Instituto Politécnico Nacional

Centro de Investigación en Computación

Laboratorio de Lenguaje Natural y Procesamiento de Texto

Un método automático para extracción de los patrones de rección en el español basado en los diccionarios explicativos y relaciones léxicas

T E S I S

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA

MCC. Noé Alejandro Castro Sánchez

DIRECTOR DE TESIS

DR. GRIGORI SIDOROV

México, D. F., 2012

Agradecimientos

Aquel de quien proviene la sabiduría, quien de Su boca emana el conocimiento y la inteligencia, quien permitió el inicio de este sueño y su materialización, quien proveyó en todo momento: gracias a Dios y a la autoridad de su Cristo.

Gracias también a todas las personas que de alguna u otra manera les pertenece también este trabajo: a mi familia, por apoyarme siempre de manera incondicional. Al *Dr. Grigori Sidorov*, por brindar una dirección puntual y acertada, por el apoyo y motivación constante. Al *Dr. Alexander Gelbukh*, *Dr. Sergio Suárez*, *Dra. Sofía Galicia Haro*, *Dr. Héctor Jiménez Salazar* que sometieron a discusión y juicio este trabajo, coadyuvando a su constante mejora. A mis amigos, que conocí durante estos estudios de doctorado dentro y fuera de la institución, por acompañarme incondicionalmente y mostrarme el "detrás de cámaras" de este proyecto.

Y en general, a todos aquellos que aportaron ideas, abrieron puertas, resolvieron dudas, inyectaron fuerza y motivación, a todos ustedes les debo también la culminación de esta tesis.

Y finalmente, a las instituciones que me apoyaron: al Centro de Investigación en Computación (CIC), al Instituto Politécnico Nacional (IPN) y al Consejo Nacional de Ciencia y Tecnología (CONACyT).

Índice

Lista de tablas

Lista de ilustraciones

INTRODUCCIÓN

El *lenguaje* constituye una actividad humana que permite a las personas comunicarse y relacionarse con las demás empleando un código lingüístico de signos orales o escritos. Estos signos ordenados y relacionados entre sí constituyen un sistema que se conoce como *lengua* [16].

Los componentes básicos de toda lengua se conforman por:

- un *léxico* o repertorio de palabras a través del cual los hablantes representan su conocimiento del mundo,
- una serie de reglas que establecen las maneras válidas de relacionar y combinar las palabras entre sí, lo que permite moldear la estructura lineal que adoptan (= enunciados).

Estos dos componentes se desenvuelven en planos diferentes: el semántico y el sintáctico respectivamente. El primero, también llamado lexicón mental, alberga los conocimientos generales o el mapa mental que se forman los seres humanos del mundo exterior e interior [25].

El plano sintáctico constituye la estructura de la expresión de los signos, lo que constituye el nivel en el que la comunicación entre personas se efectúa.

El repertorio de palabras o vocabulario existente en una lengua se divide en dos grandes clases: *palabras plenas* (también conocidas como *palabras autónomas*, *palabras autosemánticas*, *palabras lexicales*, etc.), y *palabras gramaticales* (o bien *palabras auxiliares*, *palabras sinsemánticas*, *palabras vacías*, etc.). La diferencia radica en que, en la primera clase, cada palabra designa por sí mismas un concepto léxico autónomo, es decir, hacen referencia al mundo real o abstracto, indicando personas, objetos, acciones, estados, ideas, características, propiedades, etc. Todas estas palabras plenas se clasifican en sustantivos, verbos, adjetivos y adverbios.

En la segunda clase se agrupan todas aquellas palabras que no tienen significado léxico, cumpliendo solamente funciones de tipo estructural, pues se utilizan para establecer relaciones que se dan entre las palabras que ocurren de manera secuencial en el enunciado expresado (relaciones sintagmáticas). Por lo tanto, estas palabras no pueden existir de manera autónoma en una oración, siempre deben ir acompañadas por al menos una palabra plena. Dentro de esta clase de palabras encontramos las conjunciones, las preposiciones, los numerales y los artículos.

Aunado a esta clasificación, es importante considerar que las palabras pueden poseer o carecer de *sentido propio*. Este concepto de *sentido propio* se refiere al hecho de que una palabra sea significativa, es decir, si representa o aporta conocimiento necesario que ayude a determinar el significado de la expresión emitida.

Según esto, podemos considerar que todas las clases de palabras plenas son significativas o tienen sentido propio. Por ejemplo, la palabra "*Juan*" sabemos que indica el nombre de un individuo. Decir "*Juan come*" hace referencia a un individuo que realiza la acción de ingerir alimentos. Ambas palabras, *Juan* y *come*, poseen un significado que aporta conocimiento tanto para designar a un individuo de nombre *Juan*, como a la acción que éste realiza.

Considerando la clase de palabras gramaticales, algunas de éstas pueden o no ser significativas. Es decir, en algunos contextos las palabras gramaticales llegan a operar como restricciones útiles para procesar los significados de nuevas palabras plenas. Por ejemplo, introduciendo el término inventado *gorp*, si éste es referenciado con la preposición *a*, como en "*él va* ***a*** *Gorp*", indicaría que este término se trata de un nombre propio de lugar; en cambio, si lo referenciamos con la preposición *con*, expresaría un nombre propio de persona: *"él va* ***con*** *Gorp"* [23]:

Esto nos induce a pensar que dichas preposiciones en el contexto que presentamos, son importantes para dar uno u otro sentido a la oración. Dichas preposiciones son, pues, significativas en la oración ejemplificada. Para descubrir si una palabra tiene o no sentido propio, se determina si ésta puede ser sustituida o no por otra palabra de su misma clase. Por ejemplo, en la frase:

(1) *Juan puso el libro **sobre** la mesa,*

encontramos que la preposición *sobre* puede ser sustituida por otras preposiciones como "*Juan puso el libro **en** la mesa*" o "*Juan puso el libro **bajo** la mesa*", denotando así que éstas palabras son significativas en este contexto. En contra parte, en la expresión

(2) "*depende **de** ti*",

la preposición *de* no puede ser sustituida por otras de su misma clase, como en:

(3a) *"*depende **a** ti*"

(3b) *"*depende **desde** ti*"

mostrando su carencia de sentido propio y su función meramente estructural.

De todos aquellos grupos de palabras significativas, el verbo es el vocablo por excelencia al ser el elemento indispensable de cualquier oración, pues denota tanto las relaciones entre los objetos del mundo, como el comportamiento que adoptan, los procesos que se suceden entre éstos, los estados en los que se encuentran, etc.

Con esto, podemos decir que el evento manifestado por el verbo debe *ocurrir* o *manifestarse* en *algo*. Ahora, no todos los objetos del mundo participan en los mismos eventos, o no todos los objetos pueden relacionarse de la misma manera con otros.

Esto nos lleva a pensar que los eventos designados por el verbo *seleccionan* los objetos del mundo en los cuales ocurren (sean objetos concretos o abstractos). No podemos aceptar en como válida en un sentido estrictamente literal la oración:

(4) "*la computadora lee libros*"

Pues el verbo *leer* lleva inherente el concepto de comprensión, acción que sólo puede ser ejecutada por seres humanos (y aun así debe cumplir otras tantas condiciones).

De esta manera, podemos apreciar que cada verbo sólo acepta conjuntos específicos de complementos que logren cumplir ciertas restricciones de tipo sintáctico y semántico. Estos

complementos suelen denominarse indistintamente como *argumentos*, *actantes* o *valencias*. Si estas restricciones no son cumplidas por los argumentos el resultado es una expresión agramatical en el sentido sintáctico (3a y 3b) o carente de coherencia (4) (a menos que consideramos un cuento donde objetos inanimados cobran vida y pueden competir cognitivamente con los seres humanos).

En este trabajo de investigación estudiamos relaciones sintactico-semánticas que ocurren entre verbos y argumentos bajo el enfoque teórico de la teoría *Significado↔Texto*, basándonos principalmente en las definiciones de los verbos contenidas en diccionarios explicativos, de las cuales estudiamos el denominado contorno de la definición y considerando las relaciones léxicas de polisemia e inclusión (hiponimia/homonimia), para identificar los argumentos de los verbos tanto a nivel semántico como a nivel sintáctico.

1.1. Planteamiento del problema

La adquisición de información sobre la valencia de los verbos (número de argumentos que requieren, opciones de representación que toman los argumentos a nivel sintáctico, etc.), es un tema ampliamente estudiado en idiomas como el inglés.

La gran mayoría de métodos que se utilizan para adquirir este tipo de información, se basan en el reconocimiento de patrones en Corpus. A través de estos patrones se establece una cierta probabilidad de ocurrencia entre posibles argumentos y verbos.

El éxito de acierto en este tipo de métodos depende en gran medida del tamaño de los Corpus que se utilicen. En el idioma español, los recursos no son tan variados y abundantes, situación que seguramente desalienta el desarrollo de trabajos al respecto en nuestro idioma.

Por esta razón, se debe buscar explotar otras fuentes de información y plantear la utilización de métodos alternativos que permitan aprovechar los recursos con los que actualmente contamos.

1.2. Justificación

El conocimiento de los argumentos que un verbo requiere resulta ser una parte crucial dentro del área de procesamiento del lenguaje natural. Se pueden considerar para beneficio de diversas tareas como la extracción y recuperación de información, generación automática de resúmenes, traducción automática, enseñanza de lenguas extranjeras, etc.

Sobre este tema, existen investigaciones muy avanzadas para la lengua inglesa, principalmente. Para el español, aún cuando es uno de los idiomas más utilizados en el mundo, las investigaciones al respecto son escasas, por lo cual este trabajo cobra especial importancia, contribuyendo en el avance de la investigación del Procesamiento del Lenguaje Natural.

1.3. Hipótesis

- Los diccionarios explicativos proveen información suficiente para identificar en gran medida las valencias verbales.
- Las relaciones léxicas de inclusión (hiponimia/hiperonimia) existentes entre los verbos permite identificar y precisar información desconocida de valencias a partir de la información conocida sobre la valencia de otros verbos.

1.4. Objetivos

Los objetivos que nos hemos planteado para este trabajo de investigación, son los que a continuación listamos:

1.4.1. Objetivo general

- Desarrollar un método para identificar de manera automática patrones de rección del español basado en el análisis de definiciones de verbos contenidas en diccionarios explicativos y en las relaciones de inclusión y de sinonimia establecidas entre las unidades léxicas verbales.

1.4.2. Objetivos específicos

- Procesar alguno de los principales diccionarios explicativos de la lengua española para utilizarlo para extraer verbos y sus respectivas acepciones.

- Generar heurísticas que permitan identificar los elementos que constituyen los artículos lexicográficos que conforman el diccionario.

- Generar algoritmos para identificar y separar los elementos que conforman las definiciones de verbos.

- Implementar los procesos necesarios para identificar de manera automática los verbos que participan en relaciones de inclusión y sinonimia.

- Identificar los actantes de los verbos a partir de la información de sus respectivas definiciones.

- Crear un recurso (Base de Datos de Patrones) para almacenar estos resultados de una manera clara y perfectamente estructurada.

Resumen

La gramática tradicional considera a la oración como una estructura binaria conformada por un sujeto que es el elemento principal, y un predicado que lo complementa. La gramática de dependencias, por otro lado, designa al verbo como el elemento central, teniendo la capacidad, denominada *valencia*, de abrir en torno suyo huecos que deben ser ocupadas por ciertos elementos funcionales, también llamados *argumentos*, que son requeridos para construirse en una oración gramatical e inteligible.

El conocimiento de la valencia verbal resulta ser una parte crucial dentro del área de procesamiento del lenguaje natural, ya que beneficia a diversas tareas como la extracción y recuperación de información, generación automática de resúmenes, traducción automática, enseñanza de lenguas extranjeras, etc. La descripción de valencias encuentra una solución adecuada en la teoría Significado ⇔ Texto, bajo los denominados patrones de rección.

El método más extendido para identificar tanto el número como la naturaleza gramatical de los argumentos requeridos por el verbo, es través del procesamiento de grandes volúmenes de textos utilizando métodos estadísticos. Sin embargo, en este trabajo proponemos el uso de métodos simbólicos para la extracción de las valencias verbales a partir del análisis de definiciones en diccionarios explicativos. A través del procesamiento del contorno de las definiciones y de las relaciones léxicas de inclusión y sinonimia establecidas entre los diferentes verbos dentro del diccionario, fue posible identificar con una precisión del 83% el número de argumentos que algunos de los verbos requieren junto a su especificidad semántica, además de lograr obtener algunas opciones de representación a nivel sintáctico.

Abstract

Traditional Grammar considers the sentence as a two-member structure conformed by a subject which is the main element, and a predicate which complements it. On the other hand, Dependency Grammar designates the verb as the central element, having the ability (named *valence*) to open slots around them which must be filled by certain functional elements (also called *arguments*) required to be constructed in an intelligible and grammatical sentence.

The acknowledgment of verbal valence comes out to be a crucial part within the area of Natural Language Processing, due to its benefit for diverse tasks such as Information Extraction and Retrieval Information, Automatic Text Summarization, Machine Translation, Foreign Languages Teaching, etc. The description of valences finds an appropriate solution in the Meaning ⇔ Text Theory under the so called *Government Patterns*.

The most widespread method to identify the number as well as the grammatical nature of the required arguments by the verb is through the processing of large volumes of texts using statistical methods. However, in this dissertation we propose the use of symbolic methods for the extraction of the verbal valences based on the analysis of definition in Explanatory Dictionaries. By processing the contour of the definitions and lexical inclusion and synonymy relations established among different verbs in the dictionary, it was possible to identify the number of arguments that some verbs require together with their semantic specificity with 83% precision, and achieve some representation options at the syntactic level.

CAPÍTULO 1

MARCO TEÓRICO

1.1. El Lenguaje

El *lenguaje* constituye una actividad humana que permite a las personas comunicarse y relacionarse con las demás empleando un código lingüístico de signos orales, escritos y/o visuales. Estos signos ordenados y relacionados entre sí constituyen un sistema que se conoce como *lengua* [16].

El lenguaje debe considerarse bajo dos perspectivas diferentes: una que lo considera como un fenómeno individual, que existe en la mente de cada persona y que es concerniente al habla; y otro que lo atiende como un fenómeno social, al utilizarse de manera colectiva al margen de los individuos, y que concierne a la lengua.

Ambas son interdependientes, pues el habla permite que una lengua se establezca y evolucione, y la lengua le da sentido y significado a lo expresado por el habla.

El lenguaje humano presenta una complejidad única en comparación con otro tipo de formas de comunicación existentes en el reino animal. Por muy elaborado que llegara a parecer el sistema de comunicación animal, las diferencias con el lenguaje humano son abismales. Por ejemplo, el lenguaje animal no puede superar las barreras de espacio-tiempo, es decir, la información que se comunica no hace referencia a sucesos más allá del tiempo presente ni puede ser percibida fuera del punto donde se transmite (salvo probablemente la danza de las abejas: la serie de movimientos que realizan para indicar a sus pares tanto la dirección como la distancia de una nueva fuente de alimento descubierta). En el lenguaje animal, encontramos además, que la comunicación se produce como reacción a estímulos externos (como el canto especializado que las aves emiten ante la presencia de un depredador).

Todo esto hace al lenguaje humano único, y con claras características inherentes a él, entre las que podemos mencionar:

- Se trata de un sistema complejo.
- Es una herramienta de comunicación, en el amplio sentido de la palabra.
- Es utilizado como representación de conocimiento.
- Es utilizado para discretizar al mundo.

1.1.1. Lingüística

La ciencia que estudia el lenguaje se denomina Lingüística, y dada la complejidad de éste, existen diversas ramas que atienden sus particularidades.

En primera instancia, encontramos dos grandes ramas: lingüística teórica y lingüística aplicada. La lingüística teórica trata sobre el estudio de las lenguas, ideando métodos para describirlas y clasificarlas. Por otro lado, la Lingüística aplicada se enfoca a la resolución de problemas derivados del uso del lenguaje, apoyándose en bases teóricas interdisciplinarias.

Además de estas dos vertientes, se desprende una gran variedad de áreas más especializadas que interactúan unas a otras, como se muestra en la Ilustración 1.1 Disciplinas de la lingüística. [3].

1.1.2. Procesamiento de lenguaje natural

El uso de la tecnología, en particular, las herramientas computacionales, se ha extendido a todas las áreas del conocimiento, y la lingüística no podría ser la excepción. Las técnicas computacionales se utilizan para realizar el análisis automático y la representación de modelos lingüísticos con el propósito de desarrollar aplicaciones y solucionar diversas tareas relacionadas con el lenguaje humano ya sea oral o escrito.

