



Visualización del significado de verbos en español: Estudio y experimentación escalable

Ana García Serrano, Jorge Horcas Pulido, Fernando López Ostenero
ETSI Informática – UNED, C/ Juan del Rosal 16, 28040 Madrid, España.
{agarcia, flopez}@lsi.uned.es

Abstract In this paper it is presented a study on verbs in Spanish and it's potential to display images from the Wikipedia (Wikimedia). It is designed and developed an Information Retrieval model based on linguistic structures of verbs and an environment that allows all subsequent scaling Spanish verbs. Adesse and EuroWordNet are the linguistic resources selected to bring the theoretical basis of the work. In the absence of an adequate corpus with relevant judgments to the problem, it has been recorded by the second author a subset of visual verbs sufficiently representative and enable to further work on this issue. Finally conclusions about visual verbs as well as the obtained results are provided.

Resumen En este trabajo se presenta el estudio realizado sobre verbos en español y su potencial visualización con imágenes procedentes de la Wikipedia (Wikimedia). Se diseña e implementa un entorno de trabajo que incorpora un modelo de recuperación basado en las estructuras lingüísticas de los verbos, y que permite el escalado posterior a otros verbos visuales. Se utilizan como fuentes lingüísticas externas los recursos ADESSE y EuroWordnet. Ante la inexistencia de un corpus y los juicios de relevancia adecuados al problema planteado, se ha identificado tanto un subconjunto de verbos visuales suficientemente representativo como una colección de documentos (anotados por el segundo autor), que permitirán probar el modelo planteado y trabajos posteriores. Finalmente se aportan conclusiones sobre la visualización de verbos y sobre los resultados obtenidos en la experimentación.

Keywords: Information retrieval, Linguistic arguments of verbs, Lexicon, Semantics.

Palabras clave: Recuperación de información, Argumentos lingüísticos de los verbos, Léxico, Semántica.

1 Introducción

El punto de partida de este trabajo surge de la siguiente consideración: el verbo es una unidad léxica que posee significado, ¿es posible asociar a este significado una representación visual (imagen/es)? La imagen se correspondería con el mismo significado del verbo y por lo tanto trasladaría la semántica del verbo a una semántica visual (Magariños 2001). Existen diferentes semánticas visuales (Dewes, 1985) puesto que no es lo mismo interpretar una imagen real, figurativa o abstracta. Hay tipos de verbos para los que es posible asociar una imagen con un criterio de mayor universalidad que en otros tipos, por ejemplo “correr” tiene una representación visual real, definida y unívoca de la acción pero para “pensar” o “sufrir”, con un marco semántico más abstracto, la representación visual es ambigua y difícil de determinar.

ISSN: 1988-3064 (on-line)

©IBERAMIA and the authors

La aproximación planteada en este trabajo consiste en identificar el conocimiento del verbo necesario para utilizarlo en un proceso de recuperación de imágenes en el que, dado un verbo y un significado, se obtenga una colección de imágenes que se correspondan con ese significado. En nuestro conocimiento no existe ningún recurso como el que se presenta, salvo para nombres, Imagenet (que se describe brevemente en el apartado 4). Entonces, para la visualización verbal ¿Qué tipo de información es la necesaria? Esta información de naturaleza gramatical y semántica, ¿cómo se compara con la anotación (texto) de una imagen?



Figura 1: Imágenes relacionadas con el verbo “Afeitarse”

Para responder a estas preguntas, las tareas planteadas y que se presentan en lo que sigue, son:

1. Identificar la información asociada al verbo relacionada con argumentos semánticos, construcciones sintácticas y nominalizaciones.
2. Seleccionar los recursos lingüísticos existentes que permitan extraer la citada información.
3. Diseñar las funcionalidades de un entorno que permita automatizar algunas tareas integrando los recursos seleccionados.
4. Experimentar con un conjunto de verbos objetivo y un corpus de documentos (páginas de Wikipedia con imágenes de Wikimedia) para finalmente,
5. Evaluar los resultados.

2 Marco lingüístico

El objeto de estudio en este trabajo es el significado del verbo (para su visualización) y concretamente el significado sistemático, obviando el resto de informaciones del significado que hacen referencia a la intencionalidad del emisor o aspectos contextuales. Está basado en la semántica lexicológica, y solo se consideran las raíces o lemas de los verbos.

En lingüística hay dos posiciones principales: la estructuralista y la funcionalista. La lingüística estructuralista considera el lenguaje como un sistema formal constituido por elementos diferenciados y su objeto es el estudio de las relaciones entre los elementos que forman parte del sistema lingüístico (Sausurre, 1980). La lingüística funcionalista considera el lenguaje como una investigación de las funciones desempeñadas por los elementos, las clases y los mecanismos que intervienen en ella (Halliday, 2004). Es la que se utiliza en este trabajo.

