REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2019 PUCV, Chile 52(101) 878-901

Verbo y contexto de uso: Un análisis basado en corpus con métodos cualitativos y cuantitativos[1*]

Verb and context of usage: A corpus-based analysis with quantitative and qualitative methods

Irene Renau
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
irene.renau@pucv.cl

Rogelio Nazar
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
rogelio.nazar@pucv.cl

Ana Castro
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
ana.castro.paez1@gmail.com

Benjamín López
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
benjamin.lopez.hidalgo1@gmail.com

Javier Obreque
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
CHILE
j.obrequezamora@gmail.com

Recibido: 09-V-2017 / **Aceptado:** 29-VIII-2018
DOI: 10.4067/S0718-09342019000300878

Resumen

El análisis semántico de los verbos supone un desafío teórico y metodológico debido a la complejidad de estas unidades léxicas en términos tanto semánticos como con respecto a su relación con la sintaxis oracional. El objetivo de esta investigación es identificar las estructuras léxico-sintácticas, es decir, los patrones formados a partir de la sintaxis oracional, los argumentos y los tipos semánticos de los verbos en español. El análisis de los verbos en español se realiza siguiendo la propuesta de Corpus Pattern Analysis (Hanks, 2004a). Este análisis se complementa con la automatización del procedimiento, combinando un analizador de dependencias con una serie de algoritmos basados en estadística de corpus. Como resultado del proceso, se ofrece una base de datos de patrones léxico-sintácticos de 182 verbos anotados manualmente, y una interfaz para el análisis automático que, según la evaluación realizada, muestra un 63,41% de precisión con respeto de la identificación manual. Esto sería una contribución tanto a la teoría de la semántica léxica como a la descripción del léxico del español desde una metodología basada en corpus.

Palabras Clave: Corpus Pattern Analysis, estadística de corpus, lexicografía computacional, patrón léxico, verbo.

Abstract

The semantic analysis of verbs is a theoretical and methodological challenge due to the syntactic and semantic complexity of these lexical units. The aim of this research is to identify the lexical-syntactic structures, that is, the patterns formed considering the syntactic structure, the arguments and the semantic types of the Spanish verbs. The analysis of Spanish verbs is carried out following the proposal of Corpus Pattern Analysis (Hanks, 2004a). This analysis is complemented with the automation of the procedure, combining a dependency analyser with a series of algorithms based on corpus statistics. As a result of the process, a database of lexical-syntactic patterns of 182 manually annotated verbs is offered, as well as an interface for automatic analysis which, according to the evaluation carried out, shows a 63.41% accuracy with respect to manual identification. This would be a contribution both to the theory of lexical semantics and to the description of the Spanish lexicon from a corpus-based methodology.

Key Words: Corpus Pattern Analysis, computational lexicography, corpus statistics, lexical pattern, verb.

INTRODUCCIÓN

Los verbos son unidades léxicas muy complejas gramatical y semánticamente, y ocupan un apartado preponderante en las gramáticas de cualquier lengua. Hanks (2004a) los considera '*the pivot of the clause*', alrededor del cual se organiza el resto de la oración a nivel tanto sintáctico como semántico. En efecto, en el caso de los verbos, las propiedades gramaticales parecen encontrarse especialmente imbricadas con las propiedades semánticas de sus argumentos. Las múltiples estructuras sintácticas posibles del mismo verbo, así como la tendencia a la alta polisemia de esta categoría gramatical, hacen que los verbos resulten problemáticos para los aprendices, sobre todo los de segundas lenguas (Renau, 2010), y que requieran una supervisión especialmente cuidadosa en los diccionarios (Battaner, 2011).

En la presente investigación se emplea el análisis de corpus como método que permite obtener evidencia empírica del uso de los verbos en los textos; en concreto, se basa en el análisis del contexto sintagmático de estas unidades léxicas y lo conecta con su significado en dicho contexto específico. Partiendo de la *Theory of Norms and Exploitations* (Hanks, 2013), se toma como objeto de estudio no el verbo aislado sino el patrón sintáctico-semántico de cada verbo, formado a partir de la estructura sintáctica y de valencias, así como de algunos rasgos semánticos de los argumentos. Se emplea como método el *Corpus Pattern Analysis*, CPA (Hanks, 2004a) para analizar estos patrones en el corpus.

Como se explicará a continuación, existen propuestas similares de análisis del léxico del español basado en corpus; sin embargo, existe todavía poca atención al análisis sistemático de un número suficiente de ocurrencias y de la conexión entre estas y el resultado del análisis. En particular, no encontramos propuestas que se

refieran a la detección de estructuras sintáctico-semánticas no ambiguas que sobrepasen la noción tradicional de unidad léxica y que puedan ser conectadas con los significados convencionales del verbo. Y menos adelantada aún está la automatización de este tipo de procedimientos de análisis.

El objetivo de la presente investigación es identificar las mencionadas estructuras léxico-sintácticas o patrones por medio del análisis de la sintaxis oracional, de los argumentos y sus tipos semánticos. Esto representaría un avance para la semántica léxica y para la descripción lexicográfica del español desde una metodología inductiva y basada en corpus. Se lleva a cabo, además, una automatización del procedimiento de análisis como posible ayuda al análisis manual o bien como herramienta independiente, con múltiples aplicaciones para el procesamiento del lenguaje natural como el etiquetado semántico de corpus, la creación de aplicaciones para enseñanza de las colocaciones a estudiantes de español como lengua extranjera, la detección de neología semántica o fenómenos de especialización (generación de significados especializados en significantes ya conocidos), entre otras aplicaciones posibles.

Como resultado del análisis se obtiene, por un lado, una base de datos de verbos analizados manualmente, y por otro lado, los mismos verbos analizados mediante el procedimiento automático. Los planes de trabajo futuro contemplan, lógicamente, la fusión de ambos tipos de análisis.

El artículo se organiza de la siguiente forma: tras esta introducción, se expondrá el marco teórico (apartado 1), se explicará el proceso de análisis manual y automático de los datos (apartado 2), se ofrecerán los resultados (apartado 3) y se terminará con unas conclusiones y líneas de trabajo futuro.

1. Marco teórico

El sustrato teórico de la presente propuesta está ligado a un extenso número de autores de diversos enfoques que consideran que el significado léxico está determinado por el uso que se haga de la palabra en un determinado contexto y, por tanto, que el contexto debe ser tenido en cuenta. Este término refiere tanto a aspectos netamente lingüísticos como también a aspectos sociales o pragmático-discursivos, aunque en este proyecto, por el momento, el análisis se limita al contexto sintagmático, que se considera como primer paso necesario para abordar un análisis más amplio en el futuro. En este apartado se ofrecerá una síntesis de los enfoques de dichos autores, y se prestará especial atención a la propuesta de Hanks (2013), como principal teoría ligada a nuestro trabajo.

El significado lingüístico, que es uno de los temas que más ha preocupado al ser humano desde los inicios de su historia (Yallop, 2004), se aborda, sin embargo, con escasa base empírica hasta la llegada de la lingüística de corpus en la década de 1980. No obstante, ya Malinowski (1923) advierte, desde la antropología, la necesidad de

estudiar el significado en conexión con el uso. Estudiando las culturas de Oceanía, el autor observa, por ejemplo, que las palabras equivalentes en una lengua oceánica a 'madera' y 'canoa' adquieren significados metafóricos que solo se advierten si se entiende el contexto comunicativo en el que estas voces son usadas: "*The meaning of a single word is to a very high degree dependent on its context*" (Malinowski, 1923: 306). Firth (1935) considera asimismo que el significado de una palabra está ligado al uso cotidiano de dicha palabra en sociedad y que no debe estudiarse de forma abstracta: "*The complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously*" (Firth, 1935: 37). Para dicho estudio, según el autor (Firth, 1935), debe relacionarse la semántica de la palabra con el resto de los componentes lingüísticos (fonético, morfológico, léxico y sintáctico) en la linealidad del habla en que la palabra es usada. Finalmente, además de los componentes puramente lingüísticos, es necesario incorporar aspectos relativos al contexto de uso mediante etiquetas como lenguaje coloquial, técnico, conversacional, etc., junto con datos de frecuencia (Firth, 1935).