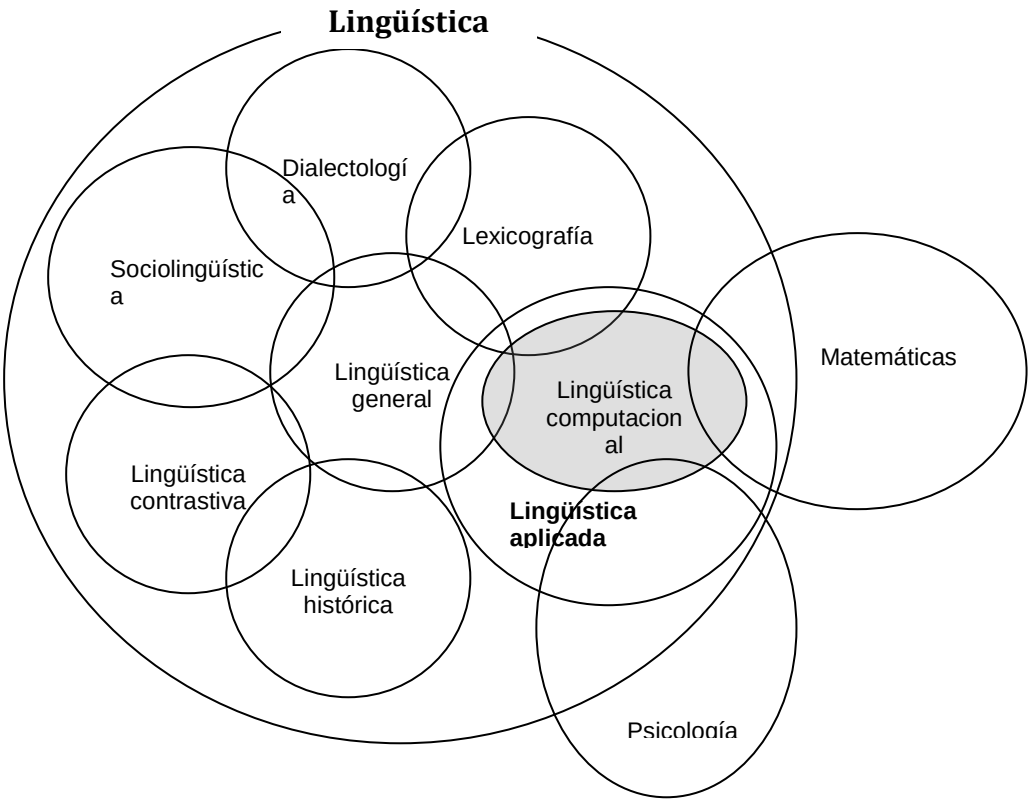


Ilustración 1.1 Disciplinas de la lingüística.

1.1.3. Niveles del lenguaje

Dada la estructura elaborada y compleja que presenta el lenguaje, es necesario abordarla segmentándola en una serie de módulos o niveles ordenados jerárquicamente que dan cuenta de su uso, funcionalidad y desarrollo.

Estos niveles son: fonética/fonología, morfología, sintaxis, semántica, pragmática y discurso. A continuación daremos una revisión breve de cada uno de estos niveles.

Fonética/fonología

Ambas disciplinas se concentran en el estudio de los sonidos de una lengua, pero desde perspectivas diferentes. La fonética se encarga de estudiar el sonido desde el punto de vista físico, atendiendo los aspectos acústicos y articulatorios, mientras que la fonología hace un estudio más abstracto del sonido, considerando las representaciones mentales de éste.

Morfología

Se encarga del estudio de la estructura de las palabras vistas de manera individual, y de los mecanismos que se siguen para la formación de nuevas palabras.

Sintaxis

Mientras la morfología atiende la forma de las palabras, en la sintaxis se observa la función que éstas desempeñan en una oración. Su objetivo es estudiar la estructura de las oraciones y las relaciones que se establecen entre las palabras.

Semántica

Estudia el significado de los signos lingüísticos en cualquiera de los posibles niveles de representación: palabras, oraciones, textos, etc. Al considerar el significado de las palabras de manera individual, se denomina:

- Lexicología: estudia el significado de las palabras.
- Lexicografía: se avoca al estudio de la formación de diccionarios.

Discurso

En todos los niveles anteriores se estudian las palabras de manera individual o relacionadas unas con otras en una oración. El discurso se enfoca en estudiar las propiedades de un texto atendiéndolo como una unidad conformada por oraciones interconectadas entre sí. Surge aquí el procesamiento de diversos recursos lingüísticos imposibles de solucionar fuera de este nivel, como la anáfora y la elipsis.

Pragmática

El objetivo en este nivel es descubrir conocimiento del mundo que aporta significado al mensaje que se emite y que no siempre se encuentra codificado en él. También le atañe considerar las motivaciones que influyen a las personas para elegir determinadas oraciones o textos en situaciones específicas.

1.2. Gramática de dependencias

La gramática tradicional considera a la oración como una estructura bimembre, formado por sujeto y predicado. De acuerdo con esta concepción, el sujeto es la palabra o conjunto de palabras que expresan un concepto, del cual se predica, es decir, se afirma o se niega algo, y predicado es aquello que se predica, es decir, lo que se afirma o se niega del sujeto.

En 1969, el lingüista francés Lucien Tesnière propone una teoría que rompe con la interpretación estructural de la oración basada en juicios lógicos. Tesnière compara la oración con un pequeño drama, u obra de teatro, donde es posible distinguir un proceso, actores y circunstancias. El verbo ocupa la posición central, el proceso. Los actantes o actores designan los personajes que participan en el proceso, siendo el sujeto de la oración un actante más, sin función privilegiada. Y finalmente los circunstantes o circunstancias expresan las características como tiempo, lugar, modo, etc., en el que se desarrolla el proceso, pudiendo o no existir.

De esta manera Tesniére concibe a la oración como una estructura jerárquica, y no binaria, donde el verbo ocupa la posición central, y por ello mismo determina los papeles que desempeñan los actantes en la oración. De acuerdo con esto, el verbo tiene la capacidad de establecer relaciones de dependencia con el resto de elementos en la oración.

Un caso particular se da en aquellas oraciones donde el verbo no viene acompañado por algún actante, lo que ocurre con los llamados *verbos metereológicos*, a lo cual Tesnière alude lo siguiente:

> *Retomando nuestra comparación de la frase con un pequeño drama, diremos que, en el caso del verbo sin actante, el telón sube sobre una escena donde cae la lluvia o la nieve, pero vacía de actores.*

Después de distinguir entre actantes y circunstantes, Tesnière introduce los conceptos de *rección* y *valencia*, que por su importancia trataremos en el siguiente apartado.

1.2.1. Rección y valencia

En las oraciones, las palabras se relacionan de manera tal que algunas establecen o determinan propiedades de otras. Esta relación de dependencia se denomina *rección*. La palabra dependiente, también llamada subordinada, se conoce como ***regida***, y aquella de la cual depende, ***regente***.

De esta manera, si decimos que un elemento es *regido* por un verbo, significa que dicho elemento constituye un complemento en la construcción del significado del verbo, pero sobre todo, que cumple las restricciones gramaticales que el elemento *regente* impone.

A este hecho de regir o exigir una o varias palabras, se le denomina ***régimen***. El *régimen* puede ser caracterizado de dos maneras, como *régimen verbal*, y como *régimen preposicional*: el primero hace referencia a la exigencia de que el verbo vaya acompañado o no por un elemento subordinado (así se habla de un régimen transitivo, y de un régimen intransitivo), y la segunda nos habla de la exigencia de una forma específica de la preposición a utilizar.

Los verbos transitivos exigen la aparición de un sintagma que funcione como complemento directo: por ejemplo, *compré ropa*, *compraron una casa*, etc. El verbo *comprar* por sí solo carece de sentido. Por otra parte, los verbos intransitivos no necesitan ser complementados por algún sintagma que funja como complemento directo: *María tose*, *Juan ríe*, etc.

En algunas ocasiones los verbos transitivos eventualmente pueden expresarse como intransitivos, aunque no por ello cambie su naturaleza: siempre podremos preguntar *qué* al verbo *(¿**qué** lees? ¿**qué** escribes?*), lo que resulta imposible para los verbos intransitivos *(¿qué descansas? ¿qué duerme María?* que sería absurdo en condiciones normales de enunciación) [30].

El régimen preposicional, por otro lado, se refiere a la obligatoriedad de uso de una determinada preposición, y no a la mera posibilidad de uso: por ejemplo, *inducir **a***, *convertirse **en***, *depender **de***, etc.

Con todo esto, podemos observar que el verbo tiene la capacidad, denominada ***valencia***, de abrir en torno suyo huecos que deben ser ocupadas por ciertos elementos funcionales, también llamados *argumentos*, que son requeridos para construirse en una oración gramatical e inteligible.

El que estos argumentos sean requeridos, no implica que deban presentarse de manera obligatoria a nivel sintáctico. Por ejemplo, *comer* es un verbo bivalente, que requiere dos argumentos: el ser vivo que realiza la acción (el que come), y aquello en lo que recae la acción (lo que es comido). Sin embargo, en la oración *Juan come mal* el segundo argumento no es indicado explícitamente, aunque en realidad éste se encuentra semánticamente implícito en la oración, pues lo que se está expresando es que *Juan come poca comida*, o que *Juan come mucha comida poco nutritiva*, y en ambas oraciones, aparece el argumento requerido [15].

Al parecer, la realización sintáctica de los argumentos requeridos dependerá de la situación y la intención comunicativa del hablante.

Lo visto anteriormente, pone de relieve el contraste existente entre *rección* y *valencia*, e introduce el concepto de *obligatoriedad*, que a su vez atrae la noción de *latencia*.

Como ya se vio, la *obligatoriedad* se refiere a la expresión u omisión a nivel superficial de ciertos argumentos requeridos por el verbo. Un argumento que puede ser omitido, se denomina *opcional*, sin embargo este puede estar implicado por el contexto, a lo cual se refiere la *latencia*.

La latencia puede detectarse cuando [11]:

a) Está dada en el contexto verbal: por ejemplo, *no quería abusar de su confianza, pero abusé*.

b) Viene dado en el contexto situacional: por ejemplo, *¡abre!* (se deduce el actante en la flexión verbal, y otro implicado situacionalmente: *el libro*, *la puerta*, etc.).

c) Ciertos actantes se pueden suponer conceptualmente: *la mujer parió [un niño/una niña]*.

d) Hablante y oyente están consientes de la falta de expresión de una valencia: por ejemplo, *el labrador ara*.

Las posibles combinaciones que pueden establecerse entre los tres conceptos antes señalados (requerimiento, obligatoriedad y rección) se indican en la siguiente tabla:

Argumentos			Ejemplos
Requeridos	**Obligatorios**	**Regidos**	
✓	✓	✓	*Carecer, encontrar,* etc.
✓	-	✓	*Comer, beber, oír,* etc.
-	-	✓	*Correr, vivir, saltar, dormir,* etc.
-	-	-	Verbos meteorológicos.

Tabla 1.1 Contrastes entre requerimiento, obligatoriedad y rección.

Con esto podemos obtener las siguientes conclusiones:

- La valencia establece el número de argumentos requeridos por el verbo.
- Algunos de estos argumentos pueden o no ser expresados a nivel sintáctico (obligatoriedad).
- Los argumentos requeridos y expresados por obligatoriedad se denominan *argumentos exigidos*.

Dado que la valencia se relaciona con los aspectos sintácticos y semánticos de la lengua, se trata de un fenómeno complejo que ha llevado a considerar que el concepto de valencia se amplía hasta distinguir tres tipos diferentes de ésta [13]:

1. *Valencia lógica*. Relación entre el verbo lógico y los argumentos (plano en el que se precisan las casillas vacías requeridas por el verbo).

2. *Valencia semántica*. Los casilleros abiertos por el verbo deben ser ocupados por elementos que lleven determinados rasgos o marcas semánticas.

3. *Valencia sintáctica*. Se encarga de estudiar la ocupación, obligatoria u opcional, de los huecos vacíos abiertos en el nivel lógico.

1.2.2. Patrones de rección

A finales de los años 60's, surgió una nueva teoría sobre el Lenguaje Natural denominada Teoría del Significado ⇔ Texto, la cual ha sido desarrollada principalmente por I. Mel'čuk.

En esta teoría, la descripción de valencias, incluyendo la relación entre valencia semántica y valencia sintáctica, encontró una solución adecuada en términos de los llamados *patrones de rección*. Estos se describen como una matriz donde se muestran todas las posibles representaciones de valencias. Esta matriz tiene las siguientes características:

1. Se genera a partir de la definición lexicográfica de una unidad léxica. La definición debe reflejar explícitamente todas las valencias del verbo a nivel semántico.
2. Se llena con el orden de las palabras de los actantes del verbo, indicando las opciones de representación en el nivel sintáctico.
3. Se especifica la obligatoriedad de los actantes (aparición obligatoria u opcional)
4. Se denotan las condiciones particulares que deben ser cumplidas por las opciones de representación [22].

Por ejemplo, considérese la definición semántica para el verbo *regalar* (ignorando los clíticos pronominales), la cual se representaría de la siguiente manera:

X ***regala*** *Y a Z*

El patrón de rección para dicho verbo, queda representado de la siguiente manera:

X = 1	Y = 2	Z = 3
1.1 ***N***	2.1 ***N***	3.1 *a* ***N***
Obligatorio	Obligatorio	Opcional

(1) $C_{1.1}$: N denota a una persona.

(2) $C_{2.1}$: N denota una cosa o un animal.

(3) $C_{3.1}$: N denota a una persona.

El primer renglón de la matriz establece las correspondencias entre los niveles sintáctico y semántico. Los símbolos *X, Y, Z* designan valencias semánticas, mientras que los números 1, 2 y 3 designan valencias sintácticas del verbo.

El segundo renglón de la matriz enumera todas las posibles opciones de representación para cada valencia sintáctica. Estas opciones se corresponden con categorías gramaticales (*N = noun* representa la categoría gramatical *Sustantivo*) y preposiciones que conectan al verbo con los argumentos.

Por debajo de la matriz se indican las condiciones que debe seguir cada categoría gramatical indicada.

En este trabajo se adoptará esta representación que se hace de la valencia verbal, incluyéndose todos los elementos que se utilizan para construir la matriz.

CAPÍTULO 2

ESTADO DEL ARTE

La noción sobre el conjunto de complementos que pueden combinarse con los verbos, es crucial para las teorías lingüísticas y es ampliamente utilizado en diversas tareas del Procesamiento del Lenguaje Natural. Estos complementos son originalmente introducidos en 1959 bajo el concepto de *actantes* por el lingüista francés Tesnière, término restringido originalmente a la sintaxis de los verbos, y que define como "*los elementos que un verbo es susceptible de regir*".

La recopilación de información sobre estos elementos (también denominados *argumentos*, *términos*, etc.) fue una idea originalmente sugerida por el lingüista Noam Chomsky, y que se ha ido implementado por las teorías sintácticas subsecuentes.

La manera de nombrar a estos elementos, así como el tipo de información que se recopila y la manera de concebirlos varía de acuerdo al formalismo teórico que los procesa. En el enfoque teórico de constituyentes, los actantes se conocen más ampliamente con el nombre de *Marcos de Subcategorización* (*Subcategorization Frames* ó *SCF*), y especifican el entorno sintáctico de un verbo. Dentro del formalismo de dependencias, en particular en la Escuela Semántica de Moscú, el concepto de *actante* siguió desarrollándose hasta llegar a distinguir entre *actantes sintácticos* y *actantes semánticos*. Más adelante, y siguiendo esta misma línea teórica, los actantes encuentran en la Teoría de Significa-Texto (*Meanning Text-Theory* ó *MTT*) una solución adecuado bajo el concepto de *Patrones de Rección* (*Government Patterns* ó *GP*).

En este capítulo, se describirán los trabajos de detección y adquisición automática de actantes que se han realizado a la fecha.

2.1. Procesamiento de Corpus

Los Corpus han sido en general una herramienta válida para el estudio de las lenguas, y que, con la aparición de las computadoras, se convirtieron en elementos imprescindibles para la realización de diversas tareas del Procesamiento de Lenguaje Natural.

El contenido de los Corpus puede variar según los objetivos y necesidades que se persiguen en cada uno de éstos. Es por ello que algunos resultan ser más convenientes para algunas tareas que otros, según el tipo de información que se desea procesar.