Los significados verbales exigen la presencia de roles semánticos (elementos lingüísticos que acompañan al verbo) necesarios para tener un significado completo, desarrollando una estructura sintagmática verbal. Por ejemplo, el verbo “comer” tiene como significado «la acción por la que se desarrolla la ingestión de alimentos» que implica necesariamente la presencia de un rol que sea «un sujeto que realiza la acción» y de otro rol que sea «objeto sobre el que se realiza la acción». Posteriormente estos roles funcionarán sintácticamente dentro de la estructura del sintagma verbal como “sujeto” y como “complemento directo”. Luego las necesidades semánticas del verbo son las generadas por una estructura sintagmática con organización funcional. Es la hipótesis principal del planteamiento presentado en este trabajo.

Desde un punto de vista sintáctico, la función principal (no es única) del verbo es el núcleo del sintagma verbal (Alarcos, 1978). El verbo genera dentro del sintagma verbal la presencia de determinados complementos que dependen de él y desarrollan un papel sintáctico. Pueden ser *argumentales*, cuando son exigidos por el verbo para precisar y complementar su significación, o *adjuntos* si son elementos opcionales que aportan otra información sobre la acción verbal. Normalmente los argumentos verbales exigidos por el verbo (valencia), lo son porque el verbo necesita expresar un objeto o tema sobre el que recae la expresión verbal o un beneficiario de la acción verbal (García Hernández, 1980).

Los complementos y la organización funcional del sintagma verbal en este estudio, se corresponde con verbos predicativos, es decir a verbos que “dicen algo del sujeto”. No obstante existen otro tipo de sintagmas verbales formados por verbos copulativos, como ser o estar, que son verbos vacíos de significación por lo que quedan descartados de este trabajo.

Con respecto al verbo como unidad semántica, se considera que el verbo está formado por un significado a nivel léxico autónomo e independiente (lexema), que puede no ser único, y cada significado desarrolla una serie de esquemas sintáctico – semánticos que expresan las necesidades básicas de la semántica del verbo (Albertuz, 2007). Se dice que una oración está completa (o saturada) cuando se han satisfecho sus requisitos argumentales.

Por otra parte, un significado del verbo puede aparecer en diferentes construcciones sintácticas. Puede ocurrir (diátesis) que en una determinada acción distintos roles temáticos desempeñen una misma posición o función sintáctica. Por ejemplo en una misma acción pueden desempeñar la función de sujeto un agente, un tema o un instrumento. También se puede dar el caso en el que en una determinada acción, un mismo rol temático desempeñe distintas posiciones sintácticas. Por ejemplo, el verbo advertir con el significado de *avisar* admite como esquemas sintácticos < SUJ CD CI> (Te lo advierto) o bien < SUJ CI CPRE> (Le advirtió del peligro)". No obstante, se debe tener presente que el significado del verbo condiciona la estructura del verbo pero no la determina. Un mismo esquema puede ser aplicado a verbos de diferentes dominios, como por ejemplo, < SUJ CD CI> aparece con verbos como *dar* «transferencia», *decir* «comunicación», *enseñar* «conocimiento, percepción», *hacer* «creación», etc.

En una visión estrictamente lexicalista, los diferentes patrones o esquemas sintácticos en los que un verbo puede participar estarán condicionados por la semántica léxica del verbo pero no estarán determinados unívocamente por ésta. En (Halliday, 2004) se definen marcos conceptuales generales que categorizan el significado verbal, y se adscriben los verbos a los marcos relacionándolos con los argumentos que desarrollan. En este trabajo se experimenta con verbos de cuatro subclases de marcos conceptuales.

Se realizó una selección no aleatoria, a partir de la frecuencia del verbo en el corpus del recurso ADESSE (descrito con cierto detalle en el siguiente apartado). Las muestras tienen una frecuencia acumulada de: <Percepción> (61,10%) <Cuidado Personal> (70,10%) y <Meteorología> (57,32). Para la subclase <Actividad> es el 36,4%.

Siguiendo la clasificación de ADESSE los tipos de los verbos seleccionados son:

- *Proceso Mental > Percepción > General*

Solo se tienen en cuenta los de percepción sensorial. Este dominio define una acción entre el individuo perceptor y el objeto percibido, que dependiendo del objeto y del sentido utilizado determinará que estímulo está recibiendo el individuo. Por ejemplo: *Ver, mirar, sentir, oír, leer, escuchar, contemplar, observar, fijar, percibir*.

- *Proceso Material > Cambio > Cuidado Personal*

Esta subclase expresa un dominio en el que un individuo experimenta algún tipo de modificación relacionada con la higiene o el cuidado del aspecto. Por ejemplo: *Vestir, bañar, desnudar, calzar, afeitarse, peinar, cepillar, ducharse, abrigarse, maquillarse*.

