR:LV:taisnība_kaķis:1.0_0_0WNTR:HI:सच_बिल्ली:1.0_0_0  WNTR:NO:sanne_katt:1.0_0_0
      WNTR:TL:totoo_cat:1.0_0_0      WNTR:KO:사실_고양이:1.0_0_0
WNTR:SQ:cat_e_vërtetë:1.0_0_0  WNTR:VI:đúng_con_mèo:1.0_0_0
WNTR:SK:mačka:0.66667_8_12     WNTR:SQ:mace:0.50000_6_12      WNTR:ZH:猫:
```

0.66667_8_12 WNTR:TR:kedi:0.83333_10_12 WNTR:BG:котка:0.66667_8_12
WNTR:LT:katė:0.83333_10_12 WNTR:LV:kaķis:0.83333_10_12
WNTR:CY:gath:0.41667_5_12 WNTR:CA:gat:0.75000_9_12
WNTR:SV:katt:0.75000_9_12 WNTR:ET:kass:0.75000_9_12
WNTR:HI:बिल्ली:0.75000_9_12 WNTR:TL:pusa:0.83333_10_12
WNTR:FR:chat:0.83333_10_12 WNTR:FA:12_7_0.58333:گربه
WNTR:SR:мачка:0.58333_7_12 WNTR:VI:mèo:0.66667_8_12
WNTR:FI:kissa:0.75000_9_12 WNTR:CS:kočka:0.66667_8_12
WNTR:SL:mačka:0.66667_8_12 WNTR:SW:paka:0.58333_7_12
WNTR:PT:gato:0.75000_9_12 WNTR:AF:kat:0.83333_10_12 WNTR:HE:0.41667:חתול
12_5_ WNTR:ID:kucing:0.75000_9_12 WNTR:MS:kucing:0.83333_10_12
WNTR:DA:kat:0.75000_9_12 WNTR:MT:qattus:0.50000_6_12
WNTR:DE:katze:0.75000_9_12 WNTR:IS:köttur:0.66667_8_12
WNTR:PL:kot:0.58333_7_12 WNTR:NL:kat:0.91667_11_12
WNTR:HR:mačka:0.58333_7_12 WNTR:EL:γάτα:0.75000_9_12
WNTR:ES:gato:0.75000_9_12 WNTR:NO:katt:0.66667_8_12
WNTR:GA:cat:0.75000_9_12 WNTR:RO:pisică:0.41667_5_12
WNTR:HU:macska:0.83333_10_12 WNTR:IT:gatto:0.83333_10_12 85
1_26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,
52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70
0_71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,
97,98,99,100,101,102,103,104,105,106,107,108,109,110 0 0 14 @
bn:00033982n ~ bn:00028153n ~ bn:00081209n gmono bn:00102882a
gmono bn:00028153n gmono bn:00114778r gmono bn:00053079n gdis
bn:00102882a gdis bn:00028153n gdis bn:00114778r gdis bn:00053079n gdis
bn:00083380v gdis bn:00081209n gdis bn:00036883n

3.6. SUMO (Suggested Upper Merged Ontology)

SUMO y sus ontologías por dominio forman la ontología formal de mayor tamaño accesible públicamente. SUMO está relacionada en su totalidad con WordNet. Vemos que la información de la palabra *cat* en inglés está íntimamente relacionada con la información de WordNet:

102121620 feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats.

- SUMO Mappings: Feline (subsuming mapping)

102127808 any of several large cats typically able to roar and living in the wild.

- SUMO Mappings: Feline (subsuming mapping)

102983507 a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work.

- SUMO Mappings: LandVehicle (subsuming mapping)

102985606 a whip with nine knotted cords; "British sailors feared the cat".

- SUMO Mappings: Whip (subsuming mapping)

103608870 the leaves of the shrub Catha edulis which are chewed like tobacco or used to make tea; has the effect of a euphoric stimulant; "in Yemen kat is used daily by 85% of adults".