2.1.1. Clasificación de Corpus

Los corpus pueden clasificarse de acuerdo a diversos criterios, como la modalidad de la lengua (Corpus textuales vs. Corpus orales), el número de lenguas (Corpus monolingües vs. Corpus bilingües o multilingües), la cantidad y distribución de textos (Corpus grandes vs. Corpus equilibrados), etc. Seguramente la clasificación más relevante en el procesamiento automático de Corpus reúne por una parte los textos sin procesar (*raw data*) y en otra los textos anotados (*annotated texts*). Como es de esperarse, los corpus sin procesar (*raw corpus*) proporcionan textos en su forma pura, libre de cualquier tipo de procesamiento e información adicional, mientras que en los segundos se proporciona información lingüística o de otro tipo. Esta información consiste en introducir una serie de códigos o etiquetas que, dentro de los a aspectos lingüísticos, proporcionan información de tipo sintáctico, semántico, contextual, etc., y en aspectos no lingüísticos, indican por ejemplo la estructura interna del texto, como el inicio y finalización de un capítulo o un párrafo, la señalización de un fragmento de texto que corresponde a una expresión oral, etc.

En la anotación de tipo lingüístico, se distinguen dos tipos: la anotación categorial o gramatical (*POS tagging*) y la anotación sintáctica (*parsed*). La primera consiste en asignar a cada unidad léxica del texto su correspondiente categoría gramatical. Un ejemplo de la forma de presentación de un texto sin procesar y uno etiquetado gramaticalmente es el que se muestra a continuación.

Texto etiquetado:

> Construir_VMN0000 cabañas_NCFP000 o_CC chozas_NCFP000 para_SPS00 guarecer_VMN0000 se_PP3CN000 de_SPS00 la_DA0FS0 intemperie_NCFS000 mientras_CS apacienta_VMIP3S0 sus_DP3CP0 ganados_NCMP000 ._Fp

Texto sin etiquetar:

> Construir cabañas o chozas para guarecerse de la intemperie mientras apacienta sus ganados.

La mayor cantidad de trabajos realizados en la adquisición de SCF, ha seguido la tendencia general de aplicar métodos estadísticos a Corpus con el fin de detectar patrones de combinación de verbos con determinados tipos de elementos léxicos.

En los siguientes apartados se dará una breve descripción de los primeros trabajos realizados y un resumen de la metodología seguida. Posteriormente se analizarán otros recursos lingüísticos tomados como entrada de datos para la adquisición de los SCF.

Corpus anotados

El trabajo en Corpus anotados categorial o gramaticalmente (*POS tagging*) supone una gran ventaja sobre los Corpus no anotados (*Raw Corpus*): destacan principalmente la facilidad de explotación y la multifuncionalidad que ofrecen.

Como ya se mencionó, la anotación gramatical consiste en asignar a cada unidad léxica del texto un código (etiqueta) que indica su categoría o parte de la oración. También suele incluir información sobre las características morfológicas (género, número, caso, persona, etc.) [32]. Aparte de este tipo de anotación, también puede incluirse información de tipo sintáctico, y con ello poder extraer información concerniente a la estructura de las oraciones.

Un ejemplo de este tipo de Corpus anotados son los *Treebanks*, Corpus cuya estructura sintáctica está representada en forma de árbol. La anotación sintáctica que sigan dependerá de la teoría lingüística que deseen adoptar.

Existe una diferencia en la representación sintáctica formal y el formato utilizado para denotarla. Por ejemplo, la representación sintáctica para la oración *I read* es:

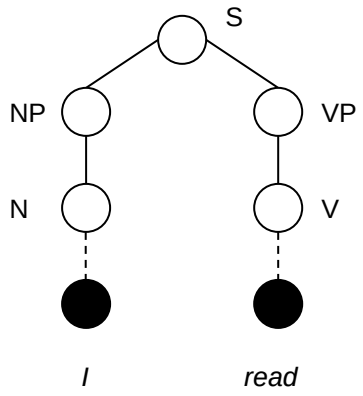


Ilustración 2.1 Representación sintáctica para una oración.

La notación en formato de texto para la oración antes mencionada, por ejemplo bajo el esquema del Corpus *Penn Treebank*, se da de la siguiente manera:

```
(S
        (NP
                (N I))
        (VP     (V read)))
```

Ilustración 2.2 Notación sintáctica en formato del corpus Penn Treebank.

2.2. Extracción automática de Marcos de Subcategorización

En general, los SCF se basan en la información sintáctica de la estructura de los actantes de un verbo: se indica tanto el número de actantes, como la categoría gramatical de cada uno y la posición que guardan respecto al verbo. En el verbo *decir*, por ejemplo, podemos observar tres actantes: (1) emisor (persona que expresa el mensaje), (2) receptor (persona que recibe el mensaje) y (3) mensaje (lo que se expresa). En la oración *Pedro* (1) *dijo*

a sus amigos (2) *muchas mentiras* (3), pueden apreciarse los actantes señalados, expresándose los dos últimos con el patrón:

Decir a <receptor> <mensaje>

El correspondiente SCF asignado a dicho patrón, es:

Decir a ***N1 N2***

En las siguientes subsecciones mostraremos los trabajos que se han realizado para extraer los actantes de verbos bajo el formalismo de SCF.

2.2.1. Primeros trabajos

Dado el avance gradual que se ha venido realizando en el Procesamiento de Lenguaje Natural, el uso de Corpus no procesados, fue la constante en los primeros trabajos de la extracción de actantes de verbos, dada la complejidad que ha implicado hacer un etiquetado manual de un Corpus, y los avances aún no del todo confiables en la automatización de esta tarea.

El diseño pionero sobre la extracción de SCF, corresponde a Michael Brent [5], quien propone el desarrollo de un programa que toma texto de un corpus no etiquetado como única entrada para identificar SCF, extrayendo primeramente los verbos contenidos en él, y a continuación, frases que representen a los argumentos de los verbos.

El enfoque que sugiere Brent radica en no analizar las oraciones de manera completa, sino utilizar pistas o claves morfosintácticas locales, por lo cual, se estaría haciendo uso de un conocimiento gramatical a priori muy básico.

En el primero de sus dos trabajos (1991), Brent identificó 5 SCF, utilizando una técnica basada en el Filtro de Casos de Rouvret y Vergnaud. A través de este filtro primero identifica los verbos potenciales, buscando, por ejemplo, palabras que contengan o carezcan del sufijo –*ing* o que sigan a un determinante o una preposición diferente a *to*. Por ejemplo, *was walking* se puede considerar como verbo, pero *a talk* no.

Para el reconocimiento de SCF, usó una gramática de estados finitos que describe sólo un pequeño fragmento del idioma inglés, y que se basa en ítems léxicos de clase cerrada, como pronombres, preposiciones, determinantes y verbos auxiliares.

En su segundo trabajo [6], donde identifica 6 marcos sintácticos (sólo uno más que el anterior), Brent incorpora un modelo estadístico en el cual se mide la frecuencia de aparición de claves con los verbos para cada uno de los marcos, así como el número de veces que cada verbo ocurre. Estos datos se van recopilando a fin de decidir en un momento dado si existe suficiente evidencia para considerar que un verbo en particular se manifiesta con un determinado marco sintáctico o no.

Los casos que no superen cierto umbral, serán explicados como:

- Apariciones circunstanciales, es decir, un determinado marco sintáctico se considera como adjunto y no como argumento del verbo en cuestión.

  *"Salí corriendo **con mucho miedo**"*

  Considerar un complemento preposicional (en este caso parametrizado por *con*) como argumento del verbo correr, es un error, ya que sólo modifica las circunstancias en las que se desarrolla la acción.

- Error del sistema de análisis.

Según los resultados obtenidos y a propio juicio de Brent, es posible aprender aspectos significativos de la sintaxis léxica inglesa utilizando regularidades gramaticales simples. Estos datos sugieren específicamente que no es necesario un gran parseador ni un gran lexicón para recuperar suficientes estructuras sintácticas para el aprendizaje de la sintaxis léxica.

El problema principal de Brent radica en que depende <solamente> de claves morfosintácticas, las cuales no siempre están presentes en muchos verbos y marcos sintácticos. Por ejemplo, algunos verbos requieren obligatoriamente un sintagma preposicional (preposición *en: él piensa **en** ti*), no obstante en la mayoría de los casos lo podemos encontrar

como un complemento adverbial (*él suele leer **en** su casa*), el cual cumple la función de adjunto.

Un segundo problema que se puede considerar como derivado del anterior, es la poca eficiencia que hace del corpus, ya que para la identificación de los verbos y de los marcos sintácticos, desprecia un muy alto porcentaje de la información potencialmente disponible.

Los problemas dados por la metodología seguida por Brent motivó que los enfoques posteriores optaran por hacer un uso más eficiente de los textos analizados. Esto llevó a la necesidad de contar con corpus etiquetados.

Siguiendo esta línea, Ushioda et al [38] establecen que lo ideal para reconocer automáticamente marcos sintácticos y sus respectivas frecuencias, sería contar con un Corpus parseado totalmente. Sin embargo, dado que los Corpus parseados manualmente son escasos además de pequeños, y los parseados automáticamente contienen muchos errores, propone hacer uso de sentencias parseadas sólo parcialmente, derivadas de un Corpus etiquetado. El parseo que propuso Ushioda produce información de frases nominales mínimas (sin frases preposicionales adjuntas u otros complementos). Para contrarrestar la falta de exactitud e insuficiencia de información obtenida, comparada con parseos manuales, su metodología garantiza generalizar y producir tamaños de muestra muy grandes.

El sistema que elaboró es capaz de reconocer y calcular las frecuencias relativas de 6 marcos de subcategorización, los mismos trabajados por Brent. El proceso consiste en extraer del Corpus etiquetado las sentencias que contienen un verbo y dividir el sintagma nominal en *chunks* utilizando un parseador de estados finitos, así como el resto de palabras usando un conjunto de 16 símbolos y categorías frasales. A estas sentencias les es aplicado un conjunto de reglas de extracción de marcos de subcategorización. Estas reglas están escritas como expresiones regulares y se obtienen a través de la extracción de ocurrencias de una pequeña muestra de verbos en un texto de entrenamiento.

A pesar de que la metodología que emplearon proporciona una aceptable medición de frecuencias de marcos de subcategorización, en algunos verbos se da la impresión de que aparecen en estructuras sintácticas que no pueden ser capturadas por su inventario de marcos

de subcategorización. Este problema lo abordan a través de un método estadístico basado en un conjunto de ejemplos de entrenamiento que posibilita al sistema para aprender patrones de error y así incrementar sustancialmente la precisión de las frecuencias de subcategorización de verbos.

Manning [26] propone un sistema más ambicioso capaz de reconocer 19 marcos sintácticos diferentes. Sugiere que es más útil extraer la mayor cantidad de información posible aún si ésta se llega a considerar como ruido, pues a cambio se obtendrían datos más completos.

Los marcos sintácticos se obtienen a través de un programa que procesa la salida de un etiquetador estocástico de partes de la oración (part-of-speech tagger) ejecutado sobre el Corpus a analizar. El programa consta de dos partes: un parseador de estados finitos que analiza el texto etiquetado buscando un verbo, y que al encontrarlo, divide toda la información que lo sigue en pequeños componentes o *chunks*, hasta encontrar algún elemento reconocido como terminador de argumentos subcategorizados. El resultado devuelto por el parseador es una lista de elementos que aparecen después de un verbo, supuestos marcos de subcategorización y estadísticas de la aparición del verbo en diferentes contextos. Toda esta información, como puede suponerse, está plagada de ruido, dado que no le es posible distinguir argumentos de adjuntos.

La segunda parte del programa, consiste en la reducción del ruido que acompaña a la información, lo cual se logra utilizando el mismo filtro estadístico usado por Brent: el ruido (o pistas falsas), puede ser eliminado observando qué marcos aparecen con un verbo en una frecuencia razonablemente superior a la que pudiera considerarse casualidad (adjuntos) o errores en la detección.

Gahl [18] inicia una propuesta diferente a las antes mencionadas consistente en detectar los marcos sintácticos a través de los sistemas de recuperación de información en corpus (*corpus query systems*). Presenta un método para extraer subcorporas que contienen diferentes marcos sintácticos de verbos, sustantivos y adjetivos del *British National Corpus* (BNC).

La herramienta de extracción de Gahl consiste en una serie de archivos batch que se usan con el procesador de consultas de Corpus (CQP por sus siglas en inglés). Esta herramienta permite a un usuario especificar en un archivo de entrada qué subcorporas deben ser creados para un lema dado. Las consultas son realizadas utilizando expresiones regulares sobre etiquetas de partes de la oración, lemas y etiquetas morfosintácticas. Por ejemplo, el usuario puede buscar en el corpus el patrón [*verb* NP VP*ing*]. La consulta devolvería "*I kept them laughing*".

Uno de los errores cometidos por este sistema es la ambigüedad en frases preposicionales adheridas. Por ejemplo, de las líneas que presumiblemente concuerdan con el patrón [_NP PP*with*] para el verbo *heal*, muchas contienen frases preposicionales incrustadas (e. g. [_NP], como en *heal [children with asthma]*, en vez de [_NP PP*with*], como en *healing [arthritis] [with a crystal ball]*.

2.2.2. Metodología de extracción

La revisión de los trabajos antes mencionados, permiten establecer una metodología de procesamiento, como la expuesta en [10] y [35], bajo la que es posible distinguir los siguientes puntos:

1. *Selección y preparación del Corpus*: indica la elección del corpus en el que se va a realizar la identificación de SCF, y, en caso de no estar anotado, el tipo de etiquetado que se le realizará (gramatical, sintáctico, etc).
2. *Detección de marcos*: establece el método computacional a seguir para identificar los SCF.
3. *Filtrado estadístico*: determina el método para eliminar el posible ruido obtenido en el paso previo.

Selección y preparación del Corpus

Es muy importante considerar tanto el tipo como el tamaño de los Corpus a procesar, pues estos factores pueden provocar variaciones en cuanto a los resultados que se obtienen. En general, los investigadores prefieren contar con la mayor cantidad de información (texto)

posible, ya que de esta manera aseguran una muestra más representativa del idioma en el que se esté trabajando.

En [33] y [34] se expone cómo diferentes géneros de Corpus provocan variaciones en las frecuencias de SCF. En [34] se estudiaron 5 Corpus diferentes, dos de los cuales fueron obtenidos de fuentes psicológicas, que se caracterizaban principalmente por contener sentencias aisladas, y los tres restantes fueron el *Brown corpus*, *Wall Street Journal corpus* y el *Switchboard corpus*. Las diferencias reportadas se encontraron tanto en los tipos de SCF como las frecuencias de los tipos de SCF. Ambas, se explica, fueron motivadas por:

- *Variación basada en el contexto*: la frecuencia de SCF observados varía entre verbos encontrados en oraciones conectadas en discursos y en oraciones separadas.

- *Variación en el sentido de palabras*: el uso de diferentes sentidos de un mismo verbo, motivado por los efectos del contexto, conlleva probabilidades diferentes de subcategorización.

La presentación del Corpus tocante a la anotación de información lingüística, determinará la manera en que se procederá para ejecutar la tarea de extracción de SCF. Brent utiliza un Corpus no anotado al cual aplica claves morfosintácticas para detectar verbos y sus posibles marcos. Ushioda propone utilizar sentencias parseadas sólo parcialmente, derivadas de un Corpus ya etiquetado, y a las cuales les es aplicado reglas escritas como expresiones regulares. Manning aplica un etiquetador estocástico sobre el Corpus a analizar y así extraer todas aquellos componentes de la oración que tengan elementos reconocidos como terminador de marcos. Gahl extrae subcorporas a través de la ejecución de expresiones regulares sobre el BNC para detectar en ellos a los posibles marcos.

Detección de marcos

La detección de SCF en general se ha realizado a través del *emparejamiento de patrones*, que consiste en definir a priori información gramatical que pudiera considerarse relevante para identificar alguna combinación de elementos léxicos como candidados a SCF. Posteriormente se busca en el Corpus información que pudiera emparejarse con los patrones predefinidos.

Filtrado estadístico

La adquisición de los posibles marcos realizada por el proceso previo, no está exenta de errores, como es de esperarse. La información obtenida contiene ruido que puede derivarse de errores en la fase de etiquetado gramatical, por ejemplo, o incluso, errores en la fase de detección de SCF provocada por una ineficiencia en la discriminación de adjuntos.

Para remover toda la información no deseada, se realiza un procesamiento estadístico. En suma, se busca determinar si un candidato a SCF de un verbo en particular debe realmente considerarse como tal o no. Los métodos estadísticos para realizar el filtrado de información se hacen usualmente con la *prueba de hipótesis* (*hypothesis test*). Esta prueba consiste en establecer una hipótesis nula H_0, como verdadera, a menos que los datos sugieran lo contrario, lo cual provoca que se rechace la hipótesis y entonces se acepta como verdadera una hipótesis alternativa H_1. En el contexto de la adquisición de SCF, H_0 se considera como una falta de asociación entre un determinado verbo y un SCF, y H_1 como la afirmación a dicha asociación. Se establece la prueba como de *una cola*, dado de que la hipótesis alternativa establece una dirección, en este caso la correlación positiva entre el verbo y el marco. En seguida se calcula el valor estadístico de prueba con los datos de la muestra, lo que sirve para decidir si H_0 es verdadera o falsa. Esto se realiza comparando la probabilidad esperada de que exista correlación si H_0 es verdadera, con la probabilidad observada de coocurrencia. Si esta última es mayor que la primera, la hipótesis H_0 es rechazada.