- *Proceso Material > Hecho > Meteorología*

Semánticamente define procesos específicos presentes en el medio natural relativos a fenómenos meteorológicos. Por ejemplo: *Llover, amanecer, nevar, anochecer, oscurecer, clarear, lloviznar, atardecer, alborear, escampar*.

- *Proceso Material > Hecho > Actividad*

Se refiere conceptualmente a procesos en los que un individuo realiza una actividad concreta y definida de tipo básicamente material, no mental ni comunicativa. Por ejemplo: *Trabajar, jugar, servir, actuar, bailar, descansar, obedecer, cultivar, operar, colaborar, dormir, llorar, respirar, soplar, toser, parir, orinar, vomitar, sudar*.

3 Recursos lingüísticos verbales

De acuerdo con los objetivos y el planteamiento lingüístico anterior, el principal recurso lingüístico es ADESSE (*Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*) (<http://adesse.uvigo.es/>). Contiene información sintáctico-semántica sobre las construcciones verbales (García Miguel et al, 2005) que permitirá modelar en este trabajo a los verbos como vectores de esquemas argumentales.

Como recurso lingüístico auxiliar se seleccionó EuroWordnet para la nominalización de los verbos, ya que mantiene relaciones a nivel de significados entre verbos y nombres. EuroWordnet y ADESSE facilitan su interoperabilidad porque mantienen un nivel de clasificación verbal similar.

3.1 Breve descripción de ADESSE

ADESSE (García-Miguel y Vaamonde, 2005) incorpora conocimiento enciclopédico del verbo con ejemplos de uso de los verbos (como *Framenet* para verbos en inglés <https://framenet.icsi.berkeley.edu/fndrupal/>). Permite el estudio empírico de las interacciones entre significado verbal y sus construcciones ya que incluye (a) los esquemas sintácticos admitidos por cada verbo, (b) una clasificación semántica de los sentidos verbales, (c) una identificación de los roles participantes y (d) datos sobre la frecuencia en el corpus de cada esquema sintáctico-semántico.

ADESSE es una ampliación de la *Base de Datos Sintácticos del Español Actual* (BDS), que contiene el análisis sintáctico manual de casi 159.000 cláusulas del corpus ARTHUS, de 1.5 millones de palabras, con textos de diferentes géneros y procedencia. En la BDS, a partir del registro de cada oración, se relacionan tres tablas: una de *lemas verbales* con 3.436 verbos diferentes, otra de *esquemas sintácticos* combinando verbo, voz y funciones sintácticas y otra de *subesquemas sintácticos*. Con esta información se constata empíricamente qué construcciones son las usadas con cada verbo y la frecuencia de cada construcción. ADESSE añade a la BDS, nuevos campos de información semántica, relativa a sentidos verbales, clases de verbos y roles semánticos, como se observa en la figura 2.

[-] Clasificación semántica y potencial valencial			
Tipo de proceso:		Percepción	
Argumentos:		Frecuencia	
A1	PCTR	PERCEPTOR	610 (98.5 %)
A2	PCDO	PERCIBIDO	596 (96.3 %)
A3	POSR	Poseedor	8 (1.3 %)
A4	BEN	Beneficiario	9 (1.5 %)

Perfil combinatorio >>

[-] Realizaciones valenciales (Esquemas sintáctico-semánticos):			
Voz	Argumentos semánticos y Funciones sintácticas		N_ejemplos
BUSCAR _{act}	A1:PCTR = SUJ	A2:PCDO = ODIR	544 >
BUSCAR _{act}	A1:PCTR = SUJ		23 >
BUSCAR _{act}	A1:PCTR = SUJ	A2:PCDO = ODIR	8 >
BUSCAR _{act}	A1:PCTR = SUJ	A2:PCDO = ODIR	8 >
BUSCARSE _{uf}	A1:PCTR = SUJ	A2:PCDO = Refl	4 >
BUSCARSE _{uf}	A1:PCTR = SUJ	A2:PCDO = ODIR	1 >

Figura 2: Esquemas en ADESSE

3.2 Los verbos en Wordnet y EWN

Wordnet (<http://wordnet.princeton.edu/>), es una base de datos léxica estructurada sobre la sinonimia (*synsets*) y sus relaciones semánticas (Miller, 1995). Incorpora cuatro categorías gramaticales, *nombres*, *verbos*, *adjetivos* y *adverbios*. Para los verbos establece una categorización en función de su aplicación o significado y la organización por *synsets* se enriquece con relaciones que se establecen entre los propios verbos.