Todos estos aspectos son antecedentes teóricos y metodológicos de la futura lingüística de corpus. En efecto, los corpus permiten, a partir de los años ochenta, obtener evidencias abundantes y analizarlas de un modo sistemático. Sinclair (1991, 2004) fue uno de los lingüistas pioneros en desarrollar teorías del significado léxico con el empleo de corpus. En concreto, Sinclair (2004) advierte que el corpus permite observar '*recurrent patterns of language'*, que son las verdaderas unidades de significado, y no las palabras por sí solas: "*By giving greater weight to the syntagmatic constraints, units of meaning can be identified that reduce the amount of meaning available to the user*" (Sinclair, 2004: 140); es decir, los patrones sintagmáticos restringen el significado de la unidad léxica. Esta concepción 'extendida' de la unidad léxica se refleja, por supuesto, en el nuevo modelo de diccionario basado en corpus y con especial atención a la combinatoria y a la estructura sintáctica, propuesto por el autor (Sinclair, 1987).

Toda esta reflexión es recogida por Hanks (2004b, 2013), que al igual que Sinclair aborda la cuestión como un problema teórico pero también relacionado con la práctica lexicográfica. La *Theory of Norms and Exploitations* (*TNE*) de Hanks (2004b, 2013) tiene como objetivo explicar cómo se construye el significado léxico, y postula que este está asociado contextualmente a la estructura sintáctica y a las combinaciones con otras unidades léxicas. De este modo, estas unidades son utilizadas de forma normal cuando existen unas relaciones estables entre el significado, la sintaxis y la combinatoria léxica. Por ejemplo, el verbo 'golpear' significa prototípicamente 'dar golpes, hacer chocar con violencia un cuerpo contra otro': para que este significado se encuentre en un texto, es necesario que el sujeto sea una persona (o, con menos frecuencia, un animal) y que el objeto directo sea un cuerpo físico; además, el complemento adverbial opcional indica aquello con lo que se golpea (una parte del cuerpo, un instrumento, etc.). Estas características reflejan uno de los usos normales

de 'golpear', encontrado en contextos de corpus como 'los vecinos condenados secuestraron y golpearon con bates de béisbol y palos a estos tres hombres'. Este verbo, empleado metafóricamente, significa también 'afectar negativamente, causar daño': en este caso, el sujeto de la oración es generalmente un suceso o acción, y el objeto directo es una persona o grupo humano, como en el ejemplo 'El corralito golpeó a la clase media'. En estos dos casos se observa, por tanto, como el contexto sintagmático activa el significado del verbo que debe emplearse en dicho contexto, según las condiciones sintácticas y semánticas de la oración.

Al mismo tiempo, Hanks (2004b, 2013) también propone una solución para la creatividad y el dinamismo del significado léxico, que permite ser usado de forma no convencional pero comprensible, como ocurre con cualquier hablante que se inventa un significado nuevo, o emplea una palabra de forma extraña como recurso expresivo, humorístico, etc. En el siguiente ejemplo, el verbo 'golpear' es utilizado de forma no convencional: 'Es un arte dramático absolutamente nuevo, un martillo con el que golpeáis sobre las cabezas huecas del público'. Esta explotación se ha construido a partir del uso también metafórico de 'martillo' y 'cabeza hueca'. Así pues, no se trata de un patrón normal de uso del verbo, sino de un uso metafórico derivado del patrón prototípico de 'golpear', que por estar ligado a dicho patrón, es fácilmente comprensible por los lectores. Este caso es una expresión lingüística de una metáfora conceptual (Lakoff & Johnson, 1980). Hanks (2009) habla de 'sistema de doble hélice' para referirse a esta alternancia en el uso de las normas y las explotaciones.

2. Marco metodológico

La perspectiva anteriormente señalada de concebir el significado léxico como fruto del uso en un determinado contexto requiere de soluciones metodológicas apropiadas que permitan un análisis sistemático y coherente del corpus. Hanks (2004a) propone el CPA como "*a technique for mapping meanings onto use in context*" (Hanks, 2004a: 87). Esta técnica requiere la creación de una muestra aleatoria de concordancias extraídas de corpus, en la que se advierten los patrones verbales a partir del análisis de la estructura sintáctica y argumental y del análisis de los tipos semánticos de los argumentos. A cada concordancia se le asigna un número de patrón, y a partir de este análisis se configuran los patrones conforme se avanza en la lectura de las concordancias. No se asume previamente qué patrones van a estar asociados a un verbo ni se consultan diccionarios. Cada patrón obtenido se conecta con el significado que se le da convencionalmente, llamado 'implicatura' siguiendo a Grice (1975). El objeto de estudio, por tanto, no es el significado propiamente, sino en el patrón léxico (Hanks, 2004a), que no es ambiguo y que puede ser formalizado en una base de datos o en un diccionario. Por esta razón, además, es más apropiado para su uso en lingüística computacional (Hanks & Pustejovsky, 2005; El Maarouf, Bradbury, Baisa & Hanks, 2014). Los siguientes son ejemplos de los dos patrones del verbo 'golpear'

mencionados anteriormente (corresponden a los patrones 1 y 6 en Verbario, *s. v.* 'golpear'):

Patrón 1	[[Humano]] golpear ([[Objeto Físico 1]]) ({con [[Objeto Físico 2 | Parte del Cuerpo]]})
Implicatura	[[Humano]] empuja violentamente [[Objeto Físico 2 | Parte del Cuerpo]] hasta que contacta con [[Objeto Físico 1]] para causarle daño o moverlo de su sitio.
Ejemplo	Los vecinos condenados, secuestraron y golpearon con bates de béisbol y palos a estos tres hombres.
Patrón 6	[[Eventualidad]] golpear [[Humano | Institución | Lugar = Poblado]]
Implicatura	[[Eventualidad]] causa un daño grave y repentino a [[Humano | Institución | Lugar = Poblado]].
Ejemplo	El corralito golpeó a la clase media.

Entre corchetes dobles se señalan los tipos semánticos, que se obtienen de una ontología que contiene los tipos más elementales (la CPA Ontology, Hanks, 2018). Cada sustantivo que se combina con el verbo como sujeto u objeto directo en este verbo se corresponde con los tipos asignados; por ejemplo, en el caso del sujeto del patrón 6, unidades como 'reforma', 'crisis', 'tragedia', 'violencia', 'estafa', etc. Herramientas como el WordSketch de Sketch Engine (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, Rychlý & Suchomel, 2014) permiten obtener la combinatoria del verbo, pero no la conectan con un tipo semántico genérico que todos los sustantivos tienen en común como hiperónimo. El análisis que se acaba de mostrar puede verse en el Pattern Dictionary of English Verbs, (PDEV) (Hanks, en proceso), un diccionario de patrones publicado en línea. El proyecto Verbario (Renau & Nazar, 2018) que se presenta en este trabajo es paralelo al PDEV y guarda la misma estructura de datos para permitir un futuro proyecto común de base de datos multilingüe.

Existen varias propuestas teórico-metodológicas para el análisis sistemático de los significados, aunque ninguna de ellas toma como objeto de estudio los patrones mencionados anteriormente, estas piezas discursivas generalizables a modo de 'plantilla' en la que colocar la combinatoria más frecuente y la estructura argumental de un verbo correspondiente a cierto significado. Es bien conocido el proyecto *WordNet* (Fellbaum, 1998), que también se ha aplicado al castellano (Atserias, Climent, Farreres, Rigau & Rodríguez, 2000). *WordNet* es en realidad una taxonomía que persigue conectar conceptos entre sí, conceptos que se representan a través de unidades léxicas sinónimas –o con una cercanía semántica estrecha– agrupadas en '*synsets*'. El proyecto *FrameNet* (Fillmore, Wooters & Baker, 2001) es uno de los antecedentes más cercanos al CPA, por su aproximación desde la lingüística de corpus

y su atención en el eje sintagmático. Sin embargo, *FrameNet* pone su foco de atención en los 'marcos', que son definidos como situaciones o aspectos de la realidad conectados a diversas estructuras en el discurso. El recorrido de CPA es inverso: parte de la unidad léxica para ir, en todo caso, al marco conceptual. Finalmente, la propuesta de Gross (1981) tiene puntos en común con la de Hanks (2004a). Influido por Harris (1985) e interesado por la compatibilidad entre el análisis lingüístico y el tratamiento informático de los datos, Gross (1981) emplea también la noción de predicado y argumento y unas mínimas categorías semánticas (del tipo 'Humano' o 'Número') para organizar el análisis de los verbos. La aproximación de Gross (1981), sin embargo, no está basada en el análisis de datos de corpus.