2.2.3. Trabajos de extracción en diversos idiomas

La mayor cantidad de trabajo realizado en la extracción de estructuras argumentales se ha realizado mayormente para el idioma inglés. Sin embargo, algunos de los modelos construidos han sido exitosamente aplicados para otros idiomas. A continuación mostramos ejemplos de estos casos:

En 1995 Monedero et al [29], inspirados en el trabajo de Brent y Manning, desarrollaron una herramienta para obtener marcos sintácticos de verbos en español. Del trabajo de Brent consideran el punto de vista de que hay muchos marcos para los cuales no existen pistas fiables, lo que implica concentrarse en un número reducido de casos que

ofrezcan pocas dudas, y de Manning adoptan el criterio de relevancia de información, es decir, proponen la descripción de estructuras relativamente sencillas, pero lo suficientemente variadas como para permitir el estudio de un amplio número de marcos de subcategorización.

El trabajo realizado, denominado SOAMAS, consistió en generar tres gramáticas: la primera de ellas encargada de identificar verbos principales y auxiliares, así como posibles conjunciones y preposiciones. La segunda realizada con el fin de reconocer sintagmas nominales, adjetivos y preposicionales. La tercera consistió en ser la encargada de identificar los complementos verbales.

El principal problema enfrentado para entonces, consistió en la carencia de corpus etiquetados para el español suficientemente extensos (dispusieron sólo de 10,000 palabras etiquetadas (Martín 94)), lo que imposibilitó llegar a resultados confiables.

En [19] se propone para el español un método estadístico para identificar SCF en grandes corpus y posteriormente utilizar un método semi-automático para hacer corresponder la información sintáctica con las valencias semánticas. También se elaboró una nueva estructura de representación de GP tomando en cuenta algunas características del idioma español, por ejemplo la conexión entre el verbo y el objeto directo, que se realiza a través de la preposición “a” cuando este último hace referencia a entidades animadas.

En [37] se utiliza una metodología para el húngaro basada en el mecanismo de aprendizaje estadístico utilizado primeramente por Brent, y se complementa con la prueba *likelihood ratio* (cociente de probabilidad) y la técnica de decisión basada en frecuencias relativas. Los métodos se probaron en dos Corpora húngaros, el Szeged Corpus, un treebank con 82000 oraciones etiquetadas morfológica y sintácticamente, y el Hungarian Webcorpus, del cual se tomaron 32000 oraciones. Dado que éste último no se encuentra etiquetado, se utilizó un parseador gratuito para extraer la información morfológica. En general obtuvieron mejores resultados procesando el Corpus Szeged, que el Corpus etiquetado automáticamente.

En [33] se presenta un trabajo para el idioma checo, lengua de orden libre de palabras. En este proyecto se tomaron datos anotados sintácticamente del *Prage Dependency Treebank*, el cual no contiene ninguna información referente a marcos de subcategorización.

Se utilizaron tres técnicas estadísticas distintas para aprender posibles marcos de subcategorización de ciertos verbos: *LRT* (*Likelihood Ratio Test*, test de razón de verosimilitudes), *T-score* y *Hypothesis testing*. Se extrajeron 19126 oraciones como datos de entrenamiento, de los cuales se obtuvieron 137 marcos de subcategorización.

En [27] se describen los avances realizados en idioma búlgaro para obtener marcos de subcategorización. Se realizaron pruebas con un extracto de datos del Bulgarian Tree Bank: 580 oraciones completamente parseadas bajo el formalismo HPSG. Se realizaron dos experimentos con los datos disponibles, el primero utilizando un corpus con etiquetado POS, y el segundo utilizando datos completamente parseados. Se implementará el sistema para aprender los marcos de subcategorización. La idea es extraer todas las posibles pistas de los verbos en el corpus, utilizando una distribución binomial para filtrar la información obtenida.

La extracción de SCF para el idioma italiano se describe en [20], donde uno de los objetivos es investigar la complejidad de aplicar experimentos reportados por la bibliografía en otros idiomas en el italiano, así como el evaluar de qué manera puede influir el utilizar datos anotados. Para ello se utilizó el Treebank italiano *Turin University Treebank*.

El conjunto de datos consiste en cerca de 2,000 oraciones, representadas bajo el enfoque de dependencias, pues éste describe mejor lenguajes como el italiano que guarda un relativo orden libre de palabras (por ejemplo la libre distribución de actantes en una oración). En este Corpus se distinguen diversas relaciones gramaticales que involucran información sobre la categoría morfológica, relaciones de dependencia dada entre las palabras, tales como sujeto y argumento, e información sintactico-semántica, como tiempo y manera.

Lo observado en sus experimentos es que la cantidad de sentencias etiquetadas con las que arrancaron los experimentos es insuficiente para producir modelos de aprendizaje robustos, sin embargo, también reportan que la cantidad de sentencias requeridas es menor cuando se trata con una representación basada en dependencias, a comparación con las requeridas con una basada en constituyentes.

2.2.4. Otras fuentes de extracción

El recurso más explotado en la extracción de marcos, ha sido el Corpus monolingüe. Sin embargo, se han realizado trabajos que también consideran otras fuentes de información, como alternativa para resolver esta tarea. En las siguientes secciones comentaremos sobre éstas fuentes y describiremos la metodología empleada en cada trabajo.

Utilización de recursos bilingües

Los recursos multilingües son ampliamente utilizados para realizar diversas tareas del Procesamiento del Lenguaje Natural, sin quedar exenta la extracción de SCF.

En [1] se parte del hecho de que para el entendimiento del lenguaje se realiza un mapeo de la estructura sintáctica hacia representaciones conceptuales (mapeo de argumentos del predicado), mientras que en la generación del lenguaje se realiza el proceso contrario. Los textos multilingües se utilizan para obtener información sobre este mapeo de manera automática. Para ello establecen cuatro tipos de mapeo de acuerdo a dos parámetros: número de frases nominales subcategorizadas y tipos de roles temáticos a los cuales se mapean los argumentos. Los tipos diferentes de mapeo que establecen son: *procesos causados*, *procesos o estados*, *acción agentiva* y *estado inverso*.

Además de estos elementos, se recopila las restricciones semánticas de los argumentos y las llamadas idiosincrasias presentes en el léxico, por ejemplo, el hecho de que una frase nominal es introducida al verbo con una determinada preposición, y no con otras (e. g., el verbo *look* es acompañado por la preposición *at*).

Tanto los datos de mapeo como las idiosincrasias son obtenidos de manera automática utilizando técnicas que dependen de algunas heurísticas sintácticas dependientes del idioma (por ejemplo, en inglés y español el objeto directo usualmente sigue al verbo). Las restricciones semánticas son definidas a priori en clases de verbos, por ejemplo, los verbos que pertenecen a la clase Evento de comunicación, que agrupa verbos como *reportar*, *confirmar*, etc., tendrán como agente un tipo *persona* u *organización*.

Otro de los recursos bilingües utilizados, son los diccionarios bilingües planos. En [17] se utilizan para incrementar el número de entradas de un diccionario japonés de valencias. La idea parte de la hipótesis de que verbos con significado similar tienen típicamente la misma estructura de valencias, aclarando que "significado similar" se refiere a la misma traducción.

Uno de los problemas con los que se enfrentan bajo este enfoque, es que una traducción producirá en la mayoría de los casos polisemia. La solución que adoptan es realizar traducción a varios idiomas, y de esta manera, considerarán las palabras *a* y *b* como similares, si éstas tienen la misma traducción en dos o más idiomas.

Uso de la Web

El Internet también ha sido aprovechado como medio a utilizar para la adquisición de SCF. Particularmente, la Web se ha utilizado para generar Corpus de los cuales se extraerán los SCF. En [24] se presenta un trabajo para el japonés, en donde se utilizó un sistema para descargar alrededor de 400 megabytes de páginas Web. Al total de páginas se les aplicaron ciertos filtros para descartar aquellas que pudieran pertenecer a otros idiomas. Al final se obtuvieron 100 megabytes de páginas exclusivamente en idioma japonés. De estas páginas se aplicó un segundo filtro para descartar oraciones escritas en otros idiomas. Para ello se extrajeron las sentencias que contuvieran caracteres específicos del japonés, como el *Katakana*, *Hiragana* y *Kanji*. Finalmente, la calidad del Corpus generado se consideró óptima, al tomar 1000 oraciones aleatorias, de las cuales, 995 correspondían al idioma japonés.

En [39] se describe un método para identificar SCF para el idioma turco. La innovación que aquí se señala es la realización automática de consultas sobre la Web para recuperar información que será utilizada para la creación de un Corpus.

El modelo que proponen está formado por cuatro módulos: (1) en primer lugar cuentan con un generador verbal, el cual realizará conjugaciones de los verbos. (2) Cada uno de estos verbos se consulta en Internet, y la información devuelta (3) es etiquetada por un etiquetador de casos, y finalmente (4) son utilizados métodos de aprendizaje (clasificador Bayesiano) para adquirir los SCF.

CAPÍTULO 3

ANÁLISIS DE LA FUENTE DE DATOS A PROCESAR

En el CAPÍTULO 2 explicamos los distintos trabajos realizados en la adquisición de actantes de verbos, y del cual podemos resumir que se basan en la aplicación de métodos estadísticos aplicados a Corpus, con el fin de analizar patrones de ocurrencia de eventos de acuerdo a la frecuencia de uso en el lenguaje.

En nuestro caso, hemos optado por tomar como fuente de información primaria los diccionarios explicativos, los cuales procesaremos empleando una serie de heurísticas basadas en observaciones a priori de la naturaleza y comportamiento de los datos contenidos en las definiciones lexicográficas. Esta información nos será útil para obtener la valencia semántica de los verbos, más no suficiente para poder distinguir actantes sobre-entendidos de los que no lo son, ni tampoco determinar la obligatoriedad de los actantes a nivel sintáctico.

Por su importancia en este trabajo y atendiendo a su elaborada estructura, iniciaremos describiendo las características de los diccionarios explicativos.

3.1. Fuente de información primaria: Diccionarios explicativos

Los diccionarios son herramientas lingüísticas muy importantes que recogen el léxico de una lengua poniéndola a disposición de los hablantes para su consulta.

Existen diferentes tipos de diccionarios, así como diferentes maneras de clasificarlos. Para fines de esta investigación, nos concentraremos en los diccionarios que van dirigidos a los hablantes nativos de una lengua (monolingües), que no tienen restricciones de dominio en el vocabulario que registran (generales) y que están encargados de la definición semántico-pragmática de la entrada léxica (explicativos o definitorios).

Este tipo de diccionarios son también llamados “pasivos” dado que están orientados a la comprensión de oraciones, mas no a su generación. Por este motivo carecen o emplean información muy reducida sobre ejemplos de uso (oraciones donde se observa el uso de la unidad léxica definida), condiciones paradigmáticas (relaciones de sinonimia, antonimia, etc.) y condiciones sintagmáticas (uso contextual de la unidad léxica definida: régimen preposicional, colocaciones y valencias verbales).

El diccionario explicativo de mayor resonancia en países hispanohablantes es el Diccionario de la Real Academia Española (DRAE), el cual también es utilizado como punto de referencia de otros diccionarios generales. Dadas estas características que le confieren autoridad de consulta, será el diccionario sobre el cuál basaremos nuestra investigación.

3.2. Secciones en un diccionario explicativo

3.2.1. Artículo lexicográfico

Las secciones textuales dispuestas ordenadamente en un diccionario se denominan *artículos*, y están conformadas por una *entrada* también denominada *unidad léxica* (términos que utilizaremos indistintamente de ahora en adelante), y la información que la define o describe. Además de estos dos elementos, se ha llegado también a considerar la categoría gramatical de la entrada como parte del artículo.

Las entradas pueden ser simples (una sola palabra) o complejas (más de una palabra), y aparecen ordenadas alfabéticamente en el diccionario en su forma lematizada.

Cuando dos o más palabras son homónimas teniendo orígenes etimológicos diferentes, se distinguen (en DRAE) unas de otras mediante el empleo de superíndices. Por ejemplo:

Trincar[1]. (Del prov. trencar). tr. Partir o desmenuzar.
Trincar[2]. (De or. inc.). tr. Atar fuertemente.
Trincar[3]. (Del al. trinken). tr. coloq. Tomar bebidas alcohólicas.

Delante de la unidad léxica se disponen información relativa de ella, de la cual puede distinguirse una serie de elementos que señalan sus restricciones y condiciones de uso, y la

información semántica, o definición, que constituye el contenido básico del artículo lexicográfico.

3.2.2. Definición

La definición constituye el elemento central del artículo lexicográfico, por lo que es necesario conocer el tipo de información que ésta maneja. Desde esta perspectiva, debemos primero subrayar que el tipo de definición que a nosotros interesa, la definición lexicográfica, da información sobre el contenido lingüístico de la unidad léxica, a diferencia de la definición enciclopédica, que informa sobre "los conocimientos sociales de la realidad concreta extralingüística" [8]. La definición lexicográfica realizará pues una función metalingüística, ya que el objeto de estudio es empleado como instrumento para llevar a cabo dicho estudio.

Las unidades léxicas pueden dividirse en dos grandes sectores: palabras de contenido léxico (sustantivos, adjetivos, verbos y adverbios) y palabras funcionales (preposiciones, pronombres, etc.). Acorde a esto se reconocerá la definición lexicográfica de dos maneras: como definición propia o perifrástica y definición impropia o funcional.

La definición propia está encargada de expresar el *significado* de las entradas en cuanto a su contenido léxico-semántico, es decir, se utiliza con las palabras de contenido léxico. Por otro lado, la definición impropia se utiliza para *describir* o *explicar* el funcionamiento y empleo de palabras funcionales, debido a su falta de un verdadero significado léxico. Característica del primero tipo de definición es que puede aplicársele la prueba de conmutabilidad.

La estructura de las definiciones propias suele seguir la norma establecida por la llamada definición aristotélica, la cual consiste de un enunciado encabezado por un término genérico o hiperónimo inmediato, seguido de una diferencia específica, o conjunto de rasgos y características que diferencian el término definido de otros que se agrupan bajo el mismo hiperónimo.

3.2.3. Contorno de la definición

Las definiciones lexicográficas van acompañadas por una serie de elementos denominados *contorno de la definición*, que no pertenecen propiamente al contenido semántico de la unidad léxica que se define, pero que se incluyen para indicar el correcto uso del término definido, pues implican ciertas restricciones contextuales de éste y, en ocasiones, algunos de sus usos sintácticos ([8]).

El contorno de una definición puede ser identificado a través de la prueba de conmutabilidad, que consiste en sustituir el término definido por su definición. Consideremos la definición del verbo "*comprar*":

Dañar. Maltratar o echar a perder algo

y tomemos una oración de ejemplo para aplicar la prueba de conmutabilidad:

Juan dañó el televisor = Juan *maltrató o echó a perder* el televisor

La frase "*maltratar o echar a perder*" tomada de la definición sustituye al término definido ("*dañar*"), y se denomina *contenido semántico*. Sin embargo, no se consideró en la sustitución la palabra "*algo*", debido a que se corresponde con el objeto directo representado en este caso por "*el televisor*". Este elemento, que alude a condiciones sintagmáticas, se denomina *contorno de la definición*.

Otras oraciones igualmente válidas pueden ser:

Juan dañó sus lentes
Juan dañó el mueble
Juan dañó las cortinas

Con estos ejemplos podemos observar que el contorno *algo* representa una categoría o clase de palabras muy amplia, las cuales satisfacen las restricciones que impone la clase, en este caso, *algo*, que denota seres inanimados.

Lo que podemos decir es que las especificaciones semánticas que el verbo requiere de sus argumentos son recogidas dentro del contorno de su definición. Dicho en otras palabras, el

contorno no representa otra cosa que en lo que ha dado en llamarse valencias o argumentos verbales, argumentos cuya indicación en la definición es imprescindible cuando deben satisfacer alguna característica o condición concreta ([31]).

En algunos casos, se puede apreciar el uso de pronombres indefinidos como *algo* y *alguien*, que designan clases de palabras muy generales, como los utilizados como elemento del contorno. En otros casos, aparecen nombres comunes que refieren palabras con características semánticas ya muy específicas. Esto puedo apreciarse en la siguiente definición:

Sainar. Engordar a los animales.