Si el principio de la herencia léxica subyace en las relaciones semánticas entre nombres, la *implicación léxica* es la base de las diferentes relaciones que organizan los verbos (roncar implica dormir). La implicación léxica se da en un único sentido. Una relación propia entre verbos es la *troponimia* (considerada un tipo de relación de implicación léxica). Relaciona tropónimos con su concepto más general a través de varias dimensiones semánticas. Los tropónimos de los verbos de comunicación a menudo incluyen la intención (confesar, examinar) del comunicante o el medio de comunicación (telefonar). Además, existe otra relación entre los verbos, la relación causal (también considerada un tipo de relación de implicación léxica) que incluye dos conceptos, causal (*dar*) y resultante (*tener*). Los verbos causales tienen el sentido de causa de *ser/llegar* a *ser/ocurrido/tener* o causa de *hacer*. La relación causal es unidireccional. Wordnet reconoce únicamente pares causativo - resultativo lexicalizados: *mostrar-ver* (mostrar hace/causa ver), *matar-morir*. Los sinónimos de estos pares heredan la relación causal: enseñar, instruir, educar causan o provocan aprender, adquirir conocimiento.

S: (v) **drive** (operate or control a vehicle) "drive a car or bus"; "Can you drive this four-wheel truck?"

- [direct troponym](#) / [full troponym](#)
- [verb group](#)
- [domain category](#)
- [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
- [derivationally related form](#)
 - [W: \(n\) drive](#) [Related to: [drive](#)] (a road leading up to a private house) "they parked in the driveway"
 - [W: \(n\) drive](#) [Related to: [drive](#)] (a wide scenic road planted with trees) "the riverside drive offers many exciting scenic views"
 - [W: \(n\) drive](#) [Related to: [drive](#)] (a journey in a vehicle (usually an automobile)) "he took the family for a drive in his new car"
 - [W: \(n\) driver](#) [Related to: [drive](#)] (the operator of a motor vehicle)
 - [W: \(n\) driving](#) [Related to: [drive](#)] (the act of controlling and steering the movement of a vehicle or animal)

Figura 3: Ejemplo de *synset* verbal

Wordnet también codifica relaciones de oposición, una forma de "antonimia" entre verbos. Como en los adjetivos, muchas de estas relaciones se basan en transformaciones morfológicas (*aparecer-desaparecer*). Por otro lado, Wordnet incluye para cada *synset* verbal uno o más patrones verbales, que especifican los tipos de sentencias que pueden formarse. Una de las relaciones establecidas entre los *synsets*, de utilidad en este trabajo, es la relación verbo-nombre. Como se observa en la Figura 3 el *synset* de "drive" está relacionado con *synsets* nominales asociados al significado.

En este trabajo se utiliza el Wordnet en español del EuroWordnet (<http://www.illc.uva.nl/EuroWordNet/>) (Vossen, 1997), recurso multilingüe con una organización similar a Wordnet.

4 Recursos multimedia

De los recursos multimedia existentes con imágenes anotadas, se valoraron los siguientes, teniendo en cuenta que se necesitaba seleccionar una fuente con imágenes anotadas con "suficiente" información textual, y no sólo una breve descripción (para su tratamiento lingüístico posterior).

- (a) Wikipedia (Wikimedia) (<http://es.wikipedia.org/>),
- (b) Europeana (<http://www.europeana.eu/>), portal del patrimonio cultural europeo de acceso libre, e
- (c) Imagenet (<http://www.image-net.org/>), una colección de imágenes organizada según la estructura de *synsets* de Wordnet.

Para conocer qué se obtiene actualmente, al buscar un verbo en estos recursos multimedia, de la lista de 500 palabras (<http://www.word-english.org/english500.html>) más usadas (inglés) se seleccionaron algunos verbos y con Wordnet se obtuvieron sus flexiones verbales (términos de la consulta), el conjunto de verbos sinónimos y sus relaciones verbo/nombre para expandir las consultas y se probaron los tres recursos multimedia indicados. En esta experiencia inicial los resultados obtenidos fueron muy pobres y con muy baja relevancia de las imágenes asociadas. En concreto se observó que:

- Wikipedia recupera sobre todo nombres propios de personas, lugares o entradas que lo incluyen, pero rara vez recupera conceptos o categorías.
- En la mayoría de las búsquedas en Wikimedia se recuperan imágenes que incluyen verbos dentro de los nombres de los archivos. Casi en ningún caso recupera categorías de imágenes.
- Europeana, recupera los verbos como elementos de los nombres de los recursos de tipo cultural y no siempre de imágenes.
- Los resultados con consultas expandidas con términos nominales mejoran los resultados, sobre todo si se asocia un synset nominal (que no sólo incluyen a la forma sustantivada).
- Wikipedia en algunos casos redirige a entradas nominales que ya incluyen imágenes relevantes.