Finalmente, existen multitud de diccionarios de valencias y de colocaciones en diversos idiomas que recogen aspectos léxico-gramaticales coincidentes con CPA. Entre ellos, deben señalarse aquí, al menos, los proyectos de Bosque (2004) y Alonso Ramos (en línea), que ofrecen información sobre las colocaciones del español.

2.1. Análisis manual de verbos del español con Corpus Pattern Analysis

2.1.1. Materiales y métodos

Para el análisis manual, se utilizaron los corpus del español disponibles en *Sketch Engine* (Kilgarriff et al., 2014). Son corpus generales panhispánicos tomados de textos descargados de Internet. El corpus esTenTen, en particular, es uno de los más grandes disponibles en castellano, con unos 9.500 millones de palabras. Siguiendo a Hanks (2004a), el procedimiento de análisis basado en CPA considera al menos seis etapas:

a) **Elaboración de una muestra aleatoria de concordancias.** Se realizan varias muestras aleatorias de 100 concordancias cada una.
b) **Análisis y anotación de la primera muestra.** Se analiza la primera muestra entera, asignando un número de patrón a cada concordancia.
c) **Redacción de los patrones.** Una vez terminada la primera muestra se realiza una primera versión, aún tentativa, de los patrones verbales; es decir, el análisis queda formalizado del modo mostrado al inicio de este apartado con el ejemplo de 'golpear'.
d) **Análisis y anotación de la segunda muestra.** Se analiza una segunda muestra de 100 concordancias para verificar la primera. Si aparecen patrones que no estaban en la primera muestra, se repite el análisis en una tercera muestra, y se continúa así hasta que ya no aparecen nuevos patrones.
e) **Revisión conjunta**. Cada análisis terminado es puesto en común y revisado por todo el equipo de lexicógrafos para obtener un acuerdo de anotadores aceptable.
f) **Publicación de los resultados**. El análisis se publica en la web.

El proyecto CPA en español emplea la base de datos de Verbario para trabajar. Esta tiene dos bloques: la interfaz de análisis de corpus (Figura 1) y la de creación de patrones (Figura 2), ambas conectadas entre sí. Está creada tomando como base la interfaz del PDEV (Baisa, El Maarouf, Rychlý & Rambousek, 2015).

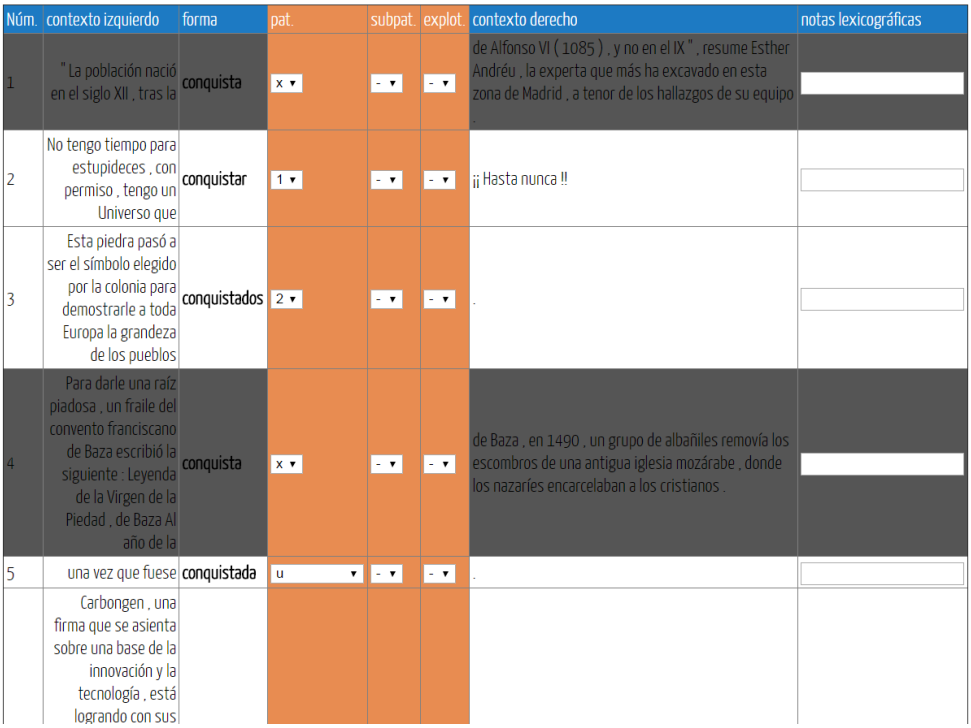

Figura 1. Interfaz de anotación de corpus de Verbario.

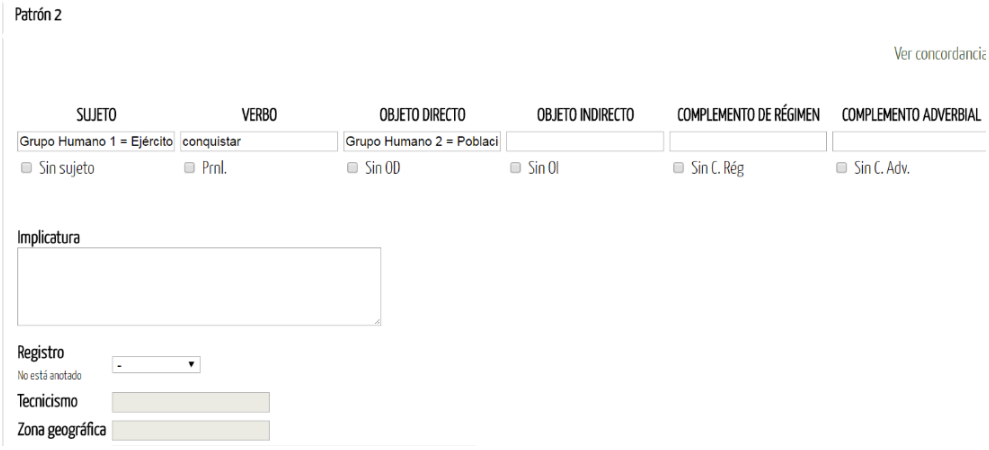

Figura 2. Interfaz de creación de patrones de Verbario.

A continuación se describirán con más detalle las fases b-d, correspondientes al análisis propiamente dicho. Estas consisten en la anotación de todas las concordancias de las muestras aleatorias creadas. Como puede verse en la Figura 1, la interfaz de anotación es igual que las usadas para consultar corpus, pero se le han añadido varios casilleros en los que anotar el análisis, que se va grabando conforme se va completando. En caso de errores de etiquetado morfosintáctico de los corpus, la concordancia se marca con una 'x' (ej. 'plantas' como sustantivo en vez de como segunda persona del singular del presente de indicativo de 'plantar'). Los casos en que no es posible determinar el patrón verbal se marcan con 'u' (de 'unspecified'[2]) (ej. ambigüedades irresolubles entre pasiva refleja y 'se' medio, contextos cortos –como títulos–, etc.). En los dos casos anteriores, las concordancias no se publican. Aparte de estos casos anómalos, la mayoría de las concordancias se anotan normalmente. Así, se anotan tres tipos de información en sendas casillas: el número de patrón, la letra del subpatrón y el tipo de explotación. La numeración del patrón, al inicio del análisis, sigue el orden en que han ido apareciendo los patrones en la muestra, y luego se reordena según criterios de prototipicidad del patrón y de frecuencia. Los llamados 'subpatrones' son los alternantes, generalmente pares activos-medios: ej. '[[Humano 1]] enamorar a [[Humano 2]]' (patrón 1a) / '[[Humano 1]] enamorarse de [[Humano 2]]' (patrón 1b). Pero como subpatrones también se incluyen a veces ciertos casos de polisemia regular (Apresjan, 1974) que no se han podido formular en un solo patrón (ej. el caso de 'sembrar', más adelante). Finalmente, se marcan como explotaciones los casos ya señalados en el apartado anterior. Igual que en el PDEV, existen tres tipos de explotaciones (Tabla 1).