El verbo *sainar* requiere como argumento palabras que satisfagan la restricción semántica denotada por el contorno *animal*.

3.2.4. Microestructura en el DRAE

La microestructura de un diccionario se define como los parámetros que se siguen para construir la información de cada artículo lexicográfico.

En general se considera que cada entrada debe ser acompañada de la siguiente disposición de elementos:

1) Información fonológica.
2) Información morfosintáctica.
3) Información semántica.
4) Índice de registro.

En el DRAE, la estructura general de los artículos se conforma por la unidad léxica, seguida algunas veces por la información etimológica, y finalmente la acepción o acepciones numeradas.

Como fue mencionado anteriormente, en caso de que exista homonimia en unidades léxicas, éstas se diferenciarán unas de otras por un superíndice colocado al final de la entrada.

Cuando se tiene información sobre la **etimología** de las entradas, ésta se colocará por delante y entre paréntesis.

La numeración de las acepciones se dará de acuerdo a la categoría gramatical de la entrada a la que corresponden, sujetándose a la frecuencia de uso de la acepción, es decir, la variante más utilizada se colocará al inicio. En verbos, que es a lo que nosotros atañe, el orden a seguir será disponiendo primero las acepciones con marca *transitiva*, seguidas de las *intransitivas*, y al final las *pronominales*.

Enseguida, las diversas **marcas gramaticales** que acompañan a la acepción, se ordenarán bajo los siguientes criterios, mostrando el siguiente orden:

1) Acepciones con marcas de niveles de lengua (“cult.”, “vulg.”, etc.) o registros de habla (“coloq.”).
2) Acepciones con marcas técnicas (“Ling.”, “Mar.”, etc.).
3) Acepciones con marcas geográficas (“Am. Mer.”, “Arg.”, “Méx.”, etc.).
4) Acepciones con marcas cronológicas (“desus.”, “ant.”, etc.).
5) Acepciones con cualquier otro tipo de marca, las cuales no tendrán una colocación fija (intensión del hablante: “despec.”, “irón.”; valoración con respecto al mensaje (“malson.”, “eufem.”, etc.).

El **texto definitorio** que es posible encontrar puede ser de tipo perifrástico, impropio y sinonímico. Este último es un recurso consiste en disponer como definición de una entrada, la unidad léxica de la cual se obtiene el significado. Cuando esta unidad consta de varias acepciones y el significado lo proporciona sólo una de ellas, se utiliza la llamada definición por remisión. Esta consiste en agregar al vocablo dispuesto como sinónimo un fragmento de la acepción que proporciona el significado a la que se debe remitir. El fragmento es colocado entre paréntesis encabezados por doble barra vertical. Por ejemplo:

Adumbrar: tr. Pint. Sombrear (|| poner sombra en un dibujo).

El **contorno** en las definiciones algunas veces es señalado explícitamente cuando hace mención al sujeto o cuando el verbo se emplea en un contexto situacional específico. Para el

primero de los casos se utiliza la fórmula *Dicho de,* y para el segundo *En* o *Entre,* ambos casos encabezando el texto definitorio. Por ejemplo:

Aclamar: tr. Dicho de la multitud: Dar voces en honor y aplauso de alguien.

Enlizar: tr. Entre tejedores, añadir lizos al telar.

Al final del texto definitorio, pueden emplearse las llamadas **notas de uso** para complementar la información proporcionada por las marcas que encabezan las acepciones. Estas indican los diferentes usos que puede recibir la unidad léxica definida, por ejemplo la zona geográfica donde tiene mayor uso, una utilización gramatical diferente, el sentido literal o figurado con el que puede emplearse, etc. Se indica a través de abreviaciones iniciando por la forma *U.*, como puede apreciarse en el siguiente ejemplo:

Abanicar. tr. Hacer aire con el abanico. U. m. c. prnl.

Sólo en algunos casos, después de la nota de uso se incluye un **ejemplo** para mostrar el uso de la unidad léxica definida. Dentro de este texto llega a indicarse con letra cursiva y mayúscula el régimen preposicional que acompaña al verbo. Por ejemplo:

Acordar. tr. Recordar (|| traer a la memoria). U. m. c. prnl. Acordarse *DE* un hijo ausente.

CAPÍTULO 4

MÉTODO PROPUESTO

En este trabajo, como se explicó anteriormente, se propone el uso del diccionario explicativo para su procesamiento, empleando una serie de heurísticas basadas en observaciones a priori de la naturaleza y comportamiento de los datos contenidos en las definiciones lexicográficas para la identificación de la valencia verbal.

En la siguiente gráfica mostramos los principales pasos que se siguieron para alcanzar este objetivo, y más adelante explicamos a detalle cada uno de ellos.

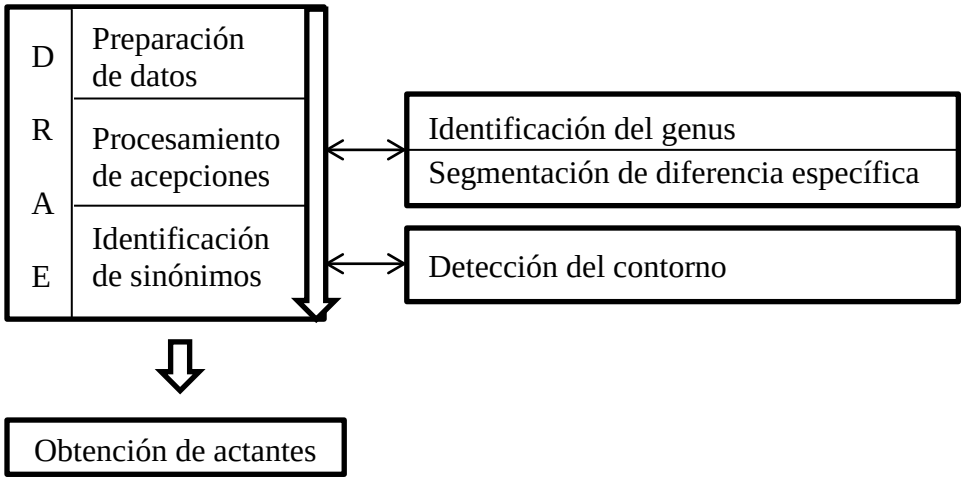


Ilustración 4.1 Principales pasos seguidos en el procesamiento del diccionario

4.1. Pre-procesamiento del diccionario

Antes de aplicar algún tipo de procesamiento automático al diccionario, es necesario manipularlo y transformarlo a un formato donde la información sea más fácilmente accesible, de lo contrario los datos podrían conducirnos a la extracción de patrones y/o reglas poco confiables. Esta tarea la dividimos en dos etapas: primero, filtrar la información que contiene

el diccionario, y segundo, convertir esta información a un formato que nos sea posible manipular.

4.1.1. Filtrado de información

El filtrado de información consiste en seleccionar sólo los datos que nos son relevantes, es decir, discriminamos del diccionario artículos lexicográficos de acuerdo a la categoría gramatical que presentan las entradas. Dado que sólo nos interesa trabajar con verbos, nos apoyaremos en la marcación gramatical que el DRAE ofrece para extraer los artículos de nuestro interés.

No existe una marca que indique explícitamente la categoría “verbo”, pero sí la que nos indica la transitividad de la entrada en cada una de sus acepciones. Siendo esta información exclusiva de los verbos, es la que utilizamos para distinguir y extraer los artículos.

El manejo del diccionario bajo estos criterios, nos permite medir a grandes rasgos su composición, mostrándose ésta en la siguiente tabla:

Elemento evaluado	Frecuencia
Unidades léxicas (UL)	89, 799
Acepciones de UL	162, 362
Unidades léxicas verbales (ULV)	12, 008
Acepciones de ULV	27, 668

Tabla 4.1 Macrostructura de DRAE.

Es de interés la manera en que se distribuyen las acepciones en las entradas verbales. Se aprecia que en promedio existen poco más de dos acepciones por entrada verbal, lo que choca con nuestro sentido común, pues fácilmente podemos asignar más de dos acepciones a la mayoría de verbos que utilizamos con regularidad en nuestra habla cotidiana. La explicación puede darse observando la siguiente gráfica (que por cierto recuerda la distribución de la ley de Zipf) donde se mide la frecuencia de verbos de acuerdo al número de

acepciones que emplean. En la gráfica encontramos que tres cuartas partes de los verbos manejan de una a dos acepciones. De estos, la gran mayoría son utilizados en un contexto situacional muy específico, y por lo tanto, de un uso más restringido.

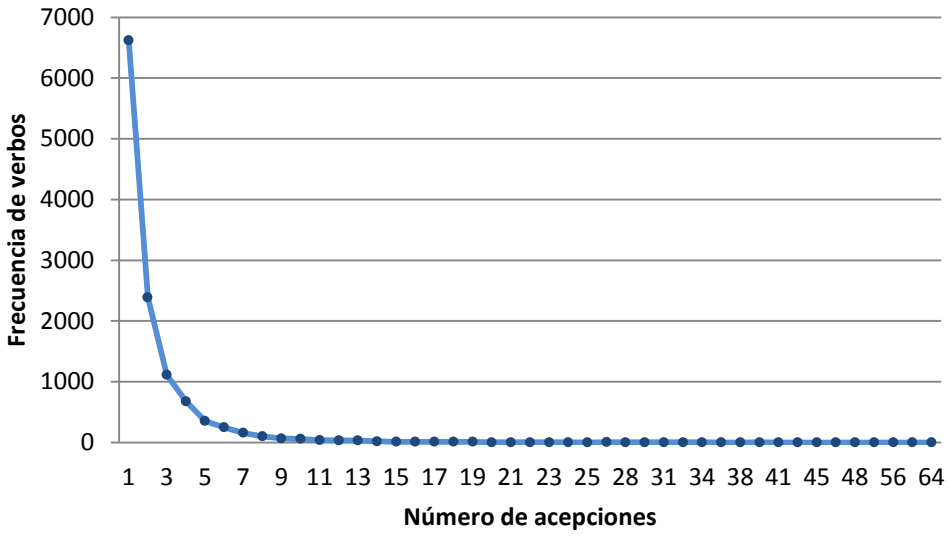


Ilustración 4.2 Distribución de acepciones en las entradas verbales

Consideramos dividir la estructura de los artículos lexicográficos en tres partes:

- Unidad léxica.
- Definición.
- Marcas gramaticales y resto de información (etimología, ejemplos, etc.).

Tener el artículo lexicográfico estructurado de esta manera nos permitirá aplicar un procesamiento especializado según el grupo que estemos manipulando.

4.1.2. Etiquetado gramatical

El siguiente paso consiste en obtener la información gramatical de cada palabra presente en las definiciones, lo que nos permitirá más adelante generar heurísticas para manipular estos datos. Esta tarea la realizamos haciendo uso del parseador Freeling, una herramienta de análisis de texto de código abierto para varios idiomas, incluido el español.

Información sin etiquetar	Etiquetado Freeling
1. Sujetar con badernas.	1 1 Z . . Fp Sujetar sujetar VMN0000 con con SPS00 badernas baderna NCFP000 . . Fp

Tabla 4.2 Oración de entrada y resultado de etiquetado POS de Freeling.

Freeling utiliza una serie de etiquetas para representar la información morfológica, las cuales están basadas en las etiquetas manejadas por el grupo Eagles, propuestas como estándar para el manejo y evaluación de recursos lingüísticos (como lexicones y corpus).

En la tabla anterior se muestra un ejemplo de la información antes y después de ser procesada por Freeling: en la primera columna se encuentra el texto de entrada, o acepción que previamente preparamos en la etapa de filtrado, y en la segunda columna se muestra la salida o etiquetado que devuelve el parseador.

Cada renglón de la segunda columna se corresponde con una palabra de la oración de entrada. Consideramos como palabra a la sucesión de caracteres ininterrumpidos separados por espacios en blanco, excepto signos de puntuación, los cuales se consideran por sí mismos como palabras.

La etiqueta que devuelve Freeling mostrando la información morfológica tiene el siguiente formato:

Forma	Lema	Gramática
badernas	baderna	NCFP000

Tabla 4.3 Formato de salida del etiquetado POS de Freeling.

La primera columna corresponde a la forma de la palabra, es decir, la palabra tal como se encuentra en la oración de entrada. La segunda indica el lema de dicha palabra, y la última proporciona la información gramatical.

En el formato que sigue la etiqueta que proporciona la información morfológica, la primera posición o carácter indica la categoría gramatical, la segunda posición indica el tipo de categoría, y del tercero en adelante otros atributos que no siempre están presentes según la categoría a la que pertenezcan. Por ejemplo, para la palabra *badernas* tenemos:

Categoría	**Tipo**	**Género**	**Número**	**Clasificación semántica**		**Grado**
N	C	F	P	0	0	0
Nombre	Común	Femenino	Plural	Sin información		

Tabla 4.4 Descripción de la etiqueta morfológica.

4.2. Procesamiento de acepciones

Lo visto hasta ahora nos muestra que las acepciones se estructuran de una manera uniforme, lo que permite utilizar heurísticas sencillas para procesar automáticamente cada uno de los elementos que las conforman.

La programación del método de procesamiento se ha realizado con el lenguaje Ruby, un lenguaje orientado a objetos de propósito general, que combina características de Perl, Smalltalk, Eiffel, Ada y Lisp. Ruby ofrece una gran cantidad de funciones para el procesamiento de texto, además de ser muy sencilla su implementación y ofrecer una sintaxis amigable que facilita el mantenimiento de código.

En primer lugar, se han extraído y almacenado los elementos que conforman cada uno de los artículos lexicográficos. Siendo la definición el elemento que más interesa, atendiendo las fórmulas que sigue el DRAE, ha sido posible separarla de las marcas gramaticales, notas de uso y demás datos que acompañan la acepción.

4.2.1. Identificación del genus y la diferencia específica

El siguiente paso consiste en identificar el genus de la diferencia específica. Considerando que el tipo de definición utilizada es la definición aristotélica, la separación entre uno y otro elemento se ha realizado extrayendo las palabras de categoría verbo que inician la definición, teniendo como remanente la diferencia específica.

Al realizar un análisis de tipo manual de una muestra aleatoria de las definiciones, encontramos diferentes maneras en que se constituye el genus, lo cual puede resumirse en lo siguiente:

1) Usando verbos individuales:
   a) Con un solo verbo. Ejemplo:

      Cotizar. *Pagar una cuota.*

   b) Con dos o más verbos enlazados por conjunciones y/o disyunciones. Ejemplo:

      Armonizar. *Escoger y escribir los acordes correspondientes a una melodía.*

      Aballar. *Amortiguar, desvanecer o esfumar las líneas y colores de una pintura.*

2) Como cláusula subordinada en infinitivo cumpliendo la función de complemento directo. Ejemplo:

   Gallear. *Pretender sobresalir entre otros con presunción o jactancia.*

3) Como Función Léxica. Ejemplo:

   Anunciar. *Dar publicidad a algo con fines de propaganda comercial.*

Cada caso particular requiere un tratamiento diferente que permita su correcta identificación. En el caso 1 y 2, todo verbo existente como cabecera de la definición se considera genus de la UL definida. En 3) se requiere un procesamiento más complejo: los verbos que vienen acompañados por un sustantivo son Funciones Léxicas (FL) potenciales. Las FL se definen ([20]) como una función que asocia una palabra denominada “base”, la cual aporta su significado literal a la expresión, a otra llamada “colocador”, que adquiere un significado diferente de su significado típico, de tal manera que el significado del conjunto

incluye el significado de una de las palabras (base), pero no del otro (colocador). De esta manera, el genus en una definición que es encabezada por una FL no puede ser el colocador.

El método que utilizamos para procesar este tipo de definiciones consiste en identificar los pares de palabras "verbo – nombre común" que encabezan las definiciones y buscar algún verbo que comparta la misma raíz que el sustantivo. De existir tal verbo, éste sustituye al par "verbo – sustantivo" y por lo tanto es tomado como genus de la definición. Con este método se realizaron alrededor de 800 sustituciones de pares "verbo – nombre común" por verbo.

Posibles funciones léxicas	Verbo usado como genus
Tener ansiedad por	Ansiar
Hacer alarde de	Alardear
Dar alojamiento a	Alojar
Echar chispas	Chispear
Dar claridad	Clarear
Tener dominio	Dominar
Causar embriaguez	Embriagar
Hacer ondas en	Ondear
Causar ardor	Arder
Hacer esclavo	Esclavizar

Tabla 4.5 Ejemplos de funciones léxicas usadas como genus en definiciones.

Siendo posible identificar el genus en la definición, el resto de elementos que la constituyen automáticamente son tomados como parte de la diferencia específica.