- SUMO Mappings: Stimulant (subsuming mapping)

109900153 a spiteful woman gossip; "what a cat she is!".

- SUMO Mappings: Female (subsuming mapping) Human (subsuming mapping)

110153414 an informal term for a youth or man; "a nice guy"; "the guy's only doing it for some doll".

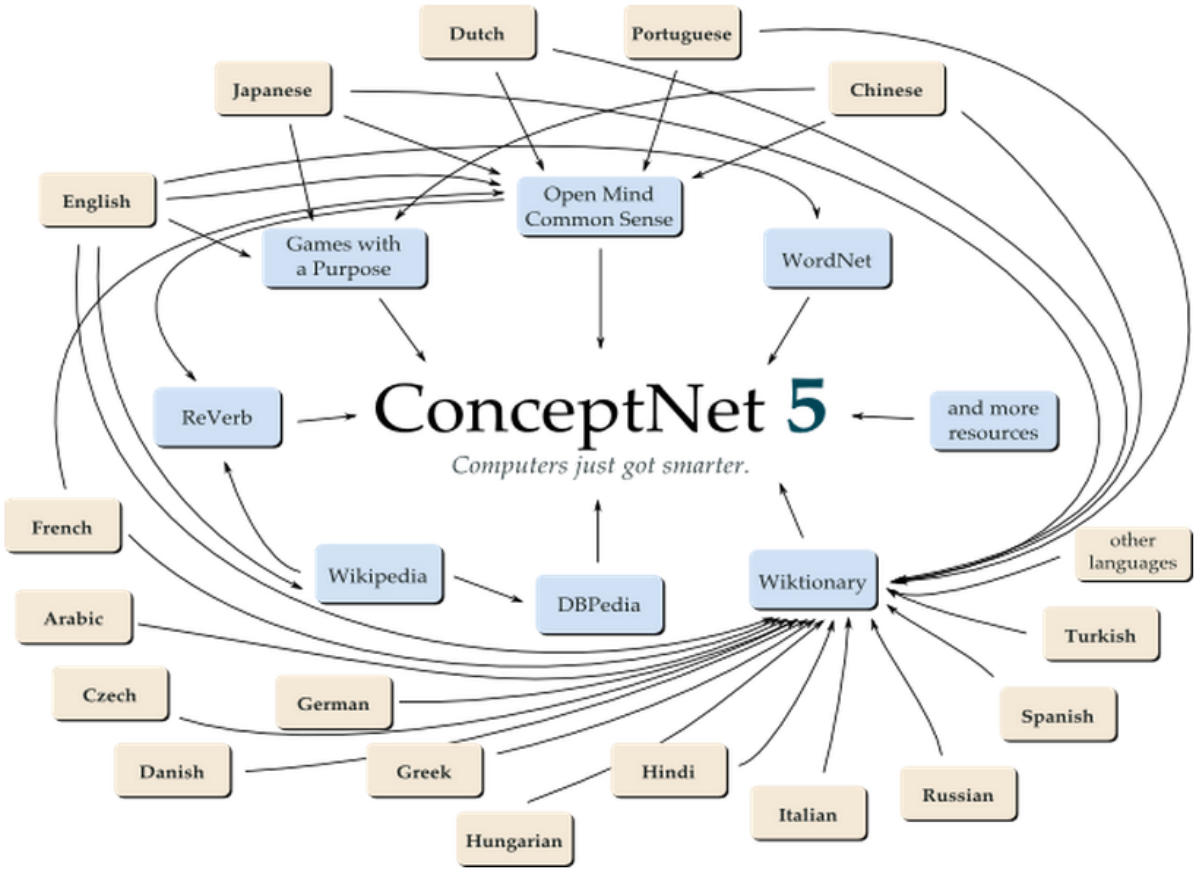
- SUMO Mappings: Male (subsuming mapping) Human (subsuming mapping)

SUMO ofrece ontologías específicas de los siguientes ámbitos:

comunicaciones
países y regiones
computación distribuida e interfaces de usuario
economía
finanzas
componentes de automoción e ingeniería
alimentos
restaurantes
deportes
catálogos comerciales y hoteles
geografía
gobiernos y justicia
taxonomía de las lenguas
medios de comunicación
militar
sistema de clasificación industrial norteamericana
las personas y sus emociones
elementos físicos
asuntos transnacionales
medios de transporte
virus
aeropuertos del mundo
armas de destrucción masiva

3.7. ConceptNet

ConceptNet es una red semántica que contiene un montón de cosas que los ordenadores tendrían que saber sobre el mundo, especialmente para la comprensión de textos escritos por personas. Está construido a partir de nodos que representan conceptos en la forma de palabras o pequeñas frases en lenguaje natural. Estos nodos están etiquetados mediante relaciones con otros nodos.