Tabla 1. Ejemplos de explotaciones del verbo 'acarrear' (Verbario, *s. v.*).

Tipo de explotación	Explicación	Ejemplo
Sintáctica	Se presenta cuando hay un uso anómalo de las estructuras sintácticas.	*El río se desbordó y acarreó con todos los materiales de la construcción.* (El uso de la preposición es anómalo).
Argumental	Se presenta cuando hay un argumento extraño combinado con el verbo usado normalmente.	*El gusano también acarrea otro anélido.* (Poco frecuentes *gusano* y *anélido* como sujeto y objeto directo de este uso del verbo).
Figurada	Se presenta cuando un verbo se utiliza metafóricamente, es decir, significa algo distinto de lo que significa cuando es usado normalmente.	*Hasta hace muy poco África seguía empezando en los Pirineos. Es una cruz que nos toca acarrear y cuyo lastre se sigue sintiendo.* (Se utiliza el verbo como parte de una metáfora más amplia construida también a partir de 'cruz' y 'lastre').

En cuanto a los criterios de definición de un patrón, que permiten observarlo como un patrón y diferenciarlo de otros, se contempla, en primer lugar, el análisis de las funciones sintácticas: sujeto, objeto directo, objeto indirecto, complemento de régimen y complemento adverbial. El análisis sintáctico es una condición necesaria pero no suficiente, porque por sí solo no discrimina semánticamente: un verbo puede usarse siempre en oraciones transitivas, que, sin embargo, se empleen para un gran

número de significados. Por ello, una vez se tiene clara la estructura sintáctica, se agrega el análisis de los tipos semánticos de los argumentos. Este análisis consiste en adjudicar un hiperónimo al sustantivo núcleo del sintagma en el sujeto o los complementos (los adverbios se clasifican en tiempo, lugar y manera). Estos hiperónimos se toman de la CPA *Ontology*, como ya se ha indicado, con el fin de proporcionar coherencia al análisis. Se van anotando los patrones procediendo por analogía: si en el siguiente patrón los mismos argumentos tienen el mismo hiperónimo que el anterior, se adjudica el mismo número de patrón. Los patrones no son polisémicos ni ambiguos. Cuando se advierte en el corpus una subespecificación recurrente de un tipo semántico, esta se anota colocándola tras dicho tipo semántico con un signo de igual entre ambos (ej. '[[Parte de Planta = Semilla]]' en el caso del verbo 'sembrar', o '[[Humano = Futbolista]]' en el caso de 'chutar'); es lo que Hanks (2004b) llama '*semantic roles*'. En el caso de las locuciones, estas se consideran patrones independientes: ej. '[[Human]] comer a dos carrillos', '[[Humano | Institución]] dar gato por liebre'.

El análisis de la primera muestra tiene un carácter exploratorio, y puede ser por supuesto modificado cuantas veces sea necesario hasta lograr la coherencia interna (entre patrones del mismo verbo) y externa (entre el análisis realizado y los datos de corpus) que requiere cualquier análisis léxico. Así, el análisis de la nueva muestra tiene tres funciones no excluyentes entre sí: confirmar los patrones extraídos en la primera muestra, agregar información a los patrones léxicos ya configurados o advertir nuevos patrones. En caso de manifestarse esta última opción, se repite el análisis en una nueva muestra aleatoria de 100 concordancias para confirmar la información aparecida, delinear de mejor manera los patrones y descartar el solapamiento de alguna otra información que se haya extraído del corpus. Los muestreos sucesivos se detienen cuando ya no aparecen nuevos patrones.

Finalmente, en cuanto a la redacción de la implicatura, esta consiste en una paráfrasis del patrón, similar a las definiciones fraseológicas (Sinclair, 2004). Esta descripción del significado del patrón debe casi todo a la tradición lexicográfica de redactar definiciones (Porto Dapena, 2014): el significado se describe de manera lo más general posible pero evitando la ambigüedad y atendiendo a la claridad y concisión siempre requeridas.

2.1.2. Un ejemplo de análisis: el verbo 'sembrar'

A modo de ejemplo, se describe en este apartado el análisis del verbo 'sembrar' (Verbario, *s. v.* 'sembrar', para el análisis completo de este verbo).

Se utilizaron 3 muestras de corpus, es decir, 300 concordancias en total. De ellas se detectaron 4 patrones, aunque en dos de ellos había alternancias, por tanto, en realidad este verbo tiene 6 patrones (Tabla 2).

Tabla 2. Análisis del verbo 'sembrar'.

Núm. patr.	Núm. conc.	Patrón	Implicatura	Ejemplo
1a	85	[[Humano]] sembrar ([[Parte de Planta = Semilla | Planta]]) ({en [[Terreno]]})	[[Humano]] esparce o hunde [[Parte de Planta = Semilla | Planta]] en [[Terreno]] para que germine.	*En Santa María Xadani hay gente con edad de 70 a 80 años que aún siembra y cosecha el maíz.*
1b	20	[[Humano]] sembrar [[Terreno]]	[[Humano]] siembra [[Terreno]] cuando esparce o hunde semillas en ella para que germinen.	*Con los primeros rayos del sol inicia su trabajo ya sea arar, sembrar, cosechar o componer el terreno.*
2a	125	[[Humano | Eventualidad]] sembrar [[Emoción | Concepto | Acción | Acto de Habla]] ({en/entre [[Grupo Humano]]})	[[Humano | Eventualidad]] causa [[Emoción | Concepto | Acción | Acto de Habla]] en [[Grupo Humano]].	*Vemos en esta fabricación y comercialización irracional de material de guerra una manera de sembrar dudas e inseguridad.*
2b	10	[[Humano | Eventualidad]] sembrar [[Lugar]] de/con [[Acción | Estado | Emoción]]	[[Humano | Eventualidad]] siembra [[Lugar]] de [[Acción | Estado | Emoción]] cuando la esparce o extiende por él en gran cantidad.	*Sembraron el país de miseria.*
3	3	[[Humano | Eventualidad]] sembrar [[Entidad Abstracta]] de/con [[Acción = Negativa | Estado = Negativo]]	[[Humano | Eventualidad]] siembra [[Entidad Abstracta]] cuando hace que surja [[Acción = Negativa | Estado = Negativo]].	*Han decidido dejar una huella horadada en mi pecho y sembrar mi mente de dudas.*
4	2	[[Humano]] sembrar {(la) cizaña}	[[Humano]] expresa o dice algo, de forma especialmente oculta o solapada, con el fin de provocar un mal dentro de un grupo de personas.	*...e intentando sembrar más cizaña, mostrarme fotos recientes de la niña.*

El patrón 1 corresponde al significado prototípico del verbo, que se expresa en dos subpatrones en los que se advierte polisemia regular: 'sembrar una semilla o planta en un terreno' / 'sembrar un terreno'. Esta estructura transitiva básica es 'heredada' por el resto de patrones, en los que cambian los tipos semánticos. En efecto, los patrones 2 y 3 corresponden a dos metáforas: la metáfora conceptual (Lakoff & Johnson, 1980) de la siembra como una acción que tiene sus consecuencias a largo plazo es una de las más antiguas de la historia de nuestra cultura, presente en múltiples refranes y locuciones. En la primera de estas metáforas (patrón 2a), 'una persona siembra una emoción, estado, acto de habla o acción', todos ellos conceptos abstractos. Por alternancia semántica (Hanks, 2013), el sujeto también puede ser una acción, proceso o estado ([[Eventualidad]]). El patrón 2b recoge muchos aspectos del 1b para construir la metáfora: 'una persona siembra un lugar de/con una emoción, estado,

acto de habla o acción'. En ambos casos (2a y 2b), el significado es muy similar, solo que al 2b se le ha agregado el lugar en el que se produce la acción del verbo. Finalmente, el patrón 3 refleja el caso de 'sembrar de dudas, incertidumbre, desconcierto'. La misma connotación negativa se advierte en la locución 'sembrar (la) cizaña', y estaba presente también en casi todas las concordancias del patrón 2a (pero no en todas, ej. 'sembrar la fe').