El recurso Imagenet fue descartado porque sólo incluye nombres de forma que es imposible realizar una recuperación por verbos, y también Europeana porque no contiene anotaciones textuales adecuadas en las imágenes que contiene. Se seleccionó el recurso Wikipedia donde los artículos se encuentran clasificados en categorías semánticas, evitando la ambigüedad de las entradas con el mismo nombre pero con distinto significado. Además las entradas de la Wikipedia están conectadas con recursos multimedia, como imágenes (Wikimedia), audio o vídeos y permite buscar artículos exigiendo que incluyan imágenes.

5 Entorno para la visualización verbal

El objetivo principal del entorno desarrollado es la gestión de la información de los verbos para la obtención de una colección de imágenes que representen su significado. El entorno se articula sobre tres tareas principales:

1. Gestión de la información lingüística extraída de ADESSE y EuroWordnet (en español),
2. Gestión de información verbal anotada en documentos de Wikipedia y
3. Recuperación de imágenes relevantes al significado de un verbo.

A partir de la siguiente información lingüística relacionada con los verbos, se elabora un modelo de recuperación de información diferente a los que se modelan con las frecuencias de los términos y que permite la visualización verbal:

- (a) los argumentos sintáctico-semánticos de cada significado del verbo,
- (b) las relaciones verbales-nominales y
- (c) la frecuencia de las construcciones en que aparece.

En lo que sigue se describe con cierto detalle el procesamiento de los verbos, la construcción del corpus (conjunto de verbos seleccionados y sus características estudiadas) y el tratamiento de los documentos seleccionados que contienen a los verbos. Finalmente se presenta el modelo de recuperación de información utilizado para la selección de imágenes asociadas a los verbos estudiados.

5.1 Procesamiento sintáctico y semántico de verbos

En primer lugar para los verbos, es necesario identificar los patrones sintáctico-semánticos asociados a cada significado y los valores de frecuencia de aparición en un corpus. Las combinaciones o realizaciones valenciales de aparición de los argumentos semánticos (tabla 1) en el corpus, cuya presencia se corresponde con la presencia de unas funciones sintácticas dentro de la oración, se almacenan con la notación E_n de cada esquema (tabla 2).

Tabla 1: Relación de roles y argumentos

ARG. SEM.	ROL SEMANTICO
A1	Agente que causa la acción verbal
A2	Agente que sufre o experimenta la acción verbal (1)
A3	Agente que sufre o experimenta la acción verbal (2)
A4	Poseedor
A5	Lugar
A6	Manera
A7	Tiempo
A8	Orientación

Tabla 2: Relación de argumentos semánticos y funciones sintácticas del verbo

ESQ.	ARGUMENTOS SEMÁNTICOS	FUNCIONES SINTÁCTICAS
E ₀		
E ₁	A1	SUJ
E ₂	A1 A2	SUJ ODIR
E ₃	A1 A2 A3	SUJ ODIR PVOD
E ₄	A1 A2 A4	SUJ ODIR OIND
E ₅	A1 (A5,A6,A7,A8)	SUJ CC
E ₆	A1 A2 (A5,A6,A7,A8)	SUJ ODIR CC
E ₇	A1 A2 A3 (A5,A6,A7,A8)	SUJ ODIR PVOD CC
E ₈	A1 A2 A4 (A5,A6,A7,A8)	SUJ ODIR OIND CC
E ₉	A2	ODIR / OIND(impersonal)
E ₁₀	A2	SUJ (pasiva)
E ₁₁	A2 A1	SUJ AGT (pasiva)
E ₁₂	A2 (A3,A4,A5,A6,A7,A8)	SUJ (pasiva) + otros complementos

Como se dispone de todos los casos en los que aparece el verbo y del número de casos correspondiente a cada combinación de argumentos semánticos, se puede expresar el valor de frecuencia en forma de probabilidad. La probabilidad para cada esquema semántico- sintáctico es el número de ejemplos en los que aparece ese esquema dividido por el número de ejemplos en los que aparece el verbo (tabla 3).

Tabla 3: Ejemplo de esquemas semántico-sintácticos, frecuencia y probabilidad

ESQUEMA E _i	FREC. (E _i)	P(E _i)
E ₀	31	0,816
E ₁	4	0,105
E ₂	2	0,053
E ₄	1	0,026

Por otra parte, para transformar un término verbal en una serie de formas nominales es necesario establecer una relación nombre-verbo a partir de su significado. Este tipo de relaciones semánticas se extraen de EuroWordnet (en español), seleccionando el *synset* que más se corresponde con cada significado y obteniendo los *synsets* nominales relacionados, de los que finalmente se extraen las formas nominales (para afeitador son afeitado, rasurado y afeitador).