4.3. Desarrollo de una gramática para la segmentación de las definiciones

Como se mostró anteriormente, las categorías gramaticales bajo las que podemos encontrar al contorno de la definición pueden ser pronombres indefinidos y nombres comunes.

La identificación de las palabras categorizadas de esta manera no sería suficiente para lograr una completa identificación del contorno, es decir, sería importante también capturar el

contexto sintáctico que delimita cada elemento del contorno, lo que ayudaría a conocer por ejemplo las preposiciones con las que pueden acompañarse. El algoritmo desarrollado para lograr esta meta se basa en una serie de reglas que reflejan la estructura básica de las definiciones, más concretamente, de la diferencia específica, lo que permite capturar fragmentos de las definiciones las cuales incluyen un solo candidato a contorno de la definición.

Símbolo utilizado	Significado
S	Símbolo inicial
Cont	Contorno
Nuc	Núcleo del contorno (PI ó NC)
EleIzq	Elementos a la izquierda
EleDer	Elementos a la derecha
EI	Elemento izquierdo
ED	Elemento derecho
DA, DI, DP, DD, CS, RG, AQ, RN, CC, FC, SP	Etiquetas utilizadas en el formato EAGLES sobre información morfosintáctica de las palabras, (ver significados en Apéndice)

Tabla 4.7 Significado de los símbolos usados en las gramáticas

Las reglas quedan definidas de la siguiente manera:

1. La nomenclatura utilizada se define en la tabla anterior.
2. El lado izquierdo de la primera producción, es el símbolo inicial
3. Las etiquetas utilizadas en el formato EAGLES se considerarán como símbolos terminales
4. Reglas:
    - 4.1. S → Cont
    - 4.2. Cont → ECont | ECont Cont
    - 4.3. ECont → Nuc | EleIzq Nuc | EleIzq Nuc EleDer | Nuc EleDer | ECont Liga ECont
    - 4.4. EleIzq → EI | EI EleIzq
    - 4.5. EleDer → ED | ED EleDer
    - 4.6. Nuc → PI | NC
    - 4.7. EI → DA | DI | DP | DD | SP | CS | RG | Z | AQ

4.8. ED → AQ | RN
4.9. Liga → CC | FC

Estas reglas no se utilizan en la producción de oraciones (pues podrían generar oraciones incoherentes como un nombre común acompañado por una sucesión ininterrumpida de preposiciones), sino en la segmentación de las definiciones, donde cada segmento está conformado por un único candidato a elemento del contorno.

4.3.1. Ejemplo de aplicación de la gramática en una definición

Consideremos la definición del verbo "poner" en su primer sentido y la secuencia de pasos para ejemplificar el funcionamiento de la gramática:

Poner (1): Colocar en un lugar a alguien o algo.

1. Etiquetación de la definición:
   colocar colocar VMN0000 1
   en en SPS00 1
   un uno DI0MS0 0.986987
   lugar lugar NCMS000 1
   a a SPS00 0.99585
   alguien alguien PI0CS000 1
   o o CC 0.998845
   algo algo PI0CS000 0.896341
   . . Fp 1
2. El análisis de la oración se realiza sobre cada palabra respetando el orden de ésta en la oración, es decir de izquierda a derecha, tomando los dos primeros caracteres de cada etiqueta asignada a cada palabra. En nuestro ejemplo, se analizarán las siguientes etiquetas en el orden de aparición:

   VM - SP - DI - NC - SP - PI - CC - PI - FP
3. Se utilizarán estructuras de tipo *cola* para almacenar sucesiones ininterrumpidas de etiquetas que existan como símbolos terminales de la gramática.

   VM: no existe como símbolo terminal, se ignora

SP: existe como símbolo terminal, se almacena en una cola

DI: existe como símbolo terminal, se almacena en la misma cola

SP: existe como símbolo terminal, se almacena en la misma cola

PI: existe como símbolo terminal, se almacena en la misma cola

CC: existe como símbolo terminal, se almacena en la misma cola

PI: existe como símbolo terminal, se almacena en la misma cola

FP: no existe como símbolo terminal, se ignora

4. Se empiezan a extraer los elementos de las estructuras sustituyendo cada uno de manera individual y también cada sucesión ininterrumpida por sus respectivas cabezas de reglas donde empatan (en una cola, el primer elemento en entrar es el primero en salir, por lo tanto el procesamiento se realiza respetando el orden en el que ingresaron los elementos). En la siguiente tabla mostramos el flujo de procesamiento en orden descendente, indicando en cada renglón el número de regla que aplica:

en	**un**	**lugar**	**a**	**alguien**	**o**	**algo**
SP	DI	NC	SP	PI	CC	PI
4.7 EI	4.7 EI	4.6 Nuc	4.7 EI	4.6 Nuc	4.9 Liga	4.6 Nuc
4.7 EI	4.4 EleIzq	4.6 Nuc	4.4 EleIzq	4.6 Nuc	4.8 Liga	4.5 Nuc
4.4 EleIzq		4.6 Nuc	4.3 ECont		4.8 Liga	4.2 ECont
4.3 ECont			4.3 ECont			
4.3 ECont			4.2 Cont			
4.2 Cont						

Tabla 4.7 Secuencia de fragmentación de una definición usando la gramática propuesta

5. Se obtiene el resultado: *en un lugar | a alguien o algo*

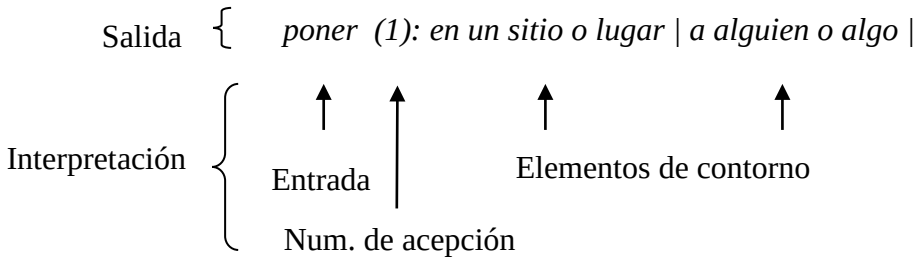
4.4. Obtención de actantes

Al estar debidamente identificada la información que contienen los artículos lexicográficos, es posible dar inicio a la extracción del contorno atendiendo únicamente el fragmento de información que los contiene, esto es, a la diferencia específica.

El proceso consiste en implementar un algoritmo para extraer los pronombres indefinidos y nombres comunes, además de otras palabras con las que se establezcan relaciones de dependencia, esto es, determinantes, preposiciones, adjetivos calificativos, etc. El algoritmo recibirá como entrada las unidades léxicas con los números de las acepciones de las que se desea extraer el contorno. Por ejemplo, sea el siguiente artículo lexicográfico:

Poner. Colocar en un sitio o lugar a alguien o algo.

El algoritmo recibe como entrada, por ejemplo, el verbo "Poner" en su número de acepción "1". Después de identificar y separar el genus de la diferencia específica, ésta última a su vez se procesa y como resultado se obtiene lo siguiente:



Cada elemento del contorno, se considera como actante semántico del verbo, y, como se vio anteriormente, informa sobre las características semánticas de éste.

Las definiciones existentes en el DRAE no siempre incluyen un contorno completo en las definiciones. Es decir, en la mayoría de los casos el contorno que se indica es incompleto. Esto lo podemos observar en el siguiente par de ejemplos, donde se aprecia que el contorno referente a algún complemento directo no está indicado:

Tapizar. *Cubrir con tapices.*
Debería ser**:** *Cubrir* ***algo*** *con tapices*

Conducir. *Llevar, transportar de una parte a otra.*
Debería ser*: Llevar, transportar* ***algo o alguien*** *de una parte a otra*

La solución que implementamos consiste en combinar *de alguna manera* las definiciones entre diferentes verbos con la idea de aumentar la probabilidad de obtener el

contorno completo para cada verbo. En primera instancia surgen las preguntas, ¿qué asepctos se deben considerar para asegurar que se están combinando las definiciones semánticamente apropiadas? ¿Cuándo podemos saber si una definición se encuentra o no incompleta? Estas preguntas las resolveremos en los siguientes capítulos.

CAPÍTULO 5

PROCESAMIENTO DE SINÓNIMOS

Para redactar las definiciones de verbos, probablemente los lexicógrafos no toman un criterio unificado sobre el uso o no del contorno asociado a los verbos, ni sobre el número de elementos del contorno que deban utilizarse en las definiciones. Es decir, nos encontramos en el diccionario con definiciones que aportan mayor información en este rubro, que otras. Sea por ejemplo, la siguiente definición:

Conducir. *Llevar, transportar de una parte a otra.*

En ella se aprecia la ausencia del objeto directo. Y como esta definición, encontramos tantas otras más que saltarán a la vista con solo abrir el diccionario en cualquier página.

Lo que hemos propuesto es utilizar las definiciones de otros verbos para complementar la información faltante en casos donde sea necesario. En primer lugar, esta selección de verbos no se realiza de manera aleatoria, sino que se basa en las relaciones semánticas dadas entre verbos, en particular la sinonimia y las relaciones de inclusión.

La razón por la cual atendemos las relaciones sinonímicas es la siguiente: si es cierto que es cuestionable la existencia de sinónimos absolutos en la lengua, al menos queda claro que la existencia de sinónimos relativos es aceptada. Esto significaría que dos o más verbos son sinónimos siempre y cuando puedan ser sustituidos entre sí en al menos un sentido a los que puedan referir.

Consideremos el uso de los verbos “llevar”, “conducir” y “transportar” en la siguiente oración:

Los mayas _____ anfibios vivos de una localidad a otra con propósitos ceremoniales

En este contexto el uso de cualquiera de los verbos sobre la línea conservaría la oración en un mismo sentido semánticamente correcto.

a. “Los mayas *llevaron* anfibios vivos de una localidad a otra con propósitos ceremoniales”.
b. “Los mayas *condujeron* anfibios vivos de una localidad a otra con propósitos ceremoniales”.
c. “Los mayas *transportaron* anfibios vivos de una localidad a otra con propósitos ceremoniales”.

Si estos verbos pueden sustituirse mutuamente sin alterar el significado de la oración, entonces podríamos concluir que los sinónimos deberían cumplir los siguientes dos supuestos:

1) El número de actantes de cada verbo es el mismo para cada uno de sus sinónimos (en al menos un sentido).
2) Las restricciones semánticas que un verbo impone a sus actantes, son las mismas que las que el resto de sus sinónimos impondría (en al menos un sentido).

De cumplirse los puntos previos, permitiría subsanar en la medida de lo posible la falta de información referente al contorno que suele existir en las definiciones de verbos en el diccionario de la RAE, combinando el contorno de las definiciones que aparecen en un conjunto de sinónimos (mas adelante explicaremos la manera de lograrlo).

5.1. Uso de definiciones sinonímicas en el diccionario

La identificación de los verbos relacionados entre sí por sinonimia, no resulta ser tan complicada debido a que el diccionario de la RAE llega a utilizar la llamada “definición sinonímica”, la cual consiste en utilizar como genus de la definición una o varias palabras con la misma categoría gramatical que la UL definida. Por ejemplo, el verbo “considerar” en su sentido 2 se define como:

Considerar (2): Juzgar, estimar

Lo que significa que el significado de “considerar” puede encontrarse en la definición de los verbos “juzgar” o “estimar”. Este tipo de definiciones puede provocar círculos viciosos, lo cual es considerado como un defecto por los lexicógrafos, pero es algo que beneficia a nuestra tarea. Un ejemplo de círculo vicioso es el conformado por los verbos “coger”, “asir”, “agarrar” y “tomar”, mostrado en la siguiente gráfica. El inicio de cada flecha indica la UL definida, y el nodo al que apunta la UL que se utiliza como sinónimo en su definición.

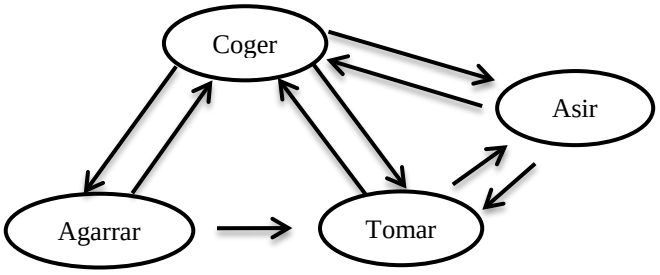


Ilustración 5.1 Ejemplo de círculo vicioso en el diccionario.

Las definiciones que componen cada verbo de la figura previa, son las siguientes:

- ***Coger***. *Asir, agarrar o tomar*
- ***Agarrar***. *Coger, tomar.*
- ***Tomar***. *Coger o asir con la mano algo.*
- ***Asir***. *Tomar o coger con la mano, y, en general, tomar, coger, prender.*

También es posible identificar círculos viciosos tomando únicamente el genus de cada verbo, sin considerar que la definición sea de tipo sinonímica. En [7] desarrollamos un grafo dirigido a través de las relaciones de hiponimia/hiperonimia dadas entre la UL y el genus de su definición. Demostramos que los círculos viciosos que se forman en el grafo están constituidos por sinónimos.

5.2. Identificación de los sentidos de verbos en las relaciones de sinonimia

Identificar qué verbos son utilizados como sinónimos, no es suficiente, pues se debe sobre todo distinguir en qué sentido en específico se logra la relación de sinonimia. Por

ejemplo, el verbo “abatir” en el sentido 6 incluye como sinónimos en su definición los verbos “desarmar” y “descomponer”. Ambos verbos disponen de varios sentidos, de entre los cuales es necesario distinguir cuáles son los que los relacionan como sinónimos. La solución que en este trabajo se implementó consiste en buscar en las definiciones algún hiperónimo común a los verbos, lo que indicaría que existe relación semántica en ese sentido en específico. Sean por ejemplo los hiperónimos de los verbos “desarmar” y “descomponer”:

Num. sentido	Hiperónimo
1	Quitar, hacer entregar
2	Desnudar o desceñir
3	Reducir
4	Dejar
5	Desunir, separar
…	…

Tabla 5.1 Genus de los primeros 5 sentidos del verbo “desarmar”

En la tabla anterior y en la siguiente, se muestran los hiperónimos de los primeros 5 sentidos de los verbos “desarmar” y “descomponer”, respectivamente. Como sabemos que ambos verbos son sinónimos (por la definición de “abatir” en el sentido 6) y además el sentido 5 de “desarmar” y el sentido 2 de “descomponer” comparten el mismo hiperónimo, entonces concluimos que la relación de sinonimia entre ambos se da en esos sentidos en específico.

Num. sentido	Hiperónimo
1	Desordenar y desbaratar
2	Separar
3	Indisponer
4	Averiar, estropear, deteriorar
5	Corromperse
…	…

Tabla 5.2 Primeros 5 sentidos del verbo “descomponer”

5.3. Combinación de información de las definiciones de sinónimos

Cuando se han identificado los sentidos relacionados semánticamente, pueden ahora combinarse los contornos de las definiciones para complementar la información faltante que exista en algunas de ellas. La ausencia de información puede darse de las siguientes maneras:

1) *No existe información alguna del contorno en alguna definición, pero sí en las otras.* Considerando las definiciones de los verbos "coger" y "tomar", observamos que la definición del verbo "coger" sólo incluye sinónimos, sin hacer mención alguna del contorno. Sin embargo, la definición del verbo "tomar" incluye dicha información. El resultado de la obtención de segmentos:

Tomar. *Coger o asir con la mano algo*

Segmentación*: con la mano | algo*

Coger. *Asir, agarrar o tomar*

Segmentación: -

Por lo tanto, el contorno del verbo "tomar" se considerará también perteneciente al verbo "coger".

2) *Algunas definiciones incluyen segmentos que no pertenecen al contorno.* Este es el caso más común, y es complicado lograr una correcta discriminación de segmentos. Por ejemplo:

Llevar. *Conducir algo desde un lugar a otro alejado de aquel en que se habla o se sitúa mentalmente la persona que emplea este verbo.*

Segmentación*: algo | desde un lugar | a otro | mentalmente la persona | este verbo*

En esta definición, los segmentos "mentalmente la persona" y "este verbo", no son elementos que puedan considerarse parte del contorno.

3) Algunas definiciones mencionan el contorno pero éste no abarca la totalidad de entidades que lo pueden conformar. Consideremos, además de la definición del verbo "llevar", las definiciones de los siguientes verbos:

Conducir*: Llevar, transportar de una parte a otra*

***Transportar**: Llevar a alguien o algo de un lugar a otro*

Atendiendo el objeto directo en las definiciones, se observa que se menciona en la definición de "llevar" ("algo") y que éste cumple la restricción semántica de "entidad inanimada". Por otro lado, el verbo "conducir" no especifica un objeto directo y "transportar" lo amplía considerando también a seres humanos ("alguien o algo"). Podemos considerar entonces que el verbo "llevar" debe reducir la restricción semántica del objeto directo al punto de abarcar también a seres humanos ("conducir *a alguien o* algo desde un lugar a otro") y que el verbo "conducir" debe incluirlo en su definición.