Las metáforas anteriores se consideran patrones porque se observan reiteradamente en el corpus (incluso si aparecen solo 3 veces como en el caso del patrón 3). Sin embargo, existen también casos de explotaciones en los que este verbo se emplea esporádicamente de otras maneras. Por ejemplo, cuando en una de las concordancias se dice que 'está sembrado el pan' (metonimia del 'pan' por el 'trigo'), o la expresión 'el lecho estaba sembrado de peñascos' una metáfora que indica el gran número de peñascos en un terreno, etc. Las explotaciones de tipo argumental y figurado (Tabla 1) de este verbo se presentan en los patrones 1a, 2a y 2b.

2.2. Automatización de la identificación de patrones mediante análisis de dependencias y estadística de corpus

En este apartado se describirán las estrategias empleadas para la automatización de la identificación de patrones con CPA. La necesidad de automatizar todo o al menos una parte del procedimiento descrito en las páginas anteriores viene motivada por el alto costo en tiempo y esfuerzo que trae aparejado el análisis del léxico basado en corpus. Hasta la fecha, se ha avanzado en soluciones computacionales que asisten y en cierta medida alivian buena parte del procedimiento manual. Este procedimiento permite vincular el análisis manual con el automático, de modo que el lexicógrafo acepte, rechace o complete la información generada por el sistema, la que estará integrada tanto en la interfaz de anotación de corpus como en la de creación de patrones.

2.2.1. Materiales y métodos

Para el procedimiento automático se utilizó el corpus esTenTen disponible en Sketch Engine. Este corpus ya tiene etiquetado gramatical con TreeTagger (Schmid, 1994); sin embargo, para el tipo de análisis que nos ocupa es necesario establecer las relaciones de dependencia sintáctica entre las palabras, por lo tanto desestimamos el etiquetado original en favor del producido por el parser sintáctico Syntaxnet, presentado recientemente por Google (Andor, Alberti, Weiss, Severyn, Presta, Ganchev, Petrov & Collins, 2016). La razón por la que se utilizó este y no otros analizadores es que es un programa de código abierto, es el más reciente y, según la evaluación presentada por sus autores, uno de los que ofrece mejores resultados.

Además del corpus y su anotación gramatical, tal como se ha explicado en el apartado anterior, el procedimiento requiere información semántica para identificar la categoría de los sustantivos etiquetados como los argumentos del verbo. Para este análisis, se desarrolló una taxonomía de sustantivos del castellano por medio de la aplicación de técnicas estadísticas para el análisis de corpus. Los detalles sobre el desarrollo de esta taxonomía se encuentran en Nazar y Renau (2016). Tanto el resultado como el código fuente del experimento de la taxonomía son de acceso público en un repositorio en línea (http://www.tecling.com/kind). Esta taxonomía, que por el momento está disponible en castellano, inglés y francés, está basada en la CPA *Ontology*, por tanto la estructuran los tipos semánticos propuestos por este autor. Cada tipo semántico de esta taxonomía está poblado con sustantivos castellanos y estos pueden, a su vez, subdividirse en otras categorías según las decisiones que va tomando el algoritmo de inducción de taxonomías.

Crear una taxonomía *ad hoc* para este proyecto era necesario por cuestiones teóricas: si se está empleando corpus para el análisis de los patrones verbales, la taxonomía debía estar también basada en corpus. Por ello, en el marco de este estudio no resultaba apropiada una taxonomía como *WordNet*, que está creada con categorías establecidas *a priori*, muchas veces tomadas de taxonomías científicas. La taxonomía descrita por Nazar y Renau (2016) permite irse renovando conforme cambia la lengua. No obstante, ello no es obstáculo para que *WordNet* u otra taxonomía se apliquen con el mismo propósito, en otro contexto de investigación (ej. podría resultar adecuada para discurso especializado).

A partir de estos materiales, en las subsecciones siguientes se definirá la cadena de procesamiento para cada verbo de entrada *i*.

2.2.1.1. Extracción de concordancias

La extracción de muestras aleatorias de concordancias por cada verbo fue abordada ya en el apartado dedicado al análisis manual. La extracción aleatoria de concordancias a partir de corpus de tamaño tan grande como los utilizados requiere de un proceso de indización, que en nuestro caso se realizó utilizando el *software* Jaguar (http://www.tecling.com/jaguar). El indexador de este programa permite extraer concordancias de cualquier verbo del esTenTen de forma instantánea, lo que sería imposible con los extractores de concordancias habituales, que no son capaces de procesar tanta información.

2.2.1.2. Análisis sintáctico de las concordancias

A partir de las concordancias de un verbo *i*, el algoritmo aplica *Syntaxnet*, cuyo demostrador en línea se ha instalado también en nuestro repositorio para facilitar la descripción de cada paso del análisis (http://www.tecling.com/syntaxnet). A modo ilustrativo, la Tabla 3 muestra el resultado del análisis sintáctico de dependencias

expresado por *Syntaxnet* para una de las concordancias que habíamos utilizado como ejemplo en la Tabla 2; el análisis se expresa en el formato CONLL (Buchholz & Marsi, 2006). La Figura 3 muestra el resultado del mismo análisis por el parser DepPattern (Gamallo & González, 2012) a modo de grafo dirigido.

Tabla 3. Ejemplo de resultado de un análisis sintáctico de dependencias con el parser *Syntaxnet*.

Posición	Forma	Categoría gramatical	Dependencia	Función gramatical
1	Sin	ADP	8	advmod
2	embargo	NOUN	1	mwe
3	,	PUNCT	1	punct
4	en	ADP	5	case
5	Santa	PROPN	8	nmod
6	María	PROPN	5	name
7	Xadani	NUM	5	nummod
8	hay	AUX	0	ROOT
9	gente	NOUN	8	dobj
10	con	ADP	11	case
11	edad	NOUN	8	nmod
12	de	ADP	13	case
13	70	NUM	11	nummod
14	a	ADP	16	case
15	80	NUM	16	nummod
16	años	NOUN	11	nmod
17	que	SCONJ	23	mark
18	aún	ADV	19	advmod
19	siembra	ADJ	23	amod
20	y	CONJ	19	cc
21	cosecha	NOUN	19	conj
22	el	DET	23	det
23	maíz	NOUN	16	nmod
24	.	PUNCT	8	punct

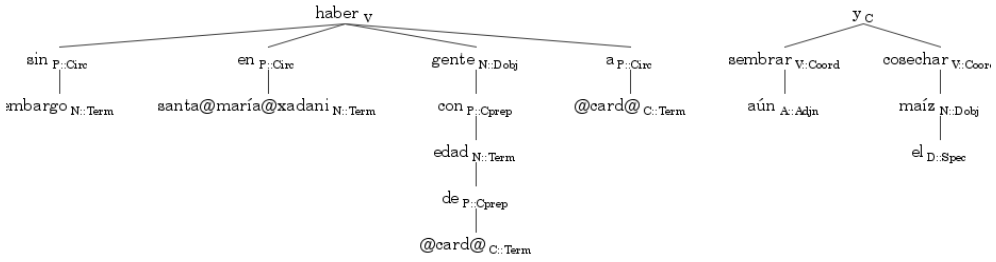

Figura 3. Ejemplo de resultado de un análisis sintáctico de dependencias con el parser DepPattern.

En la Tabla 3, la forma 'hay' en posición 8 es seleccionada como núcleo de la oración principal y 'gente', en la posición 9, se encuentra como objeto directo de este verbo sin que se encuentre ningún sustantivo marcado con función de sujeto. A su vez, la secuencia 'Santa María Xadani' es correctamente encapsulada como un mismo

sintagma nominal y como un nombre propio, tal como se indica en las columnas de categoría gramatical y la dependencia, aunque selecciona 'Santa' como núcleo del sintagma. El ejemplo también sirve para ilustrar algunos problemas del etiquetador: además de este nombre propio y el objeto directo 'gente', el sistema presenta la secuencia que comienza con 'edad' en la posición 11 como un tercer argumento del verbo principal con la etiqueta nmod ('nominal modifier') según el etiquetario '*Universal Dependency*' (Jurafsky & Martin, en preparación), que aplica cuando un sustantivo funciona como un complemento de un nombre, diferenciándose de otros complementos como el adjetivo amod ('adjectival modifier'). La etiqueta sería correcta si 'edad' dependiera de 'gente' y no del verbo núcleo de la oración. Más adelante, también, se cometen otros errores, como por ejemplo etiquetar erróneamente como nombres las formas verbales 'siembra' y 'cosecha', perjudicando naturalmente el resto del análisis a partir de allí. Errores como estos, por supuesto, repercuten en los pasos subsiguientes del análisis y perjudican los resultados.

2.2.1.3. Análisis semántico de los argumentos de los verbos

Del resultado anterior se desprende que para cada verbo será posible determinar qué sustantivos funcionan como argumentos. Como ya se indicó, en el caso de los argumentos constituidos por sintagmas nominales, el tipo semántico se adjudica al núcleo del sintagma. El propósito de esta parte de nuestro análisis es, pues, asignar dicho tipo semántico a cada uno de estos sustantivos, ya que esto es luego lo que nos va a permitir generalizar y caracterizar estos argumentos.

En esta fase del análisis se utilizan dos soluciones distintas en función de los argumentos sean nombres propios (NP) o nombres comunes (NC). Para establecer esta distinción, y para etiquetar y clasificar cada NP en las categorías de 'Persona', 'Organización' o 'Lugar', se utiliza el *software* POL disponible como *software* libre en el repositorio http://www.tecling.com/pol. Para cualquier texto de entrada sin etiquetar, POL identifica y clasifica los nombres propios, ya sean unidades simples o sintagmáticas.

Para los NC, se utiliza la mencionada taxonomía de sustantivos (Nazar & Renau, 2016). En esta fase, el algoritmo interroga la base de datos con la taxonomía para obtener la o las cadenas hiperonímicas de los NC, recorriendo ascendentemente esta estructura hasta dar con uno de los tipos semánticos de la CPA *Ontology*, que por su mayor generalidad se encuentran en las partes superiores.

Considérese, a modo de ejemplo, el resultado de cadena hiperonímica que se obtiene con esta taxonomía para la palabra 'maíz', en la Figura 4. Este grafo representa las diferentes hipótesis que el algoritmo de la taxonomía almacenó sobre las formas de conceptualizar la unidad elegida: como un tipo de alimento, como un material, como un tipo de planta y como parte de una planta. Los arcos entre los nodos, que representan las relaciones de hiperonimia, están coloreados ya que los enlaces tienen

asociado un valor de certeza en una escala de colores: verde si es existe alta certeza, azul para una certeza moderada, negro para el neutro y rojo para indicar poca certeza. En cuanto al enlace meronímico que también aparece, la CPA *Ontology* contiene un apartado para las 'partes de', aunque en general la taxonomía se compone de relaciones de hiponimia.

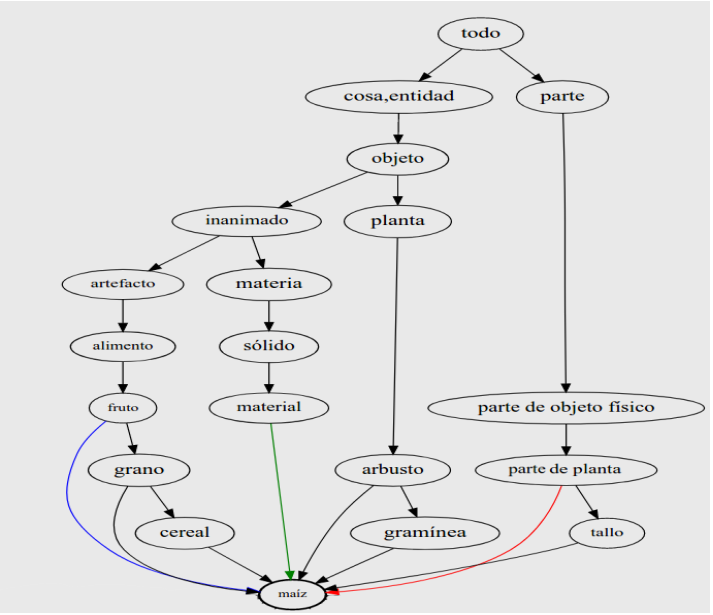

Figura 4. Fragmento de cadena hiperonímica generada automáticamente para el sustantivo 'maíz'.

2.2.1.4. Creación y ordenamiento de los patrones de cada verbo

Una vez que de cada verbo *i* se han extraído las concordancias y de estas se han identificado y clasificado semánticamente los argumentos, se procede a la construcción de los patrones. De cada concordancia se intentará generar un patrón. Por cada NC de cada concordancia del verbo *i*, el algoritmo extrae estas cadenas de la taxonomía y asciende por cada una de ellas hasta encontrar un tipo semántico o hiperónimo común. Por cada tipo semántico obtenido se ensambla un patrón con rasgos similares a los ya descritos en el caso del análisis manual, como '[[Humano]] sembrar [[Planta]]'. Sin embargo, en el caso de la palabra 'maíz', y de todas aquellas que tengan más de un hiperónimo en la taxonomía, se producirán otros patrones: '[[Humano]] sembrar [[Fruto]]', '[[Humano]] sembrar [[Material]], etc.: el primero de ellos es correcto y el segundo no, debido al error de la taxonomía al etiquetar 'maíz' como 'material'. En teoría, estas combinaciones pueden escalar rápidamente, pero ello no es un problema porque la cantidad de cadenas por palabra es muy limitada, nunca por encima de la decena.

Finalmente, en la formación del patrón también se registra la aparición de pronombres clíticos o enclíticos, y de preposiciones que preceden al complemento de régimen o adverbial.

Cada patrón que se forma se almacena en una estructura de datos que registra la frecuencia sumando 1 cada vez que se ensambla dicho patrón. Con esta estrategia, cuando se analizan grandes volúmenes de datos como los de estos corpus, es posible advertir las tendencias principales y obtener de forma correcta al menos los patrones de uso más frecuentes; la precisión va disminuyendo conforme se hallan menos instancias de cada patrón en el corpus.

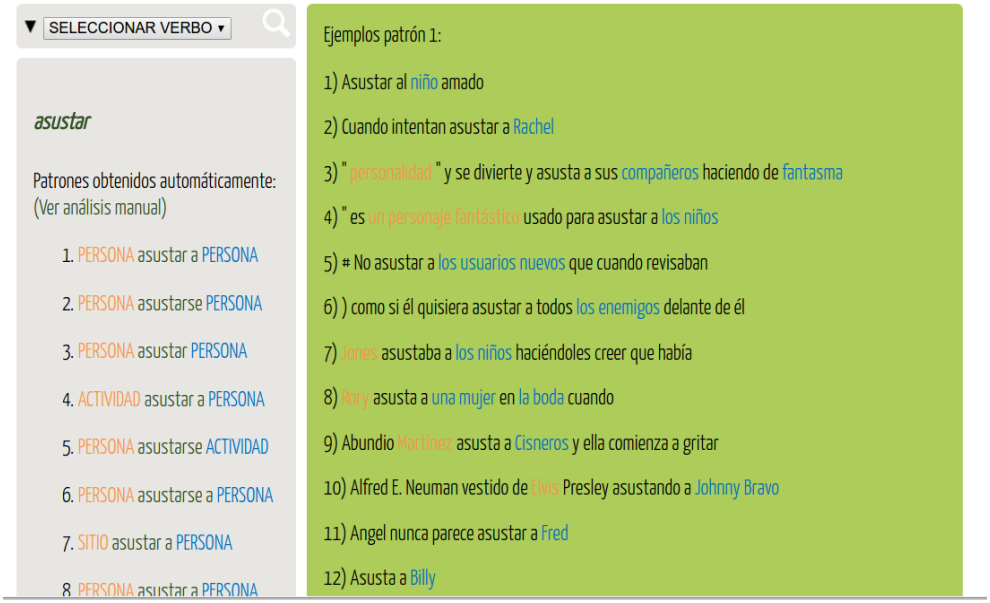

Figura 5. Resultado del ordenamiento por frecuencia decreciente de los patrones.

El último paso del análisis consiste por tanto en el ordenamiento por frecuencia decreciente de los patrones creados, que se presentan asociados a una muestra aleatoria de 10 de sus contextos de aparición. Los contextos presentan los argumentos coloreados en coincidencia con los colores elegidos en la representación del patrón, tal como se muestra en la Figura 5. Las concordancias que aparecen en la figura corresponden al patrón 1 de 'asustar': '[[Humano]] asustar a [[Humano]]'.

3. Resultados

3.1. Análisis manual

La Tabla 4 resume las cifras alcanzadas por el análisis manual hasta la fecha, disponibles en http://www.tecling.com/verbario (Renau & Nazar, 2018). Además, el número de patrones por verbo presenta un rango de 1-47 y una media de 5,48. En

cuanto al número de concordancias, el rango es de 200-2.500 y la media por verbo 384,3.

Tabla 4. Algunos datos numéricos del análisis manual.

	N
Nº verbos	182
Nº patrones	998
Nº patrones alternantes	346 (34,67%)
Nº concordancias	46.086
Nº explotaciones	2.633 (5,71%)
argumentales	881 (33,46%)
figuradas	1.451 (55,11%)
sintácticas	301 (11,43%)

Para desarrollar la base de datos se comenzó dando prioridad a verbos de alta frecuencia según el listado de Davies (2006), por lo que es normal que la mayoría de los verbos sean polisémicos. En el caso de verbos como 'comer', 'cortar', 'levantar' o 'abrir', tienen 30 patrones o más. Pese a estos verbos altamente frecuentes, la mayoría de los que figuran por el momento en Verbario, y que también se encuentran entre los más frecuentes del castellano, tienen entre 2 y 4 patrones. Del total de patrones, un 34,67% son alternantes (en su mayoría, de diátesis media con 'se'). En cuanto al número de concordancias requerido, la mayoría de los verbos no requieren más de 300 concordancias, aunque 9 verbos requieren más de 1.000. Por los motivos ya expuestos en el apartado 2.1.1, es evidente la correlación entre el número de concordancias y el de patrones. Con respecto a las concordancias que se analizaron como explotaciones, hay 2.633 (un 5,71% del total de contextos), un porcentaje muy bajo como corresponde a usos anómalos de los patrones. El tipo de explotación que predomina es la figurada, que corresponde a un 55,11% del total de explotaciones: predominan, por tanto, las creaciones metafóricas que afectan al significado del verbo. Puede consultarse en la web del proyecto Verbario cada análisis realizado. Los datos fruto de este análisis permiten observar las ventajas de un análisis de muestras de corpus a través de una metodología que, aplicada de forma sistemática, facilita la labor de lexicógrafos y lexicólogos y da cuenta de los rasgos léxicos, semánticos y gramaticales estables de la lengua, así como de su variabilidad en el uso.

3.2. Análisis automático

A continuación se examina la calidad de los resultados obtenidos en los experimentos para la automatización de CPA. Se examinan los patrones producidos y se realiza un análisis de errores y un intento de catalogación de estos de cara a futuras acciones remediales.

Para el análisis de los patrones se utilizó el análisis manual de los verbos como '*gold standard*', estableciendo una serie de reglas de conversión ya que existen diferencias

estructurales entre ambos resultados. Hay casos en que el patrón humano y el automático son idénticos, como en el caso del patrón 4 de 'llenar': '[[Humano]] llenar [[Documento]]'. No obstante, existen otros casos en que el análisis automático todavía no alcanza el grado de generalización y concreción del análisis manual, como en el siguiente caso: '[[Human | Eventuality]] llenar [[Building | Location]] {de | con [[Physical Object]]}'; por ejemplo, porque el analizador de dependencias no siempre detecta el complemento de régimen (que está ausente en muchas de las concordancias analizadas porque es un argumento opcional la mayoría de veces), o porque no advierte las alternancias semánticas. Por ese motivo, se aceptan como válidos todos los patrones en los que el análisis sintáctico-semántico se ha realizado correctamente y corresponda a un patrón manual, sin tener en cuenta los dos aspectos mencionados. En la misma línea, se aceptan también como correctos los casos de patrones automáticos que están bien creados, pero no coinciden por completo con el mismo nivel de especificidad del análisis manual. Esto se produce porque el algoritmo selecciona el primer tipo semántico de la taxonomía desde abajo, lo que puede implicar que el tipo semántico seleccionado es más específico que el elegido por el lexicógrafo. Por esta razón, se toman como equivalentes un patrón automático como '[[Human]] llenar [[Building]]' y un patrón manual como '[[Human]] llenar [[Location]]'. Finalmente, también ocurre que, dado que el sistema automático analiza muchos más contextos que el manual, el primero detecta patrones que los lexicógrafos no llegaron a encontrar en sus muestras. Cuando estos patrones son correctos, se cuentan como aciertos aunque no aparecieran en el '*gold standard*'.

Para la evaluación de los resultados se llevó a cabo un examen por parte de un grupo de 4 evaluadores entrenados para tal efecto. Se evaluó una muestra aleatoria de 13 verbos para contrastar los listados de patrones manual y automático y determinar la proporción de patrones correctos en este último. Para los 13 verbos de la muestra se generaron 164 patrones en total. De ello, un 63,41% se consideraron correctos. A modo ilustrativo, y retomando el ejemplo del verbo 'sembrar' (v. apartado 2.1.2), entre los patrones más frecuentes están 'PERSONA sembrar FALTA', donde este tipo semántico agrupa argumentos como 'desconfianza', 'dudas' o 'discordia' (el sistema automático detecta el cariz negativo de este uso, que no se había indicado en el patrón manual, el nº 2a). Le siguen 'PERSONA sembrar ESTADO' ('verdades', 'escándalo', 'confusión', etc.) y 'EVENTUALIDAD sembrar SENTIMIENTO' ('alegría', 'temor', 'xenofobia', 'racismo', etc.). Menos frecuente resulta el que corresponde al patrón prototípico: 'PERSONA sembrar PARTE DE PLANTA' ('flores', 'semilla', 'arroz', 'trigo', 'soja', 'arroz', 'hortaliza') y 'PERSONA sembrar en TERRENO' ('tierras', 'taludes', 'zonas'). Los patrones mencionados fueron considerados correctos pues ya habían sido detectados durante la anotación manual, con leves diferencias.

Entre los errores más frecuentemente encontrados destacan los problemas de lematización (patrones generados a partir de 'siembra' como sustantivo), problemas

del parser (por ejemplo, en 'Ese año sembró soja' se detecta erróneamente 'Ese año' como sujeto) y errores el etiquetador semántico ('algodón' es clasificado como 'material' y esto es erróneo como argumento de 'sembrar'). El principal problema de los resultados, sin embargo, está en la sobregeneración de patrones por ser estos demasiado específicos.

Las cifras anteriores muestran, por tanto, que casi dos tercios de los datos son correctos, lo que supone una cifra suficientemente elevada como para cumplir con el objetivo marcado de que el método pueda ayudar al análisis manual. Sin embargo, queda mucho margen aún para implementar mejoras que permitan crear una herramienta lista para ser usada por equipos de ingenieros, lingüistas o lexicógrafos. No resulta posible comparar el logro alcanzado con otras propuestas, porque hasta la fecha estas no se han elaborado. En el caso de la generación automática de patrones, resulta imposible incluso establecer un método de base o 'baseline', ya que esta solo puede ser útil para evaluar sistemas de clasificación donde existe una determinada probabilidad de éxito con selección aleatoria o con un método trivial.

CONCLUSIONES Y TRABAJO FUTURO

En este artículo se han presentado los resultados de una investigación todavía en curso, pero con lineamientos ya claramente establecidos y con resultados que alcanzan un punto de madurez y que servirán de insumo para futuras investigaciones. Por un lado, se ha presentado la aplicación de una técnica lexicográfica como el CPA a una lengua para la que no fue diseñado, y se ha demostrado la viabilidad del método para identificar patrones. Por otro lado, teniendo en cuenta uno de los problemas de esta técnica, que es la ingente tarea manual que el método requiere, se ha presentado una automatización del método con aproximadamente dos tercios de aciertos. Ambos tipos de datos se ofrecen ya en línea, en lo que constituye, incluso en el presente estado del proyecto, una base de datos léxica del español bastante completa. Ello está permitiendo –y permitirá en el trabajo futuro– contribuir a una visión del léxico en la que se observa empíricamente la tensión entre estabilidad y variabilidad, o norma y explotación en términos de Hanks (2013). La automatización de parte del proceso permite considerar que, en el trabajo futuro, será posible ofrecer una herramienta tecnológica que acompañe a los lexicógrafos y lexicólogos, facilitándoles sus labores de análisis.

Las líneas de actuación a partir de ahora se abren en distintas direcciones. En primer lugar, se pretende continuar desarrollando experimentos para disminuir la tasa de error de los patrones obtenidos automáticamente. Ello involucra un procedimiento de evaluación automática consistente en la comparación algorítmica con el resultado de la anotación manual, lo cual reducirá considerablemente los tiempos de cada experimentación. En segundo lugar, se desarrollará un protocolo para la integración

de las distintas fases del análisis manual y automático de manera de aprovechar los mejores aspectos de cada método. Se considera, por el momento, un sistema que por un lado alivie la parte más mecánica del trabajo manual, pero que por otro lado se supedite a las decisiones del lexicógrafo de aceptar, rechazar o bien corregir o completar los patrones generados de manera automática. Finalmente, se proyecta crear, junto con el equipo del PDEV, una base de datos bilingüe español-inglés que conectará los patrones de ambas lenguas.

REFERENCIAS BIBLIOGRÁFICAS

Alonso Ramos, M. (Dir.) (2018). *Diccionario de colocaciones del español* [en línea]. Disponible en: http://www.dicesp.com.

Andor, D., Alberti, C, Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S. & Collins, M. (2016). Globally normalized transition-based neural networks. Ponencia presentada en el *54th Annual Meeting of the Association for Computational Linguistics* (pp. 2442-2452). ACL: Stroudsburg, PA.

Apresjan, J. (1974). Regular polysemy. *Linguistics,* 142, 5-32.

Atserias, J., Climent, S., Farreres, J., Rigau, G. & Rodríguez, H. (2000). Combining multiple methods for the automatic construction of Multilingual WordNets. En K. Nikolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005* (pp. 143-149). Ámsterdam, Philadelphia: Benjamins Publishing.

Baisa, V., El Maarouf, I., Rychlý, P. & Rambousek, A. (2015). Software and data for Corpus Pattern Analysis. En A. Horák, P. Rychlý & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN) 2015* (pp. 75-86). Brno: Masaryk University.

Battaner, P. (2011). Los verbos de frecuencia alta y el diccionario de aprendizaje. En M. Vázquez Laslop, K. Zimmermann & F. Segovia (Eds.), *De la lengua por solo la extrañeza. Estudios de lexicología, norma lingüística, historia y literatura, en homenaje a Luis Fernando Lara* (pp. 313-332). México, D. F.: El Colegio de México, Centro de Estudios Lingüísticos y Literarios.

Bosque, I. (Dir.). (2004). *Redes. Diccionario combinatorio del español contemporáneo.* Madrid: SM.

Buchholz, S. & Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. En L. Màrquez & D. Klein (Eds.), *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)* (pp. 149-164). ACL: Nueva York.

Davies, M. (2006). *A frequency Dictionary of Spanish. Core vocabulary for learners.* Nueva York: Routledge.

El Maarouf, I., Bradbury, J., Baisa, V. & Hanks, P. (2014). Disambiguating verbs by collocation: Corpus lexicography meets Natural Language Processing. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1001-1006). Reykjavik: ELRA.

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge (Mass.): MIT Press.

Fillmore, C., Wooters, C. & Baker, C. (2001). Building a large lexical databank which provides deep semantics. En B. T'sou, O. Kwong & T. Lai (Eds.), *Proceedings of the Pacific Asian Conference on Language, Information and Computation* (pp. 3-26). Hong Kong: Language Information Sciences Research Centre.

Firth, J. (1935). The technique of semantics. *Meeting of the Philological Society*, *34*(1), 36-72.

Gamallo, P. & González, I. (2012). DepPattern: A multilingual dependency parser. Sesión de demos. Ponencia presentada en el *10th International Conference on Computational Processing of the Portuguese Language* (*PROPOR' 2012*), Coimbra, Portugal.

Grice, H. P. (1975). Logic and conversation. En P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics*, *Vol. 3*: *Speech Acts* (pp. 41-58). Nueva York: Academic Press.

Gross, M. (1981). Les bases empiriques de la notion de predicat semantique. *Langages*, 63, 7-52.

Hanks, P. (2004a). Corpus Pattern Analysis. En G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh Euralex International Congress* (pp. 87-97). Lorient: Université de Bretagne-Sud.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography, 17*(3), 245-274.

Hanks, P. (2009). The linguistic double helix: norms and exploitations. En D. Hlavácková, A. Horák, K. Osolsobě & P. Rychlý (Eds.), *After half a century of Slavonic natural language processing (Festschrift for Karel Pala)* (pp. 63-80). Brno: Masaryk University.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations.* Cambridge, Ma.: MIT Press.

Hanks, P. (2018). *CPA Ontology* [en línea]. Disponible en: http://pdev.org.uk/#onto

Hanks, P. (Ed.) (en proceso). *Pattern dictionary of English verbs* [en línea]. Disponible en: http://pdev.org.uk.

Hanks, P. & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée, 10*(2), 63-82.

Harris, Z. (1985). Distributional structure. En J. J. Katz (Ed.), *The philosophy of linguistics* (pp. 26-47). Nueva York: Oxford University Press.

Jurafsky, D. & Martin, J. (en preparación). *Dependency parsing. Speech and Language Processing* (3a ed.) [en línea]. Disponible en: https://web.stanford.edu/~jurafsky/slp3/14.pdf

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7-36.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Malinowski, B. (1923). The problem of meaning in primitive languages. En C. K. Odgen & I. A. Richards (Eds.), *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism* (pp. 296-336). Cambridge: Cambridge University Press.

Nazar, R. & Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. Ponencia presentada en *el10th International Conference on Language Resources and Evaluation (LREC'16).* European Language

Resources Association, Portorož, Eslovenia (pp. 1485-1492).Nazar, R. & Arriagada, P. (2017). POL: un nuevo sistema para la detección y clasificación de nombres propios. *Procesamiento del Lenguaje Natural,* 58, 13-20.

Porto Dapena, Á. (2014). *La definición lexicográfica.* Madrid: Arco Libros.

Renau, I. (2010). Estudio lexicográfico de verbos del español en relación con los usos de 'se' y con vistas a la confección de un diccionario para extranjeros. *Interlingüística,* XX, 1-12.

Renau, I. & Nazar, R. (2018). *Verbario. Base de datos de verbos del español* [en línea]. Disponible: http://www.verbario.com.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Ponencia presentada en el *International Conference on New Methods in Language Processing,* Manchester, United Kingdom.

Sinclair, J. (Ed.). (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing.* Glasgow: HarperCollins.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse.* Londres: Routledge.

Yallop, C. (2004). Words and meaning. En M.A.K. Halliday, A. Cermáková, W. Teubert & C. Yallop (Eds.), *Lexicology and Corpus Linguistics* (pp. 23-72). Londres: Continuum.

NOTAS

[1] La investigación contenida en el artículo fue presentada en el *Workshop* de *Procesamiento del Lenguaje Natural 2016*. El artículo fue invitado, evaluado y aprobado para su publicación en Revista Signos. Estudios de Lingüística.

[2] Como ya se indicó, el proyecto Verbario emplea la misma estructura de base de datos y la misma nomenclatura que el PDEV, por ello incluye algunas etiquetas en inglés.

* AGRADECIMIENTOS

Este trabajo se publica con el apoyo del proyecto Fondecyt 11140704, dirigido por Irene Renau. Los autores agradecen los comentarios de los evaluadores y, como siempre, la generosidad y apoyo de Patrick Hanks.