5.2 Generación del corpus

Para cada verbo del corpus, se traslada la consulta expandida con sus formas nominales al motor de búsqueda de la Wikipedia. Como resultado se obtiene un ranking de documentos (enlaces) que son filtrados mediante un crawler para seleccionar diez de ellos como sigue:

- Aquellas páginas que no contienen imágenes son descartadas. Se comprueba que la etiqueta se corresponde con un enlace de Wikimedia para garantizar que no sean imágenes tangenciales, publicitarias o meramente informativas.
- Tienen preferencia en la selección aquellas páginas que incluyen todos los términos de la consulta frente a las que sólo incluyen algunos de ellos. También tienen preferencia las que incluyen los términos como categorías de Wikipedia.
- Tienen preferencia aquellas páginas que incluyen los términos de la consulta en el título o encabezamientos frente a si aparecen en otras partes de la página.

Por ejemplo, los documentos finalmente seleccionados para el verbo “afeitar” son:

501 <http://es.wikipedia.org/wiki/Afeitado>
 502 http://es.wikipedia.org/wiki/Brocha_de_afeitar
 503 http://es.wikipedia.org/wiki/Navaja_de_afeitar
 504 http://es.wikipedia.org/wiki/Maquinilla_de_afeitar
 505 <http://es.wikipedia.org/wiki/Afeitadora>
 506 http://es.wikipedia.org/wiki/Crema_de_afeitar
 507 http://es.wikipedia.org/wiki/Jabon_de_afeitar
 508 <http://es.wikipedia.org/wiki/Barbero>
 509 <http://es.wikipedia.org/wiki/Aftershave>
 510 <http://es.wikipedia.org/wiki/Peluquero>
 501 <http://es.wikipedia.org/wiki/Afeitado>
 502 http://es.wikipedia.org/wiki/Brocha_de_afeitar
 503 http://es.wikipedia.org/wiki/Navaja_de_afeitar
 504 http://es.wikipedia.org/wiki/Maquinilla_de_afeitar
 505 <http://es.wikipedia.org/wiki/Afeitadora>
 506 http://es.wikipedia.org/wiki/Crema_de_afeitar
 507 http://es.wikipedia.org/wiki/Jabon_de_afeitar
 508 <http://es.wikipedia.org/wiki/Barbero>
 509 <http://es.wikipedia.org/wiki/Aftershave>
 510 <http://es.wikipedia.org/wiki/Peluquero>

5.3 Tratamiento lingüístico de los documentos

El tratamiento de todos los documentos obtenidos tras una consulta a Wikipedia, una vez convertidos a texto plano, consiste en su tokenización, eliminación de las Stop-Word (palabras vacías) y stemming, que resulta muy útil en el tratamiento de los verbos (siempre aparecen conjugados o en formas no personales). A continuación se añade estructura lingüística al modelo del documento con:

- Etiquetado léxico-morfológico: Se identifica al verbo de la consulta (la misma raíz verbal) y si el documento no lo incluye, este es descartado.
- Análisis sintáctico-semántico: Se extrae el patrón sintáctico de cada una de las oraciones en que aparece el verbo.

Este proceso automático utiliza recursos disponibles en código libre como son:

- Snowball (www.snowball.tartarus.org),
- el etiquetador SVMTool (www.cs.upc.edu/~nlp/SVMTool) y
- el parser (<http://nlp.stanford.edu/software/lex-parser.shtml>).

De acuerdo con la aproximación propuesta, del etiquetado léxico-morfológico no se almacena ninguna información. En la fase de reconocimiento del verbo de la consulta, se identifica su ausencia (0) o presencia (1). Por cada aparición del verbo se extrae y almacena la oración en la que se incluye el verbo (si en un documento aparece varias veces el mismo verbo, tendrá asociado varios registros). Con la fase de análisis sintáctico-semántico se añaden los argumentos del verbo (tabla 4) y se renombra el esquema (tabla 2). Si no tienen correspondencia con los esquemas semánticos-sintácticos considerados se asigna E₀.

Tabla 4: Ejemplo del índice de documentos

DOC.	VERBO	ORACIÓN	PATRÓN SINTÁCTICO DEL VERBO
01	0		
02	1	Porque CC llovía VB mucho ADV	CC V ADV
02	1	Llueve VB en PP Madrid NNP	V LOC
03	0		

A continuación se calcula, a nivel intradocumental, la frecuencia de aparición de los esquemas del verbo en el documento y a nivel interdocumental, la frecuencia de aparición del esquema del verbo en todos los documentos, como se muestra en la Tabla 5.