Lo ideal sería trabajar con conjuntos de sinónimos suficientemente grandes para aumentar la probabilidad de obtener el número correcto de actantes de las unidades léxicas. Esto finalmente lo hemos logrado uniendo grupos de sinónimos que tienen intersecciones de verbos. Los detalles de este proceso los discutimos en el siguiente capítulo.

CAPÍTULO 6

Obtención de resultados

6.1. Medición de los grupos de sinónimos identificados

El procesamiento de todas las definiciones de verbos encontramos poco más de 6000 definiciones sinonímicas. Estas definiciones se procesaron para identificar si existía algún genus común a las definiciones de los verbos agrupados y así precisar el número del sentido en que se relacionaban. Esto llevó a la identificación de un aproximado de 6500 grupos de sinónimos en donde se identificaron explícitamente los sentidos.

Por ejemplo, el verbo “amparar” en su sentido 4 se define como: “Defenderse, guarecerse”. Estos verbos usados en la definición, ambos en su sentido 2, se definen como:

***Defender (2)**: Mantener, conservar, sostener algo contra el dictamen ajeno.*
***Guarecer (2)**: Guardar, conservar y asegurar algo*

Ambas definiciones comparten el verbo conservar, por lo que en ese sentido en particular conforman un grupo de sinónimos con sentido identificado. Sin embargo, observamos también que defender en su sentido 1 y guarecer en su sentido 4 se definen como:

***Defender (1)**: Amparar, librar, proteger*
***Guarecer (4)**: Socorrer, amparar, ayudar.*

Conformarían otro grupo en dichos sentidos bajo el verbo amparar. Del total de 6 000 grupos de sinónimos, en 3000 agrupaciones no se lograron identificar los sentidos que relacionaban a los verbos siguiendo el criterio del genus común.

Elemento evaluado	Cantidad
Definiciones sinonímicas	6, 000
Grupos de sinónimos con sentidos de verbos identificados	6, 500
Grupos de sinónimos donde no se identificaron los sentidos de verbos	3, 000

Tabla 6.1 Medición de los grupos de sinónimos identificaodos

6.2. Unión de grupos de sinónimos

Varios grupos de sinónimos incluyen el mismo sentido de algún verbo. Al existir intersección entre ellos, podemos proceder a la unión de grupos, y así complementar de manera más precisa la información de los diferentes verbos y sobre todo de su contorno.

Por ejemplo, consideremos el siguiente grupo de sinónimos tomados de la definición del verbo “maliciar” en su primer sentido:

***Maliciar (1)**: Recelar, sospechar, presumir algo con malicia*

Los verbos “recelar” y “sospechar” coinciden en usar el mismo genus en sus sentidos 1 y 2 respectivamente:

***Recelar (1):** Temer, desconfiar y sospechar*
***Sospechar (2)**: Desconfiar, dudar, recelar de alguien*

Combinamos las definiciones de ambos verbos en los sentidos antes indicados y el contorno resultante es “de alguien”.

Por otro lado, “recelar” y “dudar” son también sinónimos según el segundo sentido de “sospechar”. Ambos verbos son definidos en los sentidos abajo indicados, también bajo el genus “desconfiar”:

***Recelar (1)**: Temer, desconfiar y sospechar*
***Dudar (2)**: Desconfiar, sospechar de alguien o algo*

El genus obtenido para este grupo de verbos se conformaría por la expresión “de alguien o algo”. Podemos unir ambos grupos ya que ambos incluyen el verbo “recelar” en su sentido primero, lo que nos llevaría entonces a combinar los contornos de ambos grupos.

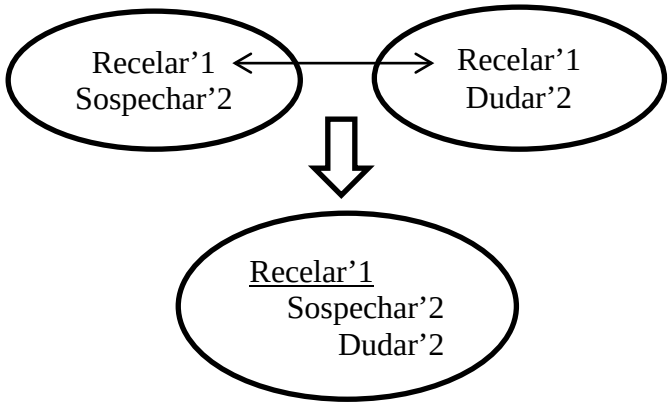


Ilustración 6.1 Representación de la unión de dos grupos de sinónimos en un nuevo conjunto.

En nuestra heurística consideramos que “los elementos de contorno que se encuentren incluidos en otros (“algo” desaparece porque existe un elementos más complejo que ya lo incluye: “de alguien o *algo*”) se eliminarán”, por lo que el contorno para “recelar” en su primer sentido se manifiesta como “de alguien o algo”.

6.3. Identificación del contorno de la definición

Considerando que no todos los sustantivos comunes y pronombres indefinidos que aparecen en una definición pueden ser catalogados como elementos del contorno (ver apartado 5.3), decidimos procesar aquellas definiciones cuyos candidatos a elementos del contorno estuvieran conformados únicamente por los pronombres indefinidos “algo, alguien”, y los sustantivos comunes “cosa, persona, animal, lugar” y “parte”, ya que al realizar una medición de las categorías gramaticales de palabras funcionales más frecuentemente utilizadas en las definiciones, las palabras antes mencionadas tuvieron mayor presencia.

Palabra	Frecuencia
Algo	3000
Alguien	2000
Otro	900
Cosa	800
Parte	500
Persona	400
Lugar	350
Cuerpo, acción, fuerza, agua, tierra, …	< 300

Tabla 6.2 Frecuencia de las palabras más utilizadas como elementos del contorno

Por otro lado, estas palabras representarían en cualquier ontología el nivel más alto o abstracto de los grupos que la componen. El procesamiento de estos datos nos arrojó un total de 420 grupos de sinónimos que contienen dichas palabras en sus definiciones.

Elemento evaluado	Cantidad
Grupos de sinónimos donde no se identificaron candidatos a contorno	500
Grupos de sinónimos con candidatos a contorno más abstractos	420

Tabla 6.3 Medición del contorno en los grupos de sinónimos

6.4. Análisis de resultados

Las intersecciones entre los grupos de sinónimos (según lo visto en 6.2) donde fue posible identificar candidatos a contorno más abstractos nos permitieron reunir un total de 397 conjuntos de grupos, cada uno formado en torno a un verbo en un sentido determinado.

Esto significa que los contornos procesados en cada conjunto correspondían al contorno del verbo en torno al cual se formaron. Dicho en otras palabras, con esta operación obtuvimos el contorno de 397 verbos en un sentido en particular.

La medición sobre la precisión de los resultados se realizó de manera manual dada la complejidad que existe para realizarlo automáticamente. Aun cuando se cuenta ya con un recurso donde se han obtenido 670 patrones de rección de 500 verbos en español recolectados manualmente ([4]), utilizarlo para validar los resultados de este trabajo resulta prácticamente inapropiado, pues dicho trabajo se tomaron sólo uno o dos sentidos por cada verbo que los autores consideraron los de uso más extendido, y en el trabajo que nosotros exponemos el procesamiento de los sentidos es guiado por los datos, es decir, no siempre existirá coincidencia entre sentidos a comparar. Y considerando que los patrones llegan a variar entre los diferentes sentidos de un verbo, la comparación no arrojaría resultados confiables.

Analizamos manualmente el total de patrones obtenidos, cotejando la información devuelta por nuestro método con la definición reportada en el diccionario para cada verbo en su sentido correspondiente. Los resultados obtenidos los exponemos en la siguiente tabla:

verbos/sentidos extraídos	Patrones evaluados	Patrones correctos	Patrones incorrectos	Porcentaje de acierto
397	397	336	61	84.63

Tabla 6.4 Análisis de resultados

El análisis de los patrones incorrectos arroja que estos pueden agruparse en las siguientes 4 categorías de errores:

- No se logró obtener suficiente información para identificar todos los actantes. Ejemplo:
  - Verbo: *Reunir'2*
  - Patrones obtenidos: *en el mismo lugar*
  - Observaciones: Ausencia de objeto directo
- El método no identifica que diferentes redacciones de un elemento del contorno pueden hacer referencia al mismo actante. Ejemplo:

    - Verbo: *Colocar'1*
    - Patrones obtenidos: *a alguien o algo* | *dentro_de otra cosa o dentro_de sus límites* | *en un lugar*
    - Observaciones: El elemento *dentro_de otra cosa o dentro_de sus límites* debería también considerarse como *en un lugar*.
- El actante obtenido no corresponde al actante sugerido en la definición. Ejemplo:
    - Verbo: *Hacer'52*
    - Patrones obtenidos: *de algo malo o perjudicial*
    - Observaciones: El actante correcto debería ser *un lugar*
- Se identifican más actantes de los que deberían existir. Ejemplo:
    - Verbo: *Abonar'4*
    - Patrones obtenidos: *por cierto y seguro algo - a alguien*
    - Observaciones: El elemento a alguien no se corresponde

CAPÍTULO 7

Recursos generados

Durante el desarrollo y como resultado de este trabajo de investigación, se generaron diversos recursos que dejamos a disposición de otros investigadores para su libre utilización. A continuación se listarán estos recursos y que consideramos son los de mayor importancia.

7.1. Listado de hipónimos-hiperónimos de verbos

Con las heurísticas implementadas, logramos separar e identificar los genus de las definiciones de los verbos mencionados en el diccionario de la RAE.

La información se encuentra en formato de archivo de texto plano ".txt" y bajo codificación UTF-8. Por cada verbo se generó un archivo. Cada línea del archivo se conforma por el par "número de sentido" y "genus" separados por una barra vertical. Por cada genus que conforme una misma definición, se agregan nuevas líneas con sus respectivos pares.

Por ejemplo, sean las siguientes definiciones que encontramos del verbo "contraer":

1. Estrechar, juntar algo con otra cosa.
2. Celebrar el contrato matrimonial.
3. Aplicar a un caso o a una proposición particular proposiciones o máximas generales.
4. Adquirir costumbres, vicios, enfermedades, resabios, deudas, etc.
5. Asumir obligaciones o compromisos.
6. Reducir el discurso a una idea, a un solo punto.
7. Reducirse a menor tamaño.

Ilustración 7.1 Definiciones del verbo "Contraer".

El procesamiento de esta información, nos devuelve un archivo de texto con el siguiente contenido:

contraer'1|estrechar
contraer'1|juntar
contraer'2|celebrar
contraer'3|aplicar
contraer'4|adquirir
contraer'5|asumir
contraer'6|reducir
contraer'7|reducir

Ilustración 7.2 Contenido del archivo generado con los genus de los sentidos del verbo “Contraer”.

Se observa como en las dos primeras líneas del archivo aparece dos veces el mismo sentido, pero cada uno con el respectivo genus que lo conforma.

7.2. Obtención de Funciones Léxicas

Muchos genus en las definiciones se encuentran conformados por una Función Léxica (FL). En estos casos, como se explicó en el capítulo 4.2.1, el genus no puede corresponderse con el colocador.

La solución propuesta consistió en identificar si algún candidato a genus venía acompañado por un nombre común. De ser así se consideraba como una FL potencial. Si el nombre común compartía su raíz con algún otro verbo, la FL potencial se sustituía por este verbo el cual se tomaba finalmente como genus en esa definición.

Las relaciones de nombres comunes y verbos compartiendo la misma raíz, se encuentran en un archivo de texto plano en formato UTF-8. El contenido de éste viene dado por una lista de verbos, seguidos por su raíz, y los nombres comunes donde también aparece.

En la siguiente ilustración mostramos un fragmento de este archivo.

abalear*abal>abaleador|abaleadura|abaleo|abalizamiento|abalorio
abanderar*abander>abanderado|abanderamiento
abandonar*abandon>abandonismo|abandono
abanicar*abanic>abanicazo|abanico
abaratar*abarat>abaratamiento
abarcar*abarc>abarca|abarcadura|abarcamiento
abarrotar*abarrot>abarrotamiento|abarrote
abastecer*abastec>abastecedor|abastecimiento

Ilustración 7.3 Fragmento del archivo de verbos y nombres comunes compartiendo la misma raíz.

Con esta información se generó otro archivo con las FL potenciales y los verbos por los cuales se sustituyeron:

finalizar
poner fin a
finalizar
dar fin a

Ilustración 7.4 FL encontradas para el verbo “finalizar”.

Aún es necesario refinar las heurísticas que generan el archivo con las FL potenciales, pues es posible encontrar algunos errores como los que se muestran a continuación:

amar
tener amor a
amar
decir amores
amar
inspirar amor

Ilustración 7.5 Algunas expresiones erróneamente tomadas como FL para el verbo “amar”.

La primera variante, que es correcta, se obtuvo directamente de la propia definición del verbo “amar” (“tener amor a”). Sin embargo, encontramos otras combinaciones de verbos con nombres comunes que no necesariamente comparten el significado del verbo “amar”. Por ejemplo, la variante dos (“decir amores”) la encontramos en el sentido segundo del verbo “enamorar”:

***Enamorar (2)**: Decir amores.*

Lo que podría hacer referencia a la emisión o expresión de mensajes de amor, que en primer lugar, no constituye la frase una FL, y en segundo, no necesariamente hace alusión a la acción de “amar”.

7.3. Diccionario de patrones

La identificación de los actantes de verbos, se dispusieron de una manera clara y ordenada en una Base de Datos (BD) apegándonos al formato de representación de patrones de rección.

La BD se implementó en SQLite, un sistema manejador de Base de Datos relacional, bajo el estándar SQL92, sin necesidad de una configuración previa para instalarse y compatible con los sistemas operativos Linux, Mac OS X y Windows. Toda la información relevante a este trabajo que se obtuvo del procesamiento del diccionario explicativo, será distribuida en las siguientes tablas, manejando la nomenclatura usada en la teoría base de este trabajo:

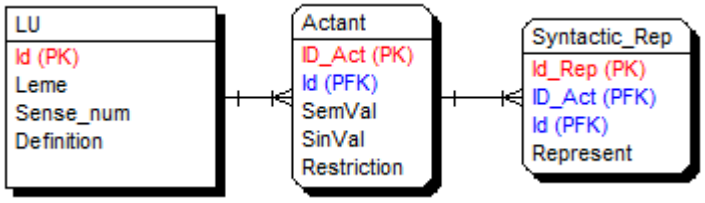


Ilustración 7.6 Esquema de la BD de Patrones de Rección

La BD se conforma por tres tablas, las que describimos a continuación

Tabla	Descripción
LU	Contiene los lemas, las definiciones de sus sentidos y el número de aparición que originalmente estos últimos tienen asignado.
Actant	Indica los actantes de cada Unidad Léxica. Muestra la variable de cada actante semántico (A, B, C, etc.), su correspondiente valor en el nivel sintáctico (1, 2, 3, etc.), y las restricciones semánticas que le corresponden (si denota una "persona", "cosa", "animal", etc.).
Syntactic_Rep	Indica las opciones de representación de los actantes.

Tabla 7.1 Listado de tablas de la Base de Datos de patrones

Con la información extraída del diccionario, todos los datos requeridos por las primeras tres tablas pueden ser asignados completamente. Sin embargo, dado que las definiciones **no proveen información sobre el agente o sujeto del verbo**, no se indica información sobre este actante.

Además de esta información, existen datos del nivel sintáctico que es imposible extraer del diccionario. En particular, la obligatoriedad de los actantes no puede ser determinada en este nivel de procesamiento, y las opciones de representación probablemente no se encuentran totalmente completadas.

Por ejemplo, consideremos los actantes correspondientes al verbo "conducir"

A alguien o algo | de un lugar | a otro lugar

Si a cada uno de estos actantes le asignamos una letra del abecedario sugiriendo su representación como un actante semántico (y omitiendo la letra "A", que correspondería al sujeto o agente), tenemos:

1. *A alguien o algo* = B
2. *De un lugar* = C
3. *A otro lugar* = D

Escribiendo con estos datos la definición semántica correspondiente para el verbo *conducir*, tendremos:

A (?) ***conduce*** *B de C a D*

Representando el patrón de rección del verbo, se obtiene lo siguiente:

A = 1 (?)	B = 2	C = 3	D = 4
?	2.1 a *N* 2.2 *N*	3.1 de *N*	4.1 a *N*
?	?	?	?