ConceptNet contiene conocimiento básico del día a día:

learn —*MotivatedByGoal*→ knowledge

You would learn because you want knowledge.

así como conocimiento cultural:

saxophone — *UsedFor* → jazz

A saxophone is used for jazz.

y conocimiento científico:

semantic role — *HasContext* → linguistics

"Semantic role" is a term in linguistics.

Y además contiene también relaciones en otras lenguas:

本 — *MadeOf* → 紙

本は紙でできている。 (A book is made of paper.)

3.8. FrameNet

FrameNet es una base de datos léxica del inglés que se ha creado mediante la anotación del uso real de las palabras en los extos. FrameNet se basa en una teoría sobre el significado denominada *frame semantics* (semántica de marcos). La idea básica de esta teoría es que el significado de la mayoría de palabras se puede representar mediante un *semantic frame* (marco semántico): una descripción de un tipo de evento, relaciones o entidades y de sus participantes. Existe también una versión de FrameNet para el español. Veamos un ejemplo de Carlos Subirats Rüggember en su artículo *La integración de la semántica de marcos y la semántica de simulación en la aplicación al procesamiento semántico automático del español:*

[El juez $_{\text{SPEAKER}}$] [le $_{\text{ADDRESSEE}}$] hizo la promesa [de que atendería su petición $_{\text{MESSAGE}}$]

El grupo nominal *el juez* es la realización del rol semántico de SPEAKER, el pronombre clítico *le* desempeña el rol de ADDRESSEE y la oración subordinada preposicional *de que atendería su petición* es el MESSAGE, mediante el que el SPEAKER le manifiesta a su ADDRESSEE que se compromete a llevar a cabo una acción determinada.

3.9. Otras fuentes de conocimiento

Existen muchas otras fuentes de conocimiento que se pueden aplicar a tareas de procesamiento del lenguaje natural. Algunas de estas fuentes son:

- VerbNet: es un diccionario de verbos del inglés con enlaces a otras fuentes como WordNet, Xtag y FrameNet. En VerbNet los verbos se organizan en clases que se caracterizan por los roles temáticos, las restricciones de selección de sus argumentos y los *frames* (marcos) que consisten

en una descripción sintáctica y en sus predicados semánticos con una función temporal.

- GeoNames: es una base de datos geográficos que cubre todos los países y que contiene más de 8 millones de nombres de lugares.
- KnowNet: es una base de datos de conocimiento que conecta grandes conjuntos de conceptos relacionados semánticamente.

4. Conclusiones

En este tema hemos visto dos aproximaciones diferenciadas para la resolución de tareas lingüísticas: las basadas en corpus y las basadas en conocimiento. Aunque existe una diferencia de base entre estas aproximaciones, no siempre se puede marcar una frontera clara entre una y otra, ya que la utilización de un corpus ya supone en sí una fuente de conocimiento y si etiquetamos con muchísima información un corpus lingüístico, su uso se aproxima más a las aproximaciones basadas en conocimiento.