Tabla 5: Ejemplo de esquemas semántico-sintácticos y su ponderación

DOC.	E _i	FR INTRAD OC. (E _i)	FR INTER DOC. (E _i)	P(E _i) _{intradocu mental}	P(E _i) _{interdocumental}	P(E _i)
01	E ₀	0	2	0	2/3 = 0,66	0
01	E ₁	0	1	0	1/3 = 0,33	0
02	E ₀	2	2	2/2 = 1	2/3 = 0,66	0,66
02	E ₁	0	1	0/2 = 0	1/3 = 0,33	0
03	E ₀	0	2	0/1 = 0	2/3 = 0,66	0
03	E ₁	1	1	1/1 = 1	1/3 = 0,33	0,33

Para convertir las frecuencias en probabilidades P(E), la frecuencia intradocumental se divide por el número total de esquemas de cada documento y para la interdocumental, se divide cada valor de frecuencia del esquema por el número total de esquemas presentes en la colección.

A su vez, cada uno de los esquemas semántico-sintácticos del documento de la Tabla 5, se puede expresar por el producto de la probabilidad intradocumental por la interdocumental, y así reducir a un único valor cada esquema semántico-sintáctico (P(E_i)).

5.4 Modelo de recuperación de información

Todo modelo de RI vectorial necesita (Salton, 1989) una *ponderación de los componentes en los documentos/consulta*, que se corresponde con la probabilidad de cada esquema semántico-sintáctico en cada documento/ consulta y una forma de *equiparación parcial*, para calcular el grado de similitud entre cada documento y la consulta.

En el modelo propuesto, el valor del espacio n-dimensional dependerá del número de esquemas identificados. Sean D_i = <P_D(E₀)_i, P_D(E₁)_i, P_D(E₂)_i, ..., P_D(E_n)_i> los vectores de los documentos de la colección y C = <P_C(E₀), P_C(E₁), P_C(E₂), ..., P_C(E_n)> el vector de la consulta, se define como medida de similitud:

$$\text{Sim} = \cos(D_i, C) = \frac{P_D(E_0)_i \cdot P_C(E_0) + P_D(E_1)_i \cdot P_C(E_1) + \dots + P_D(E_n)_i \cdot P_C(E_n)}{\sqrt{P_D(E_0)_i^2 + P_D(E_1)_i^2 + P_D(E_n)_i^2} \cdot \sqrt{P_C(E_0)^2 + P_C(E_1)^2 + P_C(E_n)^2}}$$

Figura 4: Fórmula de la similitud

En función de los valores de similitud de cada documento con la consulta se establece además un orden de relevancia (ranking).

Una vez obtenida la lista ordenada de documentos relevantes a la consulta se procede a extraer las imágenes y se automatiza con un crawler que accede al listado de enlaces del ranking, identifica y extrae las imágenes de los documentos (que proceden de la Wikimedia). Para el primer documento del ranking se tiene un conjunto de imágenes, para la segunda otro conjunto y así sucesivamente. Finalmente se obtienen imágenes asociadas al significado del verbo, en el mismo orden que el ranking de documentos (figura 1).

6 Evaluación del experimento

Para evaluar un sistema de recuperación de información siguiendo un enfoque tradicional, se deben crear los juicios de relevancia de las respuestas a determinadas consultas en un corpus concreto. En este trabajo se han construido manualmente, inspeccionando para cada consulta las imágenes de la colección de documentos.

Como criterio de relevancia se comprueba si en la imagen aparecen los argumentos semánticos exigidos por la significación verbal. Por ejemplo, el verbo “nevar” se define como la acción de “caer nieve”, luego cualquier imagen con nieve es válida independientemente de los objetos que sufren la acción verbal o el lugar

donde se desarrolla. Para otros verbos, como en el caso de “leer”, su significado exige la presencia de un agente que realiza la acción verbal (lector) y un objeto que experimenta la acción (lectura) y toda imagen relevante debe contener visualmente estos dos argumentos.

Las medidas utilizadas son la precisión, la exhaustividad (recall) y la medida F que combina a las anteriores (Baeza y Ribeiro, 1999) y se establecen para cada verbo, calculándose una media global. No obstante, estas medidas no dependen del ranking, por lo que además se comprueban las tres primeras posiciones del ranking.

La evaluación muestra un valor medio de precisión del 43%, de exhaustividad del 45% y de medida-F de 42%. No obstante, su desglose presenta un comportamiento muy heterogéneo en función del tipo de verbo y del marco conceptual semántico al que hace referencia.

Por ejemplo, en la categoría *Percepción* el sistema recupera imágenes de 9/10 verbos y las imágenes relevantes suelen aparecer en posiciones alejadas del ranking. En la categoría *Cuidado personal* se recuperan imágenes de 6/10 verbos, que suelen aparecer entre la primera y tercera posición del ranking. Verbos como “calzar” o “cepillar” recuperan imágenes que hacen referencia sólo al objeto (zapatos, cepillos), pero no la relación semántica que muestra la acción entre el sujeto y el objeto.

Para la categoría *Meteorología* se recuperan imágenes de 5/10 verbos pero con una precisión superior al 75% y con la primera imagen correcta. Es en esta categoría donde ADESSE contiene menos ejemplos del corpus y podrían ser menos “recuperables”. Para la categoría *Actividad* se recuperan imágenes de 8/10 verbos y las imágenes correctas están en el primer o segundo puesto del ranking.

7 Conclusiones

En este artículo se ha presentado un estudio teórico sobre la visualización de verbos, se ha desarrollado un modelo de visualización, implementado un entorno de trabajo, y desarrollado un corpus (libre, que se puede solicitar a la primera autora), extraído con la ayuda de los recursos lingüísticos ADESSE, EuroWordnet y Wikipedia (Wikimedia) y con el que se ha evaluado la propuesta.

Es un estudio novedoso para el que existen escasas referencias, y los autores no conocen ningún corpus con juicios de relevancia (imágenes que efectivamente se correspondan con los verbos seleccionados) como el presentado. Aunque el número de verbos estudiado no es muy elevado es suficiente para realizar la experimentación y sobre todo, este trabajo sienta las bases sobre el conocimiento lingüístico de los verbos que es necesario y que hay que formalizar e incorporar en el entorno. De esta manera, se consigue la escalabilidad del mismo, al aportar una solución a la representación de la citada información verbal.

¿Por qué es interesante este tipo de estudio y formalización? ¿En qué tipo de aplicación puede aportar beneficios? La respuesta a ambas preguntas es resultado de la disponibilidad actual de componentes, gestores y documentos multimedia en la web, entornos y los sistemas de todo tipo, con lo que visualizar aquellos verbos que lo admiten con imágenes de la web o de corpus recolectados con algún propósito, es de gran interés por ejemplo para la generación de resúmenes, obtención de imágenes alternativas a un texto en un entorno e-learning etc.

Se ha comprobado que el modelo es factible, se obtienen resultados relevantes y aunque aparentemente no existe relación entre la frecuencia del verbo en el recurso ADESSE y las imágenes encontradas en Wikimedia, se ha de estudiar en un trabajo futuro, junto con otros aspectos relacionados con la visualización de verbos.

8 Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto VOXPOPULI - Análisis eficiente de reputación, propagación y recomendación en entornos de redes sociales (TIN2013-47090-C3-1-P). Ministerio de Economía y Competitividad. Programa estatal de fomento de la investigación científica y técnica de excelencia. 2014-16.

Referencias

- [1] ALARCOS LLORACH, Emilio “*Estudios de gramática funcional del español*”. Madrid: Gredos, 1978.
- [2] ALBERTUZ CARNEIRO, F. “*Sintaxis, semántica y clases de verbos: clasificación verbal en el proyecto ADESSE*”. Universidad de Vigo, 2007

- [3] BAEZA-YATES, R.; RIBEIRO-NETO, B. “*Modern Information Retrieval*” ACM Press. Addison Wesley; 1999.
- [4] DEWES BOTUR, Ada. “*Notas sobre la semiótica visual*” Discurso vol.2, nº6, Univ. Autónoma Metropolitana Xochimilco, 1985.
- [5] GARCÍA-HERNÁNDEZ, B. “*Semántica estructural y lexemática del verbo*”. Reus: Avesta, 1980.
- [6] GARCÍA-MIGUEL, J. M.; COSTAS, L.; MARTINEZ, S. “*Diátesis verbales y esquemas construccionales en el proyecto ADESSE*”. Universidad de Vigo, 2005.
- [7] GARCÍA-MIGUEL, José M; VAAMONDE, Gael “*Semantic role annotation: From verb-specific roles to generalized semantic roles*”. Universidad de Vigo, 2010.
- [8] HALLIDAY, M. “*An Introduction to Functional Grammar*”. Hodder Education Published, 2004.
- [9] MAGARIÑOS DE MORENTIN, Juan. “*La(s) semiótica(s) de la imagen visual*” Cuadernos 17, Univ. Nacional de Jujuy; p. 295-320; 2001 <http://www.centro-de-semiotica.com.ar>
- [10] MILLER, George A. “*WordNet: a lexical database for English*”. Communications of the ACM, vol. 38, no 11, p. 39-41, 1995.
- [11] SAUSSURE, Ferdinand. “*Curso de lingüística general*” 1980.
- [12] SALTON, Gerard. “*Automatic Text Processing: The Transformation, Analysis, and Retrieval*” Addison-Wesley, 1989.
- [13] VOSSSEN, P. “*EuroWordNet: a multilingual database for information retrieval*”. Proc. DELOS Workshop on Cross-language IR. 1997.