(1) $C_{2.1}$: N denota a una persona.
(2) $C_{2.2}$: N denota una cosa
(3) $C_{3.1}$: N denota un lugar
(4) $C_{4.1}$: N denota un lugar

Los datos de los cuales carecemos de información (denotados por el signo de interrogación) son:

a) Información semántico-sintáctica del primer actante (correspondiente al sujeto de la oración).
b) Tipo de aparición (opcional u obligatoria) de los actantes en el nivel sintáctico.

CAPÍTULO 9

Conclusiones

En este trabajo propusimos un método para la extracción de los actantes de verbos para el idioma español, basándonos en el análisis de las definiciones del diccionario de la Real Academia Española.

Dado que la redacción de los artículos lexicográficos se apega a estructuras bien establecidas, es posible crear heurísticas para el análisis y extracción de información de ellos. Cada uno de los elementos que conforman estas estructuras, aportó datos relevantes para el cumplimiento de los objetivos propuestos.

En particular, el contorno de las definiciones de los verbos, al indicar condiciones sintagmáticas del verbo y recoger las restricciones de tipo semántico que sus argumentos requieren, lo llegamos a considerar como imagen de la valencia verbal. Así, el extraer el contorno se traduce en la obtención de información sobre los actantes del verbo.

La falta de una especificación rigurosa del contorno en la mayoría de las definiciones de los verbos, imposibilita conocer de manera certera sus valencias. Sin embargo, encontramos un recurso para complementar esta escasa información apoyándonos en las definiciones de otros verbos. Esto se hizo atendiendo las relaciones léxicas de inclusión (hiperonimia/hiponimia) establecidas entre los genus y los artículos léxicos y las relaciones de sinonimia entre los verbos. A través de estas relaciones fue posible agrupar verbos según sus relaciones de sinonimia para más adelante complementar la información que en sus definiciones reportaban.

Tras un análisis, se confirmó que los verbos relacionados podían sustituirse mutuamente en cualquier contexto (considerando únicamente las acepciones implicadas). Con este resultado, fue posible afirmar que bajo estas condiciones existe una coincidencia en la

valencia verbal. Esta identificación de sinónimos nos ayudó a completar la lista de actantes de cada verbo complementando la información que cada definición manejaba.

9.1. Contribuciones

En este trabajo de investigación destacan las siguientes contribuciones:

1) Proponemos una solución que consiste en procesar el contorno de las definiciones, a partir del cual obtenemos específicamente el número de actantes, las restricciones semánticas que les son impuestos por los verbos y algunas opciones de representación de éstos a nivel sintáctico.
2) Demostramos que basándonos en las relaciones léxicas de inclusión y sinonimia, es posible complementar la información referente a los contornos y así tener una mayor probabilidad de identificar la valencia verbal.
3) Una Base de Datos de Patrones (BD), en donde se recoge la información sobre la valencia verbal a nivel semántico, la cual queda a disposición para su uso y consulta. Esta BD contiene para cada verbo extraído del diccionario, el número de actantes, las restricciones semánticas de éstos, y algunas opciones de representación a nivel sintáctico.

9.2. Sugerencias para trabajo futuro

El trabajo es muy extenso y aún es posible aplicar diversas nuevas heurísticas para procesar la información que al día de hoy se ha obtenido. Los puntos que consideramos de interés para retomar a futuro son los siguientes:

1. Está abierta la opción de aplicar una heurística para procesar aquellos grupos de verbos sinónimos donde los candidatos a contorno que se utilizan no pertenecen a los niveles más abstractos de una ontología
2. Otro punto de interés, relacionado con el anterior, es analizar la posible herencia de actantes entre verbos. Es decir, ya identificados los actantes de un verbo, estudiar de qué manera estos actantes se relacionan con los actantes de todos sus verbos hipónimos (o

bien, todos los verbos que utilizan al primero como genus en sus definiciones), y de esta manera proponer una heurística que complemente a la utilizada en este trabajo.

3. Analizar la manera de no sólo extraer elementos de contornos y combinarlos entre sí, sino también lograr mejorar de manera automática las definiciones de las entradas léxicas en los diccionarios, lo que resultaría en una herramienta de apoyo al lexicógrafo.
4. En los patrones de rección confluyen datos de tipo semántico y sintáctico. A través del procesamiento del diccionario explicativo, puede obtener la información semántica de los actantes. En suma, lo que es posible obtener es:
   a) Número de actantes que los verbos requieren.
   b) Restricción semántica de los actantes.

   Los aspectos sintácticos que aún faltan por obtenerse son:

   a) La obligatoriedad de los actantes.
   b) Completar las opciones de representación.

   El medio por el cual se podría extraer esta información, sería a través del procesamiento de un corpus que contenga la suficiente cantidad de ejemplos de uso de cada verbo. Este Corpus se podría construir con oraciones que contengan los verbos a procesar a partir de páginas web obtenidas desde Internet.

Referencias

1. Aone, Ch., D. MacKee. (1996). *Acquiring Predicate-Argument Mapping Information from Multilingual Texts*. Corpus processing for lexical acquisition, pp. 191 – 202. ISBN: 0-262-02392-X.

2. Atserias, J., B. Casas, E. Comelles, Gonzáles, M., Padró. (2006). *FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library*. In: Fifth international conference on Language Resources and Evaluation, Genoa, Italy nlp/freeling, http://www.lsi.upc.edu/nlp/freeling

3. Bolshakov, Igor; A. Gelbukh. (2004). *Computational Linguistics: Models, Resources, Applications*. ISBN 970-36-0147-2.

4. Bolshakov, A. Gelbukh, S. Galicia Haro, M. Orozco Guzman. (1998). *Government patterns of 670 Spanish verbs*. Technical report. CIC, IPN

5. Brent, M. (1991). *Automatic acquisition of subcategorization frames from untagged text*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA., pp. 209-214.

6. Brent, M. (1993). *From grammar to lexicon: unsupervised learning of lexical syntax*. Computational Linguistics 19.3: 243-262.

7. Castro-Sánchez, N. A., G. Sidorov. (2011). *Automatic Acquisition of Synonyms of Verbs from an Explanatory Dictionary using Hyponym and Hyperonym Relations*. Lecture Notes in Computer Science, Volume 6718/2011, pp. 322-331.

8. Cordero, M. (2007). *"Diccionario de la lengua española secundaria" (DILES): Planta para su elaboración con algunos apuntes básicos de metalexicografía*. Káñina, Rev. Artes y Letras, Univ. Costa Rica. XXXI (1): 167-195, ISSN: 0378-0473.

9. Dagan, I., Itai A., U. Schwall. (1991). *Two Language Are More Informative Than One*. In: Proc. of the 29th annual meeting on Association for Computational Linguistics. Pp. 130-137.

10. Dale, R., H. Moisl, H. Somers. (2000) *Handbook of Natural Language Processing*. ISBN: 0-8247-9000-6.

11. De Sousa, S. (2007). *Estudio contrastivo del régimen verbal en el portugués de Brasil y el español peninsular*. ISBN: 978-84-9750-878-0.

12. De Miguel, E. (2004) *Qué significan aspectualmente algunos verbos y qué pueden llegar a significar*. Estudios de Lingüística. Anexo 2. ISSN 0212-7636, pp. 167-206.

13. Del Barrio, F. (2005). *El régimen de los verbos en español medieval*. Tesis doctoral. Universidad de Valladolid. Edición digital Biblioteca Virtual Miguel de Cervantes. ISBN 84-689-2626-4.

14. Diccionario de la Lengua Española. (2001) Edición vigésimo segunda. www.rae.es.

15. Fernández, J. (2002) *Rektion. Rección/Régimen.* http://culturitalia.uibk.ac.at. Hispanoteca.

16. Fuentes, J. (2003). *Gramática moderna de la lengua española*. Editorial Limusa, ISBN 968-18-2184-X.

17. Fujita, S., F. Bond. (2004). *An Automatic Method of Creating Valency Entries using Plain Bilingual Dictionaries*. In: The tenth conference on theoretical and methodological issues in machine translation, Baltimore, Maryland, pp. 55-64.

18. Gahl, S. (1998). *Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus*. In: Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada., pp. 428-432.

19. Galicia-Haro, S., A. Gelbukh, I. A. Bolshakov. (2001). *Acquiring syntactic information for a government pattern dictionary from large text corpora*. IEEE International Workshop on Natural Language Processing and Knowledge Engineering, NLPKE, pp. 536–542.

20. Gelbukh, A., O. Kolesnikova. (2010). Supervised Learning for Semantic Classification of Spanish Collocations. Advances in Pattern Recognition 6256: pp. 362-371.

21. Ienco, D., S. Villata., C. Bosco. (2008). *Automatic Extraction of Subcategorization Frames for Italian*. International Conference on Language Resources and Evaluation IREC.

22. Kahane, Sylvain (2003): Meaning-text theory. In: Ágel, Vilmos et al. (eds.): Dependency and Valency. An International Handbook of Contemporary Research. Berlin.

23. Karmiloff, K., A. Karmiloff-Smith. (2005). Hacia el lenguaje. ISBN 84-7112-483-1.

24. Kawahara, D. S. Kurohashi. (2006). Case frame compilation from the web using high-performance computing. In Proceedings of LREC2006.

25. Luque, D., J. de Dios. (2004). Aspectos universales y particulares del léxico de las lenguas del mundo. Volumen 21. ISSN 1139-8736.

26. Manning, C. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, pp. 235- 242.

27. Marinov, S., C. Hamming. (2004) Automatic Extraction of Subcategorization Frames from the Bulgarian Tree Bank.

28. Mendikoetxea, A. (2004). En busca de los primitivos léxicos y su realización sintáctica: del léxico a la sintaxis y viceversa. 2º Xarxa Temàtica de Gramàtica Teòrica, Barcelona, UAB.

29. Monedero, J., J. González, J. Goñi, C. Iglesias, A. Nieto. (1995). Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. Procesamiento del lenguaje natural, boletín 17.

30. Pérez, W. (2000) Manual práctico de la preposición española. ISBN: 84-7962-160-5.

31. Porto Dapena, J. A. (2002). *Manual de técnica lexicográfica*. Madrid, Arco/Libros.

32. Rojas, E. (2007). Introducción a la Lingüística de Corpus. http://www.scribd.com/doc/81823/Linguistica-de-Corpus.

33. Roland, D., D. Jurafsky. (1998). How Verb Subcategorization Frequencies Are Affected By Corpus Choice. In: Proc. of COLING/ACL-98, pp. 1122-1128.

34. Roland, D., D. Jurafsky. (2002). Verb Sense and Verb Subcategorization Probabilities. In Stevenson, Suzanne, and Paola Merlo (eds.), The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues. Amsterdam: John Benjamins, pp. 325-346.

35. Sabine, S. (2009). The Induction of Verb Frames and Verb Classes from Corpora. Corpus Linguistics. An International Handbook. Anke Lüdeling and Merja Kytö (eds). Mouton de Gruyter, Berlin, pp. 952–972. eBook ISBN: 978-3-11-021388-1. Print ISBN: 978-3-11-020733-0.

36. Sarkar, A., D. Zeman. (2000). Automatic Extraction of Subcategorization Frames for Czech. In: Proc. of the 18th International Conference on Computational Linguistics.

37. Séreny, A., Simon, E., Babarczy, A. (2008). Automatic Acquisition of Hungarian Subcategorization Frames. In: 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics CINTI 2008.

38. Ushioda, A., Evans, D., Gibson, T., Waibel, A. (1993). The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In: Boguraev, B. and Pustejovsky, J. eds. SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text. Columbus, Ohio, pp. 95-106.

39. Uzun, E., Y. Kılıçaslan, H.V. Agun, E. Uçar. (2008). Web-based Acquisition of Subcategorization Frames for Turkish. In: Computational Intelligence: Methods and Applications, IEEE Computational Intelligence Society.

40. Van Valin, R. (2004). An introduction to syntax. Cambridge University Press.

41. Van Valin, R. (2009). Lexical representation, co-composition, and linking syntax and semantics. J. Pustejovsky & P. Bullion (eds.), New Developments in the Generative Lexicon.

Publicaciones del autor

1) Noé Alejandro Castro-Sánchez and Grigori Sidorov. *Extracción automática de los patrones de rección de verbos de los diccionarios explicativos*. Research journal on Computer science and computer engineering with applications. Número 45, junio 2012, a publicarse.
2) Noé Alejandro Castro-Sánchez and Grigori Sidorov. *Automatic Acquisition of Synonyms of Verbs from an Explanatory Dictionary using Hyponym and Hyperonym Relations*. Lecture Notes in Computer Science, 2011, Volume 6718/2011, pp. 322-331.
3) Noé Alejandro Castro-Sánchez, Grigori Sidorov. *Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants Based on Detection of Patterns*. Lecture Notes in Computer Science, 2010, Volume 6177/2010, 233-239.

Ponencias impartidas

1) *Patrones de manejo sintáctico para análisis sintáctico del español*. 5to. Coloquio de Lingüística Computacional y al Primer Seminario de Lingüística Forense, UNAM, México, DF, 2011.
2) *Detección automática de patrones sintácticos*. VII Taller de Tecnologías del lenguaje humano. Tonantzintla, Puebla, 2010.
3) *Analysis of dictionary definition contour for actant extraction*. 18th International Conference on Computing. México, DF. 2009.
4) *Aprendiendo el Habla: Análisis Automático de las Relaciones entre el Significado del Verbo y la Sintaxis*. Conference in Computing CORE. México, DF. 2009.
5) *Detección automática de actantes de verbos en español utilizando diccionarios explicativos y relaciones léxicas.* 4º Coloquio de Lingüística Computacional COLICO-UNAM . México, DF, 2009.
6) Póster: *Detección automática de patrones de rección en español basada en diccionarios explicativos y relaciones léxicas*

Otros

1) Revisor adicional. Advances in Artificial Intelligence. 9th Mexican International Conference on Artificial Intelligence, MICAI 2010.
2) Miembro de comité revisor. Congreso nacional estudiantil de investigación, y 5to Congreso de investigación politécnica. Querétaro, Querétaro, 2009.
3) Miembro del comité Organizador. XVII Congreso Internacional de Computación CIC-2008. México, D. F., 2008.

Apéndice

Apéndice 1. Etiquetas Eagles

Listado de una fracción de las etiquetas “Eagles” correspondiente a las utilizadas en la gramática de segmentación de definiciones.

Código	Categoría	Tipo	Ejemplo
AQ	Adjetivo	Calificativo	Alegre, bonito, grande, malo, etc.
CC	Conjunción	Coordinada	E, i, o, u, empero, mas, ni, pero, etc.
CS	Conjunción	Subordinada	Aunque, como, conque, cuando, donde, etc.
DA	Artículo	Definido	El, la, lo, las, los
DI	Determinante	Indefinido	Alguno, ninguno, otro, etc.
DD	Determinante	Demostrativo	Aquel, ese, este, etc.
DP	Determinante	Posesivo	Mi, tu, su, etc.
FC	Signo de puntuación	Coma	,
NC	Nombre	Común	Persona, animal, planta, etc.
PI	Pronombre	Indefinido	Algo, alguien, alguno, otro, etc.
RG	Adverbio	General	Despacio, ahora, siempre, etc.
RN	Adverbio	Negativo	No
SP	Adposición	Preposición	A, ante, bajo, cabe, con, etc.

Apéndice 2. Algunos diccionarios del idioma español

Listado de algunos diccionarios del idioma de la lengua española indicando si utilizan marcas para distinguir los elementos del contorno.

Diccionario	Autor/Editorial	Marca el contorno	Descripción
Diccionario Espasa de la Lengua Española	Espasa-Calpe	Sí	Obra para usuarios cultos sin necesidades lingüísticas profesionales, y para estudiantes a partir de secundaria.
Diccionario del Español Actual (DEA)	Manuel Seco, Gabino Ramos, Andrés Olimpia	Sí	Recoge «el léxico vivo del español comprendido entre 1955 y 1993».Publicado en 1999 por Aguilar lexicografía, se trata de un diccionario que recoje términos en uso, tanto los documentadas como no documentadas, pero de uso evidente.
Diccionario VOX-Alcalá	VOX, Universidad de Alcalá de Henares	Sí	Diccionario para el aprendizaje del español como lengua extranjera, que cuenta con la homologación del Instituto Cervantes.
Diccionario de la Real Academia Española (DRAE)	Real Academia Española	No	Diccionario normativo de la lengua española. Considerado el principal diccionario y autoridad de consulta del español.
Diccionario de uso del español (DUE)	María Moliner	Sí	Considerado una obra de arte de la lexicografía del español. Su cobertura léxica puede competir con el DRAE. Contiene una abundante cantidad de datos sobre fraseología y colocaciones.
Diccionario Salamanca de la lengua española	Editorial Santillana	Sí	Diccionario para estudiantes y profesores de español, tanto como lengua materna como extranjera.