REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2022 PUCV, Chile 55(110) 732-761

Estructura argumental del nombre: Generación automática

Noun Argument Structure: Automatic Generation

María José Domínguez Vázquez
UNIVERSIDAD DE SANTIAGO DE COMPOSTELA
ESPAÑA
majo.dominguez@usc.es

Recibido: 12-V-2020 **/ Aceptado:** 04-I-2022
DOI: 10.4067/S0718-09342022000300732

Resumen

Este estudio aborda los fundamentos lingüístico-teóricos y fases metodológicas que dan sustento al desarrollo de dos prototipos para la generación automática de la estructura argumental de frases nominales simples y complejas en alemán, español y francés. El diseño de estas herramientas, Xera y Combinatoria, combina diversas aproximaciones teóricas y metodológicas, tales como la gramática de valencias, la semántica de prototipos, las ontologías de WordNet y el procesamiento y la generación del lenguaje natural. Ambos generadores aportan una nueva vía no solo para el estudio de la combinatoria del léxico, sino también para un desarrollo más automatizado de los diccionarios de valencias.

Palabras Clave: Valencia nominal multilingüe, prototipos y ontología, GLN, WordNet, *word embeddings*.

Abstract

This study deals with the linguistic, theoretical foundations and methodological phases that support the development of two prototypes for the automatic generation of the argument structure of simple and complex nominal phrases in German, Spanish, and French. The design of these tools, Xera and Combinatoria, combines several theoretical and methodological approaches, such as the valency grammar, the semantic prototypes, the WordNet ontologies, and the processing and generation of natural language. Both generators provide a new approach not only for researching the lexical combinatory patterns, but also for a more automated development of valency dictionaries.

Keywords: Multilingual noun valency, prototypes and ontology, NLG, WordNet, word embeddings.

INTRODUCCIÓN[1]

A partir de los años 90 se intensifica la investigación en el campo de la generación automática del lenguaje natural (GLN). Hoy en día, ya es posible generar de modo automático partes meteorológicos, informes deportivos, resúmenes médicos, recomendaciones o diálogos (Vicente, Barros, Agulló, Peregrino & Lloret, 2015; Nallapati, Zhou, dos Santos, Gulçehre & Xiang, 2016; Sordoni, Galley, Auli, Brockett, Ji, Mitchell, Nie, Gao & Dolan, 2015), crear textos a partir de imágenes y a la inversa (Otter, Medina & Kalita, 2020; Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár & Zitnick, 2014) y, en algunos casos, generar chistes, poemas o historias casi sin datos *input* de partida (Roemmele, 2016). A su vez, se han propuesto arquitecturas de consenso (Reiter & Dale, 2000) y se investigan y diseñan métodos para la evaluación de los textos generados automáticamente (Hashimoto, Zhang & Liang, 2019; Vicente et al., 2015), en donde tanto la calidad de lo generado como su diversidad se configuran como ejes centrales.

Encontramos menos recursos diseñados para la generación automática del lenguaje con una clara aplicación lexicográfica o desde/para la lexicografía. Es verdad que progresivamente contamos con más y mejores herramientas para el análisis lingüístico, la extracción automática de datos y, en general, para el procesamiento del lenguaje natural:

- Los corpus están cada vez mejor dotados con análisis estadísticos más precisos y con aplicaciones para analizar las propiedades distributivas y sintagmáticas del léxico, como, por ejemplo, Sketch Engine[2] o COSMAS II[3].
- Recursos de diferente tipología aportan descripciones semántico-argumentales, en especial para el inglés —como Verbnet[4], Propbank[5], PDEV[6] o VerbAtlas[7]— o, como ejemplo del español AnCora[8], y se ha avanzado en estudios relacionados con la minería de opinión o el análisis del sentimiento (SentiWordNet[9]).
- Se muestran avances en la extracción automática de datos de aplicación lexicográfica (Krek, 2019), de esquemas (*patterns*) de los diccionarios (Renau & Nazar, 2016; Renau, Nazar, Casto, López & Obreque, 2019) o de equivalentes (Gamallo & Pichel, 2007). A su vez, redes semánticas como Wordnet (2.2.2.) permiten desarrollar aplicaciones de diccionarios para dispositivos móviles. Un ejemplo es el *Dicionario Galnet* de Gómez Guinovart en Google play[10].
- Contamos también con herramientas de análisis como *Freeling* (Padró, Collado, Reese, Lloberest & Castellón, 2010) o generadores basados en reglas (por ejemplo, la traducción asistida con *Apertium*, Armentano-Oller, Corbí-Bellot, Forcada, Ginestí-Rosell, Bonev, Ortiz-Rojas, Pérez-Ortiz, Ramírez-Sánchez & Sánchez-Martínez, 2005).

- Existe, además, *software* que permite la compilación de diccionarios (TshwaneLex[11]), y diferentes propuestas para la generación automática de diccionarios (Héja & Takács, 2012; Kabashi, 2018; Delli Bovi & Navigli, 2017), de artículos lexicográficos (Geyken, Wiegand & Würzner, 2017) o bien de alguna de sus partes, como los ejemplos (Kilgarriff, Husák, McAdam, Rundell & Rychlý, 2008; Kosem, Koppel, Kuhn, Michelfeit & Tiberius, 2019).

Sin embargo, no conocemos generadores de patrones argumentales nominales mono y biargumentales para la frase nominal con fundamentación, información y acceso sintáctico-semántico valencial. Tanto el prototipo *Xera* —estructura argumental monoactancial— como *Combinatoria* —estructura biargumental— generan frases simples y complejas respectivamente, aceptables desde un punto de vista sintáctico y coherentes desde un punto de vista semántico para el español, alemán y francés[12].

Atendiendo al hecho de que los corpus manejados para las lenguas que son objeto de descripción no cuentan con anotación semántica —lo cual no posibilita una búsqueda según este criterio— y que los mismos son imprescindible para poder desambiguar diferentes significados, el principal hándicap en el desarrollo de nuestros generadores fue el diseño de una metodología que nos permitiera extraer, describir y procesar los datos lingüísticos con información sintáctico y semántica combinatoria —sintagmática y paradigmática— de modo que estos se pudieran programar y generar de modo automático para las tres lenguas.

No es posible —ni se pretende— intentar resumir aquí la importancia del conocimiento léxico para el desarrollo de nuevas herramientas —tanto como recursos independientes como integrables en otros—, pero sí que es importante recordarlo. Así, por ejemplo, desde la propia GLN se indica que, en las técnicas híbridas, en donde se asocian la estadística y el conocimiento lingüístico:

> "hay una fuerte tendencia a incorporar consideraciones pragmáticas y semánticas bajo el supuesto de que tanto el significado del discurso que conforma la salida de un sistema de GLN, como el lenguaje en tanto instrumento de comunicación, únicamente adquiere su correcto significado cuando se sustenta sobre tal entramado de conocimiento" (Vicente et al., 2015: 751).

También desde la Lexicografía se incide en la importancia del conocimiento léxico no solo para la comprensión y la desambiguación del significado, sino también porque dicho conocimiento permite el desarrollo de diferentes tareas en el campo del procesamiento del lenguaje natural (Trap-Jensen, 2018).

Nuestra metodología, en definitiva, tiene una fuerte fundamentación semántico-lingüística, pero a su vez presenta una notable recurrencia y retroalimentación de diferentes recursos para el procesamiento y generación del lenguaje. Es objeto de este

estudio mostrar los fundamentos lingüísticos y metodológicos en los que se sustentan los generadores. Para tal fin, el capítulo 1 aporta una visión de conjunto de los fundamentos teórico-lingüísticos y metodológicos generales que sirven de punto de partida para el desarrollo de los generadores. Del método combinado de análisis y de las herramientas manejadas para su diseño se ocupa el capítulo 2.

1. Del diccionario PORTLEX a los generadores del lenguaje *Xera* y *Combinatoria*

1.1. Fundamentos teóricos generales

El diccionario multilingüe en línea de valencias del nombre para el alemán, español, francés, gallego e italiano PORTLEX[13] (Domínguez Vázquez & Valcárcel, 2020) sirve de punto de partida para el diseño de los proyectos *MultiGenera* y *MultiComb*, en cuyo seno se desarrollan diferentes herramientas de generación automática, entre ellas *Xera* y *Combinatoria*. Ambos simuladores nacen en estrecha relación con las dificultades constatadas en el desarrollo de dicho diccionario y pretenden suplir la constatada carencia de recursos con información suficiente, distintiva y orientada al usuario —humano o máquina— en un contexto de lenguas extranjeras y en casos de producción de combinatoria argumental valencial nominal. Teniendo en cuenta la finalidad de dichos generadores, para su desarrollo se diseñó e implementó una metodología específica (vid. 2.), la cual se suma al marco teórico común de dichos recursos y al diccionario (vid. 1.2.) de la gramática de dependencias y valencias[14].

A diferencia de los estudios gramaticales y lexicográficos sobre la valencia del verbo, la escasez de estudios sobre el nombre y su valencia es notable[15]. Dicha situación es fácilmente atribuible a la falta de consenso sobre su estatus como portador valencial, siendo considerado avalente o bien un simple heredero del potencial argumental de las palabras de las que deriva (Ágel, 2000; Eisenberg, 2006). La literatura científica cuenta, sin embargo, con sólidas propuestas sobre el carácter valente de los sustantivos, por tanto, sobre su capacidad de abrir casillas funcionales específicas (Teubert, 1979; Zifonun, Hoffman & Strecker, 1997; Eroms, 2000; Hölzner, 2007; Engel, 2009). Nuestros recursos, que beben de estas fuentes, se fundamentan en un concepto propio de valencia multidimensional de aplicación multilingüe. Este permite describir la interfaz sintáctico-semántica de diferentes clases de palabras atendiendo a su valencia cuantitativa (número de casillas funcionales y su interacción) y cualitativa (tipología) en diferentes lenguas y niveles descriptivos (por ejemplo, conceptual, semántico o sintáctico). Dichas herramientas aportan, en definitiva, información sobre las casillas funcionales-valenciales o complementos específicos[16] del portador valencial (Engel, 2004) según la actualización concreta de sus posibles acepciones de significado.

En nuestro modelo, la aproximación a la valencia del nombre cuenta con dos vertientes: la valencia pasiva y la activa. Como valencia pasiva del nombre entendemos los contextos oracionales en los que se puede enmarcar un patrón argumental nominal, por tanto, este se analiza partiendo de unidades jerárquicamente superiores en la escala dependencial. Sirven como ejemplo (1) y (2):

(1) *Sergio observa/contempla/rechaza la discusión de los niños sobre el videojuego*.
Sujeto + Verbo + Objeto^{Nombre}(+ Argumentos)

(2) *Susana participa en/contribuye a la discusión de los representantes sindicales sobre el aumento de los impuestos*.
Sujeto + Verbo + Prepositivo^{Nombre}(+ Argumentos)

La descripción de la valencia activa supone establecer el potencial combinatorio específico del nombre, su patrón argumental, así como la combinatoria de las realizaciones argumentales en el eje sintagmático[17]. Un ejemplo de dicha aproximación es en el caso del sustantivo DISCUSIÓN: [ARG1 ('Aquel/Aquello que realiza una acción'): frase preposicional: de + ARG3 ('Aquel o aquello no afectado: Tema'): frase preposicional: sobre] (vid. Figura 3 y 4). A esta descripción le acompaña el análisis del catálogo de rasgos categoriales-ontológicos de cada argumento en el eje paradigmático. Por tanto, esta responde a la pregunta de qué entidades ontológicas pueden ocupar una casilla argumental.

La descripción de la valencia activa se articula en tres ejes centrales:

- el plano morfosintáctico: funciones sintácticas o argumentos nominales junto con sus correspondientes realizaciones formales (vid. Figura 3);
- el plano sintáctico argumental o el patrón argumental (vid. Figura 4);
- el plano semántico o el significado combinatorio, compuesto por el significado relacional y categorial-ontológico (Engel, 1996).

Siguiendo a Domínguez Vázquez, Engel y Paredes (2017) y Engel (2004) diferenciamos desde un punto de vista relacional cuatro roles focalizados o específicos[18]: el Agentivo, el Afectivo, el Locativo y el Clasificativo. Dichos roles pueden verse indexados, lo que permite mayor granularidad en la descripción semántica. Engel (1996) fundamenta dicha clasificación con 'solo' cuatro relatores o roles semánticos en el hecho de que todas las demás diferencias de significado —las cuales conducen a un catálogo variable en cantidad y calidad de roles— se deben a elementos del significado inherente del verbo, no del combinatorio. En nuestro modelo, la descripción del significado categorial, esto es, de las entidades ontológicas que pueden ocupar un determinado *slot* funcional (las denominadas 'restricciones léxicas' en REDES, 2004) parte de los inventarios de los diccionarios de valencias, tales como Domínguez Vázquez, Engel y Paredes Suárez (2017) o E-VALBU[19], y evoluciona hacia un repertorio ontológico más extenso y granular (vid. 2.2.2.).

En el siguiente apartado se desgrana de modo pormenorizado la teoría aplicada en los generadores, así como los niveles descriptivos señalados previamente.

1.2. Fundamentos teóricos específicos

Partimos de la premisa de que los hablantes nativos cuentan con una escena prototípica mental de cómo se desarrollan determinados eventos, acciones o procesos, esto es, que tienen una representación mental-cognitiva de un tipo de evento (vid. en la Figura 1 los escenarios MOVIMIENTO y LOCACIÓN). Asimismo, las escenas y escenarios funcionan también como *tertium comparationis* intra e interlingüístico. Por tanto, el nivel descriptivo de orden jerárquico superior es el escenario —una red de escenas— (Fillmore, 1977), el cual es supralexemático y cubre diversos recursos expresivos y construcciones lingüísticas, proporcionando un marco de referencia para la presentación de las escenas y subescenas a las que están ligados los diferentes recursos expresivos compatibles semántico-conceptualmente. El rol semántico específico o focalizado (vid. 1.1.) es el que permite la adjudicación de un sustantivo, actualizado en una expresión y contexto dado, a un escenario y a una escena determinados. Así, en (3) se observa un rol 'Locativo$_{directivo}$' de destino —acompañado además por un 'Afectado$_{mutativo}$'[20]— lo que permite describir al nombre VIAJE como realización del escenario MOVIMIENTO y de la escena DESPLAZAMIENTO. A diferencia de este, en (4) el rol focalizado es un 'Locativo$_{situativo}$', por lo que ESTANCIA se atribuye al escenario LOCACIÓN y a la escena LOCACIÓN SITUATIVA:

(3) *El viaje de Mario*[Afectado] *al circo*[Locativo_directivo]
(4) *La presencia de los deportistas*[Afectado] *en Berlín*[Locativo_situativo]

De este modo, el rol focalizado 'Locativo$_{directivo}$' sustenta la pertenencia de diferentes expresiones a las escenas, subescenas y al escenario MOVIMIENTO, frente, por ejemplo, al de la LOCACIÓN, que cuenta con un rol focalizado 'Locativo$_{situativo}$', tal y como resume la Figura 1:

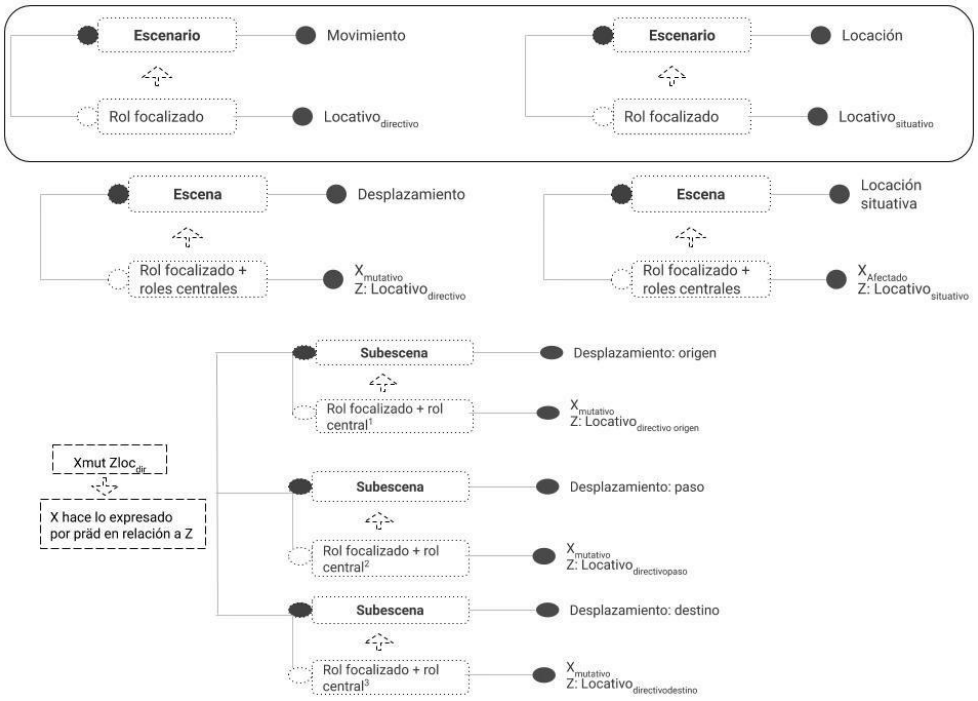

Figura 1. Escenario MOVIMIENTO frente a LOCACIÓN.

La delimitación entre diferentes escenas y entre diferentes subescenas depende del rol focalizado, así como del número y configuración de los roles centrales. Por tanto, todo sustantivo que pertenezca a una escena o escenario comparte con estos el marco conceptual-semántico en cuanto al número de argumentos centrales y roles semánticos. El modelo permite, por tanto, delimitar realizaciones como las siguientes:

(5) [Xmut Zloc]: *Pedro*[Sujeto_mutativo] *viaja a Roma*[Locativo_directivo]; *el viaje de Pedro*[Sujeto_mutativo] *a Roma*[Locativo_directivo]

(6) [Yagente Xmut Zloc]: *Juan*[Sujeto_agente] *traslada sus muebles*[Objeto_mutativo] *al piso nuevo*[Locativo_directivo]; *el traslado de los muebles*[Objeto_mutativo] *al piso nuevo*[Locativo_directivo] *por Juan*[Sujeto_agente]

Ambos ejemplos son, por tanto, representantes de la escena DESPLAZAMIENTO, dado que cuentan con un rol focalizado ZLocativo$_{directivo}$ y un rol central Xmut (vid. Figura 2). El ejemplo (5) representa una subescena del patrón [Xmut ZLocdir] con un directivo de destino: es la indexación del directivo lo que lo diferenciaría de otros trazos como 'paso' u 'origen' (vid. Figura 1). Por otra parte, la diferencia entre (5) y (6) radica en el número y en la tipología de los roles:

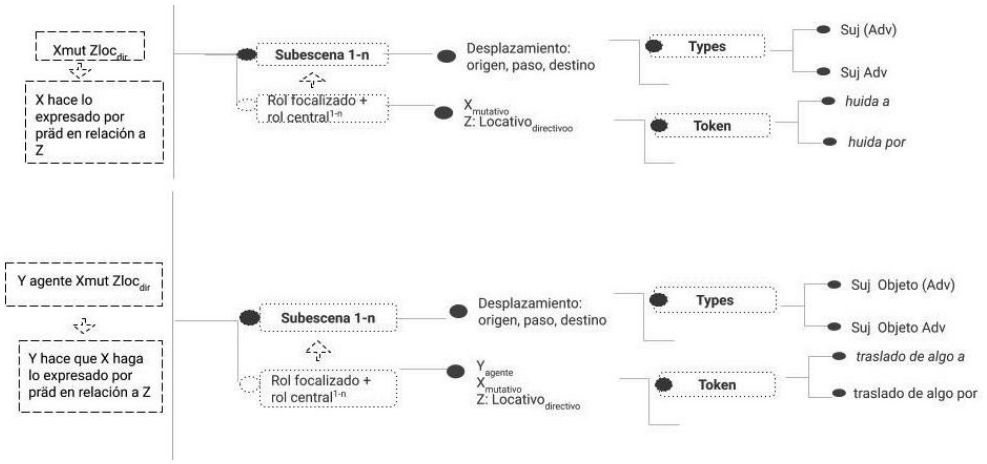

Figura 2. Ejemplo de subescenas del DESPLAZAMIENTO.

Tal y como se observa en las Figuras 1 y 2, la descripción lingüística se retroalimenta, en definitiva, del principio de recurrencia: la pertenencia o no a una clase se debe al hecho de que se comparta (o no) una serie de rasgos con los planos jerárquicamente superiores.

La Figura 2 también refleja el necesario análisis del material expresivo tanto desde un punto de vista semántico como sintáctico y distribucional (*token* y *types*). Por tanto, necesitamos analizar cuántos y qué argumentos valenciales actualiza un sustantivo en una realización concreta, qué características semánticas muestra, así como sus variedades morfosintácticas y combinatorio-distributivas. Así, tras la delimitación de los roles, es necesario un análisis sintáctico-tipológico y formal de las realizaciones expresivas de cada rol semántico, que de modo descriptivo representa la configuración argumental de DISCUSIÓN:

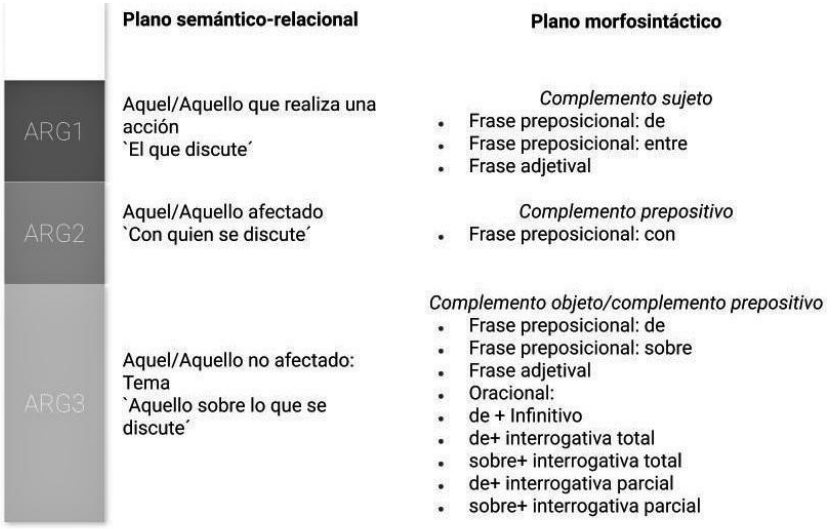

Figura 3. Argumentos del sustantivo DISCUSIÓN.

En la Figura 3 observamos que una frase preposicional, como, por ejemplo, la introducida por 'de', puede expresar diferentes roles semánticos, dependiendo de la unidad regente y de la presencia de otros argumentos y sus interacciones. Es la interfaz sintáctico-semántica la que nos permite analizar dichas expresiones y distinguirlas: 'de Pedro' en (7) funciona de complemento sujeto, mientras que la frase introducida por 'de' en (8) ejerce de complemento prepositivo o suplemento[21]. Este tipo de resultados no se alcanzan mediante los corpus estándar.

(7) *La discusión de Pedro con Juan - Pedro discute con Juan.*
(8) *La discusión de botánica - Se discute de botánica.*

A la descripción del significado relacional y sintáctico le sigue el análisis semántico categorial: en el diccionario de valencias aplicamos las jerarquías de rasgos categoriales, comúnmente manejadas en la lexicografía valencial, por ejemplo, [humano], [material], [situación], [objeto]. Dicha descripción ontológica juega un papel fundamental desde un punto de vista lingüístico, no solo porque los roles semánticos expresados son desempeñados por lexemas con determinados rasgos ontológicos-categoriales, sino por el hecho de que estos muestran paralelismos con ontologías manejadas en aplicaciones informáticas o en redes léxico-semánticas como WordNet, a las que recurriremos en el diseño de nuestros generadores (vid. 2.2.2.).

De este modo, podemos determinar el patrón sintáctico-semántico de cada argumento, pero también la combinatoria y distribución argumental, y esto para cada una de las realizaciones de cada uno de los actantes en todas sus posibles variantes y posiciones. En la Figura 4 se representa parcialmente la combinatoria y distribución del ARG1 ('Aquel/Aquello que realiza una acción'), con el ARG2 ('Aquel o aquello

afectado por la acción') y con el ARG3 ('Aquel o aquello no afectado: Tema') del sustantivo DISCUSIÓN:

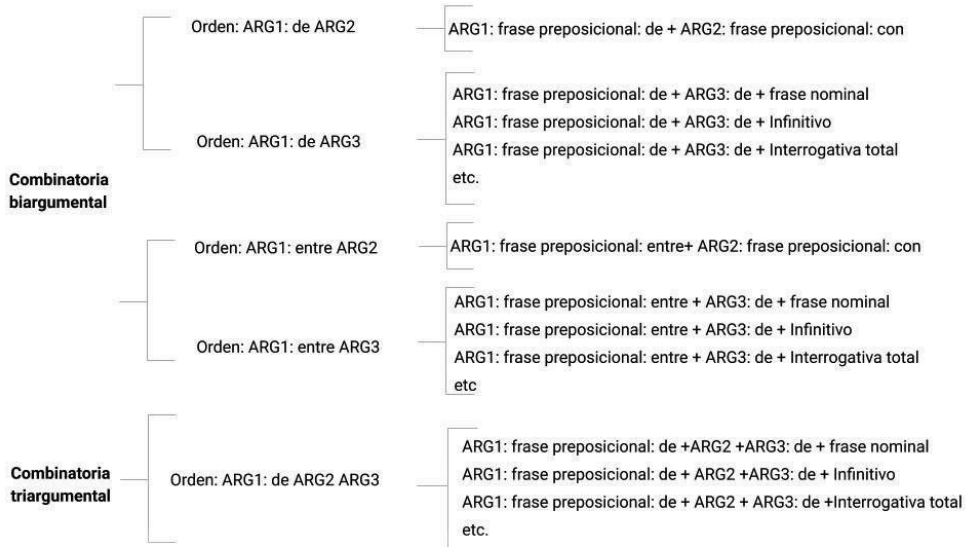

Figura 4. Muestra parcial de esquemas bi y triargumentales del sustantivo DISCUSIÓN.

Téngase en cuenta que para una completa representación de los esquemas actanciales de dicho sustantivo es necesario plasmar todas las posibles distribuciones argumentales y combinatorias interactanciales en sus diferentes realizaciones[22], en definitiva, la posible combinación de todos con todos, y para todas las lenguas. A su vez, como es esperable, las lenguas cuentan con diferentes recursos expresivos. Así, el alemán permite para la expresión de un ARG1 realizaciones con genitivo, con la preposición *von*, adjetivos y compuestos; el español frases preposicionales, adjetivales y realizaciones N+N, al igual que el francés, si bien este último recurre más habitualmente a este recurso expresivo (Valcárcel, 2017)[23]. La diversidad formal supone una dificultad añadida a la complejidad cuantitativa de un diccionario valencial multilingüe como PORTLEX, puesto que documentar mediante corpus todas las posibles realizaciones y sus combinatorias resulta ardua tarea. Así, compilar ejemplos para la realización monoargumental de algunas de las realizaciones, como el caso del ARG1 de DISCUSIÓN mediante adjetivos, compuestos o N+N es muy laborioso, ya que en la consulta de los corpus no es posibles discriminar, por ejemplo, entre argumentos específicos y no específicos. Buena prueba de ello son los datos recabados sobre los 50 primeros adjetivos en coaparición con DISCUSIÓN que arrojan las búsquedas CQL en Sketch Engine: i) circunstanciales —como 'discusión interna' (en 6. posición de frecuencia) o 'discusión seria' (en 7. posición); ii) complementos específicos que no representan el rol buscado— como 'discusión salarial' (3. posición); y iii) complementos específicos que requieren ser desambiguados —'discusión política' (1. posición) o 'discusión parlamentaria' (4. posición):

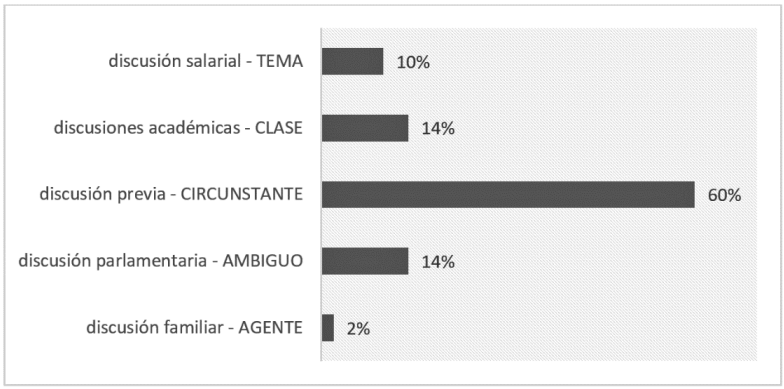

Figura 5. Análisis semántico del patrón Adjetivo + DISCUSIÓN.

Un problema semejante plantea la compilación de compuestos en alemán (Figura 6). En este caso, para obtener ejemplos del rol resultativo (por ejemplo, *LEHRERtod* - muerte DEL PROFESOR), se requiere un análisis manual de todos los datos extraídos, procedimiento costoso en tiempo a la hora de documentar el potencial combinatorio argumental:

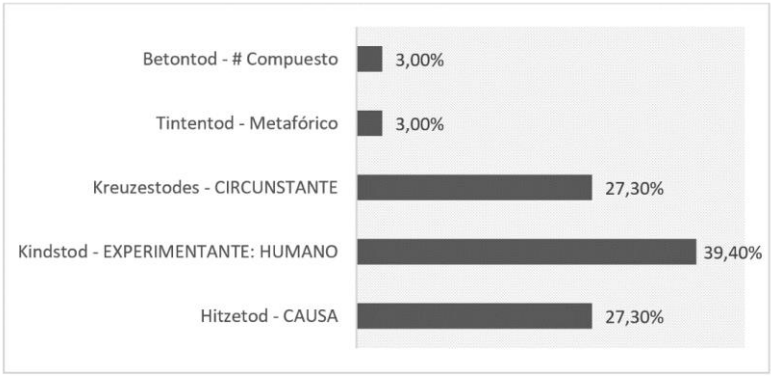

Figura 6. Análisis semántico de compuestos en alemán con TOD.

Finalmente, pero no por ello menos importante, se precisa describir una serie de restricciones de interacción entre los argumentos, como el hecho de que existen relaciones de bloqueo formal, semántico y distribucional, actantes excluyentes o condicionados, etc. Estos son de especial relevancia para una producción adecuada en una lengua extranjera. Así, siguiendo con el ejemplo de DISCUSIÓN, la realización del argumento AGENTE con la preposición 'entre' requiere la realización en plural o mediante una coordinación —entre uno y otro.

En resumen, un estudio sistemático de cada una de las lenguas objeto de análisis atendiendo a su combinatoria argumental —variedad relacional y categorial junto con la morfosintáctica— supone ya no solo analizarla y documentarla, sino además proponer ejemplos en consonancia con el tipo de diccionario y usuario. En concreto,

ejemplos sencillos que reflejen la variabilidad combinatoria y que sean manejables por parte de aprendientes de lenguas extranjeras[24]. En vista de que este procedimiento no resultaba satisfactorio a causa de la complejidad intrínseca del análisis y la necesaria adecuación de los ejemplos atendiendo al propósito del diccionario PORTLEX, decidimos diseñar generadores automáticos del lenguaje natural.

2. Prototipos de generación argumental

2.1. Cuestiones introductorias

Para el diseño de los generadores concebimos un método analítico sumatorio, esto es, que aunara los fundamentos teóricos y metodológicos aplicados en el diccionario con una metodología que comprendiera además i) la extracción automática de datos desde recursos de PLN y la interoperabilidad de recursos; ii) el análisis de corpus, bases de datos de coocurrencias y *wordnets*; iii) la semántica de prototipos; así como iv) la evaluación de datos que arrojarán los propios multigeneradores para el español, alemán y francés.

Los prototipos de generación automática monoargumental, *Xera*, y biargumental, *Combinatoria*, están en funcionamiento y constante actualización. En su calidad de prototipos han sido probados en 20 sustantivos por lengua[25]. La representatividad de dichos sustantivos atendiendo a su pertenencia a diferentes escenarios y escenas conceptuales (vid. 1.2.), su carácter de portadores valenciales y la representación mediante estos de un amplio abanico de combinatorias sintácticas y semántico-categoriales han sido criterios centrales para su selección. A su vez, se contemplan no solo sustantivos derivados de otras clases de palabras, como HUIDA, PREGUNTA[26], sino también aquellos no derivados, tales como VIDEO o TEXTO:

Tabla 1. Selección de sustantivos para los prototipos.

SUSTANTIVOS	ESCENARIO
HUIDA MUDANZA VIAJE	MOVIMIENTO
PRESENCIA AUSENCIA ESTANCIA	LOCACIÓN
DISCUSIÓN	EXPRESIÓN
TEXTO	
PREGUNTA CONVERSACIÓN RESPUESTA VIDEO	
OLOR	CLASE
SABOR COLOR ANCHO	
MUERTE	AFECCIÓN
AUMENTO DOLOR AMOR	

La Figura 7 presenta la interfaz de consulta de *Xera* y la Figura 8 el tipo de ejemplos que ofrece:

Figura 7. *Xera*: Interfaz de consulta.

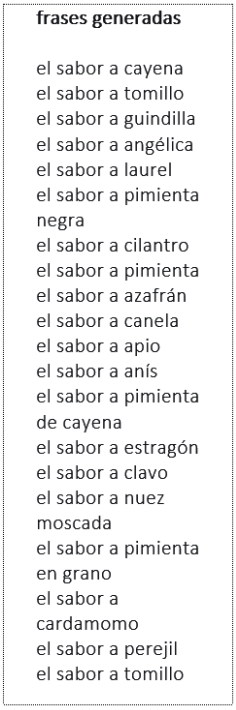

Figura 8. *Xera*: Volcado parcial de datos.

Por su parte, la herramienta *Combinatoria* permite obtener estructuras biargumentales. La Figura 9 da cuenta de la interfaz de consulta de datos, la 10 presenta las frases generadas por el prototipo con y sin el filtro de los métodos predictivos aplicados (*word2vec* y *fastText*; vid. 2.2.3.)[27]:

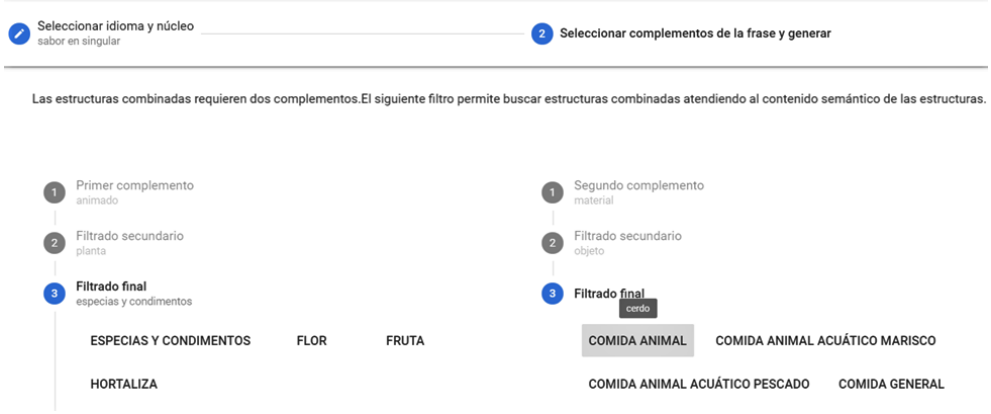

Figura 9. Herramienta *Combinatoria*: Interfaz de consulta.

frases generadas por el prototipo	frases generadas por el prototipo aplicando word2vec
el sabor a pimienta del hígado de ternera	el sabor a azafrán de los costillares
el sabor a pimienta de la carne roja	el sabor a apio de las pechugas
el sabor a tomillo de la panceta	el sabor a azafrán del lomo
el sabor a pimentón de los filetes	el sabor a canela de los berberechos
el sabor a perejil del pescado azul	el sabor a anís del pescado
el sabor a orégano del costillar	el sabor a orégano del marisco
el sabor a tomillo de las carnes de buey	el sabor a especias del pescado
el sabor a pimentón de la anguila	el sabor a perejil de los bígaros
el sabor a chile del pescado azul	el sabor a apio de los solomillos
el sabor a anís del salmonete	el sabor a orégano de los filetes
el sabor a apio de la chicharra	el sabor a cardamomo de los pescados
el sabor a nuez moscada del rape	el sabor a especias de las gambas
el sabor a comino del gallo	el sabor a guindilla del rape
el sabor a tomillo del rape	el sabor a canela de los filetes
el sabor a laurel de los besugos	el sabor a azafrán del verdel
el sabor a eneldo de la pata de cangrejo	el sabor a pimienta del abadejo
el sabor a estragón del pescado azul	el sabor a azafrán del jamón
el sabor a guindilla del congrio	el sabor a pimienta de las cigalas
el sabor a cardamomo de las gambas	el sabor a pimentón de las carnes
el sabor a pimienta de la platija	el sabor a tomillo del rodaballo
el sabor a eneldo de las carnes de búfalo	el sabor a azafrán de los besugos
el sabor a guindilla de los jamones	el sabor a perejil del salmonete

Figura 10. Herramienta *Combinatoria*: Volcado de datos.

2.2. Fundamentos teóricos y metodológicos

2.2.1. Cuestiones introductorias

El desarrollo de los prototipos se organiza en diferentes fases metodológico-analíticas, habiéndose recurrido, en cada una de ellas, a recursos ya disponibles o creados *adhoc* (Domínguez Vázquez, Solla & Valcárcel, 2019). Una visión de conjunto de dichas fases y herramientas se resume en la siguiente figura:

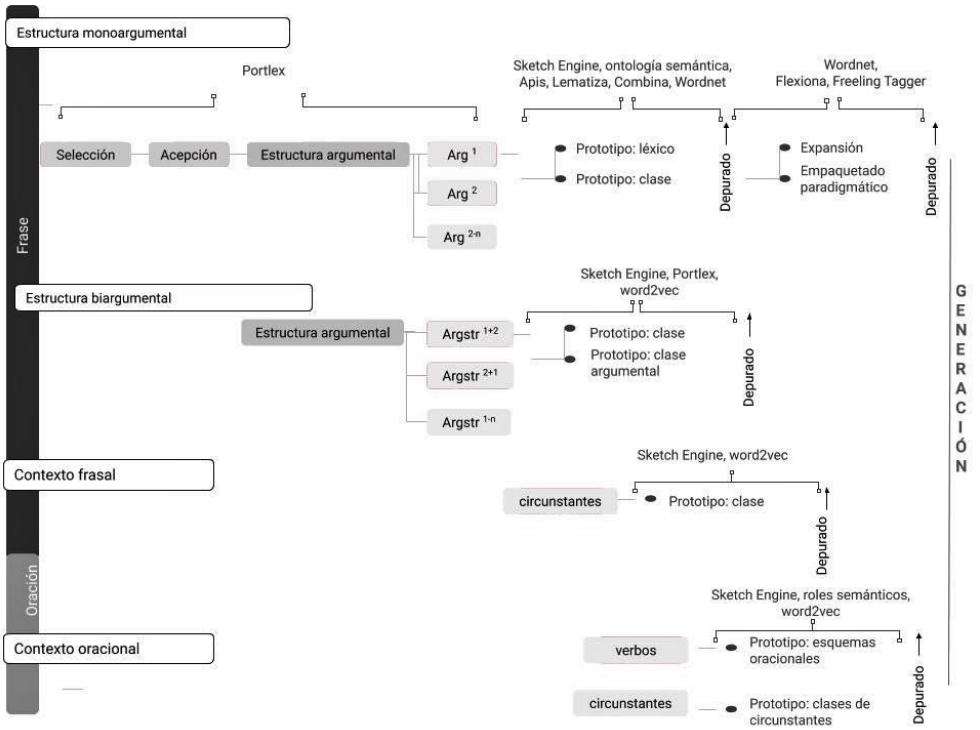

Figura 11. Metodología, fases y recursos.

Los procedimientos seguidos en el análisis y generación de los patrones monoargumentales y biargumentales, así como las herramientas que los facilitan se desglosan en los apartados siguientes. Nos detendremos, en especial, en la estructura monoargumental, pilar fundamental para la generación de los contenidos.

2.2.2. Establecimiento de patrones monoargumentales

El diccionario PORTLEX (vid. 1.2) nos proporciona los esquemas argumentales nominales —esto es la valencia cuantitativa y cualitativa de los sustantivos. Dado que el objetivo central es generar automáticamente su argumentación y potencial combinatorio —correcto, además de semánticamente coherente— resulta imprescindible saber qué candidatos léxicos pueden cubrir el eje paradigmático de las casillas argumentales para de este modo dotar a la herramienta *Xera* de esta información. En el caso de DISCUSIÓN necesitamos saber, por ejemplo, cómo se ocupan el ARG2=[con alguien] y ARG3=[sobre algo] en la siguiente estructura argumental:

Esquema argumental	"determinante", "{adjectivo o}", "nucleo", "{adjectivo o}", "con", "determinante", "actante ARG2", "sobre", "determinante", "actante ARG3"
Significado categorial	ARG2: [humano] [institución]; ARG3: [abstracto]
Modelo de ejemplo	*La acalorada discusión con la profesora sobre el resultado*

Nos encontramos, pues, en la fase de prototipado (vid. Figuras 12 y 13). El primer paso consiste en analizar semánticamente las (co)apariciones sintácticas extraídas de *Sketch Engine* según su frecuencia. Por tanto, necesitamos obtener en primer lugar candidatos prototípicos para cubrir *slots* funcionales concretos, esto es, ejemplares de una categoría en una casilla determinada de una estructura argumental de un sustantivo concreto. Continuando con el ejemplo de DISCUSIÓN, *Sketch Engine* vuelca para el *slot* ARG2=[humano] en la estructura [DISCUSIÓN + con + determinante + NOMBRE] las siguientes unidades léxicas en relación paradigmática —en este orden—: *jefe* / *agentes* / *docentes* / *representantes* / *participantes* / *asistentes* / *juez* / *periodista* / *paciente*, etc. Estos serán, a *grosso modo*, nuestros prototipos léxicos. En esta fase de prototipado (Figura 12) combinamos los inventarios de significado categorial aplicados en la lexicografía valencial (1.1.) con los datos que arroja *Lematiza*[28]. Esta herramienta nos permite analizar los documentos exportados del corpus *Sketch Engine* proporcionándonos automáticamente el lema de cada argumento con las variantes de significado recogidas en las ontologías de *WordNet*. Nuestro análisis, sin embargo, va un paso más allá: a partir de los prototipos léxicos determinamos clases semánticas prototípicas, las cuales resultan de aplicar una ontología léxica de elaboración propia (Domínguez Vázquez, Valcárcel & Bardanca, 2021). Un ejemplo simple de prototipado se observa en la siguiente figura:

	Prototipo léxico	1. nivel	2. nivel	3. nivel
DISCUSIÓN "con", "determinante", "Argumento 2",	jefe	animado	humano	cargo
	agentes	animado	humano	profesión
	docentes	animado	humano	profesión
	periodista	animado	humano	profesión
	representantes	animado	humano	condición humana
	participantes	animado	humano	condición humana
	asistentes	animado	humano	condición humana
	paciente	animado	humano	condición humana

Figura 12. Clases semánticas prototípicas.

A partir de este análisis, podemos concluir, por tanto, que los sustantivos que pertenecen a una clase semántica del tipo [ANIMADO HUMANO CARGO],

[ANIMADO HUMANO PROFESIÓN] y [ANIMADO HUMANO CONDICIÓN HUMANA] pueden conformar el eje paradigmático del argumento [con + determinante +ARG2] de DISCUSIÓN. Esta fase de prototipado (Figura 13) nos permite establecer los prototipos léxicos y las clases semánticas prototípicas, conceptos esenciales por posibilitar la expansión léxica y, de este modo, la ampliación del número de candidatos léxicos de cada *slot* funcional.

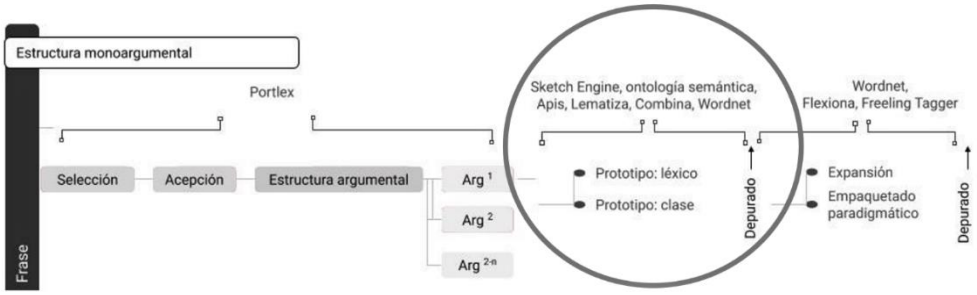

Figura 13. Prototipado léxico.

A su vez, resulta imprescindible subrayar que los rasgos categoriales desempeñan un papel fundamental en el desarrollo de los generadores, puesto que hacen las veces de nexo de unión con las clases y atributos de las categorías de *WordNet* (Gómez Guinovart & Solla Portela, 2018). Nos encontramos ahora en la fase de expansión:

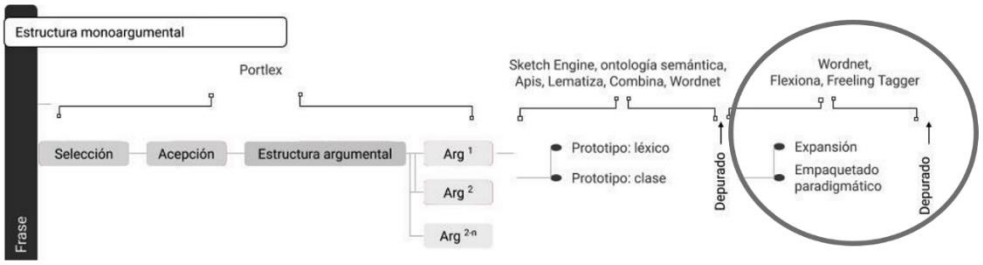

Figura 14. Expansión y empaquetado.

Domínguez Vázquez, Solla Portela y Valcárcel Riveiro (2019: 61) indican que:

> "the synsets of the wordnets following the *EuroWordNet* model of the *Multilingual Central Repository* (MCR)[29] (González-Agirre, Laparra & Rigau, 2012) are associated with semantic or cognitive features categorized in different ontologies".

Si bien es verdad que la organización cognitiva de las ontologías de *WordNet* no coincide con el inventario categorial del diccionario PORTLEX, también se constata que los rasgos categoriales de tipo general sí que conforman clases generales en las ontologías del *Multilingual Central Repository* (MCR). Por tanto, se diseñó una herramienta —*Combina*[30]— con la finalidad de combinar y cotejar los resultados de

varias consultas y, de este modo, poder extraer semiautomáticamente datos compartidos o combinados de la *Top Concept Ontology*[31] (Álvez, Atserias, Carrera, Climent, Laparra, Oliver & Rigau, 2008), de los *WordNet Domains*[32] (Bentivogli, Forner, Magnini & Pianta, 2004), de la *Suggested Upper Merged Ontology*[33] (Niles & Pease, 2001), *Basic Level Concept* (Izquierdo, Suárez & Rigau, 2007) y de los Epinónimos (Gómez Guinovart & Solla Portela, 2018), además de los primitivos semánticos de Miller, Beckwith, Fellbaum, Gross & Miller (1990)[34].

Mediante este procedimiento de expansión y posterior depurado, se consigue una selección de caudal léxico en el eje paradigmático, el cual comparte las características semánticas del prototipo léxico-semántico que tomamos como punto de partida, conformándose así una clase semántica concreta. Una vez obtenido este caudal léxico, realizamos el empaquetado paradigmático (etiquetado morfosintáctico y semántico) y la flexión. A estos paquetes léxicos (Domínguez Vázquez, Bardanca & Simões, 2021) recurre la herramienta *Xera* para la generación aleatoria automática de estructuras monoargumentales, así como la herramienta *Combinatoria,* en el caso de las biargumentales.

2.2.3. Establecimiento de patrones biargumentales

El simulador *Combinatoria* (vid. 2.1.) es el responsable de la generación de los patrones biargumentales. Cuenta con las reglas de selección programadas en *Xera* (ejemplo 9) e incorpora nuevas restricciones e incompatibilidades argumentales motivadas por la propia biargumentalidad (ejemplo 10)

(9) *el dolor en un intestino* frente a *el dolor en la pierna*: anotación de artículos determinados e indeterminados

(10) *la muerte por infarto/las muertes por infarto/las muertes por infartos* frente a *las *muertes de mi madre por infartos*: anotación de la relación singular-plural entre el núcleo y sus complementos

En esta línea, con el objetivo de afinar los resultados de la herramienta *Combinatoria* —atendiendo no solo a la (co)aparación de argumentos (lo cual se puede filtrar formalmente con consultas CQL en *Sketch Engine*), sino que a la de compatibilidad contextual de los candidatos léxicos de cada uno de los argumentos— usamos los métodos predictivos *word2vec* (Mikolov, Chen, Corrado & Dean, 2013) y *fastText* (Bojanowski, Grave, Joulin & Mikolov, 2017). De este modo, el filtrado se realiza siguiendo criterios de frecuencia de co-presencia contextual, pero no en un sentido monoargumental: según los datos de frecuencia obtenidos de *Sketch Engine* ya sabemos cuáles son los prototipos léxicos y cuáles las clases semánticas, pero lo que no podemos prever son todas las posibles incompatibilidades y restricciones que genera la combinación argumental de los ejemplares concretos de las clases léxicas. Así, por ejemplo, al combinar para el sustantivo MUERTE una clase semántica como [ANIMADO HUMANO CONDICIÓN HUMANA NEGATIVA] con [PROCESO

HUMANO MÉDICO], entre los resultados pueden aparecer ejemplos no aceptables como (11), frente a otros aceptables como (12):

(11) *la muerte del herido por ligadura de trompas* (sin aplicar *word2vec*)
(12) *la muerte del enfermo por trasplante* (aplicando *word2vec*)

Por consiguiente, nuestro empleo de *word embeddings* no va en la línea de entender o delimitar el significado de una palaba atendiendo a las palabras que habitualmente la acompañan o de analizar la similitud entre palabras (como el *Embedding Viewer* de *Sketch Engine* o el análisis paradigmático de *Derekovecs*[35]). Así pues, no tenemos como finalidad comprobar las unidades que acompañan a un núcleo o portador valencial en un contexto para con ello desambiguar su significado, sino que aplicamos estos métodos para filtrar los datos atendiendo a la compatibilidad semántico-contextual, esto es la (co)aparación de los argumentos con respecto al núcleo. De este modo, obtenemos la combinatoria de los ejemplares de las clases léxicas —por tanto, con valor argumental— respecto a un núcleo y, en definitiva, filtramos incompatibilidades del significado combinatorio (Engel, 2004). Sin lugar a dudas, esto resulta especialmente relevante, puesto que la corrección gramatical no implica automáticamente corrección o adecuación semántico-comunicativa.

La coherencia semántica, junto con la 'humanización' de los enunciados, es un parámetro de calidad de cualquier generador (Moreno Jiménez, Torres-Moreno, Wedemann & SanJuan, 2000). Con el objetivo de dotar a los datos generados de una apariencia más próxima a la comunicación humana nace nuestra herramienta más reciente, *CombiContext*. Este nuevo recurso, en fase de pruebas, bebe metodológicamente del diseño de *Xera* y *Combinatoria*, aunque requiere la aplicación de nuevos procedimientos analíticos y la implementación de nuevas reglas. El objetivo final de esta herramienta es generar contextos frasales (vid. Figura 11) y oracionales, por tanto, ejemplos de los patrones de valencia nominal pasiva más frecuentes (1.1).

CONCLUSIONES

El principal objetivo de nuestros proyectos consistía en constatar (o no) la validez del método combinado diseñado para la generación de patrones argumentales junto con un abanico variado de ejemplos, los cuales pudieran ser previamente seleccionados por el usuario siguiendo criterios formales y semánticos. De la viabilidad de las propuestas teóricas y metodológicas da cuenta no solo el actual funcionamiento de *Xera* y *Combinatoria*, sino que también su aplicación en el desarrollo de nuevos simuladores piloto para otras lenguas, como es el caso de *XeraWord*[36] para el gallego y portugués.

En su estado actual, los generadores cuentan con más de 3600 argumentos sintáctico-semánticos y más de 9000 estructuras biargumentales[37] con sus respectivos

ejemplos en ambos casos. Como era previsible, las combinatorias filtradas mediante los métodos predictivos *word2vec y fastText* presentan menos opciones de variabilidad en cuanto a sus candidatos léxicos, pero resultan siempre aceptables. Sin aplicar este filtro, los resultados obtenidos son significativamente más numerosos, esto es, obtenemos más ejemplos y más paradigmas de candidatos léxicos combinados. Todos ellos muestran corrección gramatical, pero, en algunos casos, escasa o nula aceptabilidad semántico-contextual al no concordar con nuestro conocimiento enciclopédico o cultural, como, por ejemplo, **el viaje de Cristóbal Colón a Croacia* (vid. también ejemplo 11). Esta incompatibilidad viene dada por el hecho de que los paquetes léxicos disponibles para la generación automática aleatoria son, en este caso, excesivamente generales. En favor de la optimización de los recursos en la línea señalada podemos optar por dos vías: la primera de ellas pasa por aplicar una mayor granularidad en la categorización de los paquetes léxicos, de tal modo que estos evolucionen a lo que Gross (2008: 11) denomina "*classe d'objets*"; la segunda consiste en habilitar en la interfaz de usuario un acceso doble a los datos (con o sin filtro). La opción de ofrecer ejemplos poco estándar o de dudosa aceptabilidad semántico-comunicativa es una aplicación de los generadores que también vale la pena explorar (Apresjan, Boguslavsky, Iomdin & Tsinman, 2003). Así, estudios preliminares sobre el uso de ambos generadores permitieron concluir que, entre docentes e investigadores, la obtención de dichos ejemplos resulta relevante. A este tipo de ejemplos y datos no se llega mediante el manejo de corpus, que compilan expresiones correctas y frecuentes.

El tipo de resultados que aportan las herramientas *Xera* y *Combinatoria* no son comparables con los proporcionados por recursos como los presentados en el apartado introductorio de este trabajo, pero tampoco con otras herramientas como los robots de producción de textos. Junto a las diferencias tipológicas constatables entre ambos, los mismos no persiguen los mismos objetivos: mientras que los robots crean automáticamente un texto a partir de un número reducido de palabras (para más información, vid. Simonsen, 2020), nuestros generadores ofrecen esquemas argumentales con dotación sintáctico-semántica junto con sus ejemplos, los cuales pueden ser consultados aplicando previamente filtros sintáctico-semánticos. Estos, a su vez, están diseñados para aportar una recopilación sistemática de reglas sintáctico-semánticas y de ejemplos no contemplados en otros diccionarios o, en su caso, bastante escasos o marginales.

Nuestros prototipos presentan, en definitiva, un modelo para la generación del potencial combinatorio sintáctico-semántico de la frase nominal y corporeizan, de este modo, una vía para el diseño automático de diccionarios de combinatoria que atiendan a las necesidades de consulta específica de un usuario en una situación de producción en lenguas extranjeras. En concreto, un nuevo tipo de diccionario entendido como una herramienta de información, que está concebida para "conseguir que el usuario de

estas pueda convertir los datos lexicográficos en información de la forma más rápida y fácil posible" (Fuertes Olivera, Niño Amo & Sastre Ruano, 2019: 79).

REFERENCIAS BIBLIOGRÁFICAS

Ágel, V. (2000). *Valenztheorie*. Tübingen: Narr.

Álvez, J., Atserias, J., Carrera, J., Climent, S., Laparra, E., Oliver, A. & Rigau, G. (2008). Complete and Consistent Annotation of WordNet Using the Top Concept Ontology. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds), *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08)* (pp. 1529-1234). Marrakech, Morocco: ELRA.

Apresjan, J. D., Boguslavsky, I., Iomdin, L. & Tsinman, L. (2003). Lexical Functions as a Tool of ETAP – 3. *MTT,* 16-18.

Armentano-Oller, C., Corbí Bellot, A. M., Forcada, M. L., Ginestí Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G. & Sánchez-Martínez, F. (2005). An Open-Source Shallow-Transfer Machine Translation Toolbox: Consequences of its Release and Availability. En *Proceedings of OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit,* Phuket, Thailand.

Bentivogli, L., Forner, P., Magnini, B. & Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage and Balancing. En G. Sérasset, S. Armstrong, C. Boitet, A. Popescu-Belis & D. Tufis (Eds.), *Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources* (pp. 94–101). Geneva, Switzerland: COLING.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.

Delli Bovi, C. & Navigli, R. (2017). Multilingual Semantic Dictionaries for Natural Language Processing: The Case of BabelNet. En P. C-Y. Sheu (Ed.), *World Scientific Encyclopedia with Semantic Computing and Robotic Intelligence. Semantic Computing* (pp. 149-163). Singapur: World Scientific.

DICE = Alonso Ramos, M. (2004). *Diccionario de colocaciones del español (DICE)* [en línea]. Disponible en: http://www.dicesp.com/paginas

Domínguez Vázquez, M. J. (2018). Was sind Valenzwörterbücher?. *Sprachwissenschaft*, *43*(3), 309-342.

Domínguez Vázquez, M. J., Engel, U. & Paredes Suárez, G. (2017). *Neue Wege zur Verbalenz*. Tomo I: *Theoretische und methodologische Grundlagen*. Tomo II: *Deutsch-spanisches Valenzlexikon*. Frankfurt: Peter Lang.

Domínguez Vázquez, M. J., Solla Portela, M. A. & Valcárcel Riveiro, C. (2019). Resources Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples. En I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference* (pp. 51-71). Brno: Lexical Computing CZ, s.r.o.

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2020). PORTLEX as a Multilingual and Cross-Lingual Online Dictionary. En M. J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Riveiro (Eds.), *Studies on Multilingual Lexicography* (pp. 135-158). Berlín: De Gruyter.

Domínguez Vázquez, M. J., Bardanca Outeiriño, D. & Simões, A. (2021). Automatic Lexicographic Content Creation: Automating Multilingual Resources Development for Lexicographers. En I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021* (pp. 269-287). Brno: Lexical Computing CZ, s.r.o.

Domínguez Vázquez, M. J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). *Ontología léxica***.** Santiago de Compostela [en línea]. Disponible en: http://portlex.usc.gal/ontologia/

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language, 67*(3), 547-619.

Eisenberg, P. (2006). *Grundriss der deutschen Grammatik*. 4. Edición. Stuttgart: Metzler.

Engel, U. (1996). Semantische Relatoren. Ein Entwurf für künftige Valenzwörterbücher. En N. Weber (Ed.), *Semantik, Lexikographie und Computeranwendungen* (pp. 223-236). Tübingen: Niemeyer.

Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.

Engel, U. (2009). *Syntax der deutschen Gegenwartssprache*, 4. Edición. Berlín: Erich Schmidt Verlag.

Eroms, H.-W. (2000). *Syntax der deutschen Sprache*. Berlín: de Gruyter.

Fillmore, C. J. (1977). Scenes-and-Frames Semantics. En A. Zampolli (Ed.), *Linguistic Structures Processing* (pp. 55-81). Amsterdam, New York & Oxford: North-Holland.

Foley, W. A. & Valin, R.D.V. (1984). *Functional Syntax and Universal Grammar*. Cambridge: University Press.

Fuertes Olivera, P., Niño Amo, M. & Sastre Ruano, A. (2019). Tecnología con fines lexicográficos. *Revista Internacional de Lenguas Extranjeras*, 10, 75-100.

Gamallo, P. & Pichel, J. R. (2007). Un método de extracción de equivalentes de traducción a partir de un corpus comparable castellano-gallego. *Procesamiento del Lenguaje Natural*, 39, 241-248.

Geyken, A. & Wiegand, F. & Würzner, K.-M. (2017). On-the-fly Generation of Dictionary Articles for the DWDS Website. En I. Kosem et al. (Eds.): *Electronic Lexicography in the 21st Century. Lexicography from scratch. Proceedings of eLex 2017 Conference* (pp. 560-570). Leiden, the Netherlands: Lexical Computing.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the Galician Wordnet: Methods and Applications. *Language Resources and Evaluation*, 52(1), 317-339.

González-Agirre A. & Laparra E. & Rigau G. (2012). Multilingual Central Repository Version 3.0: Upgrading a Very Large Lexical Knowledge Base. *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*, Matsue, Japan [en línea]. Disponible en: [https://adimen.si.ehu.es/~rigau/publications/gwc12-glr.pdf]

Gross, G. (2008). *Les classes d'objets*. París: Presses de l'Ecole normale supérieure.

Hashimoto, T. B., Zhang, H. & Liang, P. (2019). Unifying Human and Statistical Evaluation for Natural Language Generation. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Association for Computational Linguistics: Human Language Technologies*, 1 (pp. 1689-1701), Minneapolis, Minnesota: Association for Computational Linguistic.

Héja, E. & Takács, D. (2012). Automatically Generated Online Dictionaries. Istanbul. European Language Resources Association (ELRA), En N. Calzolari et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012*) (pp. 2487-2493) Istanbul: ELRA.

Hölzner, M. (2007). *Substantivvalenz. Korpusgestützte Untersuchungen zu Argumentrealisierungen deutscher Substantive*. Tübingen: de Gruyter.

Izquierdo, R., Suárez, A. & Rigau, G. (2007). Exploring the Automatic Selection of Basic Level Concepts. En G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov & N. Nikolov (Eds.), *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07)* (pp. 298-302). Borovetz: Incoma Ltd.

Kabashi, B. (2018). A Lexicon of Albanian for Natural Language Processing. En J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts* (pp. 855-862) Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts.

Kilgarriff, A., Husák, M., McAdam, K., Rudnell, R. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. En E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425-432) Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Krek, S. (2019). Natural Language Processing and Automatic Knowledge Extraction for Lexicography. *International Journal of Lexicography, 32*(2), 115-118.

Kipper, K., Dang, H. T. & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *Proceedings of Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 691-696). California: AAAI Press.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019). Identification and Automatic Extraction of Good Dictionary Examples: The Case(s) of GDEX. *International Journal of Lexicography, 32*(2), 119-137.

Kubczak, J. & Schumacher, H. (1998). Verbvalenz – Nominalvalenz. En D. Bresson & J. Kubczak (Eds.), *Abstrakte Nomina. Vorarbeiten zu ihrer Erfassung in einem zweisprachigen syntagmatischen Wörterbuch* (pp. 273-286). Tübingen: Niemeyer.

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. En D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Eds.), *13th European Conference on Computer Vision (ECCV)* (pp. 740-755). Zurich, Switzerland: Springer.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. En Y. Bengio & Y. LeCun (Eds.), *Proceeding of the International Conference on Learning Representations Workshop Track* (pp. 1301-3781). Arizona, USA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Wordnet: An On-Line Lexical Database. *International Journal of Lexicography*, 3, 235-244.

Moreno Jiménez, L. G., Torres-Moreno, J. M., Wedemann, R. S. & SanJuan, E. (2020). Generación automática de frases literarias. *Linguamática*, *12*(1), 15-30.

Nallapati, B., Zhou, B., dos Santos, C., Gulçehre, Ç. & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence rnns and Beyond. En S. Riezler & Y. Goldberg (Eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)* (pp. 280-290). Berlín: Association for Computational Linguistics.

Niles, I. & Pease, A. (2001). Towards a standard upper ontology. En *FOIS '01*. Ponencia presentada la *International Conference on Formal Ontology in Information Systems*. Nueva York: ACM.

Otter, D. W., Medina, J. R. & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604-624.

Padró, L., Collado, M., Reese, S., Lloberes, M. & Castellón, I. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En N. Calzolari et al. (Eds.), *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)* (pp. 931-936). La Valletta, Malta: ELRA Japan [en línea]. Disponible en: https://www.cs.upc.edu/~nlp/papers/padro10b.pdf.

Polenz, P. (1988). *Deutsche Satzsemantik: Grundbegriffe des Zwischen-den-Zeilen-Lesens*. Berlín/Nueva York: de Gruyter.

REDES = Bosque, Ignacio (Dir.) (2004). *REDES. Diccionario combinatorio del español contemporáneo*. Madrid: SM.

Reiter, E. & Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Renau, I. & Nazar, R. (2016). Automatic Extraction of Lexical Patterns from Corpora. En T. Margalitadze & G. Meladze (Eds.), *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity* (pp. 823-830). Tbilisi, Georgia: Ivane Javakhishvili Tbilisi State University.

Renau, I., Nazar, R., Casto, A., López, B. & Obreque, J. (2019). Verbo y contexto de uso: Un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos. Estudios de Lingüística, 52*(101), 878-901.

Roemmele, M. (2016). Writing Stories with Help from Recurrent Neural Networks. *AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 4311-4312). Phoenix, Arizona: AAAI Press.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R. & Scheffczyk, J (2006). *FrameNet II: Extended Theory and Practice*. Berkeley, CA: Tech. rep., International Computer Science Institute.

Simonsen, H. K. (2020). Augmented Writing Needs Lexicography: A Symbiotic Relationship? En Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, vol. I* (pp. 509-514). Ljubljana: Ljubljana University Press.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J-Y., Gao, J. & Dolan, B. (2015). A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. En R. Mihalcea, J. Chai & A. Sarkar (Eds.), *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL* (pp.196-205). Denver, Colorado: Association for Computational Linguistics.

Teubert, W. (1979). *Valenz des Substantivs. Attributive Ergänzungen und Angaben*. Düsseldorf: Schwann.

Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. En J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 25-37). Ljubljana: Ljubljana University Press.

Valcárcel, C. (2017). Las construcciones N1N2 como realizaciones actanciales del sustantivo en francés y su tratamiento en el diccionario multilingüe PORTLEX. En M. J. Domínguez Vázquez & S. Kutscher (Eds.), *Interacción entre gramática, didáctica y lexicografía: Estudios contrastivos y multicontrastivos* (pp. 193-207). Berlín: de Gruyter.

Vicente, M. E., Barros, C., Agulló, F., Peregrino, F. S. & Lloret, E. (2015*). La generación el lenguaje natural: Análisis del estado actual. Computación y Sistemas*, *19*(2), 721-756.

Zifonun, G., Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache vol. 3*. Berlín/Nueva York: de Gruyter.

NOTAS

[1] Esta investigación se desarrolla en el marco del Proyecto de investigación FFI2017-82454-P, financiado por MCIN/AEI//10.13039/501100011033/ FEDER "Una manera de hacer Europa".

[2] https://www.sketchengine.eu/

[3] https://cosmas2.ids-mannheim.de/cosmas2-web/

[4] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[5] http://verbs.colorado.edu/~mpalmer/projects/ace.html

[6] https://pdev.sketchengine.eu/

[7] http://verbatlas.org/)

[8] http://clic.ub.edu/corpus/es/ancora

[9] http://sentiwordnet.isti.cnr.it/

[10] http://sli.uvigo.gal/digalnet/

[11] http://tshwanedje.com/

[12] Ambos generadores son de acceso libre: http://portlex.usc.gal/combinatoria

[13] *PORTLEX. Diccionario multilingüe de la valencia del nombre*. [Disponible en: http://portlex.usc.gal/portlex/]

[14] La gramática de dependencias describe la relación sintáctica entre los constituyentes de un enunciado desde un punto de vista jerárquico, de tal modo que las relaciones de dependencia describen la conexión entre un regente —o unidad jerárquica superior— con un subordinado o dependiente. En su calidad de nodo superior en la estructura jerárquica oracional se le atribuye al verbo un estatus central. La rección específica de una clase concreta de portador valencial se denomina valencia (Domínguez Vázquez et al., 2017). El criterio de la especificidad argumental es aquí central (Engel, 2004).

[15] Fiel reflejo de ello es el reducido número de diccionarios de valencias monolingües, bilingües o multilingües. Para una visión de conjunto y una descripción tipológica de los diccionarios de valencias ver Domínguez Vázquez (2018).

[16] Esto los diferencia de otros diccionarios que plasman el potencial combinatorio de las unidades léxicas, como, por ejemplo, los diccionarios REDES (2004) o DICE (2004).

[17] De la valencia activa dan cuenta tanto el diccionario PORTLEX como los simuladores *Xera* y *Combinatoria*. De la valencia pasiva del nombre se ocupa el simulador *CombiContext*.

[18] Es comúnmente conocido que el número y tipo de roles semánticos difiere en la literatura científica. Así, por ejemplo, von Polenz (1988) describe 19 roles semánticos, frente a Dowty (1991) o Foley & Van Valin (1984). Estos últimos proponen dos proto-roles —Protoagente y Protopaciente— y dos macro-roles —*Actor* y *Undergoer*. En los corpus tampoco hay consenso en cuanto al tipo y número de roles semánticos. Por ejemplo, *Verbnet* (Kipper, Dang & Palmer, 2000) considera hasta 23 roles, los cuales nuevamente tampoco coindicen con las propuestas de *Framenet* (Ruppenhofer, Ellsworth, Schwarzer-Petruck, Johnson & Scheffczyk, 2016).

[19] http://hypermedia2.ids-mannheim.de/evalbu/index.html.

[20] Su paráfrasis es: 'Aquel/Aquello que experimenta un cambio', en este caso, de locación.

[21] Para mayor claridad obsérvese el análisis de la siguiente secuencia con la preposición *de*: *El viaje de Berlín a Santiago*: 'Dirección: origen'.

[22] Es, por tanto, necesario seguir contemplando posibles combinatorias argumentales, por ejemplo: *La discusión con Pedro sobre la situación actual*: ARG2 + ARG3: *sobre*; *La discusión sobre política entre los ciudadanos*: ARG3 + ARG1: *entre*; *La discusión de Pedro con Mario sobre deporte*: ARG1+ ARG2: con + ARG3: *sobre*, etc.

[23] Cabe subrayar que la inclusión de adjetivos, (miembros de) compuestos, N+N como realizaciones de casillas funcionales valenciales no plantea consenso en la literatura valencial.

[24] De esta necesidad da cuenta también la herramienta SkELL (*Sketch Engine for Language Learning*) que aporta ejemplos más adecuados para el aprendizaje de lenguas extranjeras.

[25] Atendiendo a los postulados de herencia léxica o de herencia construccional (Goldberg, 1995) la ampliación del modelo en cuanto al número de unidades descriptivas resulta viable. Esto se debe a que partimos de un modelo descriptivo recurrente, en el que las propiedades combinatorias generales de un sustantivo representante de un escenario y escena son heredadas por otros sustantivos de su misma clase semántica (vid. 1.2.).

[26] Téngase en cuenta que no todos los sustantivos derivados heredan el mismo número y tipo de argumentos de las clases palabras de las que proceden (Kubczak & Schumacher, 1998; Hölzner, 2007).

[27] Con el fin de facilitar la reutilización de los datos, su volcado y compilación puede realizarse en formato JSON o CSV.

[28] http://portlex.usc.gal/develop/lematiza/

[29] http://adimen.si.ehu.es/web/MCR

[30] http://portlex.usc.gal/develop/combina.php

[31] http://globalwordnet.org/gwa/ewn_to_bc/ewnTopOntology.htm

[32] http://wndomains.fbk.eu/

[33] http://www.adampease.org/OP/

[34] Las herramientas y el procedimiento analítico se presentan detalladamente en Domínguez Vázquez et al. (2019).

[35] http://corpora.ids-mannheim.de/openlab/derekovecs/

[36] *XeraWord*: http://ilg.usc.gal/xeraword/. En favor de la automatización de los procedimientos de análisis, este nuevo prototipo incorpora, además, la traducción automática (Domínguez Vázquez et al., 2021).

[37] Se presenta, a continuación, un modelo del tipo de información a la que refieren ambos tipos:

Argumento sintáctico-semántico	ARG1: N1:de	ARG1: N1:de: animado humano condición humana desplazamiento	"determinante", "{adjectivo o}", "nucleo", "{adjectivo o}", "de", "determinante", "actante N1"
Estructura biargumental	ARG1: N1:de ARG2: N2: con	ARG1: N1:de: animado humano familia ARG2: N2: con: animado humano nombre propio	'determinante', 'adjetivo_o', 'nucleo', 'adjetivo_o', 'de', 'determinante', 'actante N1', 'con', 'actante